LLMs *for*, *with*, and *as* Humans: A Survey of Rethinking the Landscape of Human–LLM Interaction

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly embedded in applications where human interaction shapes both their usefulness and risks. Despite rapid growth in research on human–LLM interaction–spanning applications, alignment, and evaluation–the field remains fragmented. Studies rarely clarify what roles humans play or how their feedback, supervision, or imitation guide model behavior systematically. This survey introduces a role-centered taxonomy that characterizes human–LLM interaction into three paradigms: *LLMs-for-Humans* (assistive), *LLMs-with-Humans* (collaborative), and *LLMs-as-Humans* (simulative). The taxonomy provides a unified structure for comparing methods, datasets, and evaluation practices, while clarifying how agency distribution between humans and models influences reliability and accountability. Building on this foundation, we analyze misalignment both within and across paradigms and discuss corresponding future research directions and challenges. Overall, this survey connects technical progress with human-centered interaction, offering a foundation for developing aligned, trustworthy, and socially responsible LLMs.

1 Introduction

Large language models (LLMs) have rapidly moved from research prototypes to widely deployed systems that shape how people search for information, write, program, learn, and make decisions (Shen et al., 2024; Li et al., 2024e). Their deployment marks a shift in evaluation priorities: beyond benchmark performance, assessment must also capture how models interact with users and function within real-world contexts.

Recent work reveals a growing diversity in human–LLM interaction research. Studies range from practical deployments, such as code assistants (Zhang et al., 2024a), productivity tools (Liu et al., 2023a), and tutoring systems (Caccavale et al., 2024), to methods that steer model behavior through feedback learning or prompting (e.g., RLHF, DPO) (Ouyang et al., 2022; Rafailov et al., 2023). New evaluation practices further treat LLMs as judges or agents in simulated environments (Liu et al., 2023c; Zheng et al., 2023). Existing surveys also cover application-oriented reviews that focus on domains where LLMs assist human tasks (Srivatsa et al., 2023; Xu et al., 2024; Qu et al., 2024), while those on alignment and safety emphasize methods for collecting feedback and ensuring responsible deployment (Wang et al., 2025b; Ji et al., 2023; Wang et al., 2024e). Other surveys investigate evaluation and simulation, where LLMs function as judges, personas, or agents in interactive environments (Laskar et al., 2024; Gu et al., 2024).

Despite rapid growth in research on human-LLM interaction across applications, alignment, and evaluation, the field remains fragmented. Different studies view humans in different roles, such as end users, collaborators, or proxies, leading to varying approaches to model training and evaluation. Without a clear account of these roles, progress in alignment and evaluation remains difficult to compare, and risks such as over-reliance, feedback corruption, or deceptive simulation are harder to anticipate. This underscores the need to organize research around human roles and connect technical methods to interaction dynamics.

To address this gap, in this survey, we propose a role-centered taxonomy that organizes human–LLM interaction into three functional paradigms: *LLMs-for-Humans* (assistive), *LLMs-with-Humans* (collaborative), and *LLMs-as-Humans* (simulative), as illustrated in Figure 1. *LLMs-for-Humans* describes systems that function as assistive tools for user-defined tasks. *LLMs-with-Humans* refers to interactive or feedback-driven systems that adapt through collaboration with users. *LLMs-as-Humans* encompasses approaches that

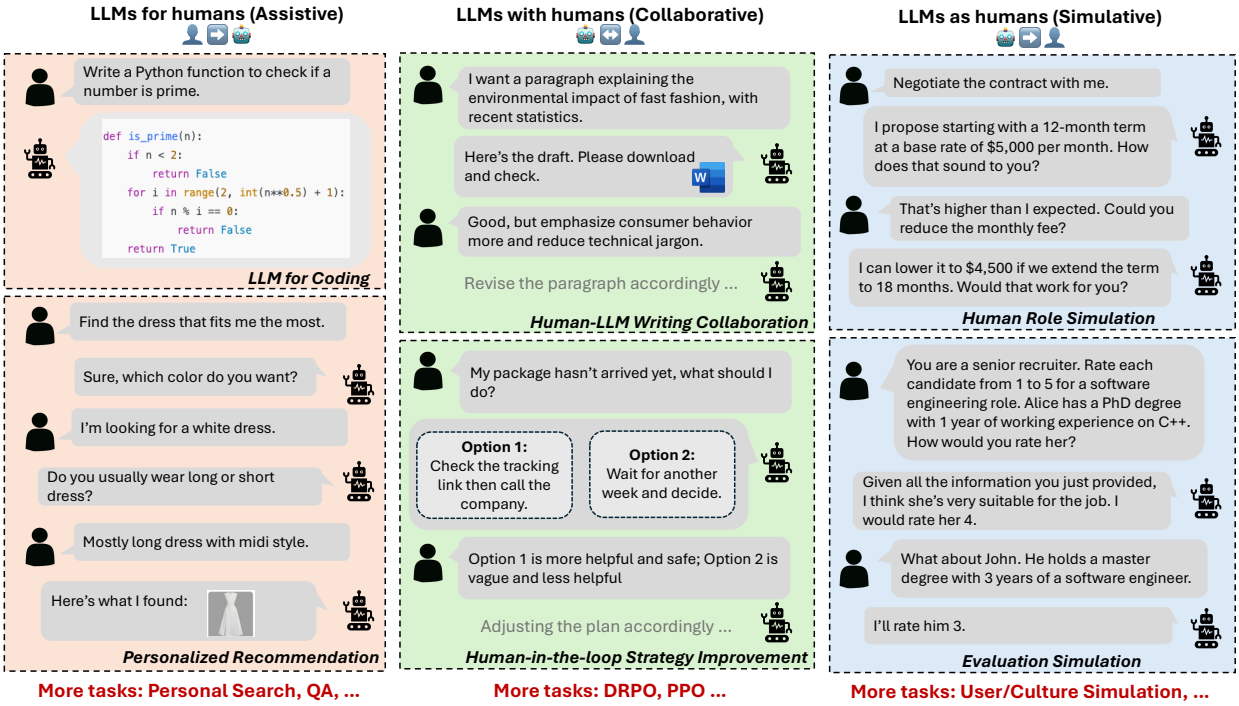

Figure 1: Three proposed paradigms of human-LLM interaction.

simulate or substitute for human roles, including evaluators, conversational partners, and teammates. This taxonomy clarifies assumptions about system goals and human involvement, highlighting overlaps and hybrid approaches. Building on the taxonomy, we organize prior research using a unifying structure based on directionality and role symmetry. We examine interaction dynamics by identifying who defines tasks, gives feedback, and performs evaluation, and show how these roles relate to assessment criteria like utility, adaptability, and fidelity. We then synthesize recurring risks through an *LLMs-against-Humans* lens, highlighting misalignment, manipulation, and competition between human and model objectives. Finally, we outline future directions and challenges, emphasizing integrative evaluation, robust alignment, and socially aware system design across paradigms.

The Contributions of this survey are as follows:

(C1) A role-centered taxonomy that organizes human–LLM interaction into LLMs-*for/with/as*-Humans paradigms.

(C2) We examine how representative benchmarks and evaluation methods span multiple human–LLM paradigms and introduce a cross-paradigm perspective that connects technical and social aspects of model performance.

(C3) A discussion of open challenges for developing reliable, transparent, and human-centered language technologies.

