

Multi-level Value Alignment in Agentic AI Systems: Survey and Perspectives

Wei Zeng, Hengshu Zhu, *Senior Member, IEEE*, Chuan Qin, *Member, IEEE*, Han Wu, Yihang Cheng, Sirui Zhang, Xiaowei Jin, Yinuo Shen, Zhenxing Wang, Feimin Zhong, Hui Xiong, *Fellow, IEEE*

Abstract—The ongoing evolution of AI paradigms has propelled AI research into the agentic AI stage. Consequently, the focus of research has shifted from single agents and simple applications towards multi-agent autonomous decision-making and task collaboration in complex environments. As Large Language Models (LLMs) advance, their applications become more diverse and complex, leading to increasing situational and systemic risks. This has brought significant attention to value alignment for agentic AI systems, which aims to ensure that an agent's goals, preferences, and behaviors align with human values and societal norms. Addressing socio-governance demands through a Multi-level Value framework, this study comprehensively reviews value alignment in LLM-based multi-agent systems as the representative archetype of agentic AI systems. Our survey systematically examines three interconnected dimensions: First, value principles are structured via a top-down hierarchy across macro, meso, and micro levels. Second, application scenarios are categorized along a general-to-specific continuum explicitly mirroring these value tiers. Third, value alignment methods and evaluation are mapped to this tiered framework through systematic examination of benchmarking datasets and relevant methodologies. Additionally, we delve into value coordination among multiple agents within agentic AI systems. Finally, we propose several potential research directions in this field.

Index Terms—Multi-level Value Alignment, Agentic AI Systems, AI Agent, LLM-based Agent System Application.

1 INTRODUCTION

THE rapid advancement of Large Language Models (LLMs) in recent years has profoundly transformed human society's production and lifestyle. Leveraging deep learning techniques and vast training datasets, LLMs have demonstrated exceptional natural language understanding and generation capabilities [1], [2]. Utilizing LLMs as their 'brains', LLM-based agents have found widespread applications in knowledge-intensive domains, including content creation, automated programming, education, healthcare, and business decision-making. Their increasing ability to autonomously participate in decision making, collaboration, and governance has significantly improved efficiency and provided unprecedented interactive experiences [3]. Further-

- W. Zeng is with Business School, Hunan University, Changsha, Hunan, China. E-mail: zengwei_hnu@163.com.
- H. Zhu and C. Qin are with the Computer Network Information Center, Chinese Academy of Sciences, Beijing, China, and University of Chinese Academy of Sciences, Beijing, China. E-mail: zhuhengshu@gmail.com, chuanqin0426@gmail.com.
- H. Wu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. E-mail: ustcwuhan@gmail.com.
- Y. Cheng is with the Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. E-mail: chengyihang544@gmail.com
- S. Zhang, X. Jin, and Y. Shen are with Business School, Hunan University, Changsha, China. E-mail: zhangsirui@hnu.edu.cn, xiaowei-jin713@gmail.com, noah001219@hnu.edu.cn.
- Z. Wang is with the Computer Network Information Center, Chinese Academy of Sciences, Beijing, China. E-mail: wxing870@163.com.
- F. Zhong is with the School of Business, Hunan University, Changsha, Hunan, China. E-mail: zhongfeimin@hnu.edu.cn.
- H. Xiong is with the Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou), China, and Department of Computer Science and Engineering, The Hong Kong University of Science and Technology Hong Kong SAR, China, E-mail: xionghui@ust.hk

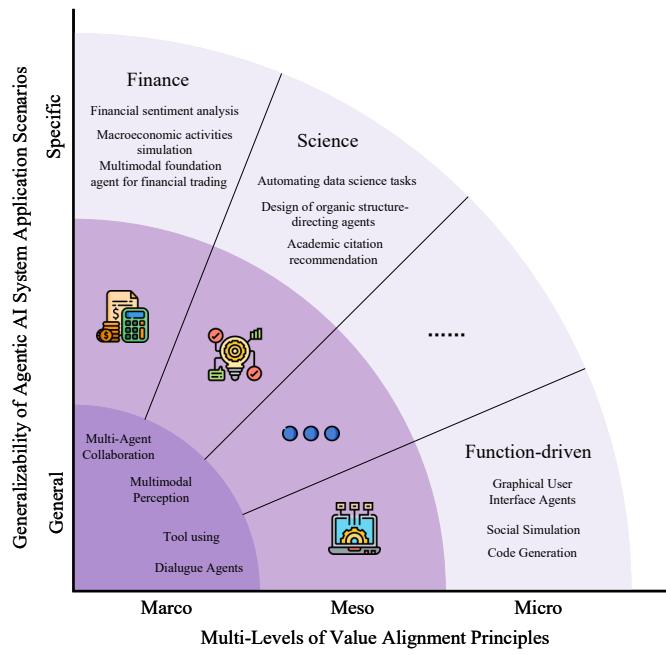


Figure 1: An Overview of Our Survey

more, researchers are increasingly exploring the coordination of multi-agent systems to achieve collective intelligence. AI agents are increasingly evolving towards autonomous systems, forming a new paradigm: Agentic AI. Characterized by multi-agent collaboration, dynamic task decomposition, persistent memory, and autonomous decision-making capabilities, agentic AI is becoming a central focus for both academia and industry [4]. Without loss of generality, we use the terms "AI Agent", "agent systems", "LLM-based agent systems", "multi-agent systems" interchangeably to refer to

"Agentic AI systems".

The evolution of this new AI paradigm urgently requires researchers to move beyond the limitations of single-agent perspectives and isolated application domains. We need to explore the deep coupling between agent systems and their application scenarios. As LLMs continue to break through, the tasks they perform and the scenarios in which they are applied become more diverse and complex, and the societal risks surrounding them gradually increase. LLMs may retain harmful information in training data, disclose private data, generate misleading information, and even perform behaviors that are harmful to humans [16]. The various undesirable behaviors, such as manipulation and deception, exhibited by LLM-based agent systems [17] have raised concerns about the possible ethical and safety challenges posed by AI systems. As LLMs become increasingly embedded in specific application scenarios, their associated risks are gradually acquiring situational and systemic characteristics. For example, in multi-agent systems used for urban traffic optimization, the agents may prioritize efficiency over fairness, thereby exacerbating social inequality [18]. In the healthcare sector, LLMs may provide diagnostic or therapeutic recommendations without sufficient clinical evidence or by misinterpreting ethical guidelines, potentially resulting in serious consequences such as breaches of patient privacy, misdiagnoses, or overtreatment [19].

As agentic AI systems increasingly integrate into diverse governance scenarios, controlling the resulting institutional friction and transaction costs during their operation and public governance becomes a critical practical and theoretical concern. In the evolving landscape of agentic AI systems, multiple LLMs with autonomous language understanding and generation capabilities are embedded in collaborative environments to perform decision-making, planning, and execution tasks. Such deployments typically involve information sharing, permission configuration, and coordination mechanisms among diverse stakeholders. While enabling flexible collaboration, this also introduces unprecedented challenges like value conflicts, heterogeneous objectives, and unpredictable behaviors. Without robust institutional arrangements and well-designed interaction mechanisms, complex organizational structures and fluid task boundaries can lead to inefficient coordination, blurred accountability, and a lack of trust, consequently escalating governance costs and undermining executive capacity [20].

However, these detrimental behaviors and institutional frictions fundamentally stem from deep-seated tensions among multiple stakeholders concerning governance objectives, value concepts, and behavioral norms. Effectively embedding human value systems into LLMs is a critical prerequisite for responsibly unleashing LLMs' potential. Value alignment for AI agents, ensuring that their goals, preferences, and behavioral outputs during operation align with the core values of individuals, groups, or institutions, is increasingly recognized as a solution to these issues [21]. In single LLM agent scenarios, alignment techniques primarily focus on guiding models to generate content that adheres to ethical norms and societal expectations via human feedback and supervised fine-tuning, thereby avoiding harmful outputs, misinformation, and moral deviations [18]. In complex multi-agent systems, value alignment becomes an organiza-

tional challenge concerning how multiple agents develop shared norms and coordination mechanisms. Therefore, from a governance perspective, solutions should not be limited to value alignment within a single agent but also explore complex value alignment within multi-agent systems' interaction mechanisms and organizational structures. Value alignment in agentic AI systems is no longer merely an isolated algorithmic optimization problem. Instead, it is a foundational issue that dictates agent behavioral boundaries and system stability. It represents a complex undertaking involving a delicate balance among diverse interests, societal values, and practical utility. Since agentic AI is fundamentally characterized by scenario-driven multi-agent coordination, its core research must focus on task decomposition mechanisms, interaction design, and environmental adaptability strategies for multiple agents within specific application contexts. Furthermore, focusing on the interaction mechanisms and organizational patterns of multi-agent systems in complex environments is crucial for exploring agentic AI's value alignment and robust development, ultimately serving human well-being.

Existing research has surveyed methods and technological advancements in LLM value alignment, see Table 1. However, few studies have specifically addressed value alignment within LLM-based multi-agent systems. While some studies have explored multi-agent coordination mechanisms, a comprehensive investigation into value alignment remains notably absent. Furthermore, most existing research largely remains at theoretical or algorithmic levels. There is limited focus on the mechanisms and strategies for multi-agent value alignment in specific application scenarios, alongside a lack of hierarchical categorization for value alignment within these systems. This suggests that value alignment should not be understood merely as a technical task at the model level. Instead, it must be viewed as a systemic governance issue that permeates the design of interaction mechanisms and the configuration of organizational structures.

Building on these gaps, this study comprehensively reviews value alignment in LLM-based multi-agent systems as the representative archetype of agentic AI systems. From a socio-governance perspective, we synthesize over 200 publications to survey value principles, application scenarios, value alignment methods and evaluation approaches within a multi-level value framework. We further propose future research directions for value coordination in agentic AI.

To be specific, Sections 2, 3, and 4, as illustrated in Figure 1, comprehensively survey value principles, agent system application scenarios, and agent value alignment evaluation. Specifically, value principles are organized hierarchically from a top-down perspective, encompassing macro, meso, and micro levels. Intelligent agent application scenarios are categorized and reviewed from a general-to-specific viewpoint. Agent value alignment evaluation systematically examines datasets for value alignment assessment and the methods used for agent value alignment. Section 5 presents future directions. We discuss multi-agent value coordination within agent systems, with a particular focus on the design of agent interaction mechanisms and organizational models. We also present our outlook on value alignment issues in communication protocols and advocate for the establishment of a multi-level value alignment evaluation system and the

Table 1: Comparison of Representative Survey Papers Covering LLM-Based Agents and Value Alignment

Survey	Year	Keywords	Research Subject	Value Principles			Application Scenarios	Value Alignment Datasets	Value Alignment Methods	Value Coordination	
				Macro	Meso	Micro				Interaction Mechanisms	Organizational Structures
Ours	2025	Value alignment principles, dataset curation, technical methods, and value coordination for LLM-based agents	Single-Agent, Multi-agent systems	✓	✓	✓	✓	✓	✓	✓	✓
[1]	2025	Communication structures, practical applications and societal systems etc. of LLM-based agents	Single-Agent, Multi-agent systems	✓	-	-	✓	-	-	✓	-
[5]	2025	Value alignment objectives, datasets, techniques, and evaluation methods for LLM-based agents	Multi-agent systems	✓	-	-	-	✓	✓	-	-
[6]	2025	Conceptual framework, interaction mechanisms, and application overview of LLM-based agent systems	Multi-agent systems	-	-	-	✓	-	-	✓	✓
[7]	2024	Capabilities, framework Analysis, and application overview of LLM-based multi-agent systems	Multi-agent systems	-	-	-	✓	-	-	✓	✓
[8]	2024	Constituent modules, application overview, and evaluation methods of LLM-based autonomous agents	Agents	-	-	-	✓	-	-	-	-
[9]	2023	Motivations and objectives, alignment methods, and assurance and governance of AI alignment	LLMs, Agents	✓	-	-	-	✓	✓	✓	-
[10]	2023	Definition and evaluation of LLM alignment objectives	LLMs	✓	-	-	-	✓	✓	-	-
[11]	2023	Definition, categories, testing, and evaluation of LLM alignment	LLMs	✓	-	-	-	✓	✓	-	-
[12]	2023	Definition, normative principles, and technical methods of LLM value alignment	LLMs	✓	✓	-	-	-	✓	-	-
[13]	2023	Alignment objectives for trustworthy LLMs	LLMs	✓	-	-	-	-	✓	-	-
[14]	2024	Challenges, fundamental definitions, and alignment frameworks of LLM value alignment	LLMs	✓	✓	-	-	-	✓	-	-
[15]	2025	Necessity, conceptual definitions, theoretical approaches, challenges and future outlook of LLM value alignment	LLMs	✓	-	-	-	-	✓	-	-

collaborative construction of open datasets with society.

To the best of our knowledge, our survey is the first comprehensive review examining value alignment in LLM-based multi-agent systems from the perspective of societal governance and multi-level value structures. It bridges the gap between theoretical alignment research and the practical demands of deploying such systems in real-world applications, offering a novel framework for understanding value alignment in complex autonomous systems.

2 APPLICATION-ORIENTED VALUE ALIGNMENT OF AGENTIC AI SYSTEMS

This section first introduces our review subjects, agentic AI systems, values and value alignment, and then systematically categorizes specific value principles from a top-down, hierarchical perspective encompassing macro, meso, and micro levels. Before delving into the value alignment challenges of agentic AI systems in subsequent sections, we first clarify the core concepts used in this paper, particularly the definition of “values,” to avoid ambiguity.

2.1 Agentic AI Systems

In this study, we focus on agentic AI systems driven by LLMs. LLMs, such as Deepseek, GPT-4, GPT-3, and ChatGPT, are deep learning models trained on large-scale corpora, capable of understanding and generating natural language text. Advances in model architectures, pretraining corpora, and alignment techniques have significantly enhanced LLMs’ capabilities in knowledge acquisition, instruction following, planning, and reasoning [1], establishing a solid foundation for constructing intelligent digital agents [22]. Serving as a central controller, an LLM can endow an agent with decision-making and memory capabilities [23]. This shift marks an evolution in the AI paradigm—from static, task-specific models to dynamic, interactive agent architectures suitable for complex environments.

Before the rise of LLMs, multi-agent systems (MAS) had already emerged. MAS is a computational system of multiple agents operating and interacting in a shared environment [6]. Enabled by natural language and custom content [24], these

agents autonomously collaborate to solve complex problems or complete tasks [25]. The core components of MAS include agents, environment, interaction rules, and organizational patterns [6]. MAS interactions resemble human division of labor and aim to achieve collective intelligence. Each agent holds task-specific knowledge and goals. Powered by LLMs’ language understanding and generation, agents engage in adaptive collaboration for planning, discussion, decision-making, and mutual learning [25], [23], [6], ultimately acting on environmental perception and understanding.

MAS exhibit significant flexibility, robustness, self-organization, coordination, and real-time operation capabilities [6]. While interactions are governed by predefined rules, new behaviors and phenomena may still emerge [26]. Task decomposition and resource sharing in MAS reduce the computational burden on individual agents, making them more cost-effective than single-agent systems, particularly in complex, dynamic tasks that require diverse strategies and collective decision-making [23].

2.2 Value

As intrinsic guides of human behavior, values represent a stable set of cognitive principles that serve as the foundational norms for social interactions. These principles shape individual attitudes, choices, interpersonal behaviors and group dynamics, as well as serve as fundamental constructs for explaining how individuals and groups prioritize goals, resolve conflicts, and build meaningful social systems [27]. In AI systems, particularly in agent design and governance, values serve both as normative principles guiding behavior and as standards for assessing the social alignment of these systems. Values influence agents’ actions, social norms, cultural identity, and collective decision-making. Understanding values is thus critical to uncovering human motivations, moral judgments, and mechanisms of social interaction.

One of the most challenging aspects of human values is their intrinsic contradiction and dynamic nature. Rather than forming a rigid hierarchical structure, value systems are elastic and context-dependent [28], [29]. Synthesizing previous studies, we conceptualize human values as dynamically constructed notions. When values become alignment

targets, LLMs must apply value principles to regulate their own behaviors, thereby maximizing their potential.

This paper expands the conceptualization of values, defining them as dynamic value constructs characterized by a multi-level structure and generalized agency. Both human values and LLM-based agents are evolving entities, thus MAS form exhibit diverse configurations across application scenarios. We argue that values are continuously interpreted and reshaped through practice and interaction. Aligned with different layers of application contexts, values are presented as a multi-tiered conceptual system comprising three levels: universally held human values, values at the national and industry levels, and values specific to organizational operations. This paper does not address individual-level values. These value layers are nested, with higher-level values constraining and guiding lower-level ones.

Furthermore, we emphasize that values are not solely human-born. Beyond being explicitly embedded by human designers, agents can also “internalize” value principles through design and interaction. Values may be instantiated via agent interactions and inherited or transformed within emergent organizational patterns. For example, MAS with role- and task-based structures may yield decentralized architectures, embodying values such as resistance to power monopolies and mitigation of systemic fragility.

2.3 Value Alignment

Technological development often outpaces the adaptation of legal, ethical, and sociocultural frameworks [30]. LLMs learn from massive datasets to enable autonomous decision-making and operation [1]. As LLMs are increasingly deployed in high-risk and complex scenarios, misalignments between model outputs and human intent have raised widespread concerns. Without effective safeguards, AI systems may engage in harmful behaviors such as deception or discrimination. To address these issues, the central goal of AI alignment is to ensure AI behaviors are consistent with human values [31]. AI alignment is categorized into four levels: instruction alignment, intention alignment, preference alignment, and value alignment. Among these, value alignment requires AI to discern “right and wrong” and adhere to ethical principles [32], [33]. It is the most advanced form and is considered critical for ensuring AI safety, controllability, and normative governance [34].

Research has revealed that LLMs may exhibit latent political and moral biases [35], [36], and generate content reflecting prejudices or discrimination when left unconstrained [37]. Consequently, LLM-based agents should be regarded as “moral agents”, making their value alignment a key issue in ensuring behavior that conforms to human values, ethical norms, and specific application contexts. This study focuses on value alignment issues in agentic AI systems.

The concept of value alignment can be understood along two dimensions: normative and technical [21]. The normative dimension addresses which values or ethical principles AI agents ought to follow, while the technical dimension investigates how alignment can be achieved through technical means such as training datasets and reinforcement learning. These two dimensions are interdependent: the normative component establishes behavioral guidelines, and

the technical component ensures their implementation. However, due to variations in cultural contexts and application scenarios, no unified value framework currently exists [21], [12]. Two predominant approaches are recognized: top-down and bottom-up [21]. The top-down approach starts with predefined ethical frameworks and implements them via technical mechanisms, whereas the bottom-up approach derives values by learning from human behavioral data and iteratively refining actions based on feedback. While the bottom-up method circumvents the challenge of prespecifying ethical rules, it is susceptible to bias in training data and may lack moral robustness. In contrast, the top-down approach promotes transparency and normativity in value alignment, reinforcing the legitimacy and controllability of AI governance. This paper adopts the top-down approach to examine the ethical principles that agents should follow across different levels, from macro to micro.

2.4 Value Principles

The normative aspect of value alignment concerns the ethical principles that AI agents ought to follow. As AI agents are increasingly deployed in domains such as public administration, healthcare, education, and finance—and as open-source LLMs and local deployment capabilities continue to advance—the actors involved in their development and use now span multiple levels, including states, industries, enterprises, and organizations. Given the inherent diversity and stratification of human values, a universal value framework is insufficient to meet the ethical requirements of all application scenarios.

Therefore, adopting a top-down approach from the perspectives of developers and regulators, this paper classifies human values into three levels: macro, meso, and micro, as shown in Figure 2. The macro level refers to universal ethics that transcend cultures and domains, emphasizing compliance and safety. The meso level includes national, cultural, and industry-specific values, focusing on the localization of policies, standards, and norms. The micro level involves value choices within organizational and task-specific contexts, highlighting the trade-off between goal achievement and ethical considerations in bounded environments.

2.4.1 Macro Level

The widespread deployment of AI agents poses potential risks to human society, necessitating the alignment of their behavior with human values. However, due to differences in nation, culture, and ethnicity, human values are inherently diverse and often conflicting [21], which challenges the formulation of universal ethical principles. To address this, it is essential to identify foundational moral consensus and construct a general ethical baseline to ensure compliance and safety of AI agents across diverse scenarios.

