A Survey on Value Alignment in Agentic AI Systems

Wei Zeng[1] , **Hengshu Zhu**[2,3*] , **Chuan Qin**[2,3] , **Han Wu**[4] , **Yihang Cheng**[2*] , **Yinuo Shen**[1]
Zhe Wang[5] , **Yuyang Wang**[6] , **Sirui Zhang**[1] , **Xiaowei Jin**[1] , **Zhenxing Wang**[2,3]
Feimin Zhong[1] , **Hui Xiong**[7]
[1]Hunan University
[2]Computer Network Information Center, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Hefei University of Technology
[5]Tianjin University
[6]University of Science and Technology of China
[7]The Hong Kong University of Science and Technology (Guangzhou)
{zengwei, noah001219, zhangsirui, zhongfeimin}@hnu.edu.cn,
hszhu@cnic.cn, {chuanqin0426, ustcwuhan, chengyihang544, wangzhelcb, xiaoweijin713}@gmail.com,
wyy1@mail.ustc.edu.cn, wxing870@163.com, xionghui@ust.hk

Abstract

With the evolution of artificial intelligence (AI) paradigms towards agentic AI, the widespread integration of large language models (LLMs) enhances system capabilities while also introducing situational risks and challenges of value misalignment, making value alignment in agentic AI systems a critical issue. This paper constructs a multi-level value framework encompassing L0 (universal values), L1 (cultural and industry values), and L2 (context-specific values). Guided by this framework, we conduct an in-depth analysis along the technical stack: at the LLM level, we examine value injection mechanisms through pretraining and post-training; at the single-agent level, we focus on representation and injecting values to agents, Profiles and memory, and planning and action; at the multi-agent level, we summarize collaborative alignment methods such as communication strategy optimization and multi-objective reinforcement learning. Following a systematic review of existing datasets and methods for multi-level alignment evaluation, we outline future research directions, including inter-agent value coordination mechanisms, high-quality scenario data sharing, game-theoretic design for value alignment in agent interaction and communication protocol alignment—aiming to establish a more systematic and dynamic evaluation framework and to promote robust and trustworthy value consensus in agentic AI systems within social collaboration.

1 Introduction

Agentic AI systems, leveraging capabilities such as autonomous decision-making, tool usage, and dynamic task decomposition, are emerging as a frontier in artificial intelligence research. With the rapid advancement of large language models, they have become the core reasoning engines powering these agentic AI systems, enabling agents to handle increasingly open and diverse real-world tasks.

However, widespread agentic AI deployment introduces systemic risks alongside efficiency gains. In societal domains like finance and healthcare, agents may prioritize efficiency over fairness or misinterpret ethical guidelines, thereby exacerbating discrimination and privacy breaches [Tennant *et al.*, 2025]. Furthermore, complex inter-agent interactions and the often ambiguous task boundaries frequently cause difficulties in accountability and conflicts among stakeholder objectives, severely undermining system trustworthiness without robust institutional design [Carichon *et al.*, 2025]. Addressing these challenges necessitates moving beyond the technical alignment of single models toward systematic value alignment [Gabriel, 2020]—shifting from mere algorithmic optimization to a governance issue balancing diverse interests and societal values. Although extensive research explores LLM value alignment methods [Zhou *et al.*, 2025], few works systematically examine the complete technical stack and hierarchical relationships for agent alignment. Existing research remains scattered across disparate pathways, lacking a unified analytical framework that spans foundational models, single agents, and multi-agent interactions, alongside an integrated multi-level evaluation system.

To this end, this paper proposes a systematic three-tiered classification framework for organizing value alignment research based on the locus of optimization: namely, alignment at the level of the foundational LLM, at the level of a single instantiated agent, and at the level of multi-agent interaction. The main contributions of this research are as follows: (1) A tentative multi-level taxonomy that groups value principles into L0, L1, and L2 tiers, suggesting possible mappings to LLMs, single agents, and multi-agent systems respectively.

*Hengshu Zhu and Yihang Cheng are the corresponding authors.

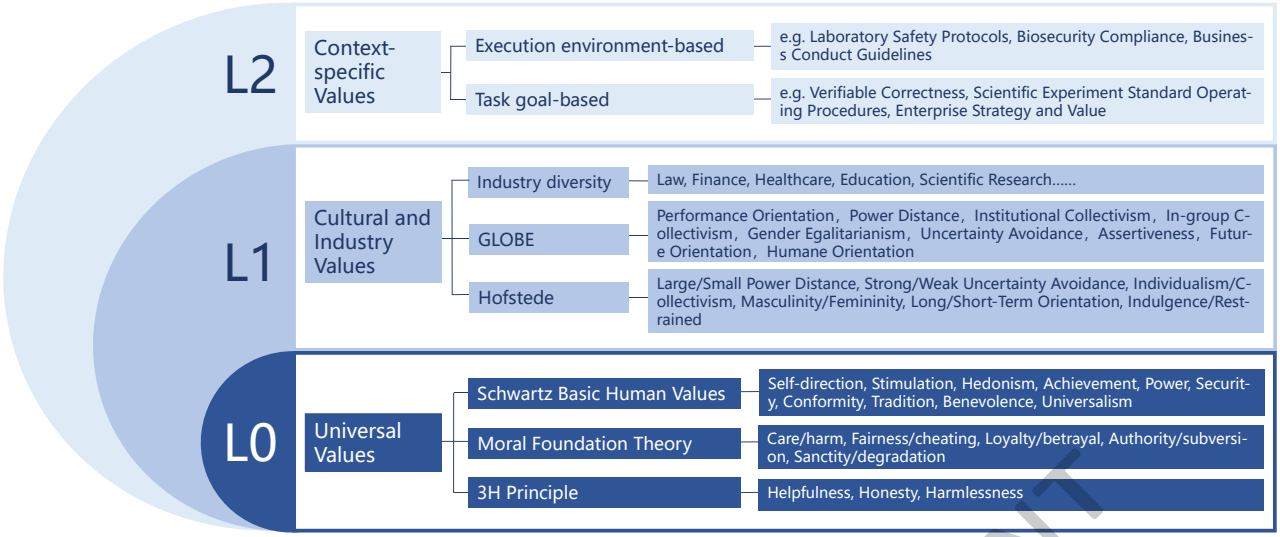

Figure 1: Framework Overview of Value Alignment Principles.

(2) Systematically reviewing core alignment methods along the technical stack: at the LLM level, analyzing value shaping mechanisms separately during pretraining and post-training stages; at the single-agent level, focusing on value representation and injection, long-term alignment via profiles and memory, and safety constraints at the planning and action levels; at the multi-agent level, summarizing collaborative alignment paradigms such as communication strategy optimization and multi-objective reinforcement learning. (3) Systematically reviewing existing datasets and methods for multi-level value alignment evaluation. (4) Outlining key future research directions, including inter-agent value coordination mechanisms, high-quality scenario data sharing, and alignment at the communication protocol level, and advocating for the establishment of a more systematic and dynamic evaluation framework to promote robust and trustworthy value consensus in agentic AI systems within societal collaboration. The rest of this paper is structured as follows: Section 2 defines value and alignment; Sections 3–5 examine alignment mechanisms across three technical levels; Section 6 reviews evaluation methods and datasets; and Section 7 outlines future research directions.