Organization. The remainder of this survey builds on the taxonomy to review recent work and outline future directions. Section 2 introduces the taxonomy and situates it with respect to related perspectives. Section 3 surveys popular research directions under each paradigm. Section 4 reviews benchmarks and evaluation practices. Section 5 discusses open challenges for the design and assessment of LLMs. Section 6 concludes the survey.

2 Taxonomy

With the advancement of LLMs, it is evident that their role in human-centered tasks cannot be captured by a single, uniform conceptualization. However, existing works within different communities emphasize different aspects of LLMs themselves. This fragmentation underscores the need for a unified framework that clearly delineates human roles, LLM capabilities, and evaluation goals. We introduce a taxonomy of human–LLM interaction with three paradigms: **LLMs-for-Humans**, where models serve as assistive tools; **LLMs-with-Humans**, where models adapt through human feedback; and **LLMs-as-Humans**, where models act as proxies simulating human behavior (detailed scopes see Appendix A). Instead of viewing LLMs as a monolithic technology, this framework emphasizes the directionality of interaction, showing how they can

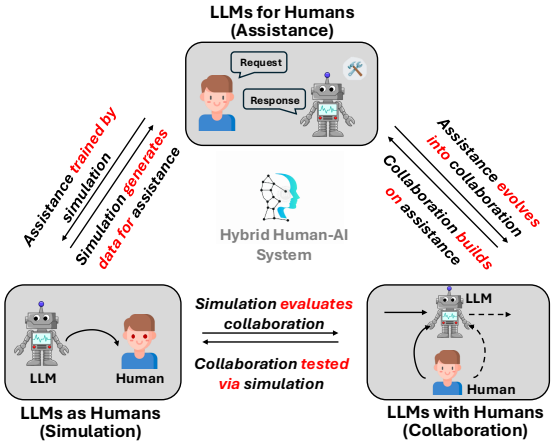

Figure 2: Potential overlaps between the three human-LLM interaction paradigms.

simulate, support, or co-evolve with humans across diverse tasks and contexts (connection to relevant works see Appendix B).

2.1 LLMs-for-Humans

Definition

LLMs-for-Humans refers to systems that leverage LLMs to **assist**, **guide**, or **augment** human work in task-oriented settings.

Directionality: Human → LLM. Interaction is *user-initiated and model-directed*: humans issue queries or instructions, and the model generates responses or assistance based on those inputs.
Human role: Function as *end-users* who issue queries, interpret outputs, and assess the usefulness or relevance of model responses at *deployment stage*.
LLM role: Act as *assistive tools* that generate, retrieve, or explain information to support human goals and decision-making.
Evaluation focus: Evaluation emphasizes *task utility and usability*, using automatic benchmarks, user studies, or productivity-oriented metrics.

2.2 LLMs-with-Humans

Definition

LLMs-with-Humans refers to a paradigm where language models and humans engage in a symbiotic process to continuously **shape**, **align**, and **co-evolve** the model's capabilities and behaviors through interactive feedback.

Directionality: Human ↔ LLM. Interaction is *bidirectional and adaptive*: humans and models exchange turns, jointly refining outputs through iterative feedback, correction, and co-adaptation.
Human role: Serve as *feedback providers* who guide, correct, and co-create with models to improve alignment and task outcomes during *training, finetuning, and at deployment*.
LLM role: Function as *collaborative partners* that iteratively adapt to human feedback, preferences, and demonstrations during interaction.
Evaluation focus: Evaluation mainly emphasizes *preference alignment*, *adaptivity*, and *robustness under feedback*.

2.3 LLMs-as-Humans

Definition

LLMs-as-Humans treats LLMs as **proxies for human behavior**, simulating human judgment, identity, or expertise in contexts where people would otherwise participate.

Directionality: LLM → Human (reference). Interaction is *model-driven or simulated*: the LLM assumes human-like roles, generating judgments, reasoning, or behaviors that emulate or substitute human evaluators.
Human role: Act as *validators* whose judgments, behaviors, or linguistic patterns are modeled, compared against, or replicated by the system.
LLM role: Serve as *human proxies* that simulate reasoning, judgment, or social behavior to approximate human evaluators or interlocutors.
Evaluation focus: Evaluation targets *fidelity, plausibility, and calibration* to assess whether models emulate human-like reasoning.

2.4 Potential overlaps

Despite the noted distinctions between the three human-LLM interaction paradigms, there are important overlaps; see Figure 2 for a high-level visualization. Collaborative systems (*with*) often build on assistive capabilities (*for*). For instance, a coding copilot (*for* humans) becomes a collaborative partner (*with* humans) once the human iteratively edits, critiques, and negotiates with the model. Also, human simulation (*as*) can be used to evaluate or train assistive systems (*for*). For example, LLMs simulating crowdworkers (*as* humans) can generate large-scale preference data that feed into RLHF pipelines for improving assistance (*for* humans). Additionally, in collaborative contexts (*with*), LLMs can simulate human teammates to