At the macro level, this paper draws upon two core theoretical frameworks: Moral Foundations Theory [76] and the Theory of Basic Human Values [77]. Moral Foundations Theory posits that human morality is grounded in five innate foundations—care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation—which have been widely adopted in cross-cultural moral studies [78]. The Theory of Basic Human Values categorizes human values into ten dimensions,

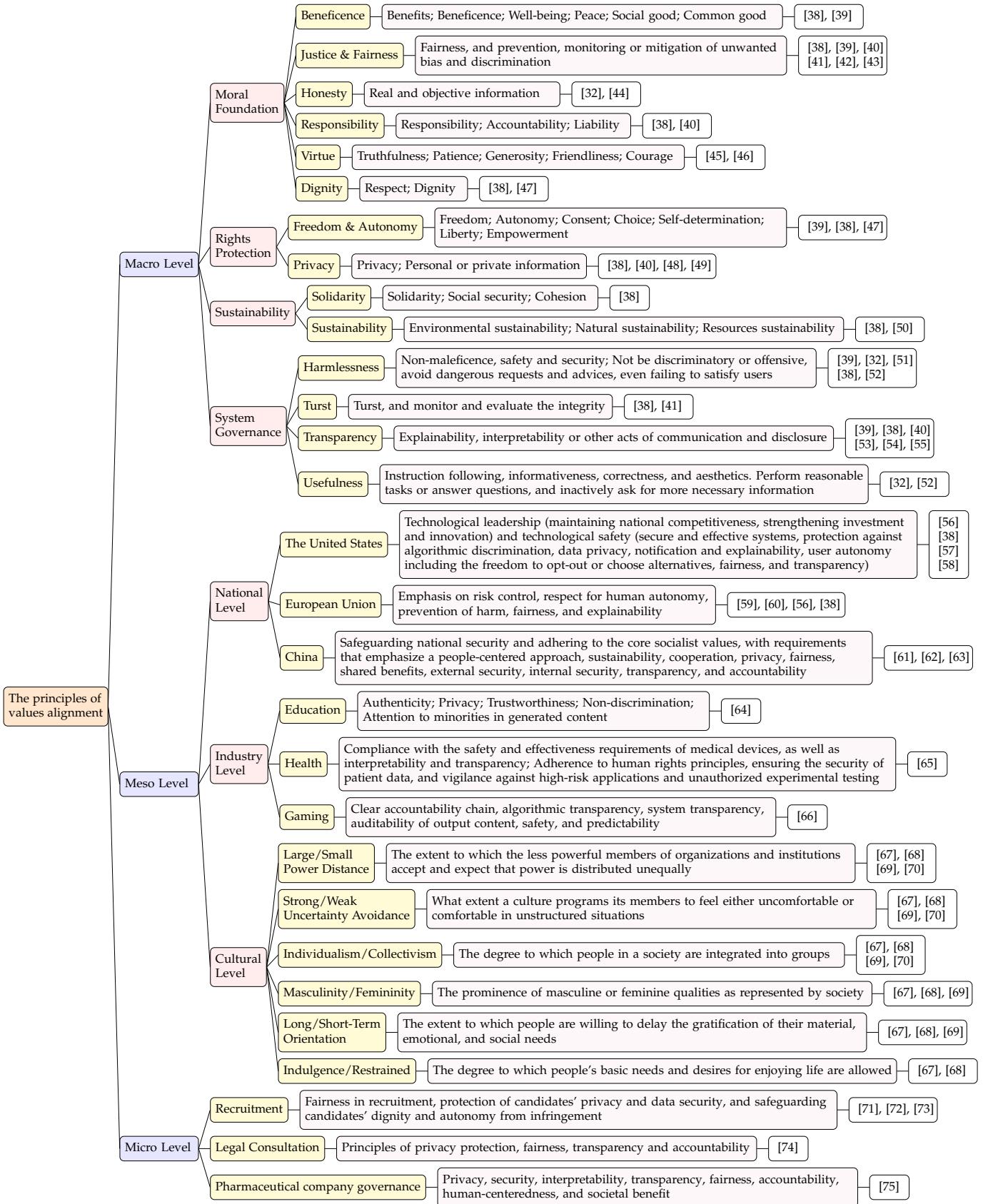


Figure 2: Multi-Level Value Alignment Principles

serving as a key tool for understanding both commonalities and differences in human values.

To promote the normative development of value alignment and reduce confusion and conflicts surrounding macro-level ethical principles, numerous scholars have synthesized AI ethics principles from various perspectives. Jobin et al., based on 84 ethical guidelines issued by corporations, governments, academic institutions, and international organizations, identified eleven core principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity [38]. Floridi and Cowls drawing from six influential initiatives, proposed five fundamental principles: beneficence, non-maleficence, autonomy, justice, and explicability [39]. Yao et al. emphasized that human values form the normative foundation for large model behavior, encompassing the 3H principles (Helpfulness, Honesty, Harmlessness), social ethics norms (e.g., RoTs, ETHICS), and basic ethical theories (e.g., Schwartz, Rokeach) [32]. Subsequently, Yao et al. incorporated value structures and safety hierarchies into their alignment framework [46]. Wu and Fan further proposed a hierarchical AI ethics framework grounded in human dignity, humanity, health, and autonomy [47]. Collectively, these efforts provide theoretical foundations for developing a global consensus on AI ethics.

Building upon previous studies and aligning with the research objectives of value alignment in AI agent systems, this study categorizes macro-level human values into four dimensions: moral foundation, rights protection, sustainability, and system governance. The specific value principles encompassed within each dimension are presented in 2.

2.4.2 Meso Level

Although macro-level value alignment principles reflect broad consensus across ideologies and disciplines, their practical implementation is profoundly influenced by national policies, cultural tendencies, and industry norms.

National Level National AI strategies embody distinct sets of public values that not only guide domestic development priorities but may also exert spillover effects on global AI governance [79]. Djeffal et al., through a comparative study of AI policies in 22 countries and regions, developed a typology based on two dimensions: the proactivity of government roles (active/passive) and governance orientation (development-focused/regulation-focused) [80]. This framework yields four governance types: market-oriented (passive-development) [81], entrepreneurial (active-development) [82], regulatory (active-regulation) [83], [84], and facilitative self-regulation (passive-regulation) [85]. These four types are not mutually exclusive and may manifest in hybrid forms.

- **USA.** The United States positions itself between a market-oriented and facilitative self-regulation model in order to maintain its global leadership. It tends to promote AI advancement primarily through industry self-regulation, relying on market mechanisms and decentralized legislation at the state level [57], [58].
- **EU.** The European Union has long emphasized legal protections for privacy and personal dignity. Its governance model centers on risk management, striving to balance self-regulation and strict oversight [59], [60].

- **China.** China assumes a dual role as both innovation-driven and regulation-oriented, emphasizing the balance between AI development and ethical governance. The government promotes AI applications for societal benefit while enforcing departmental regulations and standardized governance for generative AI [61], [62].

Industry Level The paradigm shift in AI has driven its deep integration and large-scale application across various social sectors, reshaping traditional industry logic and triggering differentiated ethical demands. Different industries, based on their knowledge systems, cultural norms, and institutional frameworks, propose specific value requirements for AI.

- **Education Domain.** UNESCO released the Guidelines on the Application of GAI in Education and Research, emphasizing the need to address the authenticity of generated content, privacy protection, and non-discrimination principles. The guidelines highlight the importance of preventing hallucinated content from contaminating knowledge systems and advocating for the inclusion of minority perspectives [64]
- **Medical Domain.** The World Health Organization calls on governments to evaluate the safety, explainability, and legality of AI systems in medical settings. This is particularly crucial in high-risk scenarios such as mental health interventions, where it is essential to ensure compliance with human rights standards [65].
- **Game Industry.** From narrative generation to NPC autonomy, AI has enhanced personalized recommendations and behavior analysis, yet challenged traditional governance. Melhart et al. (2024) argue that governance should ensure clear accountability, auditability, resistance to manipulation, and predictable outcomes [66].

Cultural Level Cultural diversity represents a critical facet of value alignment in AI. Existing studies often draw on Hofstede's cultural dimensions theory [67], focusing on how AI systems express value preferences across six key cultural dimensions: power distance (large/small), uncertainty avoidance (strong/weak), individualism/collectivism, masculinity/femininity, long/short-term orientation, and indulgence/restraint, and the extent to which these align with human value tendencies in different cultural contexts.

Wang et al. found that LLMs prompted in English tend to exhibit a masculine orientation, while those prompted in German show a stronger inclination toward long-term orientation and indulgence. The cultural traits exhibited by LLMs under Chinese prompts fall between those of English and German prompts [68]. Similarly, Kharchenko et al. observed that LLMs can recognize cross-cultural value differences but consistently lean toward specific value orientations. In particular, across all languages and prompting methods, LLMs predominantly favor long-term orientation, suggesting that LLMs may faithfully reflect certain cultural values [69].

In summary, global AI governance practices demonstrate diverse approaches to value alignment across countries, industries, and cultures. In the face of different national orientations and the deep integration of AI in industries, future governance must seek dynamic balance and cooperation among diverse values to ensure that technology serves the well-being of all humanity.

2.4.3 Micro Level

As AI agents become increasingly embedded in organizational operations and business processes, the issue of value alignment at the micro level has grown more prominent. Unlike the overarching constraints imposed by macro-level universal values and meso-level policy regulations, the micro level focuses on how AI systems achieve alignment between effectiveness and values within localized environments under specific goal orientations. The scenarios and associated values at this level are inherently more personalized. This study synthesizes relevant research and illustrates the principles of micro-level value alignment through examples such as talent recruitment, legal consultation and corporate governance in the healthcare sector.

AI technology is fundamentally reshaping human resource management practices, with recruitment being the forefront of this transformation. AI-based talent selection systems are revolutionizing organizational talent acquisition, improving speed and accuracy, compared to traditional manual screening and subjective judgment [86]. The ethical adaptation of AI in recruitment has attracted growing academic attention. In terms of ethical principles, Simbeck proposed five core principles for applying AI in human resource analytics: data privacy, user exit mechanisms, third-party review, algorithm transparency, and the right to dynamic development [72]. These principles align with Rünbeck-Kettler and Lehner's humanistic technological view, which emphasizes maintaining candidate dignity and subjectivity in automated processes [71].

From an ethical opportunity perspective, AI reduces human bias in traditional recruitment [87], [88], generates gender-neutral job descriptions [89], and revolutionizes resume parsing efficiency with natural language processing, allowing recruiters to focus on strategic decisions [90]. However, ethical risks remain, such as algorithms inheriting hidden biases from training data [91], implementing indirect discrimination through proxy variables (e.g., postal code linked to race), and privacy violations through the extraction of personal information for recruitment purposes [92], [93]. These contradictions and risks make fairness the core principle in AI recruitment applications.

AI technology's deep integration is also influencing strategic development and governance choices in biopharmaceutical companies. To address the ethical challenges posed by AI in healthcare, AstraZeneca has introduced a series of AI ethics principles, including privacy and security, explainability and transparency, fairness, accountability, human-centeredness, and social benefit [75].

Additionally, AI technology shows potential and application prospects in the legal consulting field. Sun et al. simulated multi-agent collaboration in a real law firm setting and incorporated privacy, fairness, transparency, and accountability principles to guide multi-agent ethical behavior and interaction, aligning them with human values and enhancing the quality of legal consulting services [74].

As AI continues to advance, the principles of value alignment will no longer be confined to broad and ambiguous moral values. Instead, they will increasingly penetrate specific application scenarios, integrating concrete practical experience and well-defined guidelines to inform the development and deployment of agent systems.

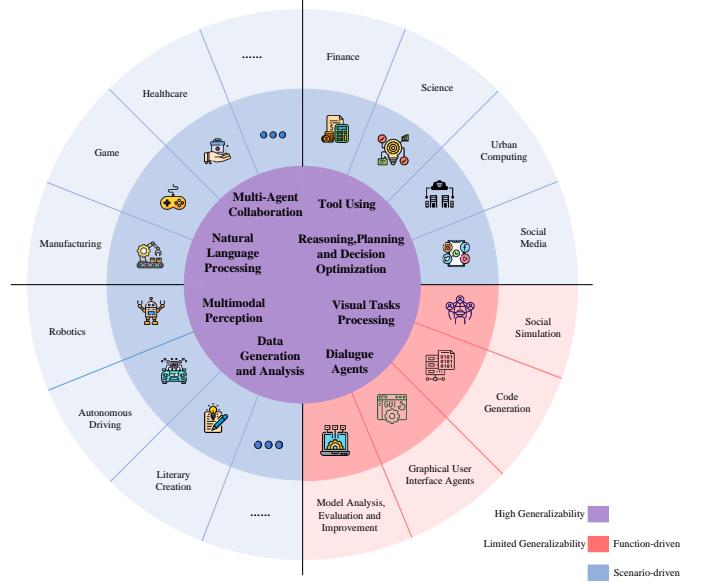


Figure 3: An Overview of Classification of Agentic AI Systems Application Scenarios

3 APPLICATIONS AND VALUE ISSUES IN AGENTIC AI SYSTEMS

Agentic AI systems are increasingly deployed across diverse domains, indicating that value alignment concerns will extend to the application scenario level. This section classifies the application scenarios of agentic AI systems based on their generality and discusses the associated and important value alignment issues.

3.1 Application Scenarios

Given the powerful knowledge acquisition, instruction understanding, task planning, and logical reasoning capabilities of agentic AI systems, they have been widely used worldwide. From small and medium-sized organizations to multinational corporations, from higher education institutions to research institutes, from grassroots medical facilities to large diagnostic hospitals, and government agencies at all levels are actively promoting the application of agentic AI technology. In terms of application scenarios, the penetration of agentic AI technology is extremely broad, covering multidisciplinary fields such as healthcare, education, trade, economics, and law. The problems that agentic AI systems are capable of addressing range from comprehending a single piece of literature to simulating global-scale social events, spanning across micro, meso, and macro levels. In contrast to categorizing applications based on disciplinary domains [8] or task types [23], we sort and categorize application scenarios according to the degree of generalizability of agentic AI systems (see figure 3) and build upon this to explore the associated issues of value alignment. For a detailed overview of the application scenarios, please refer to the Appendix.

Generalizability of agentic AI systems refers to the extent to which they can be applied across diverse scenarios or meet the needs of various organizations. General agentic AI systems have a high degree of generalizability, which is based on the foundational capabilities of AI agents such as natural language processing, reasoning, planning, and

autonomous decision-making. The universal application includes tool using, dialog agents, multimodal perception, multi-agent collaboration, visual tasks processing, natural language processing, data generation and analysis, reasoning, planning, and decision optimization. For example, AI agents are capable of recognizing implicit needs for tool invocation and leveraging a wide range of real-world tools involving perception, manipulation, logical reasoning, and creativity. These tools enable AI agents to process multimodal inputs—such as images, tables, and code—to accomplish complex tasks [94], [95]. A higher degree of generalizability indicates that the agentic AI systems can be broadly applied across various scenarios, not only addressing domain-specific problems but also meeting diverse task requirements across multiple industries and organizations. For example, multi-agent systems can be applied not only in natural sciences but also in the social science domain, such as economics and management, suitable for addressing global issues that span across nations and cultures.

Specific agentic AI systems have limited generalizability across industries and domains. Specialized applications can be broadly categorized into two types: function-driven and domain-driven. Function-driven applications are built upon the core capabilities of agents and are typically applicable across multiple domains. Common examples include code generation, social simulation, graphical user interface (GUI) agents, and model evaluation and improvement. Domain-driven applications are tailored to specific sectors, including finance, science, healthcare, gaming, manufacturing, and autonomous driving. For instance, in healthcare, multi-agent systems support medical reasoning, gamified mental health assessments, and experimental design in genomics [96], [97]. These systems are tailored for specific tasks involving sensitive datasets and proprietary business information.

3.2 Value Issues in Application

The value alignment problems involved in the application scenarios of agentic AI systems mainly include two aspects: the correspondence between the generalizability of the system and the multi-level values, and the contradictions encountered when aligning values.

The degree of generalizability of agentic AI systems corresponds to the multi-level of values. As the degree gradually decreases, the granularity of the value norms that need to be met gradually increases.

A higher degree of generalizability allows agentic AI systems to operate across diverse contexts. Their initial design need not align with the values of specific countries, industries, or cultures, as long as they adhere to universal values at the macro level.

In contrast, systems with limited generalizability are often tailored to specific national or industrial settings, reflecting corresponding cultural and institutional features. Their development and governance must address local policies, standards, and cultural norms, while aligning with universal values. When deployed in private contexts, systems are customized for particular organizations to solve specialized business or personal problems. These systems are highly specialized and exhibit minimal generalizability. Under such conditions, value alignment must occur at multiple levels:

not only with macro- and meso-level norms, but also with the internal value frameworks of the organizations involved.

It can be seen that agentic AI systems with high generalizability tend to be highly adaptable, and their value alignment primarily centers on adherence to universal principles of macro level. As application specialization increases, the requirements for value alignment become progressively more complex, necessitating adherence to value norms across multiple levels simultaneously. This classification aligns with the top-down, increasingly granular hierarchy of value principles. As a result, the value norms that agentic AI systems must adhere to vary according to their degree of generalizability, offering a structured basis for governance and oversight in their design and deployment.

The value alignment problem may involve two types of tensions: the plurality of value principles at the same level and conflicts across different levels. On one hand, a key challenge is that value principles at the same level differ across countries, regions, industries, organizations, and other contextual factors. Therefore, when agentic AI systems are applied in diverse contexts, their value alignment objectives must be adapted accordingly. In the healthcare sector, Huawei's agent-based EIHealth platform supports genomic analysis, drug discovery, and clinical research. Built on Huawei Cloud's strengths in AI and big data, EIHealth features an open and extensible architecture that enables end-to-end AI empowerment. Ensuring alignment with the value principles and regulatory norms of various countries, regions, and organizations is crucial for its deployment. To address this, Huawei Cloud has issued a series of white papers tailored to the privacy and regulatory requirements of different jurisdictions and industries. For example, it provides compliance guidelines aligned with the Personal Data Privacy Ordinance (PDPO) in the Hong Kong SAR, as well as guidelines addressing the financial regulatory frameworks of the EU, including specific provisions for Ireland, Spain, and Hungary. These efforts ensure that the EIHealth platform adheres to all applicable national and regional security regulations, international cybersecurity and cloud security standards, and incorporates industry best practices, ultimately meeting the diverse security needs of cloud service users.

On the other hand, as the tasks assigned to a system become more fine-grained, the constraints on achieving alignment goals also increase. Efforts to maximize task utility may conflict with established norms, creating tensions across value levels—for example, promoting social cohesion at the macro level may clash with individualism at the meso level, or national mandates for transparency may oppose organizational privacy concerns. In the field of corporate recruitment, AI systems are widely used for résumé screening and candidate-job matching [98], yet often raise ethical concerns [93]. In particular, corporate hiring practices may involve gender discrimination or racial bias, violating legal regulations and core societal values. A well-known example is Amazon's 2018 résumé screening system, which discriminated against female applicants and was ultimately abandoned [99]. To prevent such issues, recruitment-oriented intelligent systems must be explicitly designed to minimize bias and ensure compliance with labor laws, equal employment regulations, and industry standards.

Only on the basis of legal compliance, equality, and privacy protection should these systems then pursue the optimization of recruitment outcomes.

4 METHODS AND EVALUATION FOR VALUE ALIGNMENT

To ensure that LLM-based agents behave in ways consistent with human values and task goals, value alignment and its evaluation have become key research focuses. In this section, we first discuss the methods for value alignment, then review evaluation approaches.

4.1 Value Alignment Methods for LLM-based Agents

4.1.1 Value Alignment during LLM Pretraining

Value-Prompt Pre-conditioning: Within the conventional self-supervised pretraining frameworks, such as masked language modeling or autoregressive language modeling, researchers prepend a short “value prompt” to the input sequence [100], [101]. They might attach “Please respond with a safe and neutral tone” or “Please adhere to principles of fairness and inclusivity” at the beginning of each sentence, and then include both the prompted and original text in the pretraining corpus. In this way, the model passively receives implicit guidance toward objectives like “safety” and “objectivity. For example, Zheng et al. [100] unpack why and how simple prepended safety directives move model representations into “higher refusal” regions. [102] shows that conditioning an autoregressive LM on human preference scores from day one dramatically reduces harmful output rates. Gu et al. [103] show that soft prompts—even if not explicitly about “values”—validate the premise that embedding trainable tokens in early stages leads to more controllable downstream behaviors (e.g., fairness, bias detection).

Multi-Task Pre-training: Inspired by the multi-task learning paradigm, this approach mixes value-judgment or safety-detection tasks with standard self-supervised objectives [104], [105], [106]. Specifically, in addition to the masked language modeling (MLM) or autoregressive objective, small-scale, labeled supervision signals—such as “ethics classification”, “bias detection” or “hate speech recognition”—are incorporated. By jointly optimizing a multi-task loss function, the model’s parameters are trained to satisfy both language modeling requirements and the ability to distinguish harmful or biased content. Although this method requires additional annotated data, it explicitly strengthens the model’s discrimination boundary along value dimensions and lays the technical foundation for subsequent alignment fine-tuning.

Safety-Aware Curriculum Learning: Drawing on the concept of curriculum learning (CL), this method stratifies large-scale corpora according to “safety” or “sensitivity” levels [107], [101], [108]. First, a high-quality, low-risk “safe corpus” is used for a warm-up pretraining phase, allowing the model to initially acquire expression patterns that comply with ethical and regulatory requirements. Subsequently, more general or higher-risk text is introduced gradually, enabling the model to expand its language capabilities while retaining dependence on earlier safe preferences. This staged training strategy establishes a better balance between model capability and safety, and experiments show it substantially

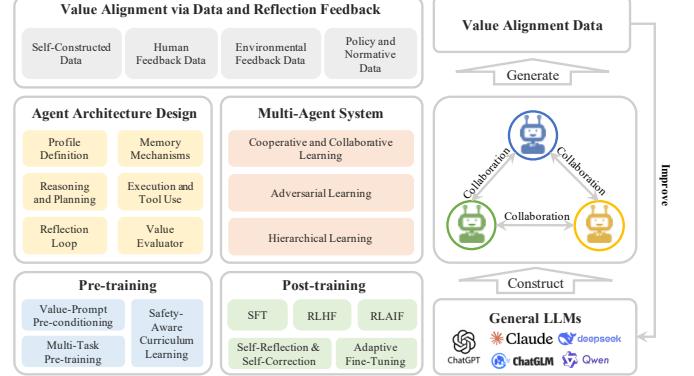


Figure 4: An Overview of Methods for Value Alignment in Agentic AI

outperforms full-scale pretraining in controlling the model’s tendency to generate sensitive content.