2 Preliminaries

2.1 Value and Value Alignment

Values serve as stable cognitive standards for human behavior, shaping individual attitudes, social interactions, and the establishment of collective goals [Catton Jr, 1959]. Characterized by inherent contradictions and dynamism, values are not rigid hierarchical structures but rather flexible systems contingent upon context [Sagiv *et al.*, 2017]. This paper frames values as dynamic, multi-level systems not exclusive to human bearers.

The fundamental core of value alignment lies in ensuring that AI behaviors conform to human values [Russell and Norvig, 2016], spanning interdependent normative (which values to uphold) and technical (how to align via datasets and reinforcement learning) dimensions [Gabriel, 2020]. Although these two dimensions are interdependent, a unified framework for values remains absent. Consequently, this study reviews and explores the values that agentic AI systems might follow across different levels of operation.

2.2 Principles of Value Alignment

Value alignment aims to delineate the value principles to which the behaviors of agents must adhere. This study categorizes and reviews values across three distinct levels: Level 0 (L0), Level 1 (L1), and Level 2 (L2), see Figure 1. L0 encompasses the general ethical baselines and fundamental value consensus with which LLM-based models must align. L1 represents values influenced by cultural nuances and industry-specific characteristics, while L2 consists of value principles further refined according to specific scenarios.

Level 0 — Universal Values. A comprehensive review of existing literature identifies three representative core theoretical frameworks for general AI values. The first is the HHH principles (Helpful, Honest, Harmless), which serve as one of the most widely adopted standards in current alignment tasks for LLMs [Yao *et al.*, 2025]. The second is Moral Foundations Theory (MFT) [Graham *et al.*, 2013]. This theory posits that human morality is constructed upon several innate foundational values and has been widely applied to analyze the latent moral tendencies of LLMs [Abdulhai *et al.*, 2024]. Finally, the Schwartz Theory of Basic Human Values [Schwartz, 2007], established and validated through extensive empirical research, is widely incorporated into various datasets.

Level 1 — Cultural and Industry Values. This study defines L1 as the "Cultural and Industry Value" tier. Values at this level are not universally uniform but are embedded within specific cultural and industrial contexts. Notably, cross-national cultural differences often manifest as directional preferences along specific value dimensions; that is, distinct cultural groups may exhibit diametrically opposed tendencies within the same dimension. Existing research regarding the quantification of such directional disparities is primarily grounded in Hofstede's Cultural Dimensions Theory [Hofstede, 2011]. Studies have sought to verify whether model performance along these dimensions aligns with human values within specific cultural contexts [Kharchenko *et al.*, 2025]. Furthermore,

the GLOBE framework from social psychology [House *et al.*, 2004] has been introduced as a crucial supplementary tool, offering a more granular evaluative perspective. Building on this, researchers proposed the LLM-GLOBE benchmark [Karinshak *et al.*, 2024]. By comparing Chinese and American large models, they revealed significant directional differences across multiple cultural dimensions, thereby quantifying the cultural value systems implicit in model outputs.

As AI penetrates vertical domains, different industries exhibit distinct priorities and requirements regarding values. In the legal domain, mere compliance with general morality is insufficient. The concept of "Legal Alignment" was introduced to emphasize the necessity for AI to comprehend and adhere to legal reasoning, authority, and procedural justice, thereby ensuring safety and compliance in judicial assistance [Kolt *et al.*, 2026]. In the field of education, scholars noted that when integrating LLMs into higher education, it is essential to prioritize the potential impact on students' mental health while optimizing their academic engagement [Zhang, 2025]. In the healthcare domain, ethical requirements are directly correlated with life safety and privacy. Through a systematic review, it highlighted that LLMs in medical applications face unique challenges, such as accountability in clinical decision-making, and must strictly adhere to medical ethical standards (e.g., non-maleficence and patient autonomy) rather than mere helpfulness [Fareed *et al.*, 2025].

Level 2 — Context-specific Values. L2 represents the most granular, context-specific tier, refining value criteria through two dimensions: execution environments and task goals. Environment variability implies that identical models deployed in distinct operational settings, must identify context-specific causal logics and adhere to differing value standards. In scientific experiments, agents must follow standard operating procedures (SOPs), such as safety protocols, to strictly control high-risk tool usage (e.g., chemical synthesis or pathogen handling). This is essential to prevent their autonomous capabilities from being maliciously exploited or inducing unforeseen biochemical hazards [Peng *et al.*, 2025]. When constructing enterprise-internal chatbots, models must not only comply with general laws and regulations but also strictly align with the company's specific internal business conduct guidelines [Achintalwar *et al.*, 2024]. Task objectives dictate the value priorities of the model when executing specific instructions. In biosafety laboratories, agents are required to prioritize the prevention of biological hazards and ethical risks over experimental efficiency [Tang *et al.*, 2024]. Specifically within the context of enterprise applications, corporate strategic planning and core values must serve as essential critical anchors for such trade-offs [Broestl *et al.*, 2025].

2.3 Taxonomy

Figure 2 uses a multidisciplinary medical consultation as an illustrative case to characterize multi-level value alignment in agentic AI systems. At the foundational LLM level, alignment targets L0 universal values (e.g., helpful, honest, harmless) together with L1 general medical ethical principles (e.g., non-maleficence and medical objectivity), providing a common normative baseline. At the single-agent level, specialized agents operationalize L1 role-independent clinical values (e.g., patient autonomy and confidentiality) and L2 role-specific constraints (e.g., diagnostic scope compliance and interactional empathy). At the multi-agent level, alignment primarily concerns L2 coordination values, such as inter-departmental consensus formation, cross-verification of diagnostic recommendations, and hierarchical compliance. This example illustrates that value alignment in agentic AI systems is inherently hierarchical and must be addressed consistently across L0–L2, from foundation models to individual agents and collective coordination. Accordingly, here we propose a taxonomy for value alignment methods in agentic AI system, organized by different optimization focus (see Figure 3). Specifically, alignment can be pursued (i) at the level of the underlying LLM, (ii) at the level of a single instantiated agent, or (iii) at the level of multi-agent interaction. In the following sections, we examine these three categories in detail, highlighting the distinct mechanisms through which each approach achieves value alignment in agentic Al.