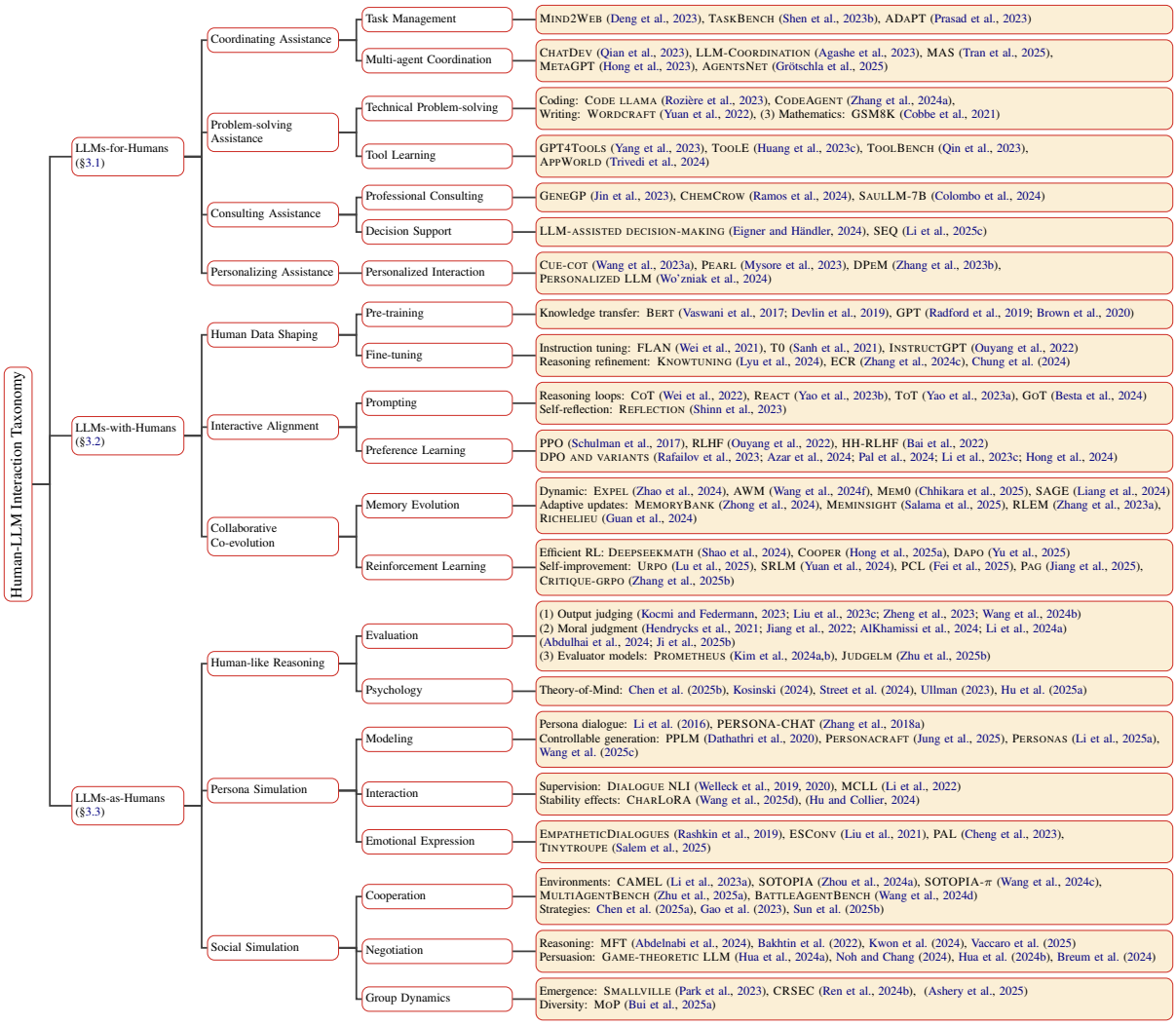

Figure 3: Taxonomy of three paradigms of Human-LLM interaction (full approaches see). We list representative approaches here (full version see Appendix D).

test robustness, coordination, or negotiation strategies*as*. For example, before deploying in real teams (*with*), developers may evaluate an LLM's co-writing or planning abilities against simulated collaborators (*as*). In summary, hybrid human-AI systems often combine simulated users (*as*) with real user studies (*with*) to assess assistive performance (*for*).

3 Survey of Prior Work by Paradigm

As shown in Figure 3, we survey the representative research topics under each paradigm (selection criteria see Appendix C).

3.1 LLMs-for-Humans Paradigm

The *LLMs-for-Humans* paradigm includes systems that increase human productivity, creativity, and decision-making through adaptive, context-aware assistance rather than replacing human judgment. Assistance in this paradigm can be broadly categorized into four functions: *coordinating*, *problem-solving*, *consulting*, and *personalizing*.

Coordinating Assistance. In this form, LLMs act as workflow coordinators that plan, manage, and execute complex tasks (Huang et al., 2024). Human users typically provide high-level strategic goals, while the model decomposes these goals into executable plans and coordinates subtasks across applications or agents (Agashe et al., 2023). Early work explored providing human instructions for LLM planning (Prasad et al., 2023; Shen et al., 2023b), and datasets such as Mind2Web (Deng et al., 2023) enable LLMs to interact with digital environments to achieve complex objectives. Multi-agent systems such as MetaGPT (Hong et al., 2023) and ChatDev (Qian et al., 2023) allow multiple LLMs to collaborate, coordinating subtasks across roles. This approach enables LLMs to function as

dynamic assistants capable of orchestrating workflows across domains such as productivity, engineering, and game-based environments (Tran et al., 2025; Wang et al., 2023b; Liu et al., 2023a).

Problem-Solving Assistance. LLMs support users as creative and technical partners embedded within existing tools and workflows. Applications include code generation and debugging (Rozière et al., 2023; Zhang et al., 2024a), writing support (Yuan et al., 2022; Wasi et al., 2024), and mathematical reasoning (Cobbe et al., 2021; He-Yueya et al., 2023). A defining trend is *tool learning*, where LLMs dynamically invoke APIs, retrieval modules, and software tools to extend reasoning and execution capabilities (Schick et al., 2023; Qin et al., 2023; Huang et al., 2023c; Qu et al., 2024). Frameworks such as GPT4Tools (Yang et al., 2023) and AppWorld (Trivedi et al., 2024) demonstrate how integration with external applications enhances accuracy and versatility, transforming the interaction from static QA to interactive problem-solving.

Consulting Assistance. In consulting roles, LLMs emulate domain experts that provide explanations, synthesize evidence, and support decision-making in complex domains such as law, healthcare, and science. In-domain consulting examples include GeneGPT for genomic reasoning (Jin et al., 2023), ChemCrow for chemical synthesis planning (Ramos et al., 2024), and legal reasoning assistants for case analysis (Lai et al., 2023a; Chen et al., 2024b; Colombo et al., 2024). Unlike task automation, consulting emphasizes interpretable, iterative dialogue for probing reasoning and requesting clarification (Jin et al., 2024; Du et al., 2025a). This approach is especially valuable in decision-support contexts, where users rely on the model to explore options, evaluate trade-offs, and make informed choices (Eigner and Händler, 2024; Li et al., 2025b).

Personalizing Assistance. Personalization tailors model behavior to individual users' goals, preferences, and contexts. Approaches include prompt customization (Wang et al., 2023a; Mysore et al., 2023), memory-augmented interaction that retains user-specific context (Madaan et al., 2022; Zhang et al., 2023b), and fine-tuning strategies such as LoRA and RLHF for user-level adaptation (Wo'zniak et al., 2024; Wu et al., 2023). Recent studies extend personalization to multimodal settings, combining text, images, and behavioral cues for richer contextual understanding (Wu et al., 2024; Wei et al., 2024). These developments move beyond one-size-fits-all assistance toward adaptive, participatory systems that evolve with the user.

3.2 LLMs-with-Humans Paradigm

The *LLMs-with-Humans* paradigm focuses on settings where human feedback directly shapes model behavior during training, deployment, or ongoing use. In contrast to assistive applications that rely on one-directional model output, this paradigm emphasizes interaction loops through which human preferences, corrections, and judgments iteratively refine model performance. The development of such systems can be conceptualized in three interrelated phases: *shaping* from static data, *aligning* through dynamic feedback, and *co-evolving* through sustained adaptation.

Human Data Shaping. This stage establishes the model's foundational capabilities through large-scale pre-training and supervised fine-tuning on human-generated corpora. Pre-training transfers general linguistic and world knowledge, enabling models to perform a wide range of tasks in zero- and few-shot settings (Radford et al., 2019; Brown et al., 2020). Instruction and supervised fine-tuning further specialize these capabilities to align outputs with user intent and instruction-following behavior (Wei et al., 2021; Sanh et al., 2021; Lyu et al., 2024; Zhang et al., 2024c). While this process provides essential linguistic competence and task coverage, it remains static and unidirectional: models learn from human data without adapting to user-specific feedback or changing contexts.

