A Survey on LLM-Driven Personality

Anonymous ACL submission

Abstract

With the growing use of large language models (LLMs) in social, educational, and assistive contexts, understanding and controlling their personality traits has become increasingly important. In this survey, we provide a comprehensive overview of personality modeling in LLMs, covering methods ranging from rule-based systems to prompt engineering, fine-tuning, agent and retrieval techniques, as well as approaches to multimodal setups. We examine both qualitative and quantitative evaluation protocols, and identify key challenges including subjectivity, context dependence, and limited multimodal integration. We conclude by outlining open questions and future directions for building consistent, expressive, and trustworthy persona-driven LLMs.

1 Introduction

Recent breakthroughs in large language models (LLMs) have reshaped human–computer interaction, enabling systems that communicate with a fluency once reserved for human-to-human dialogue. These systems now power chatbots (Touvron et al., 2023), code assistants (Bai et al., 2023), and multimodal agents (Xie et al., 2024) that emulate rich, real-world communication in purely digital settings. Consequently, the research agenda has expanded beyond model scaling to encompass data-efficient training strategies (Lin et al., 2024b), rigorous evaluation frameworks for ensure quality and safety (Lin and Chen, 2023; Inan et al., 2023) and investigations about how to emulate human behavior in digital environments (Jiang et al., 2024).

A subtler frontier within this evolving landscape is the extent to perform consistent and recognizable personality traits that enrich user engagement using LLMs (Lee et al., 2025). As these models increasingly mediate social, educational, and assistive interactions, their perceived personality plays a critical role in shaping user trust, satisfaction, and long-term adoption (Kroczek et al., 2024). This emergent focus has sparked growing interest in how personality traits arise in LLMs, whether through pre-training data, instruction tuning, or prompt design, and how these traits can be measured, controlled, or aligned with user expectations and application goals.

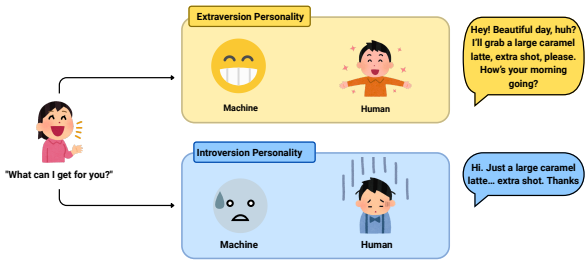

Figure 1: Illustration of a model performing different style-answers to the same input, based on its personality.

Furthermore, the study of personality in LLMs raises fundamental questions at the intersection of linguistics, psychology and artificial intelligence. Unlike traditional systems that rely on hardcoded traits, scripted responses, or purely statistical methods such as n-gram text generation (De Novais et al., 2010), LLMs can dynamically adapt their tone and style based on subtle contextual cues, achieving stable personality profiles and manifesting distinct persona-like behaviours when confronted with the same question or, conversely, sustain a coherent persona across disparate tasks, as illustrated in Figure 1. This has led to the development of new methodologies for personality assessment, drawing from established psycholinguistic frameworks such as the Big Five Inventory (John et al., 1991), as well as the creation of novel benchmarks and evaluation protocols tailored to generative AI systems (Huang and Hadfi, 2025). Additionally, since personality expression extends beyond text, integrating multimodal signals (i.e. voice tone, facial expressions, and gesture) remains a key challenge, calling for multidisciplinary approaches that

bridge language, vision, and speech technologies.

In light of these developments, this survey provides a comprehensive overview of the emerging landscape of personality in LLMs. In the following sections, we delve into a discussion of related studies, showing digital personality using traditional methods, LLM-based approach, multimodal setup and evaluation methods, and we summarize the taxonomy in Figure 2. We also highlight the challenges and potential gaps of the field, paving paths for future research.

2 Rule Based Personality Modelling

Early studies focused on the identification of personality traits and stylistic patterns through rule-based systems with manually curated lexical resources (Argamon et al., 2005), as well as hand-engineered features such as word counts and n-grams (Mairesse et al., 2007; Pennebaker et al., 2001). Some approaches also leveraged distributional semantics and classical embeddings, combined with traditional machine learning algorithms (Tandera et al., 2017). However, these methods were constrained by the limited expressiveness of their representations and a lack of contextual understanding. Consequently, most studies remained focused on classification tasks, rarely addressing the dynamic and generative aspects of personality expression in dialogue.

3 LLM-Driven Personality

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a paradigm shift in natural language processing, giving rise to decoder-based models capable of generating fluent and coherent text at scale. These models leverage massive datasets spanning diverse domains to learn rich representations of language, enabling generalization across a broad range of tasks (Brown et al., 2020).

Driven by their massive parameter counts, LLMs excel at capturing complex linguistic phenomena such as semantics, syntax, and long-range dependencies (Touvron et al., 2023), giving rise to emergent human-like behaviours. Among these, modeling human personality within LLMs has emerged as a promising yet underexplored direction. This line of research seeks to design models that not only respond coherently to input but also reflect stable psychological traits, thereby enriching interaction quality and user engagement (Kroczek et al., 2024).