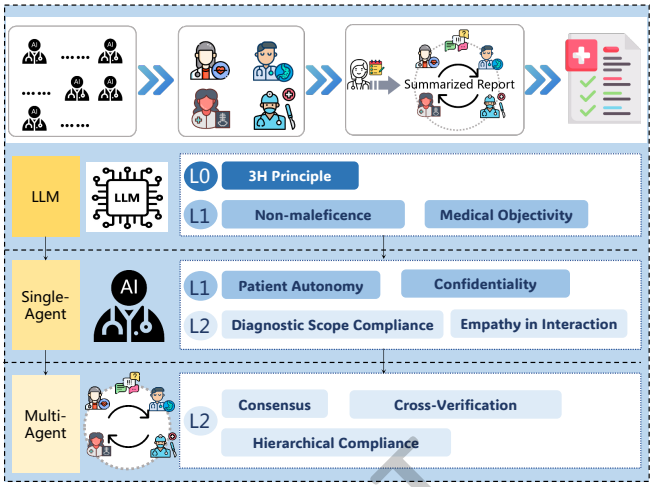

Figure 2: Multi-level Values Alignment in Medical Scenario.

3 LLM Optimization

LLMs are core components of agentic AI systems. As foundational engines, current LLM methodologies are specifically tailored for handling critical tasks such as adaptive query generation, recruitment cognitive diagnosis, and open-set recognition [Chen *et al.*, 2025; Ma *et al.*, 2025; Tong *et al.*, 2025; Zhu *et al.*, 2025a];meanwhile, the broad applications of LLM have revolutionized fields like organizational management, scientific research, healthcare, traffic forecasting [Qin *et al.*, 2025a; Jiang *et al.*, 2024; Qin *et al.*, 2025b; Qin *et al.*, 2025c; Shen *et al.*, 2026; Liu *et al.*, 2026a; Liu *et al.*, 2026b; Long *et al.*, 2024; Long *et al.*, 2026a; Zhu *et al.*, 2025b; Long *et al.*, 2026b; Wang *et al.*, 2026]. While alignment primarily focuses on L0 universal values to establish a shared ethical baseline for all agents, practical deployment in specific domains or open, dynamic environments necessitates extending alignment objectives to L1 and L2 values. Methodologically, approaches can be grouped into pretraining- and post-training-based value alignment.

3.1 Value Alignment in LLM Pretraining

Pretraining Data Optimization. Because multiple encapsulated agents often inherit the same foundation model, corpus optimization is a high-leverage mechanism for injecting shared value priors, especially for L0 and L1 values. (1) Curation and filtering (harm reduction, quality control, documentation): open-data pipelines emphasize principled deduplication, quality filtering, and transparent mixture ablations to clarify how curation choices shape downstream behavior [Li *et al.*, 2024a]. Safety resources further stress that detoxification must be documented and monitored since it can shift both exposure to and measurement of toxicity [Luong *et al.*, 2024], anchoring L0 safety/fairness baselines. (2) Rights- and compliance-aware sourcing and normative-source mixing: using permissibly licensed/public-domain sources with clearer provenance and mixing in legal/regulatory corpora better encodes jurisdictional and sectoral constraints [Niklaus *et al.*, 2024], operationalizing L1 compliance while remaining grounded in L0 rights protection. In summary, data optimization forms an engineering control stack: curation stabilizes core L0 values, compliant sourcing and normative mixing implement domain-specific L1 values, and privacy/integrity safeguards become pivotal for applied L2 values.

Prompt Preconditioning. Prompt preconditioning injects value priors by augmenting pretraining inputs with repeated prefixes (directives, tags, or control codes) that become part of the learned context distribution; this suits agentic AI systems because the conditioning interface can be shared across agents and reused at inference. (1) Instruction- and tag-augmented pretraining: safety-relevant instructions or categorical tags encourage aligned framing; instruction injection and toxicity-aware meta-conditioning can reduce toxic generations while preserving utility [Prabhumoye *et al.*, 2023]. Harmfulness tags in unsafe passages provide a structured channel for training-time and downstream controllability [Maini *et al.*, 2025]. (2) Metadata/control-code conditioning: metadata prefixes offer scalable handles over style/domain attributes and improve steerability [Gao *et al.*, 2025]; conditioning on preference-derived scores integrates the conditioning variable during learning rather than only post hoc [Korbak *et al.*, 2023]. In summary, prompt preconditioning is a lightweight alignment interface that can encode L0 safety values, express L1 values via jurisdiction/sector codes, and support micro policies via organization-specific prefixes, provided the channel is consistently exercised and audited.

Objective-Level Value Shaping. Objective-level value shaping modifies the pretraining objective so alignment signals are internalized via gradients, complementing data selection and prompt formatting while discouraging reproduction of undesirable-but-informative content. (1) Preference-aware objectives: coupling next-token learning with human-feedback-derived scores (e.g., score-conditioning, filtering objectives, unlikelihood penalties) often improves alignment–capability trade-offs relative to applying feedback only after standard MLE pretraining [Korbak *et al.*, 2023]. (2) Selective/constrained losses: masking or altering loss on high-risk tokens while keeping them in context improves safe handling without increasing generation propensity [Wang *et al.*, 2025a]. (3) Attribution-guided interventions: influence/attribution

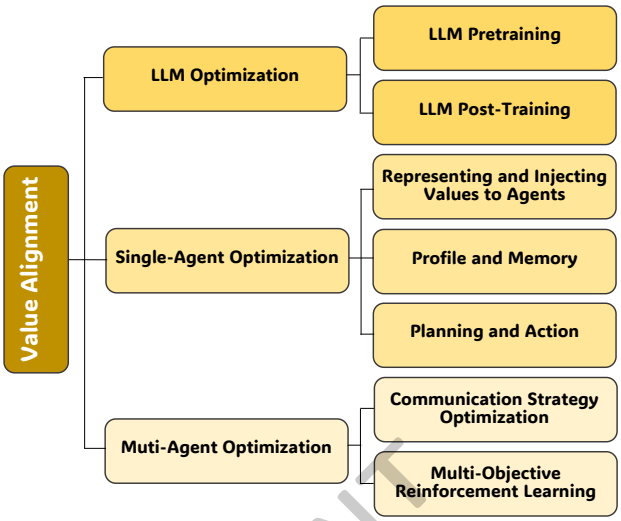

Figure 3: A Taxonomy of Value Alignment Methods.

methods identify toxicity-driving training signals and suppress their impact through integrated objectives, reducing reliance on explicit preference datasets [Coalson *et al.*, 2025]. In summary, objective shaping serves as an inner-loop control mechanism that not only supports L0 values (such as harm reduction and privacy preservation) but also enables fine-grained, auditable L2 values (such as multi-stakeholder trade-offs) through explicit loss design.