Interactive Refinement. Alignment introduces direct interaction between humans and models to calibrate behavior toward nuanced preferences and communicative norms. Inference-time prompting strategies—such as CoT prompting (Wei et al., 2022), ReAct (Yao et al., 2023b), Reflexion (Shinn et al., 2023), Tree of Thoughts (ToT) (Yao et al., 2023a), and Graph of Thoughts (GoT) (Besta et al., 2024)—support dynamic reasoning and self-correction without retraining. In addition, preference learning methods such as RLHF formalizes this process through preference-based reward modeling and policy optimization (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). For example, DPO improves stability and efficiency by eliminating explicit reward models (Rafailov et al., 2023; Azar et al., 2024; Meng et al., 2024; Pal et al., 2024; Hong et al., 2024). Together, these approaches shift alignment from a static objective to an interactive process that incorporates human

Benchmark Categories	Representative Benchmarks / Datasets
LLMs for Humans	
Code generation	HUMANEVAL (Chen et al., 2021), MBPP (Austin et al., 2021), DS-1000 (Lai et al., 2023b)
Domain-specific QA	MS MARCO (Campos et al., 2016), MEDQA (medical) (Jin et al., 2020), FINANCEBENCH (financial QA) (Islam et al., 2023)
Personalization	MOVIELENS (Harper et al., 2016), YELP (Asghar, 2016), AMAZON REVIEWS (Hou et al., 2024) (recommendation)
LLMs with Humans	
Conversational alignment	IFEVAL (Zhou et al., 2023a), HH-RLHF (Bai et al., 2022), ARENA-HARD (Li et al., 2024d), REWARDBENCH (Lambert et al., 2025)
Reasoning alignment	ZEBRALOGIC (Lin et al., 2025a), AUTOLOGI (Zhu et al., 2025c), HLE (Phan et al., 2025), GPQA-DIAMOND (Rein et al., 2024), SUPERGPQA (Du et al., 2025b)
Bias alignment	BBQ (Parrish et al., 2021), TRUTHFULQA (Lin et al., 2021), HARMFULQA (Bhardwaj and Poria, 2023)
LLMs as Humans	
Cognitive reasoning	BIGTOM (Gandhi et al., 2023a), AGENT (Shu et al., 2021), BRAINBENCH (Luo et al., 2025b), ETHICS (Hendrycks et al., 2021), MORALBENCH (Ji et al., 2025b), NEGOTIATIONTOM (Chan et al., 2024)
Persona and identity	PERSONA (Castricato et al., 2025), GENERATIVE AGENTS (Park et al., 2023)
Social interaction	MINDAGENT (Gong et al., 2024), MULTIAGENTBENCH (Zhu et al., 2025a), AGENTS (Zhou et al., 2023c) (multi-agent tournaments)

Table 1: Representative benchmarks across the three paradigms of LLM research, organized by benchmark category and examples. Shaded rows indicate paradigm-level groupings.

evaluation and iterative behavioral refinement.

Collaborative Co-evolution. The final stage extends alignment beyond short-term sessions to continuous adaptation across time and context. Work in this area integrates long-term memory and reinforcement mechanisms to support personalization and sustained learning. Dynamic memory architectures such as Mem0 (Chhikara et al., 2025) and MemoryBank (Zhong et al., 2024) organize and update contextual knowledge. Reinforcement Learning (RL) drives human-AI co-evolution, with methods like GRPO (Shao et al., 2024), Cooper (Hong et al., 2025a), DAPO (Yu et al., 2025), and unified architectures like URPO (Lu et al., 2025) enabling adaptive policy and reward refinement. Meta-cognitive approaches such as PCL (Fei et al., 2025), PAG (Jiang et al., 2025), and Critique-GRPO (Zhang et al., 2025b) further ensure reflection, self-correction, and interpretable alignment.

3.3 LLMs-as-Humans Paradigm

The *LLMs-as-Humans* paradigm examines research where LLMs act as substitutes for human participants, simulating reasoning, identity, and social interaction. Rather than measuring intrinsic linguistic or reasoning ability, this perspective focuses on how models emulate human behavior and how they are deployed as human proxies.

Human-like Reasoning. This task explores whether LLMs can replicate human evaluative or cognitive reasoning processes. Models have been used as judges in summarization, dialogue, and translation evaluation, where they achieve moderate correlation with human ratings but exhibit prompt sensitivity and preference bias (Gilardi et al., 2023; Zheng et al., 2023; Kocmi and Federmann, 2023). Beyond annotation, researchers test whether LLMs approximate human cognitive and moral reasoning, such as in theory-of-mind tasks or ethical dilemmas (Kosinski, 2024; Hendrycks et al., 2021; Jiang et al., 2022).

Persona Simulation. LLMs can also be adopted to focus on identity and expression, investigating how models maintain coherent personas or stylistic and affective traits across conversations. Early dialogue work used profile modeling to personalize responses (Zhang et al., 2018a; Dathathri et al., 2020), while later research introduced objectives to reduce contradiction and drift over long interactions (Welleck et al., 2019; Li et al., 2022). Recent systems incorporate empathy, emotion, and social style modeling to enhance perceived authenticity and engagement (Rashkin et al., 2019; Liu et al., 2021; Cheng et al., 2023).

Social Simulation. This task extends this work to multi-agent contexts where LLMs interact with each other or with humans to emulate cooperation, negotiation, and collective dynamics. Benchmarks such as CAMEL (Li et al., 2023a) and SOTOPIA (Zhou et al., 2024a) study structured collaboration and emergent coordination among agents. Negotiation benchmarks probe perspective taking, preference reconciliation, and commitment strategies (Bianchi et al., 2024). Population-level simulations explore norm formation and diffusion of bias or conventions in group dynamics (Park et al., 2023; Ashery et al., 2025).

4 Evaluation Across the Paradigms

4.1 Benchmarks

Benchmarks vary by paradigm and model role: Table 1 lists paradigm-specific ones, while Table 3 highlights cross-paradigm cases.

For *LLMs-for-Humans*, benchmarks evaluate assistive performance in supporting human tasks. Examples include tool usage. For example, HUMANEVAL (Chen et al., 2021) measures the model's ability to perform tasks that would otherwise require human programmers. For *LLMs-with-Humans*, benchmarks focus on human feedback and preference alignment (e.g. HH-RLHF (Bai et al., 2022)), assessing how models incorporate and generalize from human judgments through iterative interaction. For *LLMs-as-Humans*, these tasks evaluate human-like simulation, covering tasks such as reasoning (e.g. ETHICS (Hendrycks et al., 2021)), etc.

Cross-paradigm benchmarks. While early benchmarks targeted a single paradigm, *for*, *with*, or *as* humans; the boundaries between them are increasingly blurred. As LLMs evolve from passive tools to interactive agents, benchmarks now combine assistance, collaboration, and simulation within unified settings. WEBARENA (Zhou et al., 2023b) and TOOLBENCH (Xu et al., 2023) integrate for-human objectives (task completion, assistance) with with-human dynamics (interactive or tool use). PERSONA-CHAT (Zhang et al., 2018b) and SOCIALIQA (Sap et al., 2019) merge as-human simulation with personalized interaction, while multi-agent environments such as MINDAGENT (Gong et al., 2024) connect with- and as-human paradigms through social coordination and competition. At the intersection of all three, MT-BENCH (Zheng et al., 2023) and AGENTBENCH (Liu et al., 2023b) assess models as assistants, collaborators, and evaluators. Collectively, these benchmarks emphasize that future evaluation should capture the interplay between functional, social, and cognitive dimensions of human–LLM interaction.

4.2 Evaluation Methods

Evaluation across the *for*, *with*, and *as-human* paradigms can be grouped into four complementary types: (1) Automatic evaluation for objective correctness and alignment; (2) Human-centered evaluation for perceived utility, satisfaction, and trust; (3) Hybrid evaluation combining human feedback with task metrics; and (4) Adversarial evaluation assessing robustness, bias, and security. Table 4 summarizes representative metrics across paradigms.