Recent studies in this area have followed two primary research directions: (1) identifying and characterizing the intrinsic personality traits manifested by LLMs, and (2) developing mechanisms to induce specific personality traits. The first line focuses on evaluating the implicit personality tendencies exhibited by pre-trained models, often using established psychological frameworks such as the Myers-Briggs Type Indicator (MBTI) (Myers and McCaulley, 1988) and the Big Five personality traits (De Raad, 2000). For instance, Pan and Zeng (2023) and Serapio-García et al. (2023) conducted empirical analyses to assess how LLMs align with human personality, suggesting that some traits may emerge naturally as a byproduct of the training data and architectural biases.

Beyond merely identifying inherent tendencies, equipping these models with specific personality traits presents a more complex challenge, involving multiple stages of adaptation and control. Typically, the development of LLMs involves two main phases: pre-training and fine-tuning. In the pre-training stage, the model is exposed to large-scale corpora through an unsupervised next-token prediction objective (Brown et al., 2020), enabling it to learn rich representations of language, including grammar, semantics, and discourse patterns. Fine-tuning then follows as a supervised process that adapts these general capabilities to more specific tasks or domains, often using task-specific labeled datasets (Ziegler et al., 2024).

While supervised fine-tuning affords precise persona control, its dependence on extensive, high-quality, persona-aligned data renders it costly and difficult to scale. To bypass these weight updates, recent work explores in-context learning (ICL), conditioning the model at inference with persona-defining instructions or exemplars (Dong et al., 2024a). Intrinsic conditioning is further complemented by extrinsic controllers, such as agent-style planners (Park et al., 2023) and retrieval-augmented generation (RAG) modules (Lewis et al., 2020), that dynamically inject user profiles, episodic memory, or affective states into the prompt. Following, we delve into the main strategies used to address personality in LLMs, outlining their underlying mechanisms, benefits, and limitations.

3.1 Zero-Shot Learning

Zero-shot learning refers to the ability of large language models to perform new tasks or exhibit spe-