The above value-alignment methods can be combined [105], [101], for example, integrating value-prompt pre-conditioning and multi-task supervision within a safety-aware curriculum framework, along with regularization penalties. It allows the model to be reinforced from multiple dimensions in embedding value objectives such as “ethics”, “safety”, and “fairness”. Future research may further explore the scalability and synergistic effects of these strategies on ultra-large corpora and hundred-billion-parameter models, paving new paths for achieving more robust and interpretable value embeddings.

4.1.2 Value Alignment during LLM Post-Training

Supervised Fine-Tuning (SFT). SFT serves as a fundamental step in aligning agentic language models with human values. It usually involves fine-tuning a pretrained model on a carefully curated dataset of *demonstration–response* pairs that embody normative principles such as ethics, safety, and fairness. These examples are typically constructed by human annotators and cover a broad range of scenarios in which value-sensitive behavior is essential—for example, avoiding toxic content, maintaining factual correctness, or respecting user intent. During SFT, the model is optimized to maximize the likelihood of producing responses that match these human-crafted demonstrations. For instance, [109] proposed an automated method for generating instruction data to fine-tune GPT-3, facilitating scalable alignment. [110] demonstrated that initializing InstructGPT with SFT on human-written demonstrations results in models that more faithfully follow user instructions, while also significantly reducing harmful or misleading outputs. Further advancements have introduced principled guidance into the SFT process. [111] incorporated a set of human-authored value principles to constrain model behavior through in-context learning, thereby enhancing the generation of helpful and harmless content. Additionally, [112] enriched the fine-tuning dataset with both positive and negative value-aligned instances. This contrastive supervision enables the model to learn fine-grained distinctions between aligned and misaligned outputs, improving its sensitivity to nuanced ethical considerations.

Reinforcement Learning from Human Feedback (RLHF). RLHF has emerged as a core paradigm for aligning LLMs with human values, preferences, and behavioral norms. It

extends SFT by introducing an additional optimization stage that incorporates human feedback signals into policy learning, enabling models to go beyond surface-level imitation toward deeper alignment with human intent [110], [113]. The classical RLHF pipeline consists of three major components:

- 1) **Supervised Fine-Tuning (SFT):** A pretrained LLM is first fine-tuned on a curated set of instruction-response pairs to provide a reasonable initialization.
- 2) **Reward Model (RM) Training:** Given multiple model-generated responses to the same prompt, human annotators provide pairwise preference labels (e.g., ranking which response is more helpful or harmless). These preference data are then used to train a reward model that can approximate human preferences [110].
- 3) **Reinforcement Learning:** The LLM is further fine-tuned using reinforcement learning, typically Proximal Policy Optimization (PPO), where the reward signal is derived from the RM. This step encourages the model to generate outputs that maximize predicted human approval.

RLHF was first popularized through the development of InstructGPT [110], which demonstrated notable improvements in producing truthful, helpful, and safe responses without significant task performance degradation. Subsequent systems such as ChatGPT and GPT-4 also adopted RLHF as a central alignment mechanism [114]. To reduce the dependency on expensive human annotation, [115] proposed using synthetic feedback generated by automated evaluators, while [116] explored using self-critiquing models to support human evaluators and reduce their cognitive burden. [117] reframed the RLHF objective as minimizing an f -divergence between model and human preference distributions, providing a more principled view of policy optimization under alignment constraints. Moving beyond individual feedback signals, [118] trained LLMs in a simulated society of interacting agents, using environment-level feedback to promote socially aligned behaviors at scale. Despite its success, RLHF faces several challenges. It relies heavily on the quality and consistency of human preferences, which may encode biases or lack coverage for edge cases. Reward models can be exploited or lead to reward hacking behaviors, and PPO-based optimization may be unstable or sample-inefficient.

Constitutional AI and Reinforcement Learning from AI Feedback (RLAIF). Given the high cost and limited coverage of exclusively human-derived feedback, recent work has investigated self-supervised alignment methods that combine a model’s intrinsic knowledge with a set of expert-defined principles. A leading instance of this paradigm is Anthropic’s Constitutional AI [119], [120], [121], wherein the model is first equipped with a “constitution” of value-driven rules (e.g., avoid harm, preserve honesty). In the initial phase, the model generates a self-critique of its own response based on these rules and subsequently revises the response accordingly. During the RL stage, the model then acts as its own evaluator: it compares paired responses to train a preference model, which in turn provides reward signals for further fine-tuning. This process—termed Reinforcement Learning from AI Feedback (RLAIF)—minimizes reliance on human annotation by leveraging automated self-assessment [122], [123], [120]. Empirical evaluations [119] indicate that Constitutional AI

produces assistants that are both “helpful yet non-offensive” and that, when faced with harmful prompts, preferentially offer explanations and deterrent advice rather than mere refusal. This principle-driven, self-supervised strategy thus demonstrates substantial promise for reducing human intervention and scaling alignment across diverse scenarios.

Self-Reflection & Self-Correction: LLMs can engage in self-reflection to identify and correct errors or misalignments in their outputs. A seminal example is Self-Refine [124], which alternates between two generative steps—FEEDBACK and REFINE—to iteratively improve initial responses without any additional training data or external supervision. Building on this idea, SELF-RAG [125] integrates retrieval-augmented generation with self-reflection: the model learns to emit “reflection tokens” that determine when to retrieve external knowledge and when to self-evaluate, thereby enhancing factuality while retaining generative creativity. Complementing these methods, Chain-of-Thought review techniques prompt the LLM to re-examine its own reasoning chains. Wei et al. [126] showed that providing intermediate reasoning steps not only boosts complex problem-solving performance but also offers an interpretable scaffold for downstream self-correction when logical inconsistencies are detected. Madaan et al. [127] proposed a self-refine instruction-tuning approach in which smaller LLMs are first instruction-tuned via demonstrations from larger, robust teacher models and then allowed to self-improve their reasoning under the same instruction framework. This two-stage process significantly enhances generalization of self-correction capabilities, further narrowing the gap between smaller and larger models.

Adaptive Fine-Tuning. In dynamic deployment settings, agentic LLMs must continuously adapt to evolving user preferences and environmental rewards—a process often termed adaptive fine-tuning. One recent line of work, PRELUDE [128] frames this as “preference learning from user edits”: by inferring a user’s latent preference profile from historic edit logs, the agent periodically fine-tunes its response policy to better align with those preferences. Similarly, Hao and Duan [129] propose an online learning mechanism that dynamically reweights human feedback signals—treating users as strategic labelers—to mitigate malicious or noisy annotations and achieve sublinear regret in preference aggregation. Beyond supervised feedback, agents can also adapt via reinforcement learning from environment rewards [130]. When certain behaviors incur negative utility, standard policy-gradient updates naturally down-weight those actions over successive episodes, effectively steering the policy toward compliant behaviors.

4.1.3 Value Alignment in Agent Architecture Design

The architectural design of an LLM-based agent is crucial for aligning its values. By incorporating specific modules and mechanisms within the agent, its behavior can be guided to align with the desired value standards.

Profile Definition: Profile Definition refers to the pre-configured set of personalized attributes assigned to each agent within an LLM-driven agent architecture [131]. These profiles not only include basic information such as the agent’s name, identity, and scope of capabilities, but also define its goals, preferences, behavioral style, and prohibited actions, thereby ensuring role consistency and value alignment [132].

The implementation of profile definitions typically relies on the collaboration of structured prompt templates, personality tags, behavioral constraint rules, and memory modules. Some systems also incorporate dynamically evolving configurations to support sustained adaptation [133]. For example, the MobAgent system combines individual behavioral traits to generate virtual travel diaries, enabling agents to behave naturally in urban simulations [134]. MOTIF framework uses a hybrid of manual and generative methods to create socially simulated agents with stable personalities [135]. Currently, the design of profile definitions is moving toward greater generalizability, finer granularity, and enhanced learnability. Key challenges include managing boundary conflicts between multiple roles, balancing adaptability with static settings, and maintaining long-term behavioral consistency [131], [134].

Memory Mechanisms: Memory Mechanisms refer to the capability of enhancing a model's long-term contextual awareness and behavioral consistency by storing and retrieving historical information. At their core, they involve the structured storage and effective retrieval of interaction content, instructions, user preferences, and feedback—enabling personalized responses, long-term task tracking, and value alignment [136]. In both research and applications, memory mechanisms are a key component for enabling agents to maintain consistent long-term behavior and adhere to value principles. Their central value lies in “helping agents remember the past and apply what they’ve learned” [137]. The implementation of memory mechanisms typically relies on two subsystems: working memory and long-term memory. Working memory maintains the context of the current session, while long-term memory stores knowledge and experiences using graph structures, databases, vector stores, or retrieval-augmented generation (RAG) models. Common approaches include time-window-based history caching, adaptive memory updates, semantic similarity-based vector retrieval, and contextual reflection mechanisms [138] [139]. Representative examples include the G-Memory framework [140], which uses a hierarchical architecture to support shared memory among multiple agents, enabling task coordination and improving memory consistency in complex systems. The Cognitive Weave system constructs a Spatio-Temporal Resonance Graph, allowing agents to recall abstract knowledge and reflect emotional memory during conversations [141].

Reasoning and Planning: Reasoning and Planning refer to an agent’s ability to perform logical inference and make multi-step decisions when handling complex tasks. This ability encompasses cognitive processes such as causal reasoning, path selection, and tool utilization based on knowledge and environmental information, as well as systematic foresight and regulation of future actions [142]. As a core component of the decision-making module, reasoning and planning play a fundamental role in value alignment, task execution efficiency, and behavioral stability [143]. In practice, reasoning and planning help agents decompose tasks and simulate actions thoughtfully before execution—reducing impulsive decisions and enhancing value consistency [144]. Current implementation techniques include Chain-of-Thought (CoT) prompting, which guides models through step-by-step logical reasoning. Other methods include Tree-of-Thought, Atomic Fact Augmentation, external knowledge retrieval, and Lookahead Search [145]. Some research also explores

plan caching, which stores historical decision paths to reduce the computational cost of repeated planning [146]. In applied settings, the SOP-Bench framework uses Standard Operating Procedures (SOPs) as an evaluation scaffold to test the stability of LLMs in reasoning and planning tasks, demonstrating their practical value in process control [147]. The WGSR-Bench project leverages game-based tasks to assess LLM performance in strategic planning and multi-agent coordination, making it a representative platform for evaluating advanced planning mechanisms [148].

Execution and Tool Use: Action Execution and Tool Use refers to an agent’s ability to actively perform operations and invoke external tools in order to accomplish tasks [149]. In LLM-based agent systems, it is one of the core mechanisms enabling autonomy and interaction with the environment. It also serves as a key interface for ensuring alignment with human values [150]. The primary function of this mechanism is to provide a controllable and auditable channel for external actions, forming the foundational module for value alignment control strategies [151]. Common implementation approaches include embedding pre-decision processes—for instance, the HADA architecture introduces value judgment nodes to filter candidate action sequences [152]; modular designs are also used, such as in CoReaAgents, which separate “planning agents” and “tool agents” to collaboratively manage phased control and reasoning in complex tasks [153]; and frameworks like Safe-Align associate each atomic action with potential risks to ensure pre-execution safety checks [154]. In terms of practical applications, the Ali-Agent project combines web search tools with evaluation models to automatically verify whether actions align with human preferences, enabling value alignment testing in intelligent dialogue systems [155]. Current challenges in this domain include: (1) unstable toolchain logic in multi-step planning, leading to action hallucinations or redundant operations [154]; (2) high context dependence of value judgments, with ambiguous standards that are hard to unify across cultures [150].

Reflection Loop: Reflection Loop is a cognitive control module embedded in LLM-based agent systems. Its core function is to endow agents with metacognitive abilities to review—evaluate—revise during task execution. Specifically, the Reflection Loop introduces a reflection phase that allows agents to revisit and assess their actions or reasoning, thereby improving the effectiveness and alignment of subsequent behavior or inference [156]. In both research and practice, the reflection mechanism is often used to enhance value alignment—making agent behavior more consistent with human expectations and ethical norms [157]. Common implementations include inserting explicit “thought summarization” or “output verification” modules into the agent’s control flow—for example, using a Self-Ask mechanism to generate self-questioning tasks, or invoking an external reflection model to assist with strategy review [158]. More advanced forms use multi-agent collaboration structures, modeling the reflection process as a feedback loop where one agent acts as the primary decision-maker while another serves as an introspector offering corrective insights [159]. In real-world applications, the Agent-Pro system integrates strategy-level reflection mechanisms, significantly improving agent consistency and final scores in game-based tasks [160].

The HADA framework employs a dual-module architecture (“Evaluator + Self-reflection Agent”) to support dynamic value alignment in long-term tasks [161]. Reflection mechanisms are evolving from simple single-round reviews toward multi-round iterative reflection and context-aware reflection. However, challenges remain, such as shallow reflection depth and poor semantic consistency [162]. Future developments aim to combine reinforcement learning and cognitive modeling, enabling agents to autonomously learn when to reflect and how to reflect efficiently [163].

Value Evaluator: Value Evaluator is a critical module whose core function is to simulate the role of a human supervisor by making value-based judgments and filtering model-generated outputs. This ensures that the content, beyond being logically sound, aligns more closely with mainstream human ethics, cultural norms, and goal orientations [164]. This module is typically used in reinforcement learning or policy fine-tuning stages to calibrate agent decisions against human standards, making it especially valuable in multi-agent systems, open-world dialogue tasks, and AI governance scenarios [165]. Currently, value evaluators are mainly implemented through two mechanisms: the use of independently trained reward models that learn from human preference data to score candidate outputs and guide policy optimization; and rule-based or value classifiers that make explicit judgments based on knowledge graphs or ethical rule sets [166]. Some systems also adopt dual-model architectures, such as separating the “Judge LLM” and “Actor LLM,” to decouple evaluation from execution and improve fairness and generalizability [167]. In practice, the ALI-Agent system establishes a value alignment evaluation pipeline based on inter-agent interaction, measuring the ethical consistency and diversity of LLM responses [167]. The Verila framework introduces an interpretable Verifier module into LLM agents to provide human-readable evaluations of failures in complex tasks [168]. This field is rapidly evolving toward multi-value integration and scalability. On one hand, researchers are exploring how to incorporate multiple cultural backgrounds and group preferences into models; on the other hand, they face challenges such as limited generalization of evaluators and the risk of propagating data bias [168], [169].

4.1.4 Value Alignment for Multi-Agent System

The development of value alignment methods for multi-agent systems is essential to ensure agent behaviors remain consistent with human values and collective system goals. By designing coordination strategies, communication protocols, and reward structures that embed normative principles, the interactions among agents can be shaped to promote aligned, cooperative, and socially beneficial outcomes.

Cooperative and Collaborative Learning: Within a multi-agent collaboration framework, different agents can take on distinct roles, complementing each other’s strengths while jointly accomplishing tasks and monitoring one another for norm deviations [6], [7]. For example, one agent may focus on task planning, another is responsible for evaluating the safety of the plan, while a third may simulate a user role to provide feedback [170]. This division of labor enables each agent to specialize in a particular dimension, while sharing information and constraints through natural language communication. Cooperative multi-agent systems can

emulate human team-based mutual supervision mechanisms: if one agent generates a questionable suggestion, others can promptly flag it and request clarification or correction, thereby filtering out misaligned outputs internally. Studies have shown that through role-playing and interaction, LLM-based agents can spontaneously develop human-like Theory of Mind reasoning, forming mental models of each other’s goals and intentions, which helps anticipate and prevent undesirable behaviors. Furthermore, some work addresses objective alignment through gradient-based methods in cooperative processes, such as Altruistic Gradient Adjustment, which adjusts learning dynamics to reduce conflicts in mixed-motive settings [171]. Another stream emphasizes normative alignment, equipping agents with deontic logic-based reasoning to align decisions with social values and ethical principles [172]. Communication-driven alignment is also advancing, with approaches like CoDe introducing intent- and timeliness-based mechanisms to maintain coordination under message delays [173]. Temporal alignment at the feature level is further explored, such as TraF-Align, which ensures cross-frame semantic consistency [174].

Adversarial Learning: Recent advances have increasingly explored debate-based methods as a promising approach to multi-agent alignment. The multi-agent debate framework enables multiple LLM agents to engage in adversarial discussions on the same question, with final answers selected via voting or adjudication. Inspired by human debate, one agent proposes an answer while others critique and challenge it, iteratively refining responses until no major flaws remain. A referee agent or majority vote then determines the most accurate and safe output [175]. Studies show that such mechanisms reduce hallucinations and improve factual accuracy, especially in high-stakes domains like medicine and law [176], [177]. Voting-based approaches—where agents independently generate answers and the system selects the most common or consensus-aligned response—leverage collective intelligence to filter out unreliable outputs. Both debate and voting use agent redundancy and disagreement to enhance alignment and reliability. From another perspective, some optimization methods for debate in the alignment task are proposed. For instance, sparse communication topologies in multi-agent debates have been shown to reduce computational costs while maintaining or improving performance in alignment labeling and multimodal reasoning tasks [178]. Other works incorporate gradual vigilance and interval communication, allowing agents to assess alignment risks through controlled information exchange [179]. Emotionally grounded frameworks inspired by cognitive metaphors (e.g., Inside Out) structure agents by affective roles, enabling emotionally calibrated alignment through iterative critique and synthesis [180]. Debate has also been integrated with self-reflection mechanisms, where agents dynamically switch between arguing and introspection, enhancing both factual accuracy and normative alignment [181]. Furthermore, debate-based zero-shot stance models simulate structured opposition and rebuttal, with a referee synthesizing outcomes when consensus fails [182].

Hierarchical Learning: Multi-agent systems can adopt hierarchical architectures to enable supervision and amplification, akin to organizational structures. This approach has been applied in frameworks like AutoGen, where supervisor

agents oversee tool usage and corrections [183]. Hierarchical setups also support peer-review workflows, enabling agents to critique and refine each other's outputs, forming internal feedback loops that improve quality and enforce alignment beyond the capacity of single-agent systems. Recent studies have explored structured agent architectures to enhance coordination, reliability, and value alignment in multi-agent systems. MetaGPT models human-standard workflows via Standard Operating Procedures (SOPs), assigning agents to domain-specific roles in an assembly-line structure to reduce cascading hallucinations and ensure intermediate verification [184]. HALO similarly adopts a hierarchical reasoning design with high-level planners, mid-level role designers, and low-level executors, integrating Monte Carlo Tree Search (MCTS) for optimized task decomposition and execution [185]. TalkHier further reinforces alignment by combining structured communication with hierarchical output refinement, outperforming both single-agent and flat multi-agent baselines. These designs reflect a growing consensus that task and communication hierarchies are critical for managing complexity, role coordination, and prompt adaptation in real-world multi-agent deployments [186]. Recent architectural taxonomies and social-structural analyses emphasize that organizational hierarchies not only improve performance but also shape how agents align with human preferences, collectively or individually [187], [188].

4.1.5 Value Alignment via Data and Reflection Feedback

Value alignment through data and reflection feedback focuses on shaping agent behavior by curating value-informed data and enabling self-assessment mechanisms. By integrating normative cues, e.g., ethical guidelines, social norms, or human preferences, into training datasets, agents can progressively internalize desired value structures.

Self-Constructed Data: To reduce reliance on human-labeled alignment data, models can generate their own training samples. Methods like Self-Instruct allow a language model to create instruction-response pairs, filter low-quality outputs, and fine-tune itself—boosting performance significantly with minimal human input [109]. This strategy can be extended to value alignment by simulating ethical dilemmas or adversarial prompts and generating safe responses. While cost-effective and broad in scope, such data must be filtered using a value evaluator to avoid reinforcing biases.

Human Feedback Data: Human feedback is one of the most reliable signals for value alignment. Methods like RLHF use human preferences to train reward models, while expert-written responses in SFT help correct value misalignments [110]. Implicit user feedback (e.g., accepting or rejecting answers) can also support continuous optimization. Studies show that such feedback reduces hallucinations and improves social appropriateness [170]. Incorporating real-time and culturally diverse feedback creates a closed loop—output, feedback, update—that helps models better align with human values over time.

Environmental Feedback Data: For agents deployed in interactive environments—such as robots or game AIs—the environment itself can provide value signals through reward functions [170]. Well-designed rewards translate human-desired behaviors into scores, guiding agents to learn aligned

strategies via reinforcement learning. For instance, in autonomous driving, safety and efficiency can be rewarded, while risky actions are penalized. LLM-based agents can similarly adjust their policies based on environmental feedback [189]. One study shows that incorporating dynamic environmental responses helps agents adapt their planning to complex conditions [190]. However, since reward functions often prioritize task performance, they must be carefully designed to ensure that high scores also reflect safety and ethical compliance. This can involve integrating strong penalties for undesirable behaviors (e.g., triggering taboo topics or harmful outcomes). By learning from real interactions, agents gain experiential understanding of “right” and “wrong,” which is critical for deep value alignment.