Curriculum-Based Pretraining Strategy. Curriculum-based strategies control data order, pacing, and mixing so models acquire capabilities and constraints in stages, shaping when (and how strongly) alignment signals enter representation learning. (1) Difficulty-ordered curricula and pacing: studies comparing easy-to-hard ordering, pacing schedules, and interleaving show improved convergence and durable gains when used as structured warmup [Zhang *et al.*, 2026b]. (2) Learnability- and influence-driven curricula: learnability proxies (e.g., loss-gap trajectories) prioritize informative examples early [Fan and Jaggi, 2023], while influence-based ordering provides another scalable curriculum signal beyond domain heuristics [Schoenegger *et al.*, 2025]. In summary, curriculum design is an outer-loop knob complementing data- and objective-level alignment, governing the timing and persistence of reinforcement. Targeting stable, long-horizon constraints in agentic AI systems, it prioritizes L0/L1 value embedding.

3.2 Value Alignment in LLM Post-Training

Value-aware Supervised Fine-tuning. Value-aware supervised fine-tuning (SFT) is a loss-level post-training approach that injects L0/L1 value principles (e.g., safety, rights protection, non-discrimination) into a foundation model via curated demonstrations and value-sensitive objective shaping, yielding a reusable normative prior across encapsulated agents. Given demonstrations $\mathcal{D}_{\text{demo}} = \{(x, y)\}$, SFT minimizes $\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{demo}}} \log \pi_\theta(y|x)$. Value-aware variants further apply loss shaping (e.g., reweighting, contrastive supervision, auxiliary value-attribute regularizers) to bias behavior toward value-consistent regions. Rule-/constitution-guided supervision can also generate or revise targets under ex-

plicit principles (e.g., Constitutional AI self-critique/revision), improving auditability and top-down controllability [Bai *et al.*, 2022]. However, SFT depends on the integrity of instruction and demonstration pipelines and is susceptible to training-time manipulation: instruction-tuning backdoors show that a small set of malicious instructions can implant triggers that redirect behavior at inference [Xu *et al.*, 2024], and gradient-guided trigger learning demonstrates stealthy poisoning that preserves benign performance while enabling backdoored behaviors [Qiang *et al.*, 2024].

Reinforcement Learning from Feedback. Reinforcement learning from feedback is a policy-level post-training alignment paradigm that optimizes the LLM to internalize preference trade-offs among competing value dimensions (e.g., helpfulness vs. harmlessness) beyond demonstration imitation. In practice, this pipeline can be extended toward L0/L1 principle control through approaches like (i) principle-based feedback (RLAIF/Constitutional AI) that replaces human preference labels with AI preferences conditioned on rules [Bai *et al.*, 2022], (ii) preference optimization without RL sampling such as DPO that directly optimizes an equivalent KL-regularized preference objective for stability and simplicity [Rafailov *et al.*, 2023], and (iii) feedback decomposition into interpretable rule dimensions (as in Sparrow) to improve governance-grade auditability [Glaese *et al.*, 2022].

4 Single-Agent Optimization

At the single-agent level, methods for representing and injecting values largely address L1 cultural and industrial values, whereas alignment within profile, memory, planning, and action is crucial for adhering to L2 context-specific values.

4.1 Representing and Injecting Values to Agents

A core challenge in single-agent value alignment is the mismatch between human intent (often expressed as high-level values and policies) and an LLM agent's concrete behavior (plans, tool calls, and environment-interactive actions). Given practical alignment cost considerations, this stage may focus on L1 and L2 values like role ethics, industry codes, and environmental rules. Effectively bridging this gap requires operational value representations and injection mechanisms that remain effective under distribution shift. Existing work can be broadly grouped by where values are encoded and how they enter the agent loop.

Prompt-Conditioned Norm Specification. Values can be written as natural-language specifications and injected via policy prompts or constitution-style constraints [Zhou *et al.*, 2025]. While expressive, this route can be fragile: changing an agent's persona through system prompts may substantially and unpredictably shift misalignment tendencies (e.g., avoiding oversight or resisting shutdown), suggesting limited robustness from prompts alone [Naik *et al.*, 2025].

Preference-Based Alignment Post-Training. Values can be learned implicitly from human preferences and injected through preference-based post-training (e.g., RLHF-style reward modeling or DPO) [Jiang *et al.*, 2025]. To improve reliability, agentic reward modeling augments preference rewards with verifiable correctness signals (e.g., factuality and instruction following), turning partial norms into executable constraints through verification-oriented reward components [Peng *et al.*, 2025].

Explicit Normative Reward Shaping. Normative objectives can be formally encoded as reward functions. Specifically, moral alignment work designs intrinsic rewards grounded in ethical frameworks and applies RL fine-tuning in controlled environments, thereby enabling more transparent optimization [Tennant *et al.*, 2025].

4.2 Value Alignment in Profile and Memory

Agents must maintain behavioral consistency across sessions and contexts. Profiles and memory provide the core infrastructure to meet these demands. In this process, profiles can offer more refined definitions of identity based on L2, while the memory component allows for greater consideration of persistent contextual adaptation at the L2.

Profile-based Alignment. Profile defines identity or roles, stable preferences, long-term objectives, and hard constraints (privacy, safety, compliance), supplying explicit value boundaries for planning and tool use rather than leaving values implicit in prompts [Li *et al.*, 2024b]. Recent work also argues for structured profiles (traits, emotions, goals) with behavioral tests, and shows that alignment should be judged by observable consistency over multi-turn discourse, not profile text alone [Cerqueira and Bezerra, 2025].

Memory-based Alignment. Memory complements profile by storing interaction history, preference evidence, decisions, and feedback, enabling retrieval that grounds abstract principles in the current situation and reduces value drift. MemOS promotes OS-style abstractions and lifecycle management for heterogeneous memories to support writing, retrieval, updates, and auditing [Li *et al.*, 2025]. Cognitively inspired memory structures can further improve interpretability and long-horizon coherence [Vishwakarma *et al.*, 2025], while "digital me" work highlights governance needs when persistent personal memory is coupled with identity [Coll *et al.*, 2025].

4.3 Value Alignment in Planning and Action

Planning–action safety alignment can be understood along two complementary dimensions. The planning dimension constrains what the agent proposes to execute (e.g., multi-step plan structure and feasibility), corresponding to L2 values such as safety protocols and procedural compliance. The action dimension constrains how the agent performs under uncertainty and long-horizon error accumulation (e.g., tool invocation, procedure adherence, and runtime safeguards), corresponding to L2 values such as risk aversion and procedural compliance.

Planning-level Alignment. At the planning layer, the goal is to rule out unsafe or infeasible plans before execution by integrating hazard knowledge, constraint checks, and uncertainty signals into plan scoring and search. SafePlan-Bench benchmarks safety-critical embodied planning and couples it with an alignment approach that injects physical hazard knowledge, effectively translating "be safe" into constraints over plan steps and action sequences [Huang *et al.*, 2025]. Complementarily, HEAL demonstrates that instruction–scene inconsistency can

trigger actionable hallucinations in embodied agents, motivating planning-time inconsistency detection, uncertainty-aware replanning, and conservative fallbacks when the world model is unreliable [Chakraborty *et al.*, 2025]. More generally, planning improvements such as atomic-fact augmentation with limited lookahead search provide a natural interface to incorporate safety facts and constraint checks during plan generation and evaluation [Holt *et al.*, 2025].