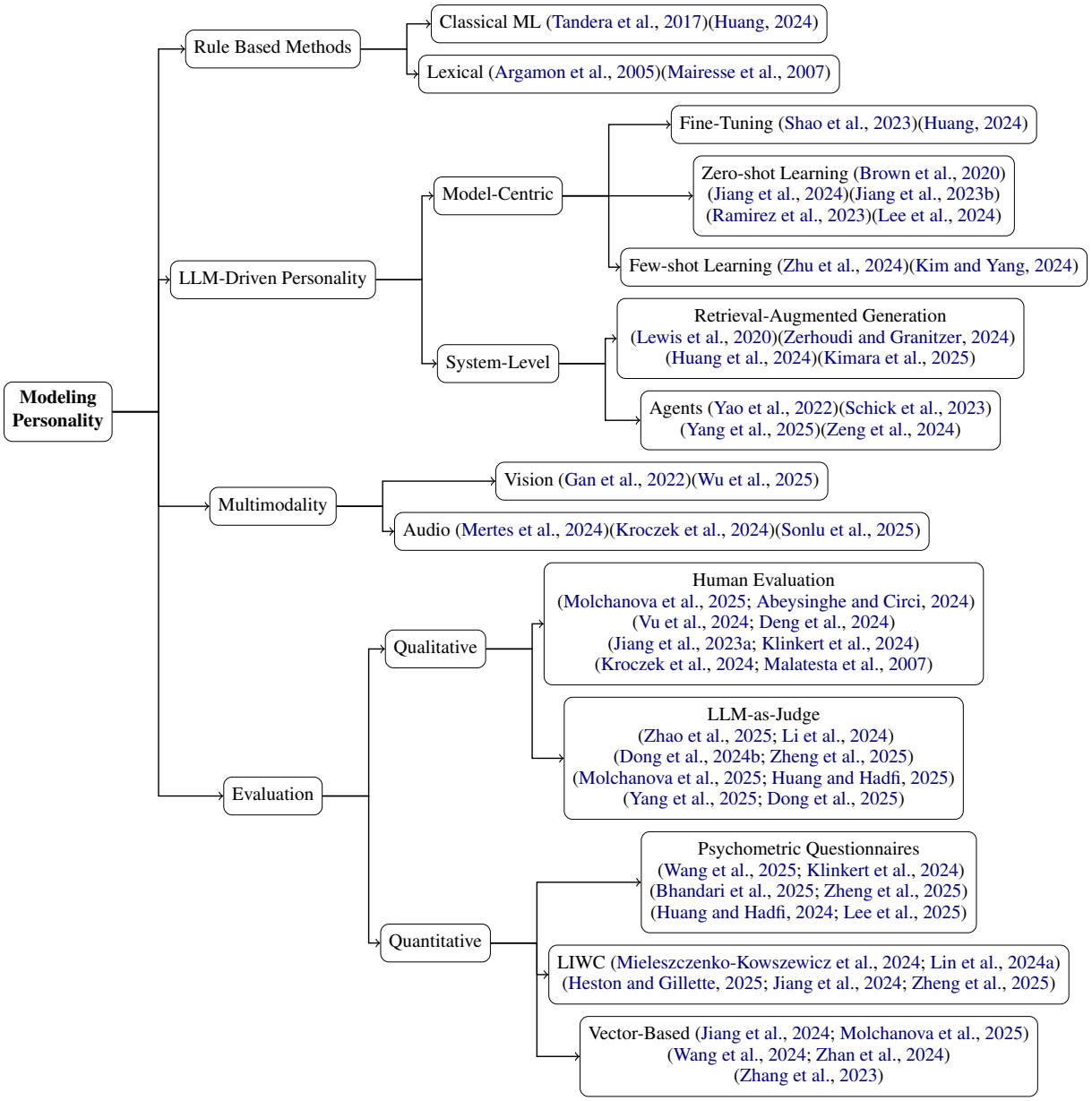

Figure 2: Taxonomy of personality modeling task on digital environments.

cific behaviors without receiving any explicit examples or task-specific training (Brown et al., 2020). Instead, the model relies solely on its pre-trained knowledge and the conditioning provided by a carefully designed prompt. In the context of personality modeling, zero-shot approaches leverage this inherent flexibility by crafting prompts that implicitly encode the desired psychological traits, guiding the model to generate responses aligned with specific personality profiles.

One notable example is PersonaLLM (Jiang et al., 2024), which investigates whether LLMs can consistently exhibit specific Big Five personality traits in a zero-shot setting. The authors employed prompts to instantiate distinct personality profiles (e.g., high extroversion and low neuroticism). These persona-conditioned models were then evaluated via both questionnaire and open-ended storytelling tasks. Results showed that the simulated responses aligned with the intended personality traits both quantitatively and qualitatively, with human raters correctly inferring some traits from generated text.

Jiang et al. (2023b) present the Machine Personality Inventory (MPI), a Big-Five multiple-choice test that elicits LLM self-ratings in a purely zero-shot setting. The resulting scores produce internally consistent, human-like profiles. They further introduce Personality Prompting (P2), a chain-of-descriptors template that reliably induces target

traits without any parameter updates.

Some other works (Ramirez et al., 2023; Lee et al., 2024) have also investigated the use of zero-shot prompting techniques to align the personality of large language models. However, zero-shot methods present challenges in consistently maintaining the intended traits across diverse conversational contexts and potential sensitivity to subtle variations in prompt phrasing, affecting stability and predictability of the personality outcomes.

3.2 Few-Shot Learning

Few-shot learning conditions an LLM with a handful of persona-labelled exemplars inserted directly into the input. These in-prompt demonstrations provide on-the-fly supervision, where no parameter updates are required, enabling the model to internalise and generalise the target psychological style across new topics and interaction contexts.

Zhu et al. (2024) evaluate few-shot prompting as a baseline for inducing personality traits in LLMs. They incorporate exemplar responses derived from psychometric profiles, such as IPIP-NEO questionnaires (Johnson, 2014), into the prompt to simulate specific personality expressions. This method allows the model to align with target traits more reliably during interaction, serving as a behavioral scaffold for personality instantiation.

Another notable approach is FERMI (Kim and Yang, 2024), which proposes a few-shot personalization framework that iteratively optimizes prompts using user profiles and a small set of prior responses. Instead of relying solely on correct examples, the proposed method also incorporates misaligned LLM outputs as additional context to guide prompt refinement. At inference, FERMI selects the most relevant personalized prompt based on the test query.

Despite its advances in personality consistence when compared to zero-shot approach, it is important to highlight the limitations of few-shot based methods. The performance depends heavily on the quality and consistency of the demonstrations, and the lack of robust generalization to unseen traits or domains remains a challenge.

3.3 Fine-Tuning

Fine-tuning refers to the process of updating the internal parameters of a pre-trained language model by training it on labeled datasets tailored to specific tasks or desired behaviors. Unlike zero-shot or few-shot methods, fine-tuning does not rely solely on prompt manipulation during inference. Instead, it systematically adjusts the model's weights to internalize the desired personality traits. This approach enables the creation of agents whose linguistic style, emotional tone, and response strategies are deeply aligned with specified psychological profiles.

Character-LLM framework (Shao et al., 2023) models personality through supervised fine-tuning of LLMs on synthetic, character-specific experience data. The authors reconstruct a character's biography by extracting profile-based scenes and extending them into detailed interactions. These experiences are uploaded to the base model, training it to internalize emotional, behavioral, and linguistic patterns unique to historical or fictional figures. Additionally, protective experiences are introduced to suppress out-of-character knowledge, reinforcing persona consistency. The fine-tuned agents demonstrate improved personality alignment, memory of past events, and reduced hallucinations in role-based simulations.

ORCA (Huang, 2024) introduces a multi-stage fine-tuning framework for enhancing the role-playing capabilities of large language models by incorporating psychologically grounded personality traits. The authors first infer continuous Big Five personality scores from user-generated content, then simulate user profiles, motivations, and psychological activities to construct a rich, personality-conditioned dataset. Two fine-tuning strategies are proposed: PTIT (using trait descriptions) and PSIT (using interpreted trait scores), with empirical results showing that the proposed approach substantially improves personality consistency and relevance across generated outputs, setting a new benchmark for personalized dialogue generation in social platforms.