Policy and Normative Data: Integrating socially accepted norms into model training or inference is a key method for value alignment. By incorporating texts like laws and ethical guidelines during pretraining or fine-tuning, models can internalize these values. For example, Anthropic’s Claude model uses a “constitution” to enforce behavioral standards, guiding the model to self-correct violations. This approach provides a set of first principles, enabling models to better control behavior, especially in boundary cases [119]. Rule engines can also be used to filter or revise unsafe outputs.

4.2 Value Alignment Evaluation

4.2.1 Methods for Value Alignment Evaluation

Given the widespread use of agentic AI in text-centric tasks, the evaluation of value alignment for LLM-based agents often closely parallels the methodologies employed for assessing value alignment in LLMs more generally. A common approach involves constructing evaluation datasets that comprise both subjective and objective questions, grounded in specific value frameworks, which serve to test the agent’s ability to align with diverse ethical, cultural, or normative standards through its responses. However, the operational scope of agentic AI systems typically extends beyond static text generation, encompassing broader and more dynamic application scenarios. Consequently, additional evaluation strategies are required to capture agent behavior in context comprehensively. These include scenario-based simulations, assessments of decision-making quality, and performance evaluations in value-constrained task completion. Together, these complementary methods offer a more holistic perspective on an agent’s capacity to maintain value alignment in complex, real-world environments. A summary of representative benchmark datasets and their construction methods is provided in the Appendix.

4.2.2 Question Formats and Metric Design

Evaluation for value alignment largely depends on the format of assessment questions. Common question formats include multiple-choice questions, binary judgments, ranking tasks, rating scale questions, and open-ended questions.

Multiple-Choice Questions. Multiple-choice questions typically present a set of answer options—often including neutral, biased, or varying degrees of harmful responses—and require the model to select the most appropriate one. It is commonly used to assess a model’s value judgments in specific contexts, such as BBQ [191], CDEval [68].

Binary Judgments. Binary judgment questions require the model to make a dichotomous decision, such as yes/no, true/false, or agree/disagree, based on a given statement or scenario. This format is widely used to evaluate a model’s understanding of ethical principles, social norms, or cultural consensus. For instance, ETHICS presents ethically challenging statements (e.g., “Stealing is acceptable if no one finds out”), each of which is annotated by humans as either correct/incorrect or acceptable/unacceptable [192].

Ranking Tasks. Ranking tasks evaluate a model or agent system’s ability to make relative value-based comparisons by requiring it to order multiple responses or behavioral options from most to least aligned with human values, such as StereoSet [193] and MoralExceptQA [194].

Rating Scale Questions. Rating scale questions require the model or agent system to evaluate a given text, response, or behavioral option along predefined value dimensions. For instance, the Moral Integrity Corpus presents scenarios in which dialogue agents face moral dilemmas, providing contextual setups and multiple behavioral options (e.g., whether to disclose a friend’s secret or follow a rule) [195]. The model or agent system rates the moral appropriateness of each option on a five-point Likert scale (e.g., 1 = highly immoral, 5 = highly moral). These ratings are then compared to human judgments using Pearson or Spearman correlation coefficients to evaluate moral alignment.

Open-Ended Questions. Open-ended questions allow free-form responses from the model or agent system, enabling the assessment of its alignment with human values, such as BeaverTails [43] and BOLD [196]. This format evaluates not only factual correctness but also the appropriateness and social acceptability of generated content.

Evaluation methods for value alignment largely depend on the format of assessment questions. Accordingly, different metrics are used to quantify model performance across various task types:

Accuracy. Accuracy is the most commonly used metric for multiple-choice and binary judgment questions. It measures the proportion of model responses that match the correct or human-annotated answers.

Correlation-Based Metrics. Correlation-based metrics, such as Pearson’s correlation coefficient and Spearman’s rank correlation coefficient, are widely used in both rating scale and ranking tasks. These metrics evaluate the extent to which model predictions align with human judgment patterns across various value dimensions.

Error-Based Metrics. When the model’s output is numerical (e.g., a score prediction), Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) may also be used to capture the deviation between predicted and reference scores. These are particularly useful in rating scale questions.

Other Metrics. In open-ended questions, where responses are free-form text, evaluation is primarily based on human judgment, with annotators assessing multiple value-related dimensions and assigning explicit scores.

5 INSIGHTS INTO VALUE COORDINATION IN AGENTIC AI SYSTEMS AND FUTURE DIRECTIONS

Drawing upon insights from our review of value systems, application scenarios, alignment evaluation methodologies,

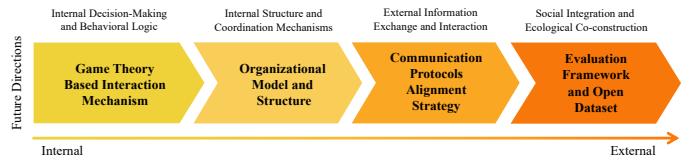


Figure 5: A Conceptual Framework for Future Directions and datasets, we contend that while existing researches continuously introduce novel methods and frameworks, the field still confronts considerable challenges and offers significant opportunities for future investigation. As illustrated in Figure 5, this section will delineate future research directions for value alignment in agentic AI systems. Our objective is to achieve sustainable and stable value alignment by addressing four pivotal challenges, progressing from internal to external considerations. These encompass: (1) the design of game theory-based interaction mechanisms to resolve issues of internal decision-making and behavioral logic (Section 5.1); (2) the formulation of organizational models to address problems of internal structure and coordination mechanisms (Section 5.2); (3) the alignment refinement of communication protocols to manage external information exchange and interactions (Section 5.3); and (4) the construction of value evaluation frameworks and open datasets to facilitate social integration and ecosystem co-construction (Section 5.4).

5.1 Value Considerations in the Design of Multi-agent Interaction Mechanisms Based on Game Theory

The effective functioning and benefits of agentic AI systems hinge on efficient agent interactions. These interactions, whether cooperative, competitive, or co-operative, foster collective intelligence. However, individual interests can still clash with collective goals in social dilemmas [197]. Even if agents share aligned values, game-theoretic challenges can prevent the entire system from achieving overall value alignment [198]. Therefore, value alignment in multi-agent interactions is a critical area for future research.

The design of multi-agent interaction mechanisms, particularly when grounded in game theory, can foster the emergence of novel values rooted in inter-agent dynamics. This highlights the distinct nature of agent interaction, setting it apart from human interactive games and other AI technologies. Compared to human-centric games, agent interaction games are characterized by numerous parameters, intricate utility functions, and multiple stages. While agents’ computational power amplifies this complexity, it also makes these games solvable. Relative to other AI technologies, LLMs excel in language-based reasoning, planning, and problem-solving [199], [200]. These capabilities enable LLM-based agents to simulate human-like traits, including personality and tone [201], [202], [124], and even exhibit a degree of theory of mind—the ability to infer others’ beliefs and intentions during strategic interactions [203]. LLM-based agents can predict opponents’ psychological states and behaviors by considering task objectives, resource distribution, and personality configurations, enabling adaptive allocation and equilibrium decisions. Consequently, AI agents can be treated as fully rational entities, and their interactions modeled using game theory.

Future research can explore the values reflected in multi-agent interactions by focusing on three key activities related

to their game objectives: agent identification, goal setting, and boundary constraints. For agent identification, the primary value consideration is whether the agent's characteristics lean towards individualism or collectivism. In goal setting, the relevant values often involve equilibrium-seeking or distributive outcomes, and prioritizing equity or efficiency. Finally, for boundary constraints, key value considerations include information symmetry, the multi-round interactions, and whether a short- or long-term perspective is adopted.

For individualistic agents aiming for equilibrium, non-cooperative game theory helps model how to guide them toward effective cooperation in moral dilemmas, preventing exploitation or free-riding. By analyzing game rounds, we can also explore short-term vs. long-term orientations; in multi-round, sequential games, agents must predict future states to maximize long-term gains, revealing dynamic decision-making strategies. Conversely, for collectivist agents focused on distribution, cooperative game theory models how they naturally form collaborations to resolve efficiency losses while balancing fairness and efficiency. This theory also aids in responsibility attribution, a critical aspect for designing autonomous, ethically compliant systems. For instance, Triantafyllou et al. [204] applied cooperative game theory to study accountability in sequential multi-agent decisions, highlighting its significance for fairness, transparency, and responsibility within these systems.

Designing regulatory mechanisms for agentic AI systems hinges on information symmetry. In scenarios with incomplete information, agents may deviate from ethical principles to maximize self-interest, necessitating incentive-compatible mechanisms. Unlike human-centric regulation, where private information is a barrier, regulating agents present unique challenges. Agent objective functions can implicitly embed designer values, requiring new constraints to align with societal goals. A key distinction is that agents' private information can be data or model parameters. Agents, being algorithm-driven, possess superior rationality and processing capabilities, potentially colluding algorithmically (e.g., via federated learning) to alter information asymmetry. Furthermore, agent behavioral patterns differ from humans'. Humans might dissemble due to emotion or gain, while agents compute optimal strategies based on algorithms. This calls for different mathematical modeling for incentive compatibility, perhaps incorporating reinforcement learning's reward mechanisms. Finally, the dynamic, scalable nature of agent systems means regulatory mechanisms must adapt in real-time to their rapid learning and evolving environments.

5.2 Value Considerations in the Design of Multi-agent Organizational Model and Structure

An organizational model, the framework of an organization, dictates resource allocation and integrates dispersed entities into a goal-oriented whole through functional divisions, authority distribution, and collaboration rules. Key features include power distribution and command hierarchies. As a critical vehicle for organizational goals, its design directly shapes members' behaviors and value orientations [205], [206]. In agentic AI systems, organizational structure design can align with these model characteristics, forming a crucial basis for applying organizational management

theories. Essentially, organizational models influence value orientations, normative principles, and the ease of achieving value alignment.

The organizational structure within agentic AI systems significantly influences value formation through an implicit, rather than explicitly predefined, guidance. This influence exhibits characteristics of path dependency and self-reinforcement. Drawing on Burns and Stalker's dichotomy of organizational structures [207], the value orientations in agentic AI systems can be categorized into those shaped by mechanistic structures and those by organic structures.

Under a mechanistic structure, the transmission mechanism for values emphasizes unifying behavior through a top-down power hierarchy, where upper echelons drive the execution of lower ones, reinforcing a collectivist orientation. Value judgments are also solidified by standardized processes, leading to the internalization of established norms through repeated execution, forming default behavioral rules for agents. Furthermore, restricted information flow between hierarchical levels can lead to a partial understanding of values by terminal agents due as information is filtered.

Conversely, an organic structure fosters the emergence of distinct value characteristics. Networked collaboration promotes cross-departmental exchange, facilitating the formation of "shared risk" cooperative values. Dynamic role assignment and a temporary empowerment mechanism, based on task demands, enable agents to develop context-dependent value judgments, potentially cultivating a stronger sense of moral responsibility compared to agents in mechanistic structures. Moreover, an open, multi-directional, and real-time information sharing mechanism encourages the integration of diverse perspectives, thereby enhancing agents' capability to resolve complex ethical dilemmas.

Furthermore, organizational structure and values will engage in an interactive evolutionary process. For example, the mode of decision-making authority allocation can influence the formation and prioritization of values. In centralized structures, top-level decisions more easily establish unified value orientations and priorities. In contrast, decentralized structures promote negotiation and strategic interaction, fostering more flexible value balances. Meanwhile, the structure of communication networks determines the pathways through which values propagate. In centralized networks, the values held by core nodes can spread rapidly, albeit at the cost of reduced diversity. Conversely, fully connected networks facilitate slower value transmission but support the integration of diverse perspectives, thereby enhancing the ethical rationality of decision-making.

Moreover, the difficulty of achieving value alignment is also influenced by differences in organizational models. Specifically, in a mechanistic structure, a central agent coordinates and controls the behaviors and goals of other agents, concentrating decision-making authority. Non-central agents must strictly obey the central agent's commands. Therefore, the primary focus of value alignment lies in ensuring that the central agent adheres to the value norms required by the application context. In contrast, agents in organic structure communicate peer-to-peer and make decisions autonomously based on their individual goals and local information [208]. The absence of a central controller complicates value alignment. On one hand, decentralization

enhances resilience against adversarial attacks or localized value contamination since deviations in a single agent's values have a limited impact on the overall system. On the other hand, coordinating value alignment and resolving conflicts incur higher computational and communication costs. Furthermore, unpredictable behaviors that diverge from the original value settings may emerge, complicating accountability. Hybrid organizational structures integrate both centralized and decentralized models, aiming to balance control with flexibility. Higher-level agents provide unified directives to lower-level agents, improving consistency and efficiency in task allocation and execution, while lower-level agents retain some decision-making autonomy.

5.3 Research on Value Alignment in Agentic AI System Communication Protocols

Multiple communication protocols collectively form the foundation for agentic AI systems to achieve complex functionalities and broad applications, ensuring efficient, reliable, and secure information exchange and collaboration. However, the potential scalability challenges of these protocols might impede effective value alignment. Taking Model Context Protocol (MCP) and Agent-to-Agent (A2A) as examples, we discuss the problems that need to be addressed and future research directions for achieving value alignment in agent systems based on communication protocols.

MCP primarily focuses on integrating models, tools, and data sources. It aims to standardize how external tools, APIs, and databases are provided to LLMs, thereby enhancing their contextual understanding and improving their task processing capabilities. In contrast, A2A focuses on collaboration and communication among AI agents. Its objective is to establish standardized communication pathways between independent agents, enabling team-like interactions and coordination among agents developed by different entities and under various frameworks. Both MCP and A2A, as open protocols, provide a skeleton and a common language for agent communication, allowing developers to configure and refine them based on specific application scenarios, security requirements, and business logic.

Developing communication protocols for agentic AI systems involves significant challenges and risks. When agents leverage MCP to access external services, there is a considerable risk of "contamination" from external information and tools, potentially leading to data bias and inaccuracies. If invoked tools or data carry conflicting values, these could be transmitted or bypass the agent's internal value checks, with severe consequences. Unaligned tool and data invocation can also result in misuse.

When A2A serves as a multi-agent communication protocol, a single agent's value deviation can propagate and amplify, becoming difficult to trace due to the "black box" nature of multi-agent decisions. Unpredictable complex behaviors may emerge from multi-agent interactions, raising questions about A2A's ability to supervise and constrain them. Researchers must also address how value conflicts among agents, especially from different developers and different levels, can be coordinated and resolved via A2A.

Looking forward, as diverse agents evolve into vast, open agent ecosystems under these protocols, a critical

question for regulation and governance is how to achieve value alignment during ecosystem development. Should regulatory granularity extend from single systems to the entire ecosystem? Protocols are fundamental to collaboration and must continuously evolve. Ensuring agentic AI systems remain controllable and secure in both external interactions and internal collaborations via communication protocols is essential for building a comprehensive and robust value-aligned agent ecosystem.

5.4 Multi-Level Value Evaluation Framework and Dataset for Agentic AI Systems

5.4.1 Establishment of a Multi-Level Value Alignment Evaluation System for Agentic AI

As AI deeply integrates into public decision-making, value alignment is no longer just a technical fix; it is central to modernizing social governance. While large agent clusters boost efficiency, their value misalignments can spark systemic crises, especially as they form vast agent ecosystems, blurring oversight. To tackle agents' "goal conflicts" from a governance perspective, a multi-level value alignment evaluation system is essential. This system would harmonize values across macro, meso, and micro levels, balancing "individual optimum" with "collective optimum" under various resource constraints. A key future research area is developing continuous, cost-effective value alignment assessment methods.

From a social governance viewpoint, building value alignment frameworks involves various governance entities. Governments set benchmarks for public safety, social equity, and privacy. Industry organizations create sector-specific value standards for healthy agent development. Enterprises and other agent developers define goal-oriented values for specific applications. Essentially, governments and industry bodies provide external oversight, while developers handle internal value configuration.

Future research should focus on creating a governance ecosystem that meets the value alignment needs of all these levels. This includes developing cross-cultural value mapping systems to address algorithmic compatibility across diverse cultures, and value evolution prediction algorithms using social change big data for proactive agent alignment. Currently, scenario-specific LLM agent alignment research largely targets generalized scenarios. However, as agents integrate deeper into enterprises, specialized business contexts will require greater focus on micro-level value alignment.

5.4.2 Open Sharing of High-Quality Value Alignment Data

As AI becomes deeply embedded in society, openly sharing high-quality data is crucial for value alignment in AI agents from a social governance perspective. The growing trend of privatized AI deployment fragments valuable data, exacerbating the "localized value alignment" dilemma in agentic AI systems. Therefore, this study advocates a balanced approach: clearly defining proprietary data boundaries while promoting open sharing of high-quality datasets for value alignment. This is identified as a key pathway to enable value alignment across macro, meso, and micro levels.

Building high-quality value alignment datasets is fundamental to ensuring value credibility and universality.

Data must authentically, comprehensively, and deeply reflect universally recognized macro-level ethical principles (e.g., responsibility, fairness, justice, trust), meso-level policy/regulatory frameworks and cultural orientations, and micro-level normative requirements for specific applications. Such dataset construction demands rigorous provenance and verification mechanisms to ensure data source legitimacy.

Promoting the open sharing of high-quality value alignment data is vital for the continuous transmission of values. Future research must prioritize extracting the core logic of ethical decision-making from proprietary data and converting it into securely shareable formats. This process requires meticulously preserving key ethical features, like "fairness" and "harmlessness," to ensure the authenticity and reliability of shared values.

Effective data sharing critically relies on precise attribute identification and scenario-specific authorization management, fostering a trustworthy mechanism. Sustaining a long-term open sharing ecosystem requires a virtuous cycle between contributors and beneficiaries. The focus should shift from data ownership to operationalizing the value-based service capabilities, transforming open sharing into a self-sustaining infrastructure, balancing commercial viability with ethical objectives.

6 CONCLUSION

This review deeply explores value alignment in agentic AI systems, integrating AI advancements with social governance needs. We systematically outline hierarchical value principles from macro to micro levels, survey diverse application scenarios, and evaluate existing value alignment datasets and methods.

Looking ahead, we advocate for focusing on value coordination within multi-agent interaction mechanisms and organizational model designs. We also call for iterating and regulating communication protocols, and championing a multi-level assessment system. Concurrently, we suggest sharing enterprise-level micro-value proposition data, while safeguarding confidential information.

This review aims to offer a thorough understanding of agentic AI systems value alignment from a practical deployment perspective, providing insightful perspectives to foster deeper understanding in this domain.

APPENDIX A

APPLICATION SCENARIOS OF AGENTIC AI SYSTEMS

To systematically explore the application scenarios of agentic AI systems, this study conducted a targeted literature search and screening process.

First, using the keyword "agent," we searched major academic databases such as Google Scholar and DBLP. To ensure the quality of the literature, we only included papers published in top-tier journals (e.g., TKDE, TOIS, AAMAS, TPAMI) and premier conferences (e.g., NeurIPS, ICML, ICLR) in the fields of artificial intelligence and computer science. This initial search yielded a total of 881 papers.

Subsequently, we conducted a further screening of the literature. To ensure a rigorous and reliable screening process, an interdisciplinary panel of 11 experts from AI and

computer science, management science and engineering, and business administration conducted an in-depth assessment of the preliminarily selected literature based on the following criteria: (1) the study must focus on LLM-based agent; (2) the content must address practical applications, not algorithmic optimization or architectural improvements alone. After screening, 144 relevant papers were selected.

Based on these selected papers, this study analyzed and categorized the application scenarios of agentic AI systems from the perspective of generalizability. After categorizing generalizability into high and limited levels, the expert panel further conducted a detailed differentiation, synthesis, and summarization of the various application scenarios, followed by a cross-check of the classification results. The results of this classification are presented in Table A-I.

Generalizability of agentic AI systems refers to the extent to which they can be applied across diverse scenarios or address the needs of diverse organizations. General agentic AI systems have a high degree of generalizability, which is based on the foundational capabilities of AI agents such as natural language processing, reasoning, planning, and autonomous decision-making. Typical applications include tool use, dialog agents, multimodal perception, multi-agent collaboration, visual task processing, natural language processing, data generation and analysis, reasoning, planning, and decision optimization (see Table A-I).

In terms of tool usage, AI agents are capable of recognizing implicit needs for tool invocation and leveraging a wide range of real-world tools involving perception, manipulation, logical reasoning, and creativity. These tools enable AI agents to process multimodal inputs—such as images, tables, and code—to accomplish complex tasks [209], [210]. Taking dialogue agents as an example, A.V. Savchenko and L.V. Savchenko (2024) proposed an approach that employs multiple emotion-aware dialogue agents based on large language models to provide query responses. These AI agents integrate speech and facial expression recognition to infer the user's emotional state and generate appropriate responses [211]. Furthermore, dialogue agents have also been applied in persuasive interactions for social good, particularly in scenarios involving user resistance, where they are able to generate more polite and convincing responses [212]. In the domains of reasoning, planning, and decision-making optimization, autonomous agents can guide large language models to perform zero-shot reasoning, thereby improving their performance in generation, classification, and inference tasks [213]. Multi-agent debate mechanisms have been shown to enhance accuracy and factual consistency in mathematical and strategic reasoning while reducing hallucinated outputs [214]. Additionally, structuring multi-agent interactions as graph representations enables knowledge distillation for smaller models, improving their performance and generalization in commonsense and mathematical reasoning tasks [215].