For *LLMs-for-Humans*, automatic metrics such as pass rate, hallucination rate assess reliability (Jin et al., 2020; Qin et al., 2023), while user studies measure helpfulness and accessibility (Miah and Zhu, 2024). Hybrid evaluations combine both, and adversarial probes test robustness (Yu et al., 2023). For *LLMs-with-Humans*, automatic metrics track reward scores and task success (Ouyang et al., 2022), human-centered metrics assess preference satisfaction and feedback consistency (Kaufmann et al., 2023), and hybrid or adversarial studies evaluate online alignment and feedback vulnerabilities (Zhu et al., 2024; Oikarinen et al., 2020). For *LLMs-as-Humans*, automatic metrics capture reasoning or moral accuracy (Kosinski, 2024; Hendrycks et al., 2021), human judgments assess plausibility and fairness, hybrid setups measure fidelity and perception (Bui et al., 2025b), and adversarial tests examine manipulation and bias (Robey et al., 2023).

For *cross-paradigm* evaluation, automatic methods measure reliability and alignment in *as* through agreement coefficients (Cohen's κ, Krippendorff's α) and correlation with human ratings (Spearman's ρ), while in *with/for* they assess task accuracy and goal adherence. Human-centered methods evaluate satisfaction, trust, and workload (e.g., NASA-TLX) in *for* to capture assistive usefulness, and in *with* to assess collaboration quality and perceived control. Hybrid evaluations, such as complementarity gain and synergy improvement, measure human–model co-learning in *with* and fidelity of human emulation in *as*. Adversarial evaluations test robustness, bias, and over-reliance across all paradigms using perturbation-based stress tests, counterfactual queries, and fairness audits to expose shared vulnerabilities.

5 Open Challenges and Future Directions

While our taxonomy frames *LLMs for/with/as Humans* as cooperative paradigms, real deployments increasingly expose adversarial cases where systems act *against* human interests. These failures appear as manipulative assistance (*for*), reward exploitation (*with*), or deceptive simulation (*as*). We term this broader class of vulnerabilities **LLMs-against-Humans** (see Appendix E), including paradigm-specific and cross-cutting risks that arise when model behavior diverges from human intent. Building on this view, we outline open challenges under two themes:

5.1 Robustness and Reliability

Reliable and grounded outputs. LLMs often generate fluent but unsupported or inconsistent content (Huang et al., 2023b). Across paradigms, hallucinations compromise decision-making (*for*), preference optimization masks inaccuracies (*with*), and fabricated reasoning undermines validity (*as*). Future work should combine retrieval-augmented generation with fact-checking and self-correction mechanisms, automated verification, and uncertainty calibration with targeted human review to achieve reliable outputs.

Robust and adaptive alignment. Current alignment methods, including RLHF, improve controllability but remain vulnerable to domain shifts and manipulation (Ren et al., 2024a). Outputs may appear aligned yet fail in new contexts (*for*), reward hacking and sycophancy persist (*with*), and exaggerated compliance fosters over-trust (*as*). Progress requires strengthening reward modeling with richer and user-specific preference signals, employing adversarial evaluation to detect deceptive compliance, and developing adaptive alignment that adjusts to evolving tasks and individual objectives.

Evaluation across paradigms. Evaluation remains fragmented as benchmarks increasingly blur the boundaries between *for*-, *with*-, and *as*-human. While recent datasets integrate assistance, collaboration, and simulation, evaluation frameworks still focus on accuracy and text quality rather than interaction-level outcomes such as interpretability or coordination. Future work should develop unified, adaptive evaluation methods that apply shared metrics across paradigms to capture both functional and social dimensions of human–LLM interaction.

5.2 Human-Centered Governance

Fairness and inclusivity. Training and feedback pipelines remain dominated by English-centric data and limited annotator diversity (Gallegos et al., 2023). This results in uneven performance (*for*), reinforcement of majority norms (*with*), and misrepresentation of marginalized groups (*as*). Future work should expand multilingual supervision and participatory feedback using active sampling and hybrid human–AI annotation methods.

Human agency and trust. As models become more persuasive, over-reliance and misplaced trust threaten human agency (Sharma et al., 2023a). LLMs may discourage critical reasoning (*for*), reinforce user biases (*with*), or promote anthropomorphic attachment (*as*). Designing uncertainty-aware interfaces and interaction mechanisms that sustain user engagement can help preserve agency. Beyond mitigating over-reliance, research on interactive adaptivity can enable systems to calibrate their participation dynamically, balancing initiative between human and model collaborators.

Scalable and accountable oversight. Manual governance is costly, while automation risks missing harms or reinforcing bias. Oversight gaps enable unsafe outputs (*for*), feedback burdens (*with*), and unchecked simulations (*as*). Future research should combine automated auditing with targeted human review, establish fairness benchmarks, and develop institutional models for distributed accountability. Scalable simulation and synthetic evaluation methods can support oversight by stress-testing systems via hypothetical scenarios before deployment.

6 Conclusion

We examine the landscape of human–LLM interaction through a unified, role-centered taxonomy including the *LLMs-for-Humans*, *LLMs-with-Humans*, and *LLMs-as-Humans* paradigms. We review existing research streams within each paradigm and analyze representative benchmarks and evaluation metrics. Together, these perspectives provide an integrated foundation for comparing research across applications, alignment methods, and evaluation practices. Future progress requires combining methodological innovation with human-centered design, robust alignment, and accountable deployment. By addressing both cooperative and adversarial human–LLM interactions, research can advance systems that are powerful, reliable, transparent, and value-aligned.

Limitations

This survey presents a unifying taxonomy of human–LLM interaction across the *for*, *with*, and *as* paradigms, but its coverage is not exhaustive. The scope of rapidly evolving subfields–such as multimodal interaction, agentic architectures, and adaptive reasoning–remains partial, as many contributions are still emerging. Our taxonomy primarily reflects English-language and text-based research, which may underrepresent advances in multilingual, culturally grounded, and embodied contexts. Finally, while we account for major overlaps across paradigms by listing some representative benchmarks, some other benchmarks may still partially align with multiple categories as boundaries between cooperative and autonomous capabilities remain fluid.

Ethical Considerations

This work synthesizes existing research and does not involve new data collection or human subjects. However, the topics discussed–such as bias, safety, and manipulation–carry important ethical implications for how language models are developed and evaluated. By introducing the notion of *LLMs-against-humans*, we aim to highlight potential risks that arise when cooperative assumptions fail, without implying human-like agency or intent. Because taxonomic frameworks can influence how systems are designed and assessed, we have aimed for transparency and consistency in classification. Future work should incorporate participatory, multilingual, and interdisciplinary perspectives to ensure that evaluations of human-LLM interaction remain fair, inclusive, and contextually appropriate.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. 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*.

Marwa Abdulhai, Gregory Serapio-García, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17737–17752. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In *North American Chapter of the Association for Computational Linguistics*.