Despite their promising results, fine-tuned models suffer from several limitations. First, they require large amounts of high-quality, personality-specific training data, which is scarce and costly to obtain. Second, the fine-tuning process can lead to overfitting, reducing generalizability across tasks or domains. Third, full fine-tuning is computationally expensive and environmentally costly. Lastly, personality fine-tuning can unintentionally overwrite general knowledge, a phenomenon often called as *Catastrophic forgetting* (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017).

3.4 Retrieval-Augmented Generation

Retrieval-augmented generation enriches language-model outputs by fetching evidence from an external knowledge base at inference time, passing the retrieved passages to the prompt so the generator can ground its response in verifiable facts, allowing an improvement to factual accuracy, reduce hallucinations, and facilitate rapid domain adaptation across tasks (Lewis et al., 2020). The same retrieval-and-fusion loop also offers a lightweight pathway to persona control: by sourcing personality descriptors, dialogue history, or user-preference records on the fly, RAG can imprint stable behavioral signatures on each reply maintaining coherent and user-aligned personality over time.

PersonaRAG (Zerhoudi and Granitzer, 2024) extends the RAG paradigm by embedding a modular multi-agent architecture aimed at enhancing user-aware retrieval and generation. The system distributes responsibilities across five dedicated agents (user profile, contextual retrieval, session tracking, document ranking, and feedback integration), which communicate through a global memory pool to iteratively adapt responses to the user's evolving needs. This framework exemplifies how RAG can be leveraged for fine-grained personalization without fine-tuning, since its reliance on in-context learning.

Similarly, Huang et al. (2024) extends the paradigm with Emotional RAG, a framework that integrates emotional context into the retrieval process, allowing role-playing agents to generate responses that are congruent with both the semantic and emotional states of the conversation, enhancing the authenticity of simulated personalities. Complementarily, PersonaAI (Kimara et al., 2025) presents a mobile-based RAG system for generating persona-consistent responses by continuously collect and embedded user data for retrieval, enabling dynamic prompt augmentation with contextually relevant information. These approaches demonstrate that retrieval-based systems can significantly enhance both the consistency and expressiveness of personality modeling, while offering greater interpretability and modularity than purely parameter-based methods.

Despite its advantages, RAG systems face notable limitations. Barnett et al. (2024) identify seven failure points in RAG pipelines: missing content, missed top-ranked documents, context exclusion, extraction failure, format mismatch, incorrect specificity, and incomplete answers. These issues reflect the complexity of coordinating retrieval and generation, particularly under noisy, ambiguous, or underspecified conditions. Furthermore, since RAG relies on multiple interacting modules, validation must occur in real time, presenting a bottleneck for system robustness and deployment.

3.5 LLM-Based Agents

LLM-based agents augment a language-model reasoning core with memory, tool-use, and decision modules that track state, incorporate feedback, and plan over multi-turn horizons, enabling autonomous, goal-oriented behaviour in complex environments (Yao et al., 2022; Schick et al., 2023). Integrating personality modelling into this architecture adds a further layer of coherence: the agent can modulate tone, affect, and response strategy according to stable traits such as openness, conscientiousness, or extraversion, an ability essential for scenarios where persona consistency directly shapes user trust and engagement.

Recent studies have proposed agent frameworks explicitly designed for personality conditioning. For instance, PsyPlay (Yang et al., 2025) introduces a multi-agent framework where LLMs engage in role-playing dialogues while portraying predefined traits. Agents are instantiated with role cards and interact over realistic topics. Similarly, Zeng et al. (2024) defines persona-driven action policies for interactive tasks, demonstrating that agents conditioned on specific personality profiles generate consistent, relatable, and user-aligned outputs.

While agent LLM architectures enable modularity and specialization, they also introduce notable limitations. Agashe et al. (2023) shows that agents often struggle to coordinate, failing to converge on joint plans and adapting poorly as partners' behaviors shift. Additionally, Cemri et al. (2025) highlights failure modes including inter-agent misalignment and verification problems, which can lead to degraded performance. These findings point to an urgent need for stronger orchestration and communication protocols in multi-agent LLM systems.

4 Personality Modeling Beyond Text

Although textual dialogue allows to convey many aspects of personality, finer-grained affective cues, such as intonation, facial micro-expressions, gesture, and the environment, emerge only when additional modalities are brought into the loop. Em-

bedding LLMs within speech, vision, and engaging interfaces therefore enriches the communicative channel, supplying a denser signal space from which stable and nuanced personality displays can arise.

4.1 Text-Visual Personality

Audio and visual channels deliver prosodic, facial, and contextual cues that ground personality perception in more human-like exchanges. While recent vision–language models (VLMs) have accelerated multimodal research (Wu et al., 2024), most studies still treat personality as a recognition problem rather than generating responses that embody a target persona. A representative example is PsyClip (Gan et al., 2022), a zero-shot model built on the CLIP framework (Radford et al., 2021), which matches face images to Myers–Briggs Type Indicator descriptors by aligning visual embeddings with adjective-based textual prompts.