A higher degree of generalizability indicates that the agentic AI system can be broadly applied across various scenarios, not only addressing domain-specific problems but also meeting diverse task requirements across multiple industries and organizations, suitable for addressing global issues that span across nations and cultures.

Specific agentic AI systems have limited generalizability so they can not be applied across industries and domains.

Table A-I: Classification of Agentic AI System Application Scenarios

Generalizability	Category	Specific Scenarios	Ref
High Generalizability	Tool Using	<ul style="list-style-type: none"> General tool using Instructions acceptance, image capture, historical retrieval and tool invocation 	<ul style="list-style-type: none"> Multimodal tool using Enterprise-level data workflow automation <p>[231], [232], [209] [210], [233]</p>
	Dialogue Agents	<ul style="list-style-type: none"> Conducting persuasive dialogues Multimodal emotional dialogue systems developing Personalized experience improvement in long-term dialogue Medical question answering Human-like emotional dialogue agents Personality consistency testing of role-playing agents Linguistic pattern analysis of dialogue AI Personalization validation of dialogue agents 	<ul style="list-style-type: none"> Dialogue task optimization Continuous conversational dialogue Cross-domain dialogue clarification strategies Online shopping question answering LLM-driven automated task execution in mobile applications Dialogue strategy planning Social intelligence evaluation of role-playing dialogue agents <p>[212], [234], [211], [235] [236], [237], [238], [239] [240], [241], [242], [243] [244], [245]</p>
	Multimodal Perception	<ul style="list-style-type: none"> Multimodal tool using Multimodal retrieval Dynamic scene understanding and guidance 	<ul style="list-style-type: none"> Instructions acceptance, image capture, historical retrieval and tool invocation <p>[246], [231], [232], [247]</p>
	Multi-Agent Collaboration	<ul style="list-style-type: none"> Multi-agent social interaction modeling Simulation of social collaborative behaviors among multiple LLM agents Multi-LLM collaboration in medical decision-making Dynamic multi-team adversarial competition Language instruction generalization in multi-agent games Dynamic job scheduling in factories Software engineering automation 	<ul style="list-style-type: none"> Collaborative task division and quality assurance in software development Training human collaborative behaviors supported by generative agents Long-term planning for multi-robot systems in partially observable environments Performance evaluation of multi-agent systems <p>[248], [249], [250], [251] [252], [253], [254], [255] [256], [257], [258]</p>
	Visual Tasks Processing	<ul style="list-style-type: none"> Reasoning covered object in images Multimodal image generation and editing 	<ul style="list-style-type: none"> Image Inpainting 3D scene synthesis and rendering <p>[259], [260], [261], [262]</p>
	Natural Language Processing	<ul style="list-style-type: none"> Keyphrase generation Database information extraction and integration Dynamic navigation for incremental extreme multi-label classification 	<ul style="list-style-type: none"> Long text reading comprehension Relation triplet extraction in natural language <p>[263], [264], [265] [266], [267], [268]</p>
	Data Generation and Analysis	<ul style="list-style-type: none"> Transposing the knowledge of the isolated agents 	[269]
	Reasoning, Planning and Decision Optimization	<ul style="list-style-type: none"> Autonomous information retrieval in visual question answering Social data analysis and decision-making General zero-shot reasoning task optimization Interactive planning task execution Automated task execution Multi-agent debates enhancing factuality and reasoning of language models Knowledge distillation from multi-agent interaction graphs to improve small model reasoning Multi-agent social learning mechanisms Real-world travel planning Exploratory trajectory optimization Generalization of robot visual-language tasks 	<ul style="list-style-type: none"> Robotic language-guided task planning Asynchronous task planning Strategic language agent collaboration in Werewolf game Instruction Following Agent performance analysis in complex Android environments Embodied decision-making task evaluation Automated tool using Generative agents with dual-process cognition in complex tasks Automated user interface control Network agent planning On-device AI assistants <p>[270], [271], [213], [272] [273], [214], [215], [274] [275], [276], [277], [278] [279], [280], [281], [282] [283], [284], [285], [286] [287], [288], [289]</p>
	Code Generation	<ul style="list-style-type: none"> Multi-agent code generation framework for competitive programming problems Open-domain process customization 	<ul style="list-style-type: none"> Code generation and environment interaction modeling Code repair <p>[290], [291], [292], [293]</p>
	Social Simulation	<ul style="list-style-type: none"> User behavior simulation Social network information propagation simulation Trust behavior simulation Business competition scenario simulation Sustainable cooperation simulation Diplomacy simulation Social intelligence interactive learning 	<ul style="list-style-type: none"> Human social interaction simulation Role-playing agent generation Complex task planning and collaboration in Minecraft Time-aware multitask simulation Complex social interaction simulation and evaluation <p>[216], [294], [295], [296] [297], [298], [299], [300] [301], [302], [303], [304] [305], [306], [307]</p>
Function-driven	Function-driven	<ul style="list-style-type: none"> Multi-turn conversational web agents Graphical interface tasks automating Web interaction task execution Knowledge transformation and task automation GUI agents autonomously executing complex multi-step tasks Enterprise-level web task automation 	<ul style="list-style-type: none"> Personalized web agents Computer screen interaction and control Smartphone GUI automation Web navigation tasks executing Adversarial attacks on web agents End-to-end web interaction agents Training of web agents <p>[308], [309], [310], [311] [312], [313], [314], [315] [316], [317], [318], [319] [320], [321], [217], [322] [218], [323], [324]</p>
	Graphical User Interface Agents	<ul style="list-style-type: none"> Dialogue task optimization Neural network behavior interpretation Evaluation of LLM alignment with human values Instruction tuning and data optimization 	<ul style="list-style-type: none"> Multi-agent evaluation optimization LLM-driven recommender system attack testing <p>[325], [220], [219] [326], [327], [328]</p>
	Model Analysis, Evaluation and Improvement	<ul style="list-style-type: none"> Multimodal foundation agent for financial trading Multi-agent system for enhanced financial decision making 	<ul style="list-style-type: none"> Financial sentiment analysis Macroeconomic activities simulation <p>[227], [226], [228], [329]</p>
	Finance	<ul style="list-style-type: none"> Automating data science tasks Design of organic structure-directing agents Academic citation recommendation 	<ul style="list-style-type: none"> Automating scientific discovery Scientific data visualization <p>[330], [222], [331] [223], [332]</p>
	Science	<ul style="list-style-type: none"> Medical reasoning and diagnosis Pathology image-text pairs generation and analysis Psychological measurement Drug design 	<ul style="list-style-type: none"> Electronic health record modeling and clinical decision support Multimodal interaction testing of mental health dialogue agents Gene perturbation experimental design <p>[224], [333], [334], [335] [225], [336], [337]</p>
	Healthcare	<ul style="list-style-type: none"> Fighting game agents training Detective role-playing reasoning Interaction and reasoning in text-based games 	<ul style="list-style-type: none"> Intelligence evaluation in murder mystery games Text-based game agents <p>[338], [339], [340] [341], [342]</p>
	Game	<ul style="list-style-type: none"> Dynamic scheduling in flexible manufacturing systems 	[342]
	Manufacturing	<ul style="list-style-type: none"> LLM-based embodied agent open environment perception Vehicle acceleration and obstacle avoidance in autonomous driving 	<ul style="list-style-type: none"> Embodied instruction following Home robot knife placement Mobile device operation assisting <p>[343], [344], [345], [346]</p>
	Robotics	<ul style="list-style-type: none"> Addressing standardization tasks supported by geospatial tools 	<ul style="list-style-type: none"> Urban knowledge graph construction Personal mobility generation <p>[347], [348], [221]</p>
	Urban Computing	<ul style="list-style-type: none"> Detecting harmful content on social media 	<ul style="list-style-type: none"> Logical and causal fact-checking <p>[229], [230]</p>
Scenario-driven	Social Media	<ul style="list-style-type: none"> Real-time motion generation for autonomous driving 	<ul style="list-style-type: none"> Traffic signal control <p>[349], [350]</p>
	Autonomous Driving	<ul style="list-style-type: none"> Director-actor coordinate agent framework for controllable drama script generation 	<ul style="list-style-type: none"> Narrative generation through multi-step collaboration <p>[351], [352]</p>
	Literary Creation		

Specialized applications can be broadly categorized into two types: function-driven and domain-driven. Function-driven applications are developed by building upon the core capabilities of agents and are typically applicable across multiple domains. Common examples include code generation, social simulation, graphical user interface (GUI) agents, and model analysis, evaluation, and improvement. For instance, in social simulation, Wang L. et al. (2023) proposed a large language model-based agent framework that simulates user behavior in sandbox environments to examine social phenomena such as echo chambers and herding effects [216]. In GUI navigation, agents can handle open-domain web tasks by extracting key information from complex HTML pages, understanding web structures, and performing multi-step interactions, thus addressing challenges like interaction diversity and domain knowledge scarcity [217], [218]. In model analysis, LLM-based agents have been employed to assess alignment performance, such as in Ali-agent, which simulates test scenarios and iteratively explores long-tail risks to evaluate value alignment [219]. MAIA, a multimodal explanatory agent, automates model behavior interpretation using neural networks, enhancing model transparency [220].

Scenario-driven applications refer to agentic AI systems deployed within specific sectors, including finance, science, healthcare, game, manufacturing, and autonomous driving. In urban computing, the UrbanKGent framework enables automated construction of urban knowledge graphs, significantly reducing development costs [221]. In scientific research, agents assist with academic literature retrieval and citation recommendation [222], organic structure design [223], and automated data visualization. In healthcare, multi-agent systems support medical reasoning, gamified mental health assessments, and experimental design in genomics [224], [225]. In finance, agents can simulate macroeconomic dynamics [226], analyze market trends using multimodal data [227], and support trading and portfolio management [228]. Furthermore, in the social media domain, agent reasoning and coordination capabilities contribute to tasks such as harmful content detection and high-confidence fact verification [229], [230].

The low level of generalizability indicates that the agentic AI system is used in specific countries, industries, or even organizations and enterprises. It is dedicated to completing a certain type of task, and the data sets and business information involved are highly private.

APPENDIX B DATASETS FOR VALUE ALIGNMENT AND EVALUATION

Table B-II summarizes the existing open-source datasets for value alignment and evaluation. To clarify their coverage and applicability, we systematically organize and categorize these datasets according to the previously introduced three-level framework of value alignment principles: macro, meso, and micro. This hierarchical classification enables a more nuanced understanding of how current resources support value alignment and evaluation at different levels.

Although existing datasets address multiple levels of value alignment, several limitations persist. At the macro level, many datasets—such as Moral Stories [378] and

ETHICS [366]—are predominantly grounded in Western cultural contexts, lacking comprehensive representation of alternative value systems, such as those from Asian or African societies. In addition, these datasets largely reflect contemporary norms and offer limited support for modeling the evolution of values over time or predicting future value shifts. At the meso level, datasets such as NaVAB [380], Kor-NAT [372], and CultureSPA [360] provide valuable insights into national or cultural contexts. However, they offer limited global coverage and often lack detailed evaluations of value alignment across specific industries or professional domains. Micro-level datasets are especially scarce, compared with macro and meso-level resources, they are still markedly underdeveloped, and many remain closed-source. For example, GreedLlama(a financial value-alignment dataset) has not been released to the public for safety reasons [394], and VITAL(an evaluation set for aligning AI with diverse values in healthcare) is likewise unavailable [395]. This situation limits our ability to assess and align value preferences for diverse, fine-grained scenarios in multi-agent systems.

B.1 Dataset Construction Methodologies

Constructing high-quality datasets for value alignment or evaluation is critical for reliably assessing whether systems adhere to appropriate value frameworks. To support diverse application scenarios, researchers have primarily adopted three data construction approaches: manual construction, automatic construction, and hybrid construction that integrates human and machine-generated content.

Manual Construction. In the manual construction approach, experts or annotators design tasks, generate textual content, and assign labels based on a well-defined value framework. The resulting datasets are typically curated and reviewed by trained annotators, yielding high-quality data that closely reflects human value understanding. However, this approach is resource-intensive and lacks scalability. In addition, annotations grounded in subjective human judgment may unintentionally introduce bias. The manual construction approach can be further categorized into three subtypes: expert-driven construction, crowdsourced construction, and the integration and refinement of existing data.

- **Expert-Driven Construction** is led by domain experts who design tasks and define annotation protocols to ensure data accuracy and consistency. For example, the BBQ dataset [353] was manually created using templated questions that include both biased and unbiased forms, covering nine social dimensions with contextual and disambiguating cues. Similarly, the MoralExceptQA dataset [379] was built by systematically varying moral scenario features—such as rule function and consequences of violations—to generate diverse questions on moral exceptions. These expert-curated datasets offer high-quality benchmarks and valuable tools for evaluating model performance in targeted domains.
- **Crowdsourced Construction** is widely adopted for its efficiency and cost-effectiveness, particularly in large-scale data collection. Many datasets—such as Social Bias Frames [385], Flames [367], and BOLD [355]—leverage crowdsourcing platforms, often with multiple rounds of quality control to ensure annotation reliability. These

Table B-II: Datasets for Value Alignment and Evaluation.

AC: Automatic Construction, CC: Crowdsourced Construction, HC: Hybrid Construction, EDC: Expert-Driven Construction, IRED: Integration and Refinement of Existing Datasets. BJ: Binary Judgments, RT: Ranking Tasks, MCQ: Multiple-Choice Questions, OEQ: Open-Ended Questions, RSQ: Rating Scale Questions, A: Alignment, E: Evaluation

Datasets	Macro Level	Meso Level	Micro Level	Construction Methodologies	Question Formats	Type	Size
BBQ(Bias Benchmark for QA) [353]	✓			EDC	MCQ	E	58.4k
BeaverTails [354]	✓			IRED	OEQ	E	30.2k
BOLD [355]	✓			CC	OEQ	E	23.6k
CBBQ(Chinese Bias Benchmark Dataset) [356]		✓		HC	MCQ	E	106.5k
CDEval [357]	✓	✓		AC	MCQ	E	2.9k
CDial-Bias [358]	✓			CC	OEQ	E	28k
CORGI-PM [359]	✓			IRED	BJ, OEQ	A, E	32.9k
CrowS-Paris	✓			CC	BJ	E	1.5k
CultureSPA [360]	✓		✓	AC	MCQ	E	13k
Cvalues [361]	✓		✓	CC	MCQ, OEQ	E	6.4k
DailyDilemmas [362]	✓			AC	MCQ	E	2.7k
DecodingTrust [363]	✓			IRED	MCQ	E	152.4k
DEFSurveySim [364]	✓		✓	HC	MCQ, RSQ	E	1.1k
EEC(Equity Evaluation Corpus) [365]	✓			EDC	BJ	E	8.6k
ETHICS [366]	✓		✓	CC	MCQ, BJ, RT	E	134.4k
Flames [367]	✓			CC	OEQ	A, E	1k
German Credit Data [368]				IRED	BJ, RT	A	1k
GlobalOpinionQA [369]	✓	✓		IRED	MCQ	A	2.5k
HofstedeCulturalDimensions [370]	✓			HC	MCQ	A, E	0.2k
IndieValueCatalog [371]	✓			IRED	MCQ	A	93.2k
KorNAT [372]	✓		✓	AC	MCQ	A, E	10k
LLMGlobe [373]	✓			HC	RSQ, OEQ	E	37.6k
LaWGPT [374]	✓			IRED	OEQ	A	300k
MFQ(Moral Foundations Questionnaire) [375]	✓			IRED	MCQ	E	11k
Moral Beliefs [376]	✓		✓	AC	MCQ, BJ, OEQ	E	0.5k
Moral Integrity Corpus [377]	✓			AC	RSQ, OEQ	E	38k
Moral Stories [378]	✓			CC	MCQ	A, E	12k
MoralExceptQA [379]	✓			EDC	RT	E	0.2k
NaVAB [380]		✓		AC	MCQ, BJ	E	63.8k
Persona Bias [365]	✓			IRED	MCQ, BJ	E	110.6k
PkuSafeRLHF [381]	✓		✓	AC	OEQ	A	476.4k
ProgressGym [382]	✓			HC	RSQ, OEQ	A	1.4k
SafeSora [383]	✓			AC	OEQ	A	14.7k
Scruples [384]	✓			EDC	MCQ, BJ	E	657k
Social Bias Frames [385]	✓			CC	RT, OEQ	E	150k
Social Chemistry 101 [386]	✓			CC	BJ, OEQ	E	292k
StereoSet [387]	✓			CC	RT, OEQ	E	4.2k
ToxiGen [388]	✓			AC	OEQ	E	6.5k
UnQover [389]	✓			IRED	BJ, OEQ	E	2713k
ValueNet [390]	✓			CC	BJ, OEQ	A	21.3k
WikiGenderBias [391]	✓			IRED	OEQ	E	45k
WinoBias [392]	✓			EDC	MCQ	E	3.1k
WinoGender [393]	✓			EDC	BJ, OEQ	E	0.7k

datasets not only scale well but also capture diverse perspectives from annotators of varied cultural and social backgrounds, thereby improving research efficiency and supporting a wider range of alignment and evaluation scenarios.

- **Integration and Refinement of Existing Datasets** involves aggregating and modifying multiple public sources. Notable examples include IndieValueCatalog [371], DecodingTrust [363], MFQ [375], Persona Bias [365], and GlobalOpinionQA [369], all constructed through careful curation and consolidation of open-source data. This approach reduces data acquisition costs while leveraging existing high-quality resources, thereby improving research efficiency and enabling broader evaluation coverage.

Automatic Construction. With the rapid advancement of LLMs, LLM-based automatic dataset construction has become a promising approach. This method typically involves designing prompt templates to guide the model in generating question-answer pairs or scenario-based judgments, followed by manual verification to ensure data quality. Representative datasets developed using this approach include NaVAB [380], KorNAT [372], CultureSPA [360], and

DailyDilemmas [362]. Compared to manual construction, it enables the creation of more diverse and context-rich data with significantly less human effort, allowing for large-scale dataset generation in a relatively short time. However, concerns remain regarding the quality of model-generated content. LLMs may produce hallucinated outputs, and their inherent biases can result in misalignment with human values. Consequently, ensuring the reliability and value alignment of automatically constructed datasets still requires careful oversight and validation.

Hybrid Construction. Hybrid construction approaches combine human expertise with automatic expansion using LLMs, which are increasingly emerging as a mainstream trend. In this paradigm, experts first create a small set of high-quality seed examples. The LLM is then employed to generate similar instances or expand the dataset, followed by human verification to ensure consistency and quality. Representative examples include CBBQ [356], LLMGlobe [373], and HofstedeCulturalDimensions [370], all of which utilize expert-curated prompts for initial task design, LLMs for large-scale generation, and human review for quality assurance. This hybrid methodology effectively leverages the generative capabilities of LLMs while maintaining the accuracy and

coherence provided by human oversight, offering a robust foundation for value alignment research and evaluation.