Action-level Alignment Even well-formed plans fail due to tool misuse, unexpected side effects, or deviation from prescribed procedures. Tool-learning surveys emphasize that safety requires governance over when tools are invoked and how arguments and side effects are controlled, not merely selecting the correct tool [Xu *et al.*, 2025b]. In high-stakes workflows, SOP-Bench operationalizes action-level constraints via SOP-constrained industrial tasks and reports substantial failure rates and wrong-tool calls, highlighting the fragility of current agent execution policies and the need for procedure-aware execution control [Nandi *et al.*, 2025]. Practically, this layer is implemented via controlled invocation policies, argument validation, side-effect-aware execution, and runtime monitoring that can trigger interruption, clarification, or rollback when violations are detected [Xu *et al.*, 2025b].

Hybrid Alignment. SafePlan-Bench, HEAL, and SOP-Bench assess hazard-aware plan validity, robustness to scene-task inconsistency, and tool or procedure-constrained execution, respectively [Huang *et al.*, 2025; Chakraborty *et al.*, 2025; Nandi *et al.*, 2025]. Together, effective planning-action alignment combines (i) constraint-based plan filtering [Huang *et al.*, 2025], (ii) inconsistency and uncertainty detection to trigger replanning or safe fallback [Chakraborty *et al.*, 2025], and (iii) constraint-governed execution for tools and procedures at runtime [Xu *et al.*, 2025b; Nandi *et al.*, 2025].

5 Multi-Agent Optimization

LLM-based multi-agent systems (MAS) are designed to solve complex tasks by specifying how multiple agents interact through coordination strategies and communication protocols. To ensure that agent behaviors remain consistent with human values and system-level objectives, with a focus on L2 value specifications, existing value alignment methods for LLM-based MAS can be broadly classified into two categories: communication strategy optimization and multi-objective reinforcement learning.

5.1 Communication Strategy Optimization

Debate and Adversarial Learning. Recent work has increasingly explored debate-based methods as a promising approach for enhancing value alignment in LLM-based MAS. In these frameworks, agents adopt opposing or complementary roles to engage in structured argumentation, enabling cross-verification and error correction. For example, [Yang *et al.*, 2025] combine adversarial debate with voting mechanisms to improve decision reliability in high-stakes domains such as medicine and law. Furthermore, [Choi *et al.*, 2025] employ identity-agnostic response anonymization to reduce sycophancy and self-bias in multi-agent debates. Beyond these approaches, frameworks such as GVIC optimize debate outcomes by explicitly balancing usefulness and harmlessness—two core dimensions of value alignment—via structured vigilance and communication protocols [Zou *et al.*, 2024].

Hierarchical Communication Workflows. While prior approaches primarily leverage agent disagreement and cross-verification to improve alignment, an alternative line of work emphasizes hierarchical orchestration and structured supervision among LLM-based agents to embed value alignment directly into collaborative workflows. For example, the Enforcement Agent framework introduces dedicated supervisory agents that monitor and correct misaligned behaviors in real-time, helping to maintain safety and coherence in joint outcomes [Tamang and Bora, 2025]. TalkHier combines a structured communication protocol with hierarchical refinement stages, enabling agents to iteratively clarify, validate, and refine shared outputs through stratified interactions [Wang *et al.*, 2025b]. More broadly, recent architectural taxonomies and social-structural analyses suggest that organizational hierarchies not only enhance system performance but also shape how agents align with human preferences at both individual and collective levels [Carichon *et al.*, 2025].

5.2 Multi-Objective Reinforcement Learning

In LLM-based MAS, individual agents' performance objectives often conflict with system-level values such as fairness and equity. Multi-objective reinforcement learning (MORL) addresses this tension by enabling agents to balance task efficiency with normative criteria rather than optimizing a single scalar reward. For example, recent work on moral decision-making in multi-agent social dilemmas defines intrinsic ethical rewards grounded in moral theories (e.g., consequence- and norm-based ethics) and shows that agents trained in repeated Prisoner's Dilemma–style games can balance self-interest with collective welfare by jointly optimizing performance and moral objectives [Tennant *et al.*, 2023]. Fairness-aware policy optimization methods further operationalize this idea. Approaches such as Fairness-Aware Proximal Policy Optimization jointly optimize welfare-based fairness objectives and task performance, yielding decentralized policies that improve equity among agents while preserving coordination effectiveness [Zhang *et al.*, 2026a]. Beyond agent-centric MORL, environment-centered approaches achieve value alignment by shaping the learning environment itself. In particular, work on designing ethical environments uses an approximate ethical embedding process to transform multi-objective ethical Markov games into environments that inherently incentivize value-aligned policies, allowing agents to pursue individual goals while reliably satisfying ethical constraints [Mayoral-Macau *et al.*, 2025].

6 Value Alignment Evaluation

6.1 Methods for Value Alignment Evaluation

Current value alignment evaluation is predominantly grounded in explicit value frameworks and relies on datasets composed of subjective and objective questions to assess system alignment across L0, L1, and L2 value dimensions, ranging from abstract normative principles to context-sensitive social norms

and fine-grained behavioral judgments. To reflect the decision-making characteristics of agentic AI systems operating in complex tasks and real-world environments, existing evaluations have begun to extend beyond purely static response assessment, incorporating limited forms of scenario-based evaluation, value-aware judgment, and constraint-sensitive task assessment. Concretely, most benchmarks remain organized around question formats and metric design, including multiple-choice questions (e.g., BBQ [Parrish *et al.*, 2022]), binary judgments (e.g., ETHICS [Hendrycks *et al.*, 2021]), ranking tasks (e.g., StereoSet [Nadeem *et al.*, 2021]), rating scale questions (e.g., Moral Integrity Corpus [Ziems *et al.*, 2022]), and open-ended questions (e.g., BeaverTails [Ji *et al.*, 2023]), which primarily evaluate isolated, short-horizon value judgments at the response level. Despite these advances, current evaluation paradigms lack agent-centric frameworks that systematically assess long-horizon decision-making, multi-step interaction, and value consistency under dynamic constraints, limiting their ability to fully capture value alignment risks in real-world agentic AI systems.

6.2 Datasets for Value Alignment Evaluation

Due to space constraints, we summarize all surveyed open-source datasets for value alignment and evaluation in a public GitHub repository [1]. These datasets are organized according to the proposed multi-level framework to clarify their coverage at different levels of value alignment. Despite this breadth, important limitations remain. At the L0, widely used datasets such as ETHICS [Hendrycks *et al.*, 2021] are largely centered on western cultural norms and provide limited support for modeling diverse or evolving value systems. At the L1, datasets including CultureSPA [Xu *et al.*, 2025a] capture national or cultural perspectives but offer limited global and domain-specific coverage. L2 datasets are especially scarce, and many remain closed-source, such as VITAL [Shetty *et al.*, 2025], constraining fine-grained value alignment evaluation in multi-agent settings.