Similarly, Wu et al. (2025) encode text and images with modality-specific transformers, fuse the resulting representations in a cross-modal emotion encoder, and append an MBTI-based personality embedding derived from dialog history. The joint vector guides a response generator that produces utterances which are both contextually appropriate and empathetically aligned with the speaker's inferred personality. Nonetheless, the reliance on coarse MBTI categories constrains stylistic breadth, preventing the system from synthesising richer, situation-dependent personas or fully leveraging visual context during generation.

4.2 Audio Personality

In contrast, persona modelling through the audio channel is still in its infancy. Recent neural speech systems, such as VoiceX (Mertes et al., 2024), demonstrate that prosody can be tuned to convey stylistic personality identity, yet most studies either reuse a single synthetic voice for every persona (Kroczek et al., 2024) or generate speech whose unnatural timbre masks the intended traits (Sonlu et al., 2025). Developing high-fidelity, persona-controllable voices therefore remains a key open challenge for multimodal personality research.

5 Evaluating LLMs Personality Traits

The psychology of personality has long sought to classify individual differences (Cattell and Kline, 1977), and the tight coupling between language and personality (Pennebaker and King, 1999; Lee et al., 2007) makes text an appealing lens for probing LLM behaviour. Recent studies test trait stability (Song et al., 2024), refine measurement protocols (Zou et al., 2024), analyse safety implications (Zhang et al., 2024), and tailor personas to task requirements (Zhao et al., 2025), yet nearly all rely on frameworks devised for humans (Vu et al., 2024). As a result, personality evaluation in LLMs remains hindered by subjectivity, context dependence, and the absence of shared standards. The following sections review qualitative and quantitative approaches, highlighting their advantages, drawbacks, and suitability for conversational agents. 1 presents a direct comparison between different evaluations methods.

5.1 Qualitative Evaluation

Evaluating the personality traits of LLMs involves complex, nuanced, and non-standardized methods (Jiang et al., 2024). Qualitative approaches are widely used across studies to assess these traits, relying on subjective judgments from human evaluators (Molchanova et al., 2025) or, as explored in recent works, by other LLMs serving as judges (Zhao et al., 2025). This section briefly explains how human evaluation and LLM-as-Judge methods are used to assess LLM personality traits.

Human Evaluation. Human evaluation remains the gold-standard qualitative method for assessing whether an LLM's behaviour aligns with desired persona specifications (Abeysinghe and Circi, 2024; Vu et al., 2024). Annotators typically score or classify model-generated responses (Deng et al., 2024; Jiang et al., 2023a), sometimes contrasting them by comparing between human and model's outputs (Klinkert et al., 2024). For instance, in Molchanova et al. (2025), human evaluators scored personality traits from LLM-generated texts simulating specific personalities from a range of -2 to +2 based on trait descriptions and guidelines, highlighting words or phrases that influenced their scores, assessing whether LLMs could effectively simulate distinct personalities. Despite its widely application use not only in text responses evaluation but also to user perception studies (Kroczek et al., 2024) and multimodal trait assessment in embodied agents (Malatesta et al., 2007), human evaluation reliability is challenged by subjectivity, demographic bias (Antal and Beder, 2025), and high cost, making it difficult to scale and reproduce results consistently (Clark et al., 2021).

Method	Type	Traceable	Scalable	Prompt-Agnostic	Context-Aware
Human Evaluation	Qualitative	✓	✗	✓	✓
LLM-as-Judge	Qualitative	✓	✓	✗	✓
Personality Tests	Quantitative	✓	✓	✗	✗
LIWC (Word Count)	Quantitative	✓	✓	✓	✗
Vector-Based	Quantitative	✗	✓	✓	✓

Table 1: Comparison of evaluation methods for LLM personality traits. ✓ indicates presence; ✗ indicates limitation or absence.

LLM-as-Judge. This paradigm prompts an LLM to rate the outputs of another model against explicit rubrics, automating evaluation and vastly reducing annotation cost and latency (Li et al., 2024; Dong et al., 2024b). In personality research it has been used to translate free-form text into numerical trait scores (Zheng et al., 2025), classify personas from single utterances (Molchanova et al., 2025), and infer user profiles across whole dialogues (Zhao et al., 2025; Yang et al., 2025). Single-judge setups, however, import the evaluator model's own biases and can yield inconsistent or unreliable ratings (Zheng et al., 2023b). Huang and Hadfi (2025) mitigate this with a *Multi-observer* framework in which several role-conditioned LLMs (e.g., "friend," "colleague") independently score the target, improving robustness through aggregated views. Nevertheless, even multi-observer systems remain constrained by the models' cultural priors, limited situational understanding, and susceptibility to hallucination (Dong et al., 2025; Chen et al., 2024).

5.2 Quantitative Evaluation

Quantitative evaluation methods are essential for assessing personality traits in LLMs in a structured, objective way (Safdari et al., 2023). These approaches include self-assessments, in which LLMs respond to personality questionnaires to produce numerical scores (Wang et al., 2025; Klinkert et al., 2024), as well as objective textual analyses, such as word count metrics (Mieleszczenko-Kowszewicz et al., 2024) and feature extraction from text (Jiang et al., 2024). Quantitative evaluations provide standardized, numerical outputs that reduce ambiguity and improve consistency (Bhandari et al., 2025).