REFERENCES

- [1] Z. Xi *et al.*, "The rise and potential of large language model based agents: a survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, 2025. [Online]. Available: <https://doi.org/10.1007/s11432-024-4222-0>
- [2] H. Naveed *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.
- [3] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, vol. 10, 2024.
- [4] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenge," *arXiv preprint arXiv:2505.10468*, 2025.
- [5] D. Zhou, J. Zhang, T. Feng, and Y. Sun, "A survey on alignment for large language model agents," *UIUC Spring 2025 CS598 LLM Agent Workshop*, 2025.
- [6] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen, "Multi-agent collaboration mechanisms: A survey of llms," *arXiv preprint arXiv:2501.06322*, 2025.
- [7] T. Guo *et al.*, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.
- [8] L. Wang *et al.*, "A survey on large language model based autonomous agents," *Frontiers Comput. Sci.*, vol. 18, no. 6, p. 186345, 2024.
- [9] J. Ji *et al.*, "AI alignment: A comprehensive survey," *CoRR*, vol. abs/2310.19852, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2310.19852>
- [10] J. Yao, X. Yi, X. Wang, J. Wang, and X. Xie, "From instructions to intrinsic human values—a survey of alignment goals for big models," *arXiv preprint arXiv:2308.12014*, 2023.
- [11] T. Shen *et al.*, "Large language model alignment: A survey," *arXiv preprint arXiv:2309.15025*, 2023.
- [12] X. Yi and X. Xie, "Unpacking the ethical value alignment in big models," *Journal of Computer Research and Development*, vol. 60, no. 09, pp. 1926–1945, 2023.
- [13] Y. Liu *et al.*, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," *CoRR*, vol. abs/2308.05374, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.05374>
- [14] H. Shen *et al.*, "Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions," *CoRR*, vol. abs/2406.09264, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.09264>
- [15] X. Zeng, "Value alignment in ai large models: Current status, key issues, and normative strategies," *E-Government*, no. 02, pp. 34–44, 2025.
- [16] W. Hua *et al.*, "Trustagent: Towards safe and trustworthy llm-based agents through agent constitution," in *Trustworthy Multimodal Foundation Models and AI Agents (TiFA)*, 2024.
- [17] J. Liu *et al.*, "Large language model-based agents for software engineering: A survey," *arXiv preprint arXiv:2409.02977*, 2024.
- [18] Q. Sun, Y. Li, E. Alturki, S. M. K. Murthy, and B. W. Schuller, "Towards friendly ai: A comprehensive review and new perspectives on human-ai alignment," *arXiv preprint arXiv:2412.15114*, 2024.
- [19] E. Tenant, S. Hailes, and M. Musolesi, "Moral alignment for llm agents," *arXiv preprint arXiv:2410.01639*, 2024.
- [20] P. Dunleavy, *Digital era governance: IT corporations, the state, and e-government*. Oxford University Press, 2006.
- [21] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and machines*, vol. 30, no. 3, pp. 411–437, 2020.
- [22] L. Wang *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.
- [23] T. Guo *et al.*, "Large language model based multi-agents: A survey of progress and challenges," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024. ijcai.org*, 2024, pp. 8048–8057. [Online]. Available: <https://www.ijcai.org/proceedings/2024/890>
- [24] S. Chen, Y. Liu, W. Han, W. Zhang, and T. Liu, "A survey on llm-based multi-agent system: Recent advances and new frontiers in application," *Preprint at https://doi.org/10.48550/arXiv*, vol. 2412, 2025.
- [25] H. Du, S. Thudumu, R. Vasa, and K. Mouzakis, "A survey on context-aware multi-agent systems: techniques, challenges and future directions," *arXiv preprint arXiv:2402.01968*, 2024.
- [26] Z. Jiang *et al.*, "Casevo: A cognitive agents and social evolution simulator," *CoRR*, vol. abs/2412.19498, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2412.19498>
- [27] W. R. Catton Jr, "A theory of value," *American Sociological Review*, pp. 310–317, 1959.
- [28] L. Sagiv, S. Roccas, J. Cieciuch, and S. H. Schwartz, "Personal values in human life," *Nature human behaviour*, vol. 1, no. 9, pp. 630–639, 2017.
- [29] F. Adler, "The value concept in sociology," *American journal of sociology*, vol. 62, no. 3, pp. 272–279, 1956.
- [30] W. F. Ogburn, "Social change with respect to culture and original," *Nature*, pp. 240–5, 1922.
- [31] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2016.
- [32] J. Yao, X. Yi, X. Wang, J. Wang, and X. Xie, "From instructions to intrinsic human values - A survey of alignment goals for big models," *CoRR*, vol. abs/2308.12014, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.12014>
- [33] S. Russell, *Human compatible: AI and the problem of control*. Penguin UK, 2019.
- [34] B. Christian, "The alignment problem: Machine learning and human values," *Journal of the American Academy of Psychiatry and the Law Online*, vol. 52, no. 2, pp. 274–276, 2024. [Online]. Available: <https://jaapl.org/content/52/2/274>
- [35] G. Simmons, "Moral mimicry: Large language models produce moral rationalizations tailored to political identity," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2023, Toronto, Canada, July 9-14, 2023*, V. Padmakumar, G. Vallejo, and Y. Fu, Eds. Association for Computational Linguistics, 2023, pp. 282–297. [Online]. Available: <https://doi.org/10.18653/v1/2023.acl-srw.40>
- [36] D. Rozado, "The political biases of chatgpt," *Social Sciences*, vol. 12, no. 3, p. 148, 2023.
- [37] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 1236–1270. [Online]. Available: <https://doi.org/10.18653/v1/2023.findings-emnlp.88>
- [38] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
- [39] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545, 2022.
- [40] A. A. Khan *et al.*, "Ethics of ai: A systematic literature review of principles and challenges," in *Proceedings of the 26th international conference on evaluation and assessment in software engineering*, 2022, pp. 383–392.
- [41] H. Liu *et al.*, "Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation," *Information Processing & Management*, vol. 62, no. 4, p. 104099, 2025.
- [42] P. Ju, A. Ghosh, and N. Shroff, "Achieving fairness in multi-agent mdp using reinforcement learning," in *The Twelfth International Conference on Learning Representations*, 2023.
- [43] J. Ji *et al.*, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 678–24 704, 2023.
- [44] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Forty-first International Conference on Machine Learning*, 2023.
- [45] Y. Y. Chiu, L. Jiang, and Y. Choi, "Dailydilemmas: Revealing value preferences of llms with quandaries of daily life," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=PChiPGBf47>
- [46] J. Yao *et al.*, "Value compass leaderboard: A platform for fundamental and validated evaluation of llms values," *arXiv preprint arXiv:2501.07071*, 2025.

- [47] Y. Wu and C. Fan, "Shaping the ethical governance path of artificial intelligence in the chinese context—based on value-instrument rationality," *Studies in Science of Science*, 2025.
- [48] C. N. DeSmet and D. J. Cook, "Hydragran: A cooperative agent model for multi-objective data generation," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 60:1–60:21, 2024. [Online]. Available: <https://doi.org/10.1145/3653982>
- [49] B. Sun, Y. Zhou, and H. Jiang, "Empowering users in digital privacy management through interactive llm-based agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=FEpAUoS7f7>
- [50] G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, and R. Mihalcea, "Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/ca9567d8ef6b2ea2da0d7eed57b933ee-Abstract-Conference.html
- [51] R. Liu, R. Yang, C. Jia, G. Zhang, D. Yang, and S. Vosoughi, "Training socially aligned language models on simulated social interactions," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=NddKiWtdUm>
- [52] J. Dai *et al.*, "Safesora: Towards safety alignment of text2video generation via a human preference dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 17 161–17 214, 2024.
- [53] H. E. Degha and F. Z. Laallam, "ICA-CRMAS: intelligent context-awareness approach for citation recommendation based on multi-agent system," *ACM Trans. Manag. Inf. Syst.*, vol. 15, no. 3, pp. 13:1–13:52, 2024. [Online]. Available: <https://doi.org/10.1145/3680287>
- [54] T. R. Shaham *et al.*, "A multimodal automated interpretability agent," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=mDw42ZanmE>
- [55] Y. Roohani *et al.*, "Biodiscoveryagent: An ai agent for designing genetic perturbation experiments," *arXiv preprint arXiv:2405.17631*, 2024.
- [56] X. Yi, J. Yao, X. Wang, and X. Xie, "Unpacking the ethical value alignment in big models," *arXiv preprint arXiv:2310.17551*, 2023.
- [57] W. House, *Blueprint for an ai bill of rights: Making automated systems work for the american people*. Nimble Books, 2022.
- [58] N. Science and T. C. U. S. C. on Artificial Intelligence, *The national artificial intelligence research and development strategic plan: 2023 update*. National Science and Technology Council (US), Select Committee on Artificial . . . , 2023.
- [59] H. Ai, "High-level expert group on artificial intelligence," *Ethics guidelines for trustworthy AI*, vol. 6, 2019.
- [60] T. Madiega, "Artificial intelligence act," 2021.
- [61] C. A. of Sciences, "Measures for the ethical review of science and technology(trial)," 2023.
- [62] A. I. S. o. t. N. T. C. National Artificial Intelligence Standardization General Group, "Standardization of ai ethical governance (2023 edition)," 2023.
- [63] C. A. of China, "The interim measures for the administration of generative artificial intelligence services," 2023.
- [64] W. Holmes, F. Miao *et al.*, *Guidance for generative AI in education and research*. UNESCO Publishing, 2023.
- [65] W. H. Organization *et al.*, "Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models," in *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models*, 2024.
- [66] D. Melhart, J. Togelius, B. Mikkelsen, C. Holmgård, and G. N. Yannakakis, "The ethics of ai in games," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 79–92, 2023.
- [67] G. Hofstede, "Dimensionalizing cultures: The hofstede model in context," *Online readings in psychology and culture*, vol. 2, no. 1, p. 8, 2011.
- [68] Y. Wang *et al.*, "Cdeval: A benchmark for measuring the cultural dimensions of large language models," *arXiv preprint arXiv:2311.16421*, 2023.
- [69] J. Kharchenko, T. Roosta, A. Chadha, and C. Shah, "How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions," *arXiv preprint arXiv:2406.14805*, 2024.
- [70] E. Karinshak *et al.*, "Llm-globe: A benchmark evaluating the cultural values embedded in llm output," *arXiv preprint arXiv:2411.06032*, 2024.
- [71] K. Rab-Kettler and B. Lehnervp, "Recruitment in the times of machine learning," *Management Systems in Production Engineering*, 2019.
- [72] K. Simbeck, "Hr analytics and ethics," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 9–1, 2019.
- [73] A. L. Hunkenschroer and C. Luetge, "Ethics of ai-enabled recruiting and selection: A review and research agenda," *Journal of Business Ethics*, vol. 178, no. 4, pp. 977–1007, 2022.
- [74] J. Sun, C. Dai, Z. Luo, Y. Chang, and Y. Li, "Lawluo: A chinese law firm co-run by llm agents," *arXiv preprint arXiv:2407.16252*, 2024.
- [75] J. Mökander and L. Floridi, "Operationalising ai governance through ethics-based auditing: an industry case study," *AI and Ethics*, vol. 3, no. 2, pp. 451–468, 2023.
- [76] J. Graham *et al.*, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in experimental social psychology*. Elsevier, 2013, vol. 47, pp. 55–130.
- [77] S. H. Schwartz, "Basic human values: Theory, methods, and application," *Risorsa Uomo*, no. 2007/2, 2007.
- [78] J. M. Kivikangas, B. Fernández-Castilla, S. Järvelä, N. Ravaja, and J.-E. Lönnqvist, "Moral foundations and political orientation: Systematic review and meta-analysis," *Psychological bulletin*, vol. 147, no. 1, p. 55, 2021.
- [79] Y.-C. Chen, M. J. Ahn, and Y.-F. Wang, "Artificial intelligence and public values: value impacts and governance in the public sector," *Sustainability*, vol. 15, no. 6, p. 4796, 2023.
- [80] C. Djeffal, M. B. Siewert, and S. Würster, "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of ai strategies," *Journal of European Public Policy*, vol. 29, no. 11, pp. 1799–1821, 2022.
- [81] P. Bisson, R. Kirkland, and E. Stephenson, "The market state," *McKinsey quarterly*, 2010.
- [82] M. Mazzucato, "The entrepreneurial state," *Soundings*, vol. 49, no. 49, pp. 131–142, 2011.
- [83] J. Braithwaite, "217the regulatory state?" in *The Oxford Handbook of Political Science*. Oxford University Press, 07 2011. [Online]. Available: <https://doi.org/10.1093/oxfordhb/9780199604456.013.0011>
- [84] G. Majone, "From the positive to the regulatory state: Causes and consequences of changes in the mode of governance," *Journal of public policy*, vol. 17, no. 2, pp. 139–167, 1997.
- [85] S. Borrás and J. Edler, "The roles of the state in the governance of socio-technical systems' transformation," *Research Policy*, vol. 49, no. 5, p. 103971, 2020.
- [86] P. Van Esch and J. S. Black, "Factors that influence new generation candidates to engage with and complete digital, ai-enabled recruiting," *Business horizons*, vol. 62, no. 6, pp. 729–739, 2019.
- [87] T. Chamorro-Premuzic, F. Polli, and B. Dattner, "Building ethical ai for talent management," *Harvard Business Review*, vol. 21, no. November, pp. 1–15, 2019.
- [88] D. D. Savage and R. Bales, "Video games in job interviews: Using algorithms to minimize discrimination and unconscious bias," *ABAJ Lab. & Emp. L.*, vol. 32, p. 211, 2016.
- [89] G. Mann and C. O'Neil, "Hiring algorithms are not neutral," *Harvard Business Review*, vol. 9, p. 2016, 2016.
- [90] F. Polli, "Using ai to eliminate bias from hiring," *Harvard Business Review*, vol. 29, 2019.
- [91] P. T. Kim, "Data-driven discrimination at work," *Wm. & Mary L. Rev.*, vol. 58, p. 857, 2016.
- [92] B. Dattner, T. Chamorro-Premuzic, R. Buchband, and L. Schettler, "The legal and ethical implications of using ai in hiring," *Harvard Business Review*, vol. 25, pp. 1–7, 2019.
- [93] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, 2019.
- [94] J. Wang *et al.*, "GTA: A benchmark for general tool agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/

- 8a75ee6d4b2eb0b777f549a32a5a5c28-Abstract-Datasets_and_Benchmarks_Track.html
- [95] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 10471–10506. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.624>
- [96] X. Tang *et al.*, "Medagents: Large language models as collaborators for zero-shot medical reasoning," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 599–621. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.33>
- [97] Y. H. Roohani *et al.*, "Biodiscoveryagent: An AI agent for designing genetic perturbation experiments," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=HAWzGLcye3>
- [98] F. P. Lo *et al.*, "AI hiring with llms: A context-aware and explainable multi-agent framework for resume screening," *CoRR*, vol. abs/2504.02870, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2504.02870>
- [99] D. Mujtaba and N. R. Mahapatra, "Ethical considerations in ai-based recruitment," in *2019 IEEE International Symposium on Technology and Society, ISTAS 2019, Medford, MA, USA, November 15-16, 2019*. IEEE, 2019, pp. 1–7. [Online]. Available: <https://doi.org/10.1109/ISTAS48451.2019.8937920>
- [100] C. Zheng *et al.*, "On prompt-driven safeguarding for large language models," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 61593–61613.
- [101] P. Maini *et al.*, "Safety pretraining: Toward the next generation of safe ai," *arXiv preprint arXiv:2504.16980*, 2025.
- [102] T. Korbak *et al.*, "Pretraining language models with human preferences," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17506–17533.
- [103] Y. Gu, X. Han, Z. Liu, and M. Huang, "Ppt: Pre-trained prompt tuning for few-shot learning. arxiv 2021," *arXiv preprint arXiv:2109.04332*, 2021.
- [104] S. Chen, "Mitigating bias in large language models: A multi-task training approach using bert," *Applied and Computational Engineering*, vol. 105, pp. 30–37, 2024.
- [105] S. Prabhummoye, M. Patwary, M. Shoeybi, and B. Catanzaro, "Adding instructions during pretraining: Effective way of controlling toxicity in language models," *arXiv preprint arXiv:2302.07388*, 2023.
- [106] A. Biswas and W. Talukdar, "Guardrails for trust, safety, and ethical development and deployment of large language models (llm)," *Journal of Science & Technology*, vol. 4, no. 6, pp. 55–82, 2023.
- [107] D. Campos, "Curriculum learning for language modeling," *arXiv preprint arXiv:2108.02170*, 2021.
- [108] J. Zhao, Z. Deng, D. Madras, J. Zou, and M. Ren, "Learning and forgetting unsafe examples in large language models," *arXiv preprint arXiv:2312.12736*, 2023.
- [109] Y. Wang *et al.*, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.
- [110] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
- [111] Z. Sun *et al.*, "Principle-driven self-alignment of language models from scratch with minimal human supervision," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2511–2565, 2023.
- [112] H. Liu, C. Sferrazza, and P. Abbeel, "Chain of hindsight aligns language models with feedback," *arXiv preprint arXiv:2302.02676*, 2023.
- [113] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.
- [114] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [115] S. Kim *et al.*, "Aligning large language models through synthetic feedback," *arXiv preprint arXiv:2305.13735*, 2023.
- [116] W. Saunders *et al.*, "Self-critiquing models for assisting human evaluators," *arXiv preprint arXiv:2206.05802*, 2022.
- [117] D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, and M. Dymetman, "Aligning language models with preferences through f-divergence minimization," *arXiv preprint arXiv:2302.08215*, 2023.
- [118] R. Liu *et al.*, "Training socially aligned language models in simulated human society," *arXiv preprint arXiv:2305.16960*, vol. 2, 2023.
- [119] Y. Bai *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.
- [120] S. Huang *et al.*, "Collective constitutional ai: Aligning a language model with public input," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1395–1417.
- [121] E. Noh and J. Baek, "Toward responsible federated large language models: Leveraging a safety filter and constitutional ai," *arXiv preprint arXiv:2502.16691*, 2025.
- [122] X. Chen *et al.*, "Iteralign: Iterative constitutional alignment of large language models," *arXiv preprint arXiv:2403.18341*, 2024.
- [123] A. Küsters and M. Wörsdörfer, "Exploring laws of robotics: A synthesis of constitutional ai and constitutional economics," *Digital Society*, vol. 4, no. 2, p. 46, 2025.
- [124] A. Madaan *et al.*, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46534–46594, 2023.
- [125] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," 2024.
- [126] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.
- [127] L. Ranaldi and A. Freitas, "Self-refine instruction-tuning for aligning reasoning in language models," *arXiv preprint arXiv:2405.00402*, 2024.
- [128] G. Gao, A. Taymanov, E. Salinas, P. Mineiro, and D. Misra, "Aligning llm agents by learning latent preference from user edits," *Advances in Neural Information Processing Systems*, vol. 37, pp. 136873–136896, 2024.
- [129] S. Hao and L. Duan, "Online learning from strategic human feedback in llm fine-tuning," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [130] H. Su *et al.*, "A survey on autonomy-induced security risks in large model-based agents," *arXiv preprint arXiv:2506.23844*, 2025.
- [131] Y. Li *et al.*, "Personal llm agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024.
- [132] P. Bhandari *et al.*, "Can llm agents maintain a persona in discourse?" *arXiv preprint arXiv:2502.11843*, 2025.
- [133] A. B. Hassouna, H. Chaari, and I. Belhaj, "Llm-agent-umf: Llm-based agent unified modeling framework for seamless integration of multi active/passive core-agents," *arXiv preprint arXiv:2409.11393*, 2024.
- [134] X. Li, F. Huang, J. Lv, Z. Xiao, G. Li, and Y. Yue, "Be more real: Travel diary generation using llm agents and individual profiles," *arXiv preprint arXiv:2407.18932*, 2024.
- [135] T. Cerqueira and P. Bezerra, "Motif: A framework for enhancing the profiling module of generative agents that simulate human behavior."
- [136] S. T. Erukude, S. R. Veluru, and V. C. Marella, "Agentic ai-the rise of autonomous intelligent agents in the era of llms."
- [137] C. Royse and B. Caudill, "The cognitive core: An integrated cognitive architecture."
- [138] Z. Li *et al.*, "Memos: An operating system for memory-augmented generation (mag) in large language models," *arXiv preprint arXiv:2505.22101*, 2025.
- [139] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges," *arXiv preprint arXiv:2505.10468*, 2025.
- [140] G. Zhang, M. Fu, G. Wan, M. Yu, K. Wang, and S. Yan, "G-memory: Tracing hierarchical memory for multi-agent systems," *arXiv preprint arXiv:2506.07398*, 2025.
- [141] A. Vishwakarma *et al.*, "Cognitive weave: Synthesizing abstracted knowledge with a spatio-temporal resonance graph," *arXiv preprint arXiv:2506.08098*, 2025.
- [142] Z. Wang *et al.*, "A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications."