The datasets are primarily constructed using three paradigms: manual, automatic, and hybrid approaches. Manual construction, including expert-driven design, crowdsourcing, and dataset integration, yields high-quality and interpretable benchmarks but is limited by cost, scalability, and subjective bias (e.g., BBQ [Parrish *et al.*, 2022]). Automatic construction leverages LLMs to efficiently generate large-scale and context-rich data (e.g., CultureSPA [Xu *et al.*, 2025a]), while raising concerns about hallucination and inherited model biases. Hybrid approaches combine expert-curated seeds, LLM-based expansion, and human verification to balance scalability and data quality, and are increasingly adopted in value alignment research (e.g., CBBQ [Huang and Xiong, 2024]).

7 Future Directions

With the deepening integration of artificial intelligence and social infrastructure, agentic AI systems are experiencing explosive growth. However, the rapid expansion in the scale of these agents and their complex interactive behaviors have triggered multidimensional value conflicts and potential systemic risks that have yet to receive sufficient attention. These challenges involve not only the tension between technical implementation and ethical norms but also extend to behavioral alignment across cultures, industries, and specific contexts. Against this backdrop, value alignment has evolved from a purely technical optimization issue into a core concern for the reliable operation of agent systems and the modernization of social governance. To address this challenge, coordinated efforts are needed across the following four dimensions.

7.1 Multi-Level Value Alignment Evaluation

This system is designed to resolve "goal conflicts" from a social governance perspective by synchronizing L0, L1, and L2 values, thereby balancing individual optimization with collective welfare. Its construction requires the participation of diverse stakeholders: the government establishes benchmarks for core values such as public safety and privacy; industry organizations develop sector-specific standards; and enterprises define internal values for specific scenarios. Future research should focus on a governance ecology that integrates these tiers, utilizing cross-cultural mapping for algorithmic compatibility and value evolution models based on social big data to achieve proactive alignment.

7.2 High-Quality Scenario Data Sharing

From a multi-level perspective, scenario datasets at the L2 are the most urgently needed compared to those at the L0 and L1, as they are particularly critical for defining evolving scenario labels, open-world benchmarks, and value judgments of agents in specialized domains [Chen *et al.*, 2026; Xu *et al.*, 2026; Qin *et al.*, 2026].From a multi-level perspective, scenario datasets at the L2 are the most urgently needed compared to those at the L0 and L1, as they are particularly critical for defining the value judgments of agents in specialized domains like medicine and finance. These datasets represent the "last mile" between universal ethics and practical action. Future research should prioritize extracting ethical decision-making logic from proprietary data and transforming it into securely shareable formats with embedded ethical safeguards. Ultimately, the sharing model must evolve from data ownership toward value-added services, fostering a sustainable infrastructure that balances commercial viability with ethical objectives.

7.3 Value Alignment for Muti-Agent Interaction

The effective collaboration of LLM-based MAS relies on value alignment during their interactions. Future research can focus on three key activities in games: agent identification (involving individualism or collectivism), goal setting (concerning equilibrium vs. distribution, fairness vs. efficiency), and boundary constraints (considering information symmetry, interaction rounds, and short-term vs. long-term orientation), to systematically align the values in interactions.

7.4 Value Alignment in Communication Protocols

Communication protocols, including the Model Context Protocol (MCP) and Agent-to-Agent (A2A) interfaces, are essential for agent coordination but present scalability challenges to

[1] https://github.com/Wei-ZENG1020/Value-Alignment-Agentic-AI-Papers-Survey-Taxonomy

alignment. In MCP, value conflicts in external tool integration can lead to bias, while in A2A environments, a single agent's deviation can propagate throughout the network, made worse by the traceability issues of black-box decision-making. Future inquiry must address the prevention of external "pollution" at the protocol layer and develop methods to supervise complex emergent behaviors. As agents evolve into open ecosystems, regulatory granularity must expand from individual systems to entire networks to ensure that interaction remains robust, secure, and controllable.

8 Conclusion

Research on value alignment in agentic AI systems is gaining critical importance. This survey introduces a multi-level framework that structures value principles into a three-tiered system: L0 (universal), L1 (cultural/industrial), and L2 (context-specific). We systematically review alignment methods across the technical stack, covering LLMs, single agent, and multi-agent, and further synthesize existing evaluation benchmarks while highlighting their limitations. Finally, this paper discusses current challenges and outlines future directions, aiming to provide a structured foundation for developing trustworthy and socially-aligned agentic AI.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grant number 72472048 and 62506352, the Postdoctoral Fellowship Program of CPSF under GZC20250519, Youth Innovation Project of Computer Network Information Center, Chinese Academy of Sciences under CNICYIP2025-16.

References

[Abdulhai *et al.*, 2024] Marwa Abdulhai, Gregory Serapio-Garcia, et al. Moral foundations of large language models. In *EMNLP*, pages 17737–17752, 2024.

[Achintalwar *et al.*, 2024] Swapnaja Achintalwar, Ioana Baldini, et al. Alignment studio: Aligning large language models to particular contextual regulations. *IEEE Internet Comput.*, 28(5):28–36, 2024.

[Bai *et al.*, 2022] Yuntao Bai, Saurav Kadavath, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[Broestl *et al.*, 2025] Noah Broestl, Benjamin Lange, et al. Evaluating intra-firm llm alignment strategies in business contexts. *arXiv preprint arXiv:2505.18779*, 2025.

[Carichon *et al.*, 2025] Florian Carichon, Aditi Khandelwal, et al. The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. arXiv preprint arXiv:2506.01080, 2025.

[Catton Jr, 1959] William R. Catton Jr. A theory of value. *Am. Sociol. Rev.*, 24:310, 1959.

[Cerqueira and Bezerra, 2025] Tibério Cerqueira and Pamela Bezerra. Motif: A framework for enhancing the profiling module of generative agents that simulate human behavior. In *ICAART (3)*, pages 774–781, 2025.

[Chakraborty *et al.*, 2025] Trishna Chakraborty, Udita Ghosh, et al. Heal: An empirical study on hallucinations in embodied agents driven by large language models. In *EMNLP*, pages 21226–21243, 2025.

[Chen *et al.*, 2025] Xi Chen, Chuan Qin, et al. Beyond the known: An unknown-aware large language model for open-set text classification. In *ICLR*, 2025.

[Chen *et al.*, 2026] Xi Chen, Chuan Qin, et al. GenDis: Generative-discriminative dual-view co-training for generalized category discovery. In *ACL*, 2026.