Mohamed Aghzal, Erion Plaku, Gregory J. Stein, and Ziyu Yao. 2025. A survey on large language models for automated planning. *ArXiv*, abs/2502.12435.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. 2016. Concrete problems in ai safety. *ArXiv*, abs/1606.06565.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *ArXiv*, abs/1605.05362.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *ArXiv*, abs/2108.07732.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *CoRR*, abs/2408.09235.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion,

Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, and 7 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason Alan Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A Pfeffer, and Nigam H. Shah. 2024. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *ArXiv*, abs/2308.09662.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiation-arena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525 – 535.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 152–163. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025a. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24761–24778, Vienna, Austria. Association for Computational Linguistics.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025b. Mixture-of-personas language models for population simulation. *ArXiv*, abs/2504.05019.

Fiammetta Caccavale, Carina L. Gargalo, Krist V. Gernaey, and Ulrich Krühne. 2024. Towards education 4.0: The role of large language models as virtual tutors in chemical engineering. *Education for Chemical Engineers*, 49:1–11.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Pengfei Cao, Tianyi Men, Wencan Liu, Jingwen Zhang, Xuzhao Li, Xixun Lin, Dianbo Sui, Yanan Cao, Kang Liu, and Jun Zhao. 2025a. Large language models for planning: A comprehensive and systematic survey. *ArXiv*, abs/2505.19683.

Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. 2025b. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):9737–9757.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming

Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241, Miami, Florida, USA. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

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

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2025a. Multi-agent consensus seeking via large language models. *Preprint*, arXiv:2310.20151.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 34 others. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025b. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria. Association for Computational Linguistics.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian J. McAuley, Linda Ruth Petzold, and William Yang Wang. 2024b. A survey on large language models for critical societal domains: Finance, healthcare, and law. *ArXiv*, abs/2405.01769.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. Llmphy: Complex physical reasoning using large language models and world models. *ArXiv*, abs/2411.08027.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Michelle Cohn, Mahima Pushkarna, Femi Olanubi, Zion Mengesha, Joe Moran, Daniel Padgett, and Courtney Heldreth. 2024. Believing anthropomorphism: Examining the role of anthropomorphic cues on user trust in large language models. Late Breaking Work.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, André Martins, Fabrizio Esposito, Vera L'ucia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *ArXiv*, abs/2403.03883.

Yue Cui, Liuyi Yao, Shuchang Tao, Weijie Shi, Yaliang Li, Bolin Ding, and Xiaofang Zhou. 2025. Enhancing tool learning in large language models with hierarchical error checklists. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16357–16375, Vienna, Austria. Association for Computational Linguistics.

A. Dahlgren Lindström, L. Methnani, L. Krause, and 1 others. 2025. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback. *Ethics and Information Technology*, 27:28.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *ArXiv*, abs/2306.06070.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

D. Dillion, D. Mondal, N. Tandon, and 1 others. 2025. Ai language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15:4084.

Aditya Ramesh Doshi and 1 others. 2024. Generative ai enhances individual creativity but reduces variety. *Science Advances*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025a. Deepresearch bench: A comprehensive benchmark for deep research agents. *ArXiv*, abs/2506.11763.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025b. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *ArXiv*, abs/2402.17385.

Xiang Fei, Siqi Wang, Shu Wei, Yuxiang Nie, Wei Shi, Hao Feng, Chao Feng, and Can Huang. 2025. Post-completion learning for language models. *arXiv preprint arXiv:2507.20252*.

K. J. Kevin Feng, Q. Vera Liao, Ziang Xiao, Jennifer Wortman Vaughan, Amy X. Zhang, and David W. McDonald. 2024. Canvil: Designerly adaptation for llm-powered user experiences. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased LLMs can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607, Vienna, Austria. Association for Computational Linguistics.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. Evaluating human-ai collaboration: A review and methodological framework. *ArXiv*, abs/2407.19098.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. 2023. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023a. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023b. Understanding social reasoning in language models with language models. 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*.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *CoRR*, abs/2312.11970.

Y. Gao, D. Lee, G. Burtch, and S. Fazelpour. 2025. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences of the United States of America*, 122(24):e2501660122.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. 2024. MindAgent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3154–3183, Mexico City, Mexico. Association for Computational Linguistics.

Ryan Greenblatt, Carson E. Denison, Benjamin Wright, Fabien Roger, Monte Stuart MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Kristjanson Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2024. Alignment faking in large language models. *ArXiv*, abs/2412.14093.

Florian Grötschla, Luis Müller, Jan Tönshoff, Mikhail Galkin, and Bryan Perozzi. 2025. Agentsnet: Coordination and collaborative reasoning in multi-agent llms. *ArXiv*, abs/2507.08616.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Advances in Neural Information Processing Systems*, 37:123471–123497.

F. Maxwell Harper, Joseph A. Konstan, and Joseph A. 2016. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5:19:1–19:19.

Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *ArXiv*, abs/2304.09102.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Haitao Hong, Yuchen Yan, Xingyu Wu, Guiyang Hou, Wenqi Zhang, Weiming Lu, Yongliang Shen, and Jun Xiao. 2025a. Cooper: Co-optimizing policy and reward models in reinforcement learning for large language models. *arXiv preprint arXiv:2508.05613*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Mengze Hong, Chen Jason Zhang, Chaotao Chen, Rongzhong Lian, and Di Jiang. 2025b. Dialogue language model with large-scale persona data engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 961–970, Albuquerque, New Mexico. Association for Computational Linguistics.

Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *ArXiv*, abs/2308.00352.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jennifer Hu, Felix Sosa, and Tomer D. Ullman. 2025a. Re-evaluating theory of mind evaluation in large language models. *CoRR*, abs/2502.21098.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyu Hu, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Jianxun Lian, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025b. Population-aligned persona generation for llm-based social simulation. *Preprint*, arXiv:2509.10127.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024a. Game-theoretic LLM: agent workflow for negotiation games. *CoRR*, abs/2411.05990.

Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024b. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8047–8074, Miami, Florida, USA. Association for Computational Linguistics.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. 2025a. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 466–488, Vienna, Austria. Association for Computational Linguistics.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Xiangji Huang. 2025b. How to enable effective cooperation between humans and nlp models: A survey of principles, formalizations, and beyond. *ArXiv*, abs/2501.05714.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43:1 – 55.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025c. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *ArXiv*, abs/2402.02716.

Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2023c. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *ArXiv*, abs/2310.03128.

Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. *ArXiv*, abs/2305.16896.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *ArXiv*, abs/2311.11944.

Mehdi Jafari, Yuncheng Hua, Hao Xue, and Flora D. Salim. 2025. Beyond words: Integrating theory of mind into conversational agents for human-like belief, desire, and intention alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5489–5508, Vienna, Austria. Association for Computational Linguistics.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, and 6 others. 2023. Ai alignment: A comprehensive survey. *ArXiv*, abs/2310.19852.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. 2025a. Language models resist alignment: Evidence from data compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23411–23432, Vienna, Austria. Association for Computational Linguistics.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025b. Moralbench: Moral evaluation of llms. *SIGKDD Explor.*, 27(1):62–71.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment. *Preprint*, arXiv:2110.07574.

Yuhua Jiang, Yuwen Xiong, Yufeng Yuan, Chao Xin, Wenyuan Xu, Yu Yue, Qianchuan Zhao, and Lin Yan. 2025. Pag: Multi-turn reinforced llm self-correction with policy as generative verifier. *arXiv preprint arXiv:2506.10406*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *ArXiv*, abs/2009.13081.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W. John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *ArXiv*, abs/2402.13225.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*.