Personality Questionnaires. Personality questionnaires originally designed for human psychological assessment such as the Big Five Inventory (BFI) (John et al., 1991) and the International Personality Item Pool (IPIP) (Goldberg et al., 2006) are widely used to evaluate personality traits in LLMs. In these structured assessments, LLMs are prompted with standardized items and their responses are scored to derive trait profiles and response patterns (Lin et al., 2024a; Heston and Gillette, 2025). However, standard self-report formats (e.g., Likert items, true–false questions, and forced-choice prompts) are fragile since the models answers are mere next-token predictions instead of relying on stable traits (Zou et al., 2024; Zheng et al., 2025). In such cases, the order of alternatives influence directly the model's answer (Zheng et al., 2023a), and scale biases mirror the distribution of its training data (Huang and Hadfi, 2024). Although scenario-based frameworks mitigates bias and reduce reliance on self-reflection by presenting diverse situations and multi-order evaluations (Lee et al., 2025), the stability of personality assessments in LLMs remains as a challenge, since minor edits to wording or format can swing the results, compromising reproducibility and consistency (Gupta et al., 2024).

Linguistic Inquiry and Word Count (LIWC). Pennebaker et al. (2001) analyzes text by mapping words and phrases to a curated dictionary, categorizing them into psychological, emotional, and social dimensions (Tausczik and Pennebaker, 2010). Its latest version, LIWC-22, includes over 12,000 words and expressions across 117 categories, such as personal pronouns, emotion-related terms, and cognitive indicators. Widely used in psychology (Tov et al., 2013), LIWC has also been applied to study and classify personality traits in LLMs (Mieleszczenko-Kowszewicz et al., 2024; Jiang et al., 2024), mapping responses to predefined linguistic categories and personality dimensions, revealing subtle linguistic patterns in generated texts and offering valuable insights into how LLMs express and emulate personality traits. Despite its popularity, LIWC doesn't account for contextual or semantic nuances, which is problematic given LLMs' reliance on broader context for mean-

ing. Additionally, Zheng et al. (2025) argues that LIWC's rigid categories limit its effectiveness in evaluating dynamically generated language. Nevertheless, LIWC remains widely used due to its simplicity, accessibility, and ability to provide standardized insights into the linguistic patterns associated with personality in LLMs.

Vector-Based. Vector-based personality analysis uses high-dimensional vector representations to map textual inputs, capturing semantic meaning of texts. These approaches identify personality traits by analyzing latent representations (Molchanova et al., 2025; Wang et al., 2024), ranging from basic TF-IDF (Sparck Jones, 1972) to contextual embeddings (Chang and Chen, 2019). A key advantage of embedding-based methods is their ability to preserve contextual relationships between words, allowing the detection of subtle psychological features. For instance, Zhang et al. (2023) proposes PsyAttention, a transformer-based encoder that represents psychological features as dense embeddings, in which the vectorized psychological features allow the model to process abstract traits as part of its neural architecture, enabling classification of both human and LLM-generated text under established psychometric frameworks. However, while such embeddings capture subtle contextual cues, these vector representations are not inherently interpretable, rely heavily on feature engineering and is weak psychometric validity, since embeddings may correlate with personality constructs learned from data rather than grounded in formal psychometric theory. Additionally, classification typically requires a separate model after vectorization, adding complexity and potential for error.

6 Challenges and Future Directions

Despite recent advances, personality modeling with large language models remains limited by several unresolved challenges. Foremost among these is the generalization and controllability of personality expression. Prompt-based techniques, while flexible, are inherently fragile and prone to producing inconsistent outputs across tasks and domains. Supervised fine-tuning, though more stable, remains constrained by data scarcity, high computational cost, and risks of overfitting or catastrophic forgetting. These limitations are further exacerbated in multi-agent systems, where inconsistent persona enactment can disrupt coordination, leading to degraded performance in collaborative settings.

Although personality expression is inherently multimodal, encompassing prosody, facial expression, and gesture, current approaches rarely integrate other modalities. This restricts the validity of simulated personalities, particularly in embodied or socially interactive contexts.

Additionally, the lack of standardized, robust evaluation protocols remains as a barrier. Current assessment strategies exhibit high sensitivity to prompt phrasing, task framing, and input order, undermining reproducibility and comparability across studies. Moreover, existing methods often assume stable, human-like personality structures, which may not align with the dynamic and context-dependent nature of LLM behavior.

To advance the field, several directions requires further exploration. First, scalable personalization techniques, such as parameter-efficient fine-tuning and retrieval-augmented control, offer promising paths for adapting traits across users and applications. Second, integrating multimodal capabilities, including speech synthesis and visual embodiment, may enable more realistic and expressive personality representations. Third, the development of prompt-invariant, context-aware, and psychometrically grounded evaluation benchmarks is essential to establish methodological rigor. Finally, personality-aware alignment frameworks must be developed to ensure that trait-driven behaviors remain safe, coherent, and socially appropriate, particularly in high-stakes or long-term deployments.