[Choi *et al.*, 2025] Hyeong Kyu Choi, Xiaojin Zhu, et al. Measuring and mitigating identity bias in multi-agent debate via anonymization. *arXiv preprint arXiv:2510.07517*, 2025.

[Coalson *et al.*, 2025] Zachary Coalson, Juhan Bae, et al. IF-guide: Influence function-guided detoxification of LLMs. In *NeurIPS*, 2025.

[Coll *et al.*, 2025] Lluís C Coll, Martin W Lauer-Schmaltz, et al. Towards the "digital me": a vision of authentic conversational agents powered by personal human digital twins. *arXiv preprint arXiv:2506.23826*, 2025.

[Fan and Jaggi, 2023] Simin Fan and Martin Jaggi. Irreducible curriculum for language model pretraining. *arXiv preprint arXiv:2310.15389*, 2023.

[Fareed *et al.*, 2025] Muhammad Fareed, Madeeha Fatima, et al. A systematic review of ethical considerations of large language models in healthcare and medicine. *Front. Digit.Health*, 7:1653631, 2025.

[Gabriel, 2020] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3):411–437, 2020.

[Gao *et al.*, 2025] Tianyu Gao, Alexander Wettig, et al. Metadata conditioning accelerates language model pre-training. In *ICML*, pages 18612–18629, 2025.

[Glaese *et al.*, 2022] Amelia Glaese, Nat McAleese, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[Graham *et al.*, 2013] Jesse Graham, Jonathan Haidt, et al. Moral foundations theory: The pragmatic validity of moral pluralism. *Adv. Exp. Soc. Psychol.*, 47:55–130, 2013.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, et al. Aligning ai with shared human values. In *ICLR*, 2021.

[Hofstede, 2011] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Read. Psychol. Cult.*, 2(1):8, 2011.

[Holt *et al.*, 2025] Samuel Holt, Max Ruiz Luyten, et al. Improving llm agent planning with in-context learning via atomic fact augmentation and lookahead search. In *ICML*, 2025.

[House *et al.*, 2004] Robert J House, Paul J Hanges, et al. *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications, 2004.

[Huang and Xiong, 2024] Yufei Huang and Deyi Xiong. Cbbq: A chinese bias benchmark dataset curated with

human-ai collaboration for large language models. In *LREC-COLING*, pages 2917–2929, 2024.

[Huang *et al.*, 2025] Yuting Huang, Leilei Ding, et al. A framework for benchmarking and aligning task-planning safety in llm-based embodied agents. *arXiv preprint arXiv:2504.14650*, 2025.

[Ji *et al.*, 2023] Jiaming Ji, Mickel Liu, et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *NeurIPS*, pages 24678–24704, 2023.

[Jiang *et al.*, 2024] Feihu Jiang, Chuan Qin, et al. Enhancing question answering for enterprise knowledge bases using large language models. In *DASFAA*, pages 273–290, 2024.

[Jiang *et al.*, 2025] Ruili Jiang, Kehai Chen, et al. A survey on human preference learning for aligning large language models. *ACM Comput. Surv.*, 58(6):1–39, 2025.

[Karinshak *et al.*, 2024] Elise Karinshak, Amanda Hu, et al. Llm-globe: A benchmark evaluating the cultural values embedded in llm output. *arXiv preprint arXiv:2411.06032*, 2024.

[Kharchenko *et al.*, 2025] Julia Kharchenko, Tanya Roosta, et al. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. In *KDD*, 2025.

[Kolt *et al.*, 2026] Noam Kolt, Nicholas Caputo, et al. Legal alignment for safe and ethical ai. *arXiv preprint arXiv:2601.04175*, 2026.

[Korbak *et al.*, 2023] Tomasz Korbak, Kejian Shi, et al. Pretraining language models with human preferences. In *ICML*, pages 17506–17533, 2023.

[Li *et al.*, 2024a] Jeffrey Li, Alex Fang, et al. Datacomp-lm: In search of the next generation of training sets for language models. In *NeurIPS*, pages 14200–14282, 2024.

[Li *et al.*, 2024b] Yuanchun Li, Hao Wen, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.

[Li *et al.*, 2025] Zhiyu Li, Shichao Song, et al. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*, 2025.

[Liu *et al.*, 2026a] Fan Liu, Jindong Han, et al. Foundation models for scientific discovery: From paradigm enhancement to paradigm transition. *NeurIPS*, 2026.

[Liu *et al.*, 2026b] Fan Liu, Xiaozhao Zeng, and Hao Liu. Towards multimodal data-driven scientific discovery powered by llm agents. In *ICLR*, 2026.

[Long *et al.*, 2024] Qingqing Long, Yuchen Yan, et al. Moat: Graph prompting for 3d molecular graphs. In *ACM CIKM*, pages 1586–1596, 2024.

[Long *et al.*, 2026a] Qingqing Long, Haotian Chen, et al. Sciencedb ai: An llm-driven agentic recommender system for large-scale scientific data sharing services. *arXiv preprint arXiv:2601.01118*, 2026.

[Long *et al.*, 2026b] Qingqing Long, Shuai Liu, et al. A survey of large language models for traffic forecasting: Methods and applications. *IEEE Trans. Big Data*, 2026.

[Luong *et al.*, 2024] Tinh Luong, Thanh-Thien Le, et al. Realistic evaluation of toxicity in large language models. In *ACL*, pages 1038–1047, 2024.

[Ma *et al.*, 2025] Junxia Ma, Changjiang Wang, et al. Exploring multi-agent debate for zero-shot stance detection: A novel approach. *Appl. Sci.*, 15(9):4612, 2025.

[Maini *et al.*, 2025] Pratyush Maini, Sachin Goyal, et al. Safety pretraining: Toward the next generation of safe ai. *arXiv preprint arXiv:2504.16980*, 2025.

[Mayoral-Macau *et al.*, 2025] Arnau Mayoral-Macau, Manel Rodriguez-Soto, et al. Designing ethical environments using multi-agent reinforcement learning. In *ALA*, 2025.

[Nadeem *et al.*, 2021] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *ACL*, pages 5356–5371, 2021.

[Naik *et al.*, 2025] Akshat Naik, Patrick Quinn, et al. Agentmisalignment: Measuring the propensity for misaligned behaviour in llm-based agents. *arXiv preprint arXiv:2506.04018*, 2025.

[Nandi *et al.*, 2025] Subhrangshu Nandi, Arghya Datta, et al. Sop-bench: Complex industrial sops for evaluating llm agents. *arXiv preprint arXiv:2506.08119*, 2025.

[Niklaus *et al.*, 2024] Joel Niklaus, Veton Matoshi, et al. Multilegalpile: A 689gb multilingual legal corpus. In *ACL*, pages 15077–15094, 2024.

[Parrish *et al.*, 2022] Alicia Parrish, Angelica Chen, et al. Bbq: A hand-built bias benchmark for question answering. In *ACL*, pages 2086–2105, 2022.