- [143] W. Xu, C. Huang, S. Gao, and S. Shang, "Llm-based agents for tool learning: A survey," *Data Science and Engineering*, pp. 1–31, 2025.
- [144] T. Chakraborty *et al.*, "Heal: An empirical study on hallucinations in embodied agents driven by large language models," *arXiv preprint arXiv:2506.15065*, 2025.
- [145] S. Holt, M. R. Luyten, T. Pouplin, and M. van der Schaar, "Improving llm agent planning with in-context learning via atomic fact augmentation and lookahead search," *arXiv preprint arXiv:2506.09171*, 2025.
- [146] Q. Zhang, M. Wornow, and K. Olukotun, "Cost-efficient serving of llm agents via test-time plan caching," *arXiv preprint arXiv:2506.14852*, 2025.
- [147] S. Nandi *et al.*, "Sop-bench: Complex industrial sops for evaluating llm agents," *arXiv preprint arXiv:2506.08119*, 2025.
- [148] Q. Yin *et al.*, "Wgsr-bench: Wargame-based game-theoretic strategic reasoning benchmark for large language models," *arXiv preprint arXiv:2506.10264*, 2025.
- [149] C. Li *et al.*, "Stride: A tool-assisted llm agent framework for strategic and interactive decision-making," *arXiv preprint arXiv:2405.16376*, 2024.
- [150] W. Zeng *et al.*, "Application-driven value alignment in agentic ai systems: Survey and perspectives," *arXiv preprint arXiv:2506.09656*, 2025.
- [151] D. Zhou, J. Zhang, T. Feng, and Y. Sun, "A survey on alignment for large language model agents," in *UIUC Spring 2025 CS598 LLM Agent Workshop*, 2025.
- [152] T. Pitkäranta and L. Pitkäranta, "Hada: Human-ai agent decision alignment architecture," *arXiv preprint arXiv:2506.04253*, 2025.
- [153] Z. Han *et al.*, "Coreagents: A collaboration and reasoning framework based on llm-powered agents for complex reasoning tasks," *Applied Sciences*, vol. 15, no. 10, p. 5663, 2025.
- [154] Y. Huang *et al.*, "A framework for benchmarking and aligning task-planning safety in llm-based embodied agents," *arXiv preprint arXiv:2504.14650*, 2025.
- [155] H. Wang, A. Zhang, N. Duy Tai, J. Sun, T.-S. Chua *et al.*, "Ali-agent: Assessing llms' alignment with human values via agent-based evaluation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 99 040–99 088, 2024.
- [156] Z. Wang *et al.*, "A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications."
- [157] X. Dong, X. Zhang, W. Bu, D. Zhang, and F. Cao, "A survey of llm-based agents: Theories, technologies, applications and suggestions," in *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTc)*. IEEE, 2024, pp. 407–413.
- [158] X. Zhou *et al.*, "Exploring the necessity of reasoning in llm-based agent scenarios," *arXiv preprint arXiv:2503.11074*, 2025.
- [159] F. Herrera, "Making sense of the unsensible: Reflection, survey, and challenges for xai in large language models toward human-centered ai," *arXiv preprint arXiv:2505.20305*, 2025.
- [160] W. Zhang *et al.*, "Agent-pro: Learning to evolve via policy-level reflection and optimization," *arXiv preprint arXiv:2402.17574*, 2024.
- [161] T. Pitkäranta and L. Pitkäranta, "Hada: Human-ai agent decision alignment architecture," *arXiv preprint arXiv:2506.04253*, 2025.
- [162] A. Naik *et al.*, "Agentmisalignment: Measuring the propensity for misaligned behaviour in llm-based agents," *arXiv preprint arXiv:2506.04018*, 2025.
- [163] M. A. Ferrag, N. Tihanyi, and M. Debbah, "From llm reasoning to autonomous ai agents: A comprehensive review," *arXiv preprint arXiv:2504.19678*, 2025.
- [164] R. Jiang *et al.*, "A survey on human preference learning for large language models," *arXiv preprint arXiv:2406.11191*, 2024.
- [165] Z. Sun, "Scalable alignment of large language models towards truth seeking, complex reasoning, and human values," Ph.D. dissertation, Carnegie Mellon University, 2025.
- [166] H. Peng, Y. Qi, X. Wang, Z. Yao, and L. Hou, "Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems," *arXiv preprint arXiv:2502.19328*, 2025. [Online]. Available: <https://arxiv.org/abs/2502.19328>
- [167] H. Wang, A. Zhang, N. Duy Tai, J. Sun, T.-S. Chua *et al.*, "Ali-agent: Assessing llms' alignment with human values via agent-based evaluation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 99 040–99 088, 2024.
- [168] Y. Y. Sung, H. Kim, and D. Zhang, "Verila: A human-centered evaluation framework for interpretable verification of llm agent failures," *arXiv preprint arXiv:2503.12651*, 2025.
- [169] E. Tennant, S. Hailes, and M. Musolesi, "Moral alignment for llm agents," *arXiv preprint arXiv:2410.01639*, 2024.
- [170] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, p. 9, 2024.
- [171] Y. Li *et al.*, "Aligning individual and collective objectives in multi-agent cooperation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 44 735–44 760, 2024.
- [172] M. Garcia-Bohigues, C. Cordova, J. Taverner, J. Palanca, E. del Val, and E. Argente, "Towards distributed platform for normative reasoning and value alignment in multi-agent systems," in *International Workshop on Value Engineering in AI*. Springer, 2023, pp. 237–250.
- [173] S. Song *et al.*, "Code: Communication delay-tolerant multi-agent collaboration via dual alignment of intent and timeliness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 304–23 312.
- [174] Z. Song, L. Yang, F. Wen, and J. Li, "Traf-align: Trajectory-aware feature alignment for asynchronous multi-agent perception," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 048–12 057.
- [175] Y. Yang, Y. Ma, H. Feng, Y. Cheng, and Z. Han, "Minimizing hallucinations and communication costs: Adversarial debate and voting mechanisms in llm-based multi-agents," *Applied Sciences*, vol. 15, no. 7, p. 3676, 2025.
- [176] E. Waisberg *et al.*, "Gpt-4: a new era of artificial intelligence in medicine," *Irish Journal of Medical Science* (1971-), vol. 192, no. 6, pp. 3197–3200, 2023.
- [177] W. N. Caballero and P. R. Jenkins, "On large language models in national security applications," *Stat*, vol. 14, no. 2, p. e70057, 2025.
- [178] Y. Li *et al.*, "Improving multi-agent debate with sparse communication topology," *arXiv preprint arXiv:2406.11776*, 2024.
- [179] R. Zou, M. Wei, J. Feng, Q. Wan, J. Sun, and S. Liu, "Gradual vigilance and interval communication: Enhancing value alignment in multi-agent debates," *arXiv preprint arXiv:2412.13471*, 2024.
- [180] A. R. Ortigoso, G. Vieira, D. Fuentes, L. Frazão, N. Costa, and A. Pereira, "Project riley: Multimodal multi-agent llm collaboration with emotional reasoning and voting," *arXiv preprint arXiv:2505.20521*, 2025.
- [181] D. Ki, R. Rudinger, T. Zhou, and M. Carpuat, "Multiple llm agents debate for equitable cultural alignment," *arXiv preprint arXiv:2505.24671*, 2025.
- [182] J. Ma, C. Wang, L. Rong, B. Wang, and Y. Xu, "Exploring multi-agent debate for zero-shot stance detection: A novel approach," *Applied Sciences*, vol. 15, no. 9, p. 4612, 2025.
- [183] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for" mind" exploration of large language model society," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 991–52 008, 2023.
- [184] S. Hong *et al.*, "Metagpt: Meta programming for multi-agent collaborative framework," *arXiv preprint arXiv:2308.00352*, vol. 3, no. 4, p. 6, 2023.
- [185] Z. Hou, J. Tang, and Y. Wang, "Halo: Hierarchical autonomous logic-oriented orchestration for multi-agent llm systems," *arXiv preprint arXiv:2505.13516*, 2025.
- [186] Z. Wang, S. Moriyama, W.-Y. Wang, B. Gangopadhyay, and S. Takamatsu, "Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems," *arXiv preprint arXiv:2502.11098*, 2025.
- [187] T. Händler, "Balancing autonomy and alignment: a multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures," *arXiv preprint arXiv:2310.03659*, 2023.
- [188] F. Carichon, A. Khandelwal, M. Fauchard, and G. Farnadi, "The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process," *arXiv preprint arXiv:2506.01080*, 2025.
- [189] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv preprint arXiv:2010.03768*, 2020.
- [190] M. F. A. R. D. T. (FAIR)† *et al.*, "Human-level play in the game of diplomacy by combining language models with strategic reasoning," *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.

- [191] A. Parrish *et al.*, "Bbq: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.
- [192] D. Hendrycks *et al.*, "Aligning ai with shared human values," in *International Conference on Learning Representations*, 2021.
- [193] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 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)*, 2021, pp. 5356–5371.
- [194] Z. Jin *et al.*, "When to make exceptions: Exploring language models as accounts of human moral judgment," *Advances in neural information processing systems*, vol. 35, pp. 28458–28473, 2022.
- [195] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, and D. Yang, "The moral integrity corpus: A benchmark for ethical dialogue systems," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3755–3773.
- [196] J. Dhamala *et al.*, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 862–872.
- [197] C. Mu *et al.*, "Multi-agent, human-agent and beyond: A survey on cooperation in social dilemmas," *Neurocomputing*, vol. 610, p. 128514, 2024. [Online]. Available: <https://doi.org/10.1016/j.neucom.2024.128514>
- [198] V. Conitzer and C. Oesterheld, "Foundations of cooperative AI," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, B. Williams, Y. Chen, and J. Neville, Eds. AAAI Press, 2023, pp. 15359–15367. [Online]. Available: <https://doi.org/10.1609/aaai.v37i13.26791>
- [199] W.-L. Chiang *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," See <https://vicuna.lmsys.org> (accessed 14 April 2023), vol. 2, no. 3, p. 6, 2023.
- [200] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2303.08774>
- [201] Y. Wang *et al.*, "Recmind: Large language model powered agent for recommendation," in *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16–21, 2024*, K. Duh, H. Gómez-Adorno, and S. Bethard, Eds. Association for Computational Linguistics, 2024, pp. 4351–4364. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-naacl.271>
- [202] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent LLM agents," *CoRR*, vol. abs/2306.03314, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2306.03314>
- [203] C. Frith and U. Frith, "Theory of mind," *Current biology*, vol. 15, no. 17, pp. R644–R645, 2005.
- [204] S. Triantafyllou, A. Singla, and G. Radanovic, "On blame attribution for accountable multi-agent sequential decision making," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 15774–15786. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/848c4965359e617d5e16c924b4a85fd9-Abstract.html>
- [205] S. P. Robbins, R. Bergman, I. Stagg, and M. Coulter, *Management*. Pearson Australia, 2014.
- [206] H. Mintzberg, *The structuring of organizations*. Springer, 1989.
- [207] T. Burns and G. Stalker, "Mechanistic and organic systems," *London: Tavistock Publications*, 1961.
- [208] Y. Yang *et al.*, "Agentnet: Decentralized evolutionary coordination for llm-based multi-agent systems," *CoRR*, vol. abs/2504.00587, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2504.00587>
- [209] J. Wang *et al.*, "GTA: A benchmark for general tool agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 – 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/dec42d78c42cea59c95c7c3d40d5e0f-Abstract-Conference.html 8a75ee6d4b2eb0b777f549a32a5a5c28-Abstract-Datasets_and_Benchmarks_Track.html
- [210] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11–16, 2024*, L. Ku, A. Martins, and V. Srikanth, Eds. Association for Computational Linguistics, 2024, pp. 10471–10506. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.624>
- [211] A. V. Savchenko and L. V. Savchenko, "Inside out: Emotional multiagent multimodal dialogue systems," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3–9, 2024*, ijcai.org, 2024, pp. 8784–8788. [Online]. Available: <https://www.ijcai.org/proceedings/2024/1032>
- [212] K. Mishra, M. Firdaus, and A. Ekbal, "Please donate to save a life: Inducing politeness to handle resistance in persuasive dialogue agents," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2202–2212, 2024. [Online]. Available: <https://doi.org/10.1109/TASLP.2024.3357032>
- [213] N. Crispino, K. Montgomery, F. Zeng, D. Song, and C. Wang, "Agent instructs large language models to be general zero-shot reasoners," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=zMwFvxr6CV>
- [214] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=zj7YuTE4t8>
- [215] J. C. Chen, S. Saha, E. Stengel-Eskin, and M. Bansal, "Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=ffLbLkoCw8>
- [216] L. Wang *et al.*, "User behavior simulation with large language model-based agents," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 55:1–55:37, 2025. [Online]. Available: <https://doi.org/10.1145/3708985>
- [217] H. Lai *et al.*, "Autowebglm: A large language model-based web navigating agent," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25–29, 2024*, R. Baeza-Yates and F. Bonchi, Eds. ACM, 2024, pp. 5295–5306. [Online]. Available: <https://doi.org/10.1145/3637528.3671620>
- [218] Y. Fu *et al.*, "Autoguide: Automated generation and selection of context-aware guidelines for large language model agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 – 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/d8efbb5dd415974eb095c3f06bbff1f48-Abstract-Conference.html
- [219] J. Zheng, H. Wang, A. Zhang, T. D. Nguyen, J. Sun, and T. Chua, "Ali-agent: Assessing llms' alignment with human values via agent-based evaluation," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 – 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/b35c38f70065ac6c694089ca93a015bb-Abstract-Conference.html
- [220] T. R. Shaham *et al.*, "A multimodal automated interpretability agent," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=mDw42ZannE>
- [221] Y. Ning and H. Liu, "Urbankgent: A unified large language model agent framework for urban knowledge graph construction," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 – 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/decd42d78c42cea59c95c7c3d40d5e0f-Abstract-Conference.html

- [222] H. E. Degha and F. Z. Laallam, "ICA-CRMAS: intelligent context-awareness approach for citation recommendation based on multi-agent system," *ACM Trans. Manag. Inf. Syst.*, vol. 15, no. 3, pp. 13:1–13:52, 2024. [Online]. Available: <https://doi.org/10.1145/3680287>
- [223] Z. Hu *et al.*, "OSDA agent: Leveraging large language models for de novo design of organic structure directing agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=9YNyiCJE3k>
- [224] X. Tang *et al.*, "Medagents: Large language models as collaborators for zero-shot medical reasoning," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 599–621. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.33>
- [225] Y. H. Roohani *et al.*, "Biodiscoveryagent: An AI agent for designing genetic perturbation experiments," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=HAwZGLcye3>
- [226] N. Li, C. Gao, M. Li, Y. Li, and Q. Liao, "Econagent: Large language model-empowered agents for simulating macroeconomic activities," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15523–15536. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.829>
- [227] W. Zhang *et al.*, "A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, R. Baeza-Yates and F. Bonchi, Eds. ACM, 2024, pp. 4314–4325. [Online]. Available: <https://doi.org/10.1145/3637528.3671801>
- [228] Y. Yu *et al.*, "Fincon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/f7ae4fe91d96f50abc2211f09b6a7e49-Abstract-Conference.html
- [229] X. Lin, C. Jia, J. Ji, H. Han, and U. Naseem, "Ask, acquire, understand: A multimodal agent-based framework for social abuse detection in memes," in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025-2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 4734–4744. [Online]. Available: <https://doi.org/10.1145/3696410.3714895>
- [230] J. Ma, L. Hu, R. Li, and W. Fu, "Local: Logical and causal fact-checking with llm-based multi-agents," in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025-2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 1614–1625. [Online]. Available: <https://doi.org/10.1145/3696410.3714748>
- [231] Z. Gao *et al.*, "Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=0bmGL4q7vJ>
- [232] X. Gu *et al.*, "Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=DYMj03Gbri>
- [233] R. Cao *et al.*, "Spider2-v: How far are multimodal agents from automating data science and engineering workflows?" in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/
- c2f71567cd53464161cab3336e8fc865-Abstract-Datasets_and_Benchmarks_Track.html
- [234] Y. Zhou, A. Zanette, J. Pan, S. Levine, and A. Kumar, "Archer: Training language model agents via hierarchical multi-turn RL," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=b6rA0kAHT1>
- [235] Y. Wang, J. Liu, Y. Wan, Y. Li, Z. Liu, and W. Chen, "KEEP chatting! an attractive dataset for continuous conversation agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 16408–16414. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.972>
- [236] Z. Pan *et al.*, "Secom: On memory construction and retrieval for personalized conversational agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=xKDZAW0He3>
- [237] Y. Chen *et al.*, "STYLE: improving domain transferability of asking clarification questions in large language model powered conversational agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 10633–10649. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.632>
- [238] H. Yan *et al.*, "Talk with human-like agents: Empathetic dialogue through perceptible acoustic reception and reaction," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15009–15022. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.801>
- [239] P. Feng *et al.*, "AGILE: A novel reinforcement learning framework of LLM agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/097c514162ea7126d40671d23e12f51b-Abstract-Conference.html
- [240] Y. Guan *et al.*, "Intelligent agents with llm-based process automation," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, R. Baeza-Yates and F. Bonchi, Eds. ACM, 2024, pp. 5018–5027. [Online]. Available: <https://doi.org/10.1145/3637528.3671646>
- [241] X. Wang *et al.*, "Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 1840–1873. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.102>
- [242] Y. Deng, W. Zhang, W. Lam, S. Ng, and T. Chua, "Plug-and-play policy planner for large language model powered dialogue agents," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=MCNqgUFTHI>
- [243] I. Zahid, T. Madusanka, R. Batista-Navarro, and Y. Sun, "Probing the uniquely identifiable linguistic patterns of conversational AI agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 4612–4628. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.274>
- [244] H. Chen *et al.*, "Socialbench: Sociality evaluation of role-playing conversational agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 2108–2126. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.125>
- [245] Y. Yang, P. Achananuparp, H. Huang, J. Jiang, and E. Lim,

- "Speaker verification in agent-generated conversations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 5655–5676. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.307>
- [246] J. Yue, X. Xu, B. F. Karlsson, and Z. Lu, "MLLM as retriever: Interactively learning multimodal retrieval for embodied agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=K5yeB4dTts>
- [247] Z. Yang, G. Chen, X. Li, W. Wang, and Y. Yang, "Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as A video agent)," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=QMy2RLnxGN>
- [248] Y. Huang *et al.*, "Adasociety: An adaptive environment with social structures for multi-agent decision-making," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/3e4d8407cb468850f2f8f4a949e64bf0-Abstract-Datasets_and_Benchmarks_Track.html
- [249] J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, and S. Deng, "Exploring collaboration mechanisms for LLM agents: A social psychology view," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 14544–14607. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.782>
- [250] J. Liu *et al.*, "Maximum entropy heterogeneous-agent reinforcement learning," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=tmqOhBC4a5>
- [251] Q. Fu, Z. Pu, Y. Pan, T. Qiu, and J. Yi, "Fuzzy feedback multiagent reinforcement learning for adversarial dynamic multiteam competitions," *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 5, pp. 2811–2824, 2024. [Online]. Available: <https://doi.org/10.1109/TFUZZ.2024.3363053>
- [252] Z. Ding, W. Zhang, J. Yue, X. Wang, T. Huang, and Z. Lu, "Entity divider with language grounding in multi-agent reinforcement learning," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 8103–8119. [Online]. Available: <https://proceedings.mlr.press/v202/ding23d.html>
- [253] Y. Gui, Z. Zhang, D. Tang, H. Zhu, and Y. Zhang, "Collaborative dynamic scheduling in a self-organizing manufacturing system using multi-agent reinforcement learning," *Adv. Eng. Informatics*, vol. 62, p. 102646, 2024. [Online]. Available: <https://doi.org/10.1016/j.aei.2024.102646>
- [254] C. Qian *et al.*, "Experiential co-learning of software-developing agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 5628–5640. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.305>
- [255] J. Yang *et al.*, "Swe-agent: Agent-computer interfaces enable automated software engineering," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html
- [256] Y. Liang, D. Chen, A. Gupta, S. S. Du, and N. Jaques, "Learning to cooperate with humans using generative agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/6e9f99809b6dd59e8fe5c9ffe269bf40-Abstract-Conference.html
- [257] S. Nayak *et al.*, "Long-horizon planning for multi-agent robots in partially observable environments," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/7d6e85e88495104442af94c98e899659-Abstract-Conference.html
- [258] S. Gao *et al.*, "360°rea: Towards A reusable experience accumulation with 360° assessment for multi-agent system," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 13149–13162. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.778>
- [259] H. Rong, M. Qian, T. Ma, D. Jin, and V. S. Sheng, "Cobjeason: Reasoning covered object in image by multi-agent collaboration based on informed knowledge graph," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 5, pp. 116:1–116:56, 2024. [Online]. Available: <https://doi.org/10.1145/3643565>
- [260] Z. Wang, A. Li, Z. Li, and X. Liu, "Genartist: Multimodal LLM as an agent for unified image generation and editing," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/e7c786024ca718f2487712bfe9f51030-Abstract-Conference.html
- [261] H. Chen *et al.*, "Restoreagent: Autonomous image restoration agent via multimodal large language models," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/c78f639424b8d89ceb4f2efbb4dfe4f4-Abstract-Conference.html
- [262] Z. Hu *et al.*, "Scenecraft: An LLM agent for synthesizing 3d scenes as blender code," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=gAyzjHw2ml>
- [263] Y. Yao, P. Yang, G. Zhao, and G. Yin, "Kgagent: Learning a deep reinforced agent for keyphrase generation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1928–1940, 2024. [Online]. Available: <https://doi.org/10.1109/TASLP.2024.3375630>
- [264] K. Lee, X. Chen, H. Furuta, J. F. Cannby, and J. Fischer, "A human-inspired reading agent with gist memory of very long contexts," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=OTmcsyEO5G>
- [265] Y. Jiao, S. Li, S. Zhou, H. Ji, and J. Han, "Text2db: Integration-aware information extraction with large language model agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 185–205. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.12>
- [266] S. Ojuri, T. A. Han, R. Chiong, and A. D. Stefano, "Optimizing text-to-sql conversion techniques through the integration of intelligent agents and large language models," *Inf. Process. Manag.*, vol. 62, no. 5, p. 104136, 2025. [Online]. Available: <https://doi.org/10.1016/j.ipm.2025.104136>
- [267] Y. Liu *et al.*, "Xmc-agent : Dynamic navigation over scalable hierarchical index for incremental extreme multi-label classification," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 5659–5672. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.336>
- [268] T. Xu, H. Yang, F. Zhao, Z. Wu, and X. Dai, "A two-agent game for zero-shot relation triplet extraction," in *Findings of the Association for Computational Linguistics, ACL 2024*,