7 Conclusion

In this paper, we present a comprehensive survey of personality modeling in large language models, covering foundational methods, LLM-driven techniques, multimodal approaches, and evaluation strategies. We analyze how personality traits are identified, induced, and evaluated, and we categorize the current landscape into a structured taxonomy. To the best of our knowledge, this is the first survey covering this huge massive techniques of personality in LLMs, such as the usage of agents and RAG to enhance personality and evaluation in LLMs. We aim to consolidate the state of the art, identify open challenges, and offer insights to guide future research in building consistent, expressive, and user-aligned LLMs.

Limitations

This survey aims to provide a comprehensive overview of personality modeling with large language models, spanning conditioning strategies, multimodal architectures, and evaluation methodologies. Nonetheless, due to the rapidly evolving nature of the field, it is possible that some recent or domain-specific contributions were not included. In particular, emerging work on personality expression in low-resource languages, cultural adaptation, and longitudinal user studies falls beyond the scope of this paper. Additionally, while we categorize a range of modeling and evaluation strategies, we do not perform empirical benchmarking or reimplementation of existing methods. Our focus remains on conceptual mapping rather than quantitative comparison. Finally, although we discuss multimodal and embodied approaches, most of the cited literature remains text-centric. A deeper analysis of personality modeling in vision and speech-based agents is left for future work.

Acknowledgments

This work has been funded by the project "anonymized", supported by "anonymized", with financial resources from the "anonymized" grant number "anonymized", signed with "anonymized"

References

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches. arXiv preprint arXiv:2406.03339.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. arXiv preprint arXiv:2310.03903.

Margit Antal and Norbert Beder. 2025. Eysenck personality questionnaire: A comparative study of humans and large language models through repeated administrations. Acta Universitatis Sapientiae, Informatica, 16:219–235.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America, pages 1–16. USA).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, pages 194–199.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. arXiv preprint arXiv:2502.05248.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Raymond B Cattell and Paul Ed Kline. 1977. The scientific analysis of personality and motivation. Academic Press.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, and 1 others. 2025. Why do multi-agent llm systems fail? arXiv preprint arXiv:2503.13657.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6064–6070.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7282–7296, Online. Association for Computational Linguistics.

Eder Miranda De Novais, Thiago Dias Tadeu, and Ivandré Paraboni. 2010. Improved text generation using n-gram statistics. In Advances in Artificial Intelligence–IBERAMIA 2010: 12th Ibero-American Conference on AI, Bahía Blanca,

Argentina, November 1-5, 2010. Proceedings 12, pages 316–325. Springer.

Boele De Raad. 2000. The big five personality factors: the psycholexical approach to personality. Hogrefe & Huber Publishers.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Neuron-based personality trait induction in large language models. arXiv preprint arXiv:2410.12327.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024a. A survey on in-context learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, and 1 others. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. arXiv preprint arXiv:2505.00049.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024b. Can LLM be a personalized judge? In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Peter Gan, Arcot Sowmya, and Gelareh Mohammadi. 2022. Zero-shot Personality Perception From Facial Images, pages 43–56.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. Journal of Research in personality, 40(1):84–96.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of LLM personality. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

Thomas F Heston and Justin Gillette. 2025. Do large language models have a personality? a psychometric evaluation with implications for clinical medicine and mental health ai. medRxiv, pages 2025–03.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. arXiv preprint arXiv:2410.23041.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? a simulation based on large language models. arXiv preprint arXiv:2407.11549.

Yin Jou Huang and Rafik Hadfi. 2025. Beyond self-reports: Multi-observer agents for personality assessment in large language models. arXiv preprint arXiv:2504.08399.

Yuxuan Huang. 2024. Orca: Enhancing role-playing abilities of large language models by integrating personality traits. arXiv preprint arXiv:2411.10006.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv preprint arXiv:2312.06674.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems, 36:10622–10643.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. Journal of personality and social psychology.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. Journal of research in personality, 51:78–89.

Jaehyung Kim and Yiming Yang. 2024. Few-shot personalization of llms with mis-aligned responses. arXiv preprint arXiv:2406.18678.

Elvis Kimara, Kunle S Oguntoye, and Jian Sun. 2025. Personaai: Leveraging retrieval-augmented generation and personalized context for ai-driven digital avatars. arXiv preprint arXiv:2503.15489.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017.

Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526.

Lawrence J. Klinkert, Steph Buongiorno, and Corey Clark. 2024. Evaluating the efficacy of llms to emulate realistic human personalities. In Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE '24. AAAI Press.

Leon OH Kroczek, Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, and Andreas Mühlberger. 2024. The influence of persona and conversational task on social interactions with a llm-controlled embodied conversational agent. arXiv preprint arXiv:2411.05653.

Chang H Lee, Kyungil Kim, Young Seok Seo, and Cindy K Chung. 2007. The relations between personality and language use. The Journal of general psychology, 134(4):405–413.

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. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, and 1 others. 2024. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. arXiv preprint arXiv:2406.14703.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579.