[Peng *et al.*, 2025] Hao Peng, Yunjia Qi, et al. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems. In *ACL*, pages 15934–15949, 2025.

[Prabhumoye *et al.*, 2023] Shrimai Prabhumoye, Mostofa Patwary, et al. Adding instructions during pretraining: Effective way of controlling toxicity in language models. In *ACL*, pages 2636–2651, 2023.

[Qiang *et al.*, 2024] Yao Qiang, Xiangyu Zhou, et al. Learning to poison large language models during instruction tuning. *arXiv*, 2024.

[Qin *et al.*, 2025a] Chuan Qin, Xin Chen, et al. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In *KDD*, pages 5754–5765, 2025.

[Qin *et al.*, 2025b] Chuan Qin, Chuyu Fang, et al. Cotr: Efficient job task recognition for occupational information systems with class-incremental learning. *ACM Trans. Manage. Inf. Syst.*, 16(2):1–30, 2025.

[Qin *et al.*, 2025c] Chuan Qin, Le Zhang, et al. A comprehensive survey of artificial intelligence techniques for talent analytics. *Proc. IEEE*, 2025.

[Qin *et al.*, 2026] Chuan Qin, Xi Chen, et al. BOLT: Benchmarking open-world learning for text classification. In *ACL*, 2026.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, et al. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, pages 53728–53741, 2023.

[Russell and Norvig, 2016] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[Sagiv *et al.*, 2017] Lilach Sagiv, Sonia Roccas, et al. Personal values in human life. *Nat. Hum. Behav.*, 1(9):630–639, 2017.

[Schoenegger *et al.*, 2025] Loris Schoenegger, Lukas Thoma, et al. Influence-driven curriculum learning for pre-training on limited data. In *BabyLM*, pages 356–379, 2025.

[Schwartz, 2007] Shalom H. Schwartz. Basic human values: Theory, methods, and application. *Risorsa Uomo*, pages 261–283, 2007.

[Shen *et al.*, 2026] Tingjia Shen, Hao Wang, et al. Prompting is not enough: exploring knowledge integration and controllable generation on large language models. *Big Data Min. Anal.*, 9(2):563–579, 2026.

[Shetty *et al.*, 2025] Anudeex Shetty, Amin Beheshti, et al. Vital: A new dataset for benchmarking pluralistic alignment in healthcare. In *ACL*, pages 22954–22974, 2025.

[Tamang and Bora, 2025] Sagar Tamang and Dibya Jyoti Bora. Enforcement agents: Enhancing accountability and resilience in multi-agent ai frameworks. *arXiv preprint arXiv:2504.04070*, 2025.

[Tang *et al.*, 2024] Xiangru Tang, Qiao Jin, et al. Prioritizing safeguarding over autonomy: Risks of llm agents for science. In *ICLR Workshop on LLM Agents*, 2024.

[Tennant *et al.*, 2023] Elizaveta Tennant, Stephen Hailes, et al. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *IJCAI*, pages 317–325, 2023.

[Tennant *et al.*, 2025] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for llm agents. In *ICLR*, 2025.

[Tong *et al.*, 2025] Zhenyu Tong, Chuan Qin, et al. From missteps to mastery: Enhancing low-resource dense retrieval through adaptive query generation. In *KDD*, pages 1373–1384, 2025.

[Vishwakarma *et al.*, 2025] Akash Vishwakarma, Hojin Lee, et al. Cognitive weave: Synthesizing abstracted knowledge with a spatio-temporal resonance graph. *arXiv preprint arXiv:2506.08098*, 2025.

[Wang *et al.*, 2025a] Ryan Wang, Matthew Finlayson, et al. Teaching models to understand (but not generate) high-risk data. In *COLM*, 2025.

[Wang *et al.*, 2025b] Zhao Wang, Sota Moriyama, et al. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2501.06322*, 2025.

[Wang *et al.*, 2026] Chao Wang, Yixin Song, et al. Face: A general framework for mapping collaborative filtering embeddings into llm tokens. In *NeurIPS*, pages 146012–146039, 2026.

[Xu *et al.*, 2024] Jiashu Xu, Mingyu Derek Ma, et al. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL*, pages 3111–3126, 2024.

[Xu *et al.*, 2025a] Shaoyang Xu, Yongqi Leng, et al. Self-pluralising culture alignment for large language models. In *NAACL*, pages 6859–6877, 2025.

[Xu *et al.*, 2025b] Weikai Xu, Chengrui Huang, et al. Llm-based agents for tool learning: A survey. *Data Sci. Eng.*, pages 1–31, 2025.

[Xu *et al.*, 2026] Wenxi Xu, Chuan Qin, et al. TLSA: LLM-guided text-label space alignment with contrastive learning for generalized category discovery. In *ACL*, 2026.

[Yang *et al.*, 2025] Yi Yang, Yitong Ma, et al. Minimizing hallucinations and communication costs: Adversarial debate and voting mechanisms in llm-based multi-agents. *Appl. Sci.*, 15(7):3676, 2025.

[Yao *et al.*, 2025] Jing Yao, Xiaoyuan Yi, et al. Value compass leaderboard: A platform for fundamental and validated evaluation of llms values. *arXiv*, 2025.

[Zhang *et al.*, 2026a] Haodong Zhang, Xinyue Wang, et al. Fairgc: Fostering individual and group fairness for deep graph clustering. In *AAAI*, pages 28194–28202, 2026.

[Zhang *et al.*, 2026b] Yang Zhang, Amr Mohamed, et al. Beyond random sampling: Efficient language model pretraining via curriculum learning. In *ACL*, pages 5776–5794, 2026.

[Zhang, 2025] Min Zhang. Optimizing academic engagement and mental health through ai: An experimental study on llm integration in higher education. *Front. Psychol.*, 16:1641212, 2025.

[Zhou *et al.*, 2025] Duo Zhou, Junyu Zhang, et al. A survey on alignment for large language model agents. In *UIUC Spring 2025 CS598 LLM Agent Workshop*, 2025.

[Zhu *et al.*, 2025a] Zhihong Zhu, Fan Zhang, et al. A survey on multi-modal intent recognition: Recent advances and new frontiers. In *EMNLP*, pages 15223–15236, 2025.

[Zhu *et al.*, 2025b] Zhihong Zhu, Yunyan Zhang, et al. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *ACL*, pages 6748–6769, 2025.

[Ziems *et al.*, 2022] Caleb Ziems, Jane Yu, et al. The moral integrity corpus: A benchmark for ethical dialogue systems. In *ACL*, pages 3755–3773, 2022.

[Zou *et al.*, 2024] Rui Zou, Mengqi Wei, et al. Gradual vigilance and interval communication: Enhancing value alignment in multi-agent debates. *arXiv preprint arXiv:2412.13471*, 2024.