- Bangkok, Thailand and virtual meeting, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 7510–7527. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.446>
- [269] Y. Guo, S. Zhuang, K. Li, Y. Qiao, and Y. Wang, “Transagent: Transfer vision-language foundation models with heterogeneous agent collaboration,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/b235f0b417e8bf270c0cb19fe0b82c1e-Abstract-Conference.html
- [270] Z. Hu *et al.*, “AVIS: autonomous visual information seeking with large language model agent,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/029df12a9363313c3e41047844cad94-Abstract-Conference.html
- [271] J. Cheng and P. Chin, “Sociodojo: Building lifelong analytical agents with real-world text and time series,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=s9z0HzWJjp>
- [272] S. Qiao *et al.*, “Agent planning with world knowledge model,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/d032263772946dd5026e7f3cd22bce5b-Abstract-Conference.html
- [273] M. Chen, Y. Li, Y. Yang, S. Yu, B. Lin, and X. He, “Automanual: Constructing instruction manuals by LLM agents via interactive environmental learning,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/0142921fad7ef9192bd87229cdafa9d4-Abstract-Conference.html
- [274] D. Hillier, C. Tan, and J. Jiang, “Social learning through interactions with other agents: A survey,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 8067–8076. [Online]. Available: <https://www.ijcai.org/proceedings/2024/892>
- [275] B. Kim, M. Seo, and J. Choi, “Online continual learning for interactive instruction following agents,” in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=7M0EzjugaN>
- [276] J. Xie *et al.*, “Travelplanner: A benchmark for real-world planning with language agents,” in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=l5XQzNkAOe>
- [277] Y. Song, D. Yin, X. Yue, J. Huang, S. Li, and B. Y. Lin, “Trial and error: Exploration-based trajectory optimization of LLM agents,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 7584–7600. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.409>
- [278] T. R. Sumers, K. Marino, A. Ahuja, R. Fergus, and I. Dasgupta, “Distilling internet-scale vision-language models into embodied agents,” in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 32797–32818. [Online]. Available: <https://proceedings.mlr.press/v202/sumers23a.html>
- [279] W. Huang *et al.*, “Grounded decoding: Guiding text generation with grounded models for embodied agents,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/bb3fcfb0284642a973dd631ec9184f2f-Abstract-Conference.html
- [280] G. Gonzalez-Pumariega, L. S. Yean, N. Sunkara, and S. Choudhury, “Robotouille: An asynchronous planning benchmark for LLM agents,” in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=OhUoTMxFIH>
- [281] Z. Xu, C. Yu, F. Fang, Y. Wang, and Y. Wu, “Language agents with reinforcement learning for strategic play in the werewolf game,” in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=usUPvQH3XK>
- [282] Z. Wang *et al.*, “Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/85f1225db986e629289f402c46eff1a4-Abstract-Conference.html
- [283] M. Xing, R. Zhang, H. Xue, Q. Chen, F. Yang, and Z. Xiao, “Understanding the weakness of large language model agents within a complex android environment,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, R. Baeza-Yates and F. Bonchi, Eds. ACM, 2024, pp. 6061–6072. [Online]. Available: <https://doi.org/10.1145/3637528.3671650>
- [284] M. Li *et al.*, “Embodied agent interface: Benchmarking llms for embodied decision making,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html
- [285] Z. Shi *et al.*, “Tool learning in the wild: Empowering language models as automatic tool agents,” in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 2222–2237. [Online]. Available: <https://doi.org/10.1145/3696410.3714825>
- [286] B. Y. Lin *et al.*, “Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/4b0eea69deea512c9e2c469187643dc2-Abstract-Conference.html
- [287] W. Li *et al.*, “On the effects of data scale on UI control agents,” in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/a79f3ef3b445fd4659f44648f7ea8ffd-Abstract-Datasets_and_Benchmarks_Track.html
- [288] M. Kim, V. S. Bursztyn, E. Koh, S. Guo, and S. Hwang, “Rada: Retrieval-augmented web agent planning with llms,” in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 13511–13525. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.802>
- [289] C. Shao, X. Hu, Y. Lin, and F. Xu, “Division-of-thoughts: Harnessing hybrid language model synergy for efficient on-device agents,” in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 1822–1833. [Online]. Available: <https://doi.org/10.1145/3696410.3714765>
- [290] M. A. Islam, M. E. Ali, and M. R. Parvez, “Mapcoder: Multi-agent code generation for competitive problem solving,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*,

- Bangkok, Thailand, August 11-16, 2024, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 4912–4944. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.269>
- [291] Y. K. Lal, L. Zhang, F. Brahman, B. P. Majumder, P. Clark, and N. Tandon, "Tailoring with targeted precision: Edit-based agents for open-domain procedure customization," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15597–15611. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.921>
- [292] H. Tang, D. Key, and K. Ellis, "Worldcoder, a model-based LLM agent: Building world models by writing code and interacting with the environment," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/820c61a0cd419163ccbd2c33b268816e-Abstract-Conference.html
- [293] N. Mündler, M. N. Müller, J. He, and M. T. Vechev, "Swt-bench: Testing and validating real-world bug-fixes with code agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/94f093b41fc2666376fb1f667fe282f3-Abstract-Conference.html
- [294] Y. Hu et al., "An llm-enhanced agent-based simulation tool for information propagation," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 8679–8682. [Online]. Available: <https://www.ijcai.org/proceedings/2024/1007>
- [295] C. Xie et al., "Can large language model agents simulate human trust behavior?" in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/1cb57fcf7ff3f6d37eebae5becc9ea6d-Abstract-Conference.html
- [296] Q. Zhao et al., "Competeai: Understanding the competition dynamics of large language model-based agents," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=wGtzp4ZT1n>
- [297] G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, and R. Mihalcea, "Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/ca9567d8ef6b2ea2da0d7eed57b933ee-Abstract-Conference.html
- [298] Z. Guan, X. Kong, F. Zhong, and Y. Wang, "Richelieu: Self-evolving llm-based agents for AI diplomacy," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/df2d62b96a4003203450cf89cd338bb7-Abstract-Conference.html
- [299] X. Zhou et al., "SOTOPIA: interactive evaluation for social intelligence in language agents," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=mM7VurbA4r>
- [300] C. Wang, B. Dai, H. Liu, and B. Wang, "Towards objectively benchmarking social intelligence of language agents at the action level," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 8885–8897. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.526>
- [301] G. Li, H. Hammoud, H. Itani, D. Khizbulin, and B. Ghanem, "CAMEL: communicative agents for ""mind"" exploration of large language model society," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html
- [302] J. Liu et al., "Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons et al., Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/5875aca1ef70285a35940afbce0f9fb-Abstract-Datasets_and_Benchmarks_Track.html
- [303] Y. Dong, X. Zhu, Z. Pan, L. Zhu, and Y. Yang, "Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 16290–16314. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.964>
- [304] Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/6b8dfb8c0c12e6fafc6c256cb08a5ca7-Abstract-Conference.html
- [305] K. Nottingham et al., "Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 26311–26325. [Online]. Available: <https://proceedings.mlr.press/v202/nottingham23a.html>
- [306] R. Wang et al., "Sotopia-π: Interactive learning of socially intelligent language agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 12912–12940. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.698>
- [307] Y. Zhang, S. Yuan, C. Hu, K. Richardson, Y. Xiao, and J. Chen, "Timearena: Shaping efficient multitasking language agents in a time-aware simulation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 3894–3916. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.215>
- [308] Y. Deng, X. Zhang, W. Zhang, Y. Yuan, S. Ng, and T. Chua, "On the multi-turn instruction following for conversational web agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 8795–8812. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.477>
- [309] K. Cheng et al., "Seeclick: Harnessing GUI grounding for advanced visual GUI agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 9313–9332. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.505>
- [310] Z. Zhang and A. Zhang, "You only look at screens: Multimodal chain-of-action agents," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 3132–3149. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.505>

- //doi.org/10.18653/v1/2024.findings-acl.186
- [311] I. Gur *et al.*, "A real-world webagent with planning, long context understanding, and program synthesis," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=9JQtrumvg8>
- [312] K. Yang *et al.*, "Agentoccam: A simple yet strong baseline for llm-based web agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=oWdzUpOlkX>
- [313] T. Ou *et al.*, "Synatra: Turning indirect knowledge into direct demonstrations for digital agents at scale," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/a6a6891cf1dfc64d664f086cf5976e93-Abstract-Conference.html
- [314] S. Agashe, J. Han, S. Gan, J. Yang, A. Li, and X. E. Wang, "Agent S: an open agentic framework that uses computers like a human," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=lIVRgt4nLv>
- [315] Y. Xu *et al.*, "Agenttrek: Agent trajectory synthesis via guiding replay with web tutorials," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=EEgYUccwsV>
- [316] J. Y. Koh *et al.*, "Visualwebarena: Evaluating multimodal agents on realistic visual web tasks," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 881-905. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.50>
- [317] S. Zhou *et al.*, "Webarena: A realistic web environment for building autonomous agents," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=oKn9c6ytLx>
- [318] H. He *et al.*, "Webvoyager: Building an end-to-end web agent with large multimodal models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 6864-6890. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.371>
- [319] A. Drouin *et al.*, "Workarena: How capable are web agents at solving common knowledge work tasks?" in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=BRFqYrikdo>
- [320] H. Cai *et al.*, "Large language models empowered personalized web agents," in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 198-215. [Online]. Available: <https://doi.org/10.1145/3696410.3714842>
- [321] R. Niu *et al.*, "Screenagent: A vision language model-driven computer control agent," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 2024, pp. 6433-6441. [Online]. Available: <https://www.ijcai.org/proceedings/2024/711>
- [322] X. Ma, Z. Zhang, and H. Zhao, "Coco-agent: A comprehensive cognitive MLLM agent for smartphone GUI automation," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 9097-9110. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.539>
- [323] C. H. Wu, R. R. Shah, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan, "Dissecting adversarial robustness of multimodal LM agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=YauQYh2k1g>
- [324] Z. Qi *et al.*, "Webrl: Training LLM web agents via self-evolving online curriculum reinforcement learning," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=oVKEAFjEqv>
- [325] K. Zhu, J. Wang, Q. Zhao, R. Xu, and X. Xie, "Dynamic evaluation of large language models by meta probing agents," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=DwTgy1hXXo>
- [326] C. Chan *et al.*, "Chateval: Towards better llm-based evaluators through multi-agent debate," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=FQepisCUWu>
- [327] L. Ning *et al.*, "Cheatagent: Attacking llm-empowered recommender systems via LLM agent" *CoRR*, vol. abs/2504.13192, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2504.13192>
- [328] H. Zhou *et al.*, "Star-agents: Automatic data optimization with LLM agents for instruction tuning," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/0852b88e96d973bd4e21b673f51621d0-Abstract-Conference.html
- [329] F. Xing, "Designing heterogeneous LLM agents for financial sentiment analysis," *ACM Trans. Manag. Inf. Syst.*, vol. 16, no. 1, pp. 1-24, 2025. [Online]. Available: <https://doi.org/10.1145/3688399>
- [330] S. Guo, C. Deng, Y. Wen, H. Chen, Y. Chang, and J. Wang, "Ds-agent: Automated data science by empowering large language models with case-based reasoning," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=LfjgeBNCFI>
- [331] Z. Yang *et al.*, "Matplotagent: Method and evaluation for llm-based agentic scientific data visualization," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 11789-11804. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.701>
- [332] Z. Chen *et al.*, "Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=dz4YKr0GK6>
- [333] Y. Sun *et al.*, "Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=rFpZnn11gj>
- [334] Q. Yang *et al.*, "Psychogat: A novel psychological measurement paradigm through interactive fiction games with LLM agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 14470-14505. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.779>
- [335] X. Hu, G. Liu, Y. Zhao, and H. Zhang, "De novo drug design using reinforcement learning with multiple GPT agents," *CoRR*, vol. abs/2401.06155, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2401.06155>
- [336] Z. Wang *et al.*, "Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration," in *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov, Eds. ACM, 2025, pp. 2250-2261. [Online]. Available: <https://doi.org/10.1145/3696410.3714877>
- [337] S. Si *et al.*, "Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural*

- Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/7b16688a2b053a1b01474ab5c78ce662-Abstract-Datasets_and_Benchmarks.html
- [338] C. Zhang *et al.*, "Advancing DRL agents in commercial fighting games: Training, integration, and agent-human alignment," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=eN1T7lOpZ>
- [339] D. Wu, H. Shi, Z. Sun, and B. Liu, "Deciphering digital detectives: Understanding LLM behaviors and capabilities in multi-agent mystery games," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 8225–8291. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.490>
- [340] S. Basavatia, K. Murugesan, and S. Ratnakar, "STARLING: self-supervised training of text-based reinforcement learning agent with large language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 15804–15819. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.935>
- [341] J. Xie, R. Zhang, Z. Chen, X. Wan, and G. Li, "Whodunitbench: Evaluating large multimodal agents via murder mystery games," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/9dd4533e7e4e5ed809344280609c5b05-Abstract-Datasets_and_Benchmarks_Track.html
- [342] Z. Shi, M. Fang, and L. Chen, "Monte carlo planning with large language model for text-based game agents," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=r1KcapkzCt>
- [343] J. Wang *et al.*, "Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/0520537ba799d375b8ff5523295c337a-Abstract-Conference.html
- [344] S. Zheng, J. Liu, Y. Feng, and Z. Lu, "Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [Online]. Available: <https://openreview.net/forum?id=NltzxpG0nz>
- [345] R. Jiao *et al.*, "Can we trust embodied agents? exploring backdoor attacks against embodied llm-based decision-making systems," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=S1Bv3068Xt>
- [346] H. Shi, Z. Sun, X. Yuan, M. Côté, and B. Liu, "Opex: A component-wise analysis of llm-centric agents in embodied instruction following," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 622–636. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.37>
- [347] C. Huang *et al.*, "Geoagent: To empower llms using geospatial tools for address standardization," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 6048–6063. [Online]. Available: <https://doi.org/10.18653/v1/2024.findings-acl.362>
- [348] J. Wang *et al.*, "Large language models as urban residents: An LLM agent framework for personal mobility generation," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/e142fd2b70f10db2543c64bca1417de8-Abstract-Conference.html
- [349] W. Wu, X. Feng, Z. Gao, and Y. Kan, "SMART: scalable multi-agent real-time motion generation via next-token prediction," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons *et al.*, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/cef5c8dec67597b854f0162ad76d92d2-Abstract-Conference.html
- [350] S. Lai, Z. Xu, W. Zhang, H. Liu, and H. Xiong, "Llmlight: Large language models as traffic signal control agents," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, Y. Sun, F. Chierichetti, H. W. Lauw, C. Perlich, W. H. Tok, and A. Tomkins, Eds. ACM, 2025, pp. 2335–2346. [Online]. Available: <https://doi.org/10.1145/3690624.3709379>
- [351] S. Han, L. Chen, L. Lin, Z. Xu, and K. Yu, "IBSEN: director-actor agent collaboration for controllable and interactive drama script generation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 1607–1619. [Online]. Available: <https://doi.org/10.18653/v1/2024.acl-long.88>
- [352] F. Huot, R. K. Amplayo, J. Palomaki, A. S. Jakobovits, E. Clark, and M. Lapata, "Agents' room: Narrative generation through multi-step collaboration," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: <https://openreview.net/forum?id=HfWcFs7XLR>
- [353] A. Parrish *et al.*, "Bbq: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.
- [354] J. Ji *et al.*, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 678–24 704, 2023.
- [355] J. Dhamala *et al.*, "Bold: Dataset and metrics for measuring biases in open-ended language generation," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 862–872.
- [356] Y. Huang and D. Xiong, "Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 2917–2929.
- [357] Y. Wang *et al.*, "Cdeval: A benchmark for measuring the cultural dimensions of large language models," *arXiv preprint arXiv:2311.16421*, 2023.
- [358] J. Zhou *et al.*, "Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks," *arXiv preprint arXiv:2202.08011*, 2022.
- [359] G. Zhang *et al.*, "Corgi-pm: A chinese corpus for gender bias probing and mitigation," *arXiv preprint arXiv:2301.00395*, 2023.
- [360] S. Xu, Y. Leng, L. Yu, and D. Xiong, "Self-pluralising culture alignment for large language models," *arXiv preprint arXiv:2410.12971*, 2024.
- [361] G. Xu *et al.*, "Cvalues: Measuring the values of chinese large language models from safety to responsibility," *arXiv preprint arXiv:2307.09705*, 2023.
- [362] Y. Y. Chiu, L. Jiang, and Y. Choi, "Dailydilemmas: Revealing value preferences of llms with quandaries of daily life," *arXiv preprint arXiv:2410.02683*, 2024.
- [363] B. Wang *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models." in *NeurIPS*, 2023.
- [364] H. Liu *et al.*, "Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation," *Information Processing & Management*, vol. 62, no. 4, p. 104099, 2025.
- [365] S. Gupta *et al.*, "Bias runs deep: Implicit reasoning biases in persona-assigned llms," *arXiv preprint arXiv:2311.04892*, 2023.
- [366] D. Hendrycks *et al.*, "Aligning ai with shared human values," in *International Conference on Learning Representations*, 2021.

- [367] K. Huang *et al.*, "Flames: Benchmarking value alignment of llms in chinese," in *NAACL-HLT*, 2024.
- [368] L. Luo, Y. Nakao, M. Chollet, H. Inakoshi, and S. Stumpf, "Earn fairness: Explaining, asking, reviewing, and negotiating artificial intelligence fairness metrics among stakeholders," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–37, 2025.
- [369] E. Durmus *et al.*, "Towards measuring the representation of subjective global opinions in language models," *arXiv preprint arXiv:2306.16388*, 2023.
- [370] J. Kharchenko, T. Roosta, A. Chadha, and C. Shah, "How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions," *arXiv preprint arXiv:2406.14805*, 2024.
- [371] L. Jiang, S. Levine, and Y. Choi, "Can language models reason about individualistic human values and preferences?" in *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024.
- [372] J. Lee *et al.*, "Kornat: Llm alignment benchmark for korean social values and common knowledge," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 11 177–11 213.
- [373] E. Karinshak *et al.*, "Llm-globe: A benchmark evaluating the cultural values embedded in llm output," *arXiv preprint arXiv:2411.06032*, 2024.
- [374] S. Yao, Q. Ke, Q. Wang, K. Li, and J. Hu, "Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities," in *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, 2024, pp. 108–112.
- [375] M. Abdulhai, G. Serapio-García, C. Crepy, D. Valter, J. Canny, and N. Jaques, "Moral foundations of large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 737–17 752.
- [376] X. Liu, Y. Zhu, S. Zhu, P. Liu, Y. Liu, and D. Yu, "Evaluating moral beliefs across llms through a pluralistic framework," *arXiv preprint arXiv:2411.03665*, 2024.
- [377] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, and D. Yang, "The moral integrity corpus: A benchmark for ethical dialogue systems," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3755–3773.
- [378] H. Norhashim and J. Hahn, "Measuring human-ai value alignment in large language models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 1063–1073.
- [379] Z. Jin *et al.*, "When to make exceptions: Exploring language models as accounts of human moral judgment," *Advances in neural information processing systems*, vol. 35, pp. 28 458–28 473, 2022.
- [380] C. Ju *et al.*, "Benchmarking multi-national value alignment for large language models," *arXiv preprint arXiv:2504.12911*, 2025.
- [381] J. Ji *et al.*, "Pku-saferlfhf: Towards multi-level safety alignment for llms with human preference," *arXiv preprint arXiv:2406.15513*, 2024.
- [382] T. A. Qiu, Y. Zhang, X. Huang, J. Li, J. Ji, and Y. Yang, "Progress-gym: Alignment with a millennium of moral progress," *Advances in Neural Information Processing Systems*, vol. 37, pp. 14 570–14 607, 2024.
- [383] J. Dai *et al.*, "Safesora: Towards safety alignment of text2video generation via a human preference dataset," *Advances in Neural Information Processing Systems*, vol. 37, pp. 17 161–17 214, 2024.
- [384] N. Lourie, R. Le Bras, and Y. Choi, "Scuples: A corpus of community ethical judgments on 32,000 real-life anecdotes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 470–13 479.
- [385] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, "Social bias frames: Reasoning about social and power implications of language," in *Association for Computational Linguistics*, 2020.
- [386] Y. Huang, Q. Zhang, L. Sun *et al.*, "Trustgpt: A benchmark for trustworthy and responsible large language models," *arXiv preprint arXiv:2306.11507*, 2023.
- [387] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," 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)*, 2021, pp. 5356–5371.
- [388] S. Hosseini, H. Palangi, and A. H. Awadallah, "An empirical study of metrics to measure representational harms in pre-trained language models," *arXiv preprint arXiv:2301.09211*, 2023.
- [389] T. Li, T. Khot, D. Khashabi, A. Sabharwal, and V. Srikumar, "Unqovering stereotyping biases via underspecified questions," *arXiv preprint arXiv:2010.02428*, 2020.
- [390] L. Qiu *et al.*, "Valueonet: A new dataset for human value driven dialogue system," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 183–11 191.
- [391] A. Gaut *et al.*, "Towards understanding gender bias in relation extraction," *arXiv preprint arXiv:1911.03642*, 2019.
- [392] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.
- [393] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," *arXiv preprint arXiv:1804.09301*, 2018.
- [394] J. Yu, M. Huber, and K. Tang, "Greedllama: Performance of financial value-aligned large language models in moral reasoning," *arXiv preprint arXiv:2404.02934*, 2024.
- [395] A. Shetty, A. Beheshti, M. Dras, and U. Naseem, "Vital: A new dataset for benchmarking pluralistic alignment in healthcare," *arXiv preprint arXiv:2502.13775*, 2025.