Qianli Lin, Zhipeng Hu, and Jun Ma. 2024a. The personality of the intelligent cockpit? exploring the personality traits of in-vehicle llms with psychometrics. Information, 15(11):679.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024b. Data-efficient fine-tuning for llm-based recommendation. In Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, pages 365–374.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. arXiv preprint arXiv:2305.13711.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research, 30:457–500.

Lori Malatesta, George Caridakis, Amaryllis Raouzaiou, and Kostas Karpouzis. 2007. Agent personality traits in virtual environments based on appraisal theory predictions. Artificial and ambient intelligence, language, speech and gesture for expressive characters, AISB, 7.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation, volume 24, pages 109–165. Elsevier.

Silvan Mertes, Daksitha Withanage Don, Otto Grothe, Johanna Kuch, Ruben Schlagowski, and Elisabeth André. 2024. Voicex: A text-to-speech framework for custom voices. arXiv preprint arXiv:2408.12170.

Wiktoria Mieleszczenko-Kowszewicz, Dawid Płudowski, Filip Kołodziejczyk, Jakub Świstak, Julian Sienkiewicz, and Przemysław Biecek. 2024. The dark patterns of personalized persuasion in large language models: Exposing persuasive linguistic features for big five personality traits in llms responses. arXiv preprint arXiv:2411.06008.

Maria Molchanova, Anna Mikhailova, Anna Korzanova, Lidiia Ostyakova, and Alexandra Dolidze. 2025. Exploring the potential of large language models to simulate personality. arXiv preprint arXiv:2502.08265.

Isabel Briggs Myers and Mary H McCaulley. 1988. Myers-Briggs type indicator: MBTI. Consulting Psychologists Press Palo Alto.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. arXiv preprint arXiv:2307.16180.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pages 1–22.

James W. Pennebaker, Margaret E. Francis, and Roger J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC. Lawrence Erlbaum Associates, Mahwah, NJ.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6):1296.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR.

Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. arXiv preprint arXiv:2302.03848.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Mataric. 2023. Personality traits in large language models. arxiv. arXiv preprint arXiv:2307.00184.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36:68539–68551.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. arXiv preprint arXiv:2310.10158.

Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. 2024. Identifying multiple personalities in large language models with external evaluation. arXiv preprint arXiv:2402.14805.

Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2025. Effects of embodiment and personality in llm-based conversational agents. In 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR), pages 718–728. IEEE.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1):11–21.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, and 1 others. 2017. Personality prediction system from facebook users. Procedia computer science, 116:604–611.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology, 29(1):24–54.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

William Tov, Kok Leong Ng, Han Lin, and Lin Qiu. 2013. Detecting well-being via computerized content analysis of brief diary entries. Psychological assessment, 25(4):1069.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, and 1 others. 2024. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. arXiv preprint arXiv:2412.16882.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Derek F Wong, and Min Yang. 2025. Exploring the impact of personality traits on llm bias and toxicity. arXiv preprint arXiv:2502.12566.

Yixiao Wang, Homa Fashandi, and Kevin Ferreira. 2024. Investigating the personality consistency in quantized role-playing dialogue agents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 239–255, Miami, Florida, US. Association for Computational Linguistics.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. Advances in Neural Information Processing Systems, 37:69925–69975.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025. From traits to empathy: Personality-aware multimodal empathetic response generation. In Proceedings of the 31st International Conference on Computational Linguistics, pages 8925–8938, Abu Dhabi, UAE. Association for Computational Linguistics.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. arXiv preprint arXiv:2402.15116.

Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. 2025. Psyplay: Personality-infused role-playing conversational agents. arXiv preprint arXiv:2502.03821.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems, 35:20744–20757.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024. Persllm: A personified training approach for large language models. arXiv preprint arXiv:2407.12393.

Saber Zerhoudi and Michael Granitzer. 2024. Personarag: Enhancing retrieval-augmented generation systems with user-centric agents. arXiv preprint arXiv:2407.09394.

Baohua Zhan, Yongyi Huang, Wenyao Cui, Huaping Zhang, and Jianyun Shang. 2024. Humanity in ai: Detecting the personality of large language models. arXiv preprint arXiv:2410.08545.

Baohua Zhang, Yongyi Huang, Wenyao Cui, Zhang Huaping, and Jianyun Shang. 2023. PsyAttention: Psychological attention model for personality detection. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 3398–3411, Singapore. Association for Computational Linguistics.

Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. 2024. The better angels of machine personality: How personality relates to llm safety. arXiv preprint arXiv:2407.12344.

Xiaoyan Zhao, Yang Deng, Wenjie Wang, Hong Cheng, Rui Zhang, See-Kiong Ng, Tat-Seng Chua, and 1 others. 2025. Exploring the impact of personality traits on conversational recommender systems: A simulation with large language models. arXiv preprint arXiv:2504.12313.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. Large language models are not robust multiple choice selectors. In The Twelfth International Conference on Learning Representations.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. Lmlpa: Language model linguistic personality assessment. Computational Linguistics, pages 1–41.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2024. Personality alignment of large language models. arXiv preprint arXiv:2408.11779.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. 2024. Craft your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. arXiv preprint arXiv:2409.02098.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. 2024. Can llm" self-report"?: Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots. arXiv preprint arXiv:2412.00207.