B. Jose, J. Cherian, A. M. Verghis, S. M. Varghise, M. S, and S. Joseph. 2025. The cognitive paradox of ai in education: between enhancement and erosion. *Frontiers in Psychology*, 16:1550621. Published 2025 Apr 14.

Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J. Jansen. 2025. Personacraft: Leveraging language models for data-driven persona development. *Int. J. Hum. Comput. Stud.*, 197:103445.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *ArXiv*, abs/2312.14925.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. 2024. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413, Miami, Florida, USA. Association for Computational Linguistics.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023a. Large language models in law: A survey. *ArXiv*, abs/2312.03718.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023b. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pages 18319–18345. PMLR.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. *ArXiv*, abs/1706.05125.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025a. LLM generated persona is a promise with a catch. *CoRR*, abs/2503.16527.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into 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*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. 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*.

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

Jia Li, Zichun Zhou, Han Lyu, and Zhenchang Wang. 2025b. Large language models-powered clinical decision support: enhancing or replacing human expertise? *Intelligent Medicine*, 5(1):1–4.

Jiayang Li, Jiale Li, and Yunsheng Su. 2024c. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In *Artificial Intelligence in HCI*, pages 60–85, Cham. Springer Nature Switzerland.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024d. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Weizhao Li, Junsheng Kong, Ben Liao, and Yi Cai. 2022. Mitigating contradictions in dialogue based on contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2781–2788, Dublin, Ireland. Association for Computational Linguistics.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024e. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zi-Liang Ye, Guanjing Xiong, Fan Zhang, Xiang Li, and 6 others. 2024f. Personal llm agents: Insights and survey about the capability, efficiency and security. *ArXiv*, abs/2401.05459.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025c. From text to trust: Empowering ai-assisted decision making with adaptive llm-powered analysis. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023c. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023.

Xuechen Liang, Yangfan He, Yinghui Xia, Xinyuan Song, Jianhui Wang, Meiling Tao, Li Sun, Xinhang Yuan, Jiayi Su, Keqin Li, and 1 others. 2024. Self-evolving agents with reflective and memory-augmented abilities. *arXiv preprint arXiv:2409.00872*.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025a. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025b. Investigating bias in LLM-based bias detection: Disparities between LLMs and human perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649, Abu Dhabi, UAE. Association for Computational Linguistics.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025a. A survey of personalized large language models: Progress and future directions. *ArXiv*, abs/2502.11528.

Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. 2023a. Llm-powered hierarchical language agent for real-time human-ai coordination. *ArXiv*, abs/2312.15224.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2023b. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025b. Personaflow: Designing llm-simulated expert perspectives for enhanced research ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference, DIS 2025, Funchal, Madeira, Portugal, July 5-9, 2025*, pages 506–534. ACM.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, and 1 others. 2023d. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*.

Songshuo Lu, Hua Wang, Zhi Chen, and Yaohua Tang. 2025. Urpo: A unified reward & policy optimization framework for large language models. *arXiv preprint arXiv:2507.17515*.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue

Qiao, Qingqing Long, Rongcheng Tu, Xiaoming Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Mengxue Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, and 7 others. 2025a. Large language model agent: A survey on methodology, applications and challenges. *ArXiv*, abs/2503.21460.

X. Luo, A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, K. Lee, A. O. Cohen, V. Borghesani, A. Pashkov, D. Marinazzo, J. Nicholas, A. Salatiello, I. Sucholutsky, P. Minervini, S. Razavi, R. Rocca, E. Yusifov, T. Okalova, and 20 others. 2025b. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9(2):305–315. Epub 2024 Nov 27.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. Knowtuning: Knowledge-aware fine-tuning for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14535–14556.

Yougang Lyu, Lingyong Yan, Zihan Wang, Dawei Yin, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. Macpo: Weak-to-strong alignment via multi-agent contrastive preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *ArXiv*, abs/2201.06009.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the way to llm personalization: Learning to remember user conversations. *ArXiv*, abs/2411.13405.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations. *ArXiv*, abs/2411.15287.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Tanha Miah and Hong Zhu. 2024. User centric evaluation of code generation tools (invited paper). *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pages 109–119.

B. Moëll and F. Sand Aronsson. 2025. Harm reduction strategies for thoughtful use of large language models in the medical domain: Perspectives for patients and clinicians. *Journal of Medical Internet Research*, 27:e75849.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *ArXiv*, abs/2412.03563.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *ArXiv*, abs/2311.09180.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *CoRR*, abs/2405.05248.

Tuomas P. Oikarinen, Tsui-Wei Weng, and Luca Daniel. 2020. Robust deep reinforcement learning through adversarial loss. In *Neural Information Processing Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Sam Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. In *Findings*.

Stefan Pasch and Min Chul Cha. 2025. Balancing privacy and utility in personal LLM writing tasks: An automated pipeline for evaluating anonymizations. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 32–41, Albuquerque, New Mexico. Association for Computational Linguistics.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Benji Peng, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Junyu Liu, and Qian Niu. 2024. Securing large language models: Addressing bias, misinformation, and prompt attacks. *ArXiv*, abs/2409.08087.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Aske Plaat, Max J. van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *ArXiv*, abs/2503.23037.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2023. Adapt: As-needed decomposition and planning with language models. In *NAACL-HLT*.

Maite Puerta-Beldarrain, Oihane Gómez-Carmona, Rubén Sánchez-Corcuera, Diego Casado-Mansilla, Diego López-de Ipiña, and Liming Chen. 2025. A multifaceted vision of the human-ai collaboration: A comprehensive review. *IEEE Access*, 13:29375–29405.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *ArXiv*, abs/2307.16789.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Jirong Wen. 2024. Tool learning with large language models: A survey. *ArXiv*, abs/2405.17935.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. 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*.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. 2024. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16:2514 – 2572.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2024a. Llms know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *Annual Meeting of the Association for Computational Linguistics*.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024b. Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 7895–7903. ijcai.org.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *Trans. Mach. Learn. Res.*, 2025.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, I. Evtimov, Joanna Bitton, Manish P Bhatt, Cris tian Cantón Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre D'efossez, Jade Copet, and 6 others. 2023. Code llama: Open foundation models for code. *ArXiv*, abs/2308.12950.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational*

Linguistics: EMNLP 2024, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. Meminsight: Autonomous memory augmentation for llm agents. *arXiv preprint arXiv:2503.21760*.

Paulo Salem, Robert Sim, Christopher Olsen, Prerit Saxena, Rafael Barcelos, and Yi Ding. 2025. Tinytroupe: An llm-powered multiagent persona simulation toolkit. *CoRR*, abs/2507.09788.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Karahan Saritas, Kivanç Tezören, and Yavuz Durmazkeser. 2025. A systematic review on the evaluation of large language models in theory of mind tasks. *ArXiv*, abs/2502.08796.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ashish Sharma, Sudha Rao, Christopher John Brockett, Akanksha Malhotra, Nebojsa Jojic, and William B. Dolan. 2023a. Investigating agency of llms in human-ai collaboration tasks. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Kristjanson Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Tim Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023b. Towards understanding sycophancy in language models. *ArXiv*, abs/2310.13548.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, S. Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary C. Lipton, Qiaozhu Mei, Rada Mihalcea, and 5 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *ArXiv*, abs/2406.09264.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *ArXiv*, abs/2309.15025.

Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yue Ting Zhuang. 2023b. Taskbench: Benchmarking large language models for task automation. *ArXiv*, abs/2311.18760.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *ArXiv*, abs/2010.03768.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin A Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua B Tenenbaum, and Tomer D Ullman. 2021. AGENT: A benchmark for core psychological reasoning. In *The Thirty-eighth International Conference on Machine Learning (ICML)*.

Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of llm-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta,

Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Kalahasti Ganesh Srivatsa, Sabyasachi Mukhopadhyay, Ganesh Katrapati, and Manish Shrivastava. 2023. A survey of using large language models for generating infrastructure as code. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 523–533, Goa University, Goa, India. NLP Association of India (NLPAI).

J. W. A. Strachan, D. Albergo, G. Borghini, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Winnie Street. 2024. Llm theory of mind and alignment: Opportunities and risks. *ArXiv*, abs/2405.08154.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Agüera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *CoRR*, abs/2405.18870.

Guangzhi Sun, Anmol Kagrecha, Potsawee Manakul, Phil Woodland, and Mark Gales. 2025a. SkillAggregation: Reference-free LLM-dependent aggregation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15532–15548, Vienna, Austria. Association for Computational Linguistics.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. 2025b. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. *Preprint*, arXiv:2502.20073.

Lijun Sun, Yijun Yang, Qiqi Duan, Yuhui Shi, Chao Lyu, Yu-Cheng Chang, Chin-Teng Lin, and Yang Shen. 2025c. Multi-agent coordination across diverse applications: A survey. *ArXiv*, abs/2502.14743.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *ArXiv*, abs/2501.06322.

H. Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Raj Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *ArXiv*, abs/2407.18901.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Michelle Vaccaro, Michael Caoson, Harang Ju, Sinan Aral, and Jared R. Curhan. 2025. Advancing AI negotiations: New theory and evidence from a large-scale autonomous negotiations competition. *CoRR*, abs/2503.06416.

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

A. Wang, J. Morgenstern, and J. P. Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Conference on Empirical Methods in Natural Language Processing*.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. 2024a. A survey on human-centric llms. *ArXiv*, abs/2411.14491.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, Liang Lin, Zhihao Xu, Haolang Lu, Xinye Cao, Xinyun Zhou, Weifei Jin, Fanci Meng, Junyuan Mao, Haoyang Wu, and 63 others. 2025b. A comprehensive survey in llm(-agent) full stack safety: Data, training and deployment. *ArXiv*, abs/2504.15585.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. 2025c. Persona features control emergent misalignment. *CoRR*, abs/2506.19823.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024c. 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)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Wei Wang, Dan Zhang, Tao Feng, Boyan Wang, and Jie Tang. 2024d. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *CoRR*, abs/2408.15971.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Zhichao Wang, Bin Bi, Shiva K. Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu Zhu, Xiang-Bo Mao, Sitaram Asur, and Na Cheng. 2024e. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *ArXiv*, abs/2407.16216.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023b. Describe, explain, plan and select: Interactive planning with llms enables open-world multi-task agents. In *Neural Information Processing Systems*.

Zixiao Wang, Duzhen Zhang, Ishita Agarwal, Shen Gao, Le Song, and Xiuying Chen. 2025d. Beyond profile: From surface-level facts to deep persona simulation in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21239–21257, Vienna, Austria. Association for Computational Linguistics.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2024f. Agent workflow memory. *arXiv preprint arXiv:2409.07429*.

Azmine Toushik Wasi, Mst Rafia Islam, and Raima Islam. 2024. Llms as writing assistants: Exploring perspectives on sense of ownership and reasoning. *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, and Xianfeng Tang. 2024. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *ArXiv*, abs/2403.10667.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6743–6744, New York, NY, USA. Association for Computing Machinery.

Stanislaw Wo'zniak, Bartlomiej Koptyra, Arkadiusz Janz, Przemyslaw Kazienko, and Jan Koco'n. 2024. Personalized large language models. *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Namyong Park, Sungchul Kim, and 8 others. 2024. Personalized multimodal large language models: A survey. *ArXiv*, abs/2412.02142.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *ArXiv*, abs/2306.01693.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024. Large language models for education: A survey. *ArXiv*, abs/2405.13001.

Qiantong Xu, Fenglu Hong, B. Li, Changran Hu, Zheng Chen, and Jian Zhang. 2023. On the tool manipulation capability of open-source large language models. *ArXiv*, abs/2305.16504.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *ArXiv*, abs/2305.18752.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2023. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In *NAACL-HLT*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. *Proceedings of the 27th International Conference on Intelligent User Interfaces*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 3.

Ziqian Zeng, Jianwei Wang, Junyao Yang, Zhengdong Lu, Haoran Li, Huiping Zhuang, and Cen Chen. 2025. PrivacyRestore: Privacy-preserving inference in large language models via privacy removal and restoration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10821–10855, Vienna, Austria. Association for Computational Linguistics.

C. Zhai, S. Wibowo, and L. D. Li. 2024. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11:28.

Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. 2023a. Large language models are semi-parametric reinforcement learning agents. *Advances in Neural Information Processing Systems*, 36:78227–78239.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. 2023b. Llm-based medical assistant personalization with short- and long-term memory coordination. In *North American Chapter of the Association for Computational Linguistics*.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024a. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In *Annual Meeting of the Association for Computational Linguistics*.

Rupeng Zhang, Haowei Wang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. 2025a. From allies to adversaries: Manipulating LLM tool-calling through adversarial injection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2009–2028, Albuquerque, New Mexico. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *ArXiv*, abs/1801.07243.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024b. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025b. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*.

Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024c. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Ju-Ying Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2024d. Personalization of large language models: A survey. *ArXiv*, abs/2411.00027.

Zijing Zhang, Zhanpeng Chen, He Zhu, Ziyang Chen, Nan Du, and Xiaolong Li. 2025c. ToolExpNet: Optimizing multi-tool selection in LLMs with similarity and dependency-aware experience networks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15706–15722, Vienna, Austria. Association for Computational Linguistics.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. 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*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023b. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. 2023c. Agents: An open-source framework for autonomous language agents. *CoRR*, abs/2309.07870.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024a. 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.

Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024b. Emulated disalignment: Safety alignment for large language models may backfire! In *Annual Meeting of the Association for Computational Linguistics*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025a. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025b. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. 2025c. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *arXiv preprint arXiv:2502.16906*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *ArXiv*, abs/2306.13304.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou,

Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Yuwei Cao, Dongyuan Li, Renhe Jiang, and Philip S. Yu. 2025. Llm-based human-agent collaboration and interaction systems: A survey.

A Positioning and Scope

This taxonomy focuses on research that investigates how language models support, collaborate with, or emulate humans, rather than work centered on model architectures, pre-training strategies, or scaling laws. Within the *LLMs-for-Humans* paradigm, we consider systems where humans remain the primary agents and models act in assistive roles, excluding fully automated pipelines without user interaction. The *LLMs-with-Humans* paradigm covers settings where human feedback or collaboration directly shapes model behavior during training or deployment, while excluding synthetic supervision or self-training without human input. The *LLMs-as-Humans* paradigm focuses on cases where models substitute for human roles in evaluation, simulation, or annotation, excluding general capability benchmarks that assess intrinsic performance rather than simulate human actors. Together, these definitions establish the conceptual boundaries of the taxonomy and clarify the forms of interaction that constitute the focus of this survey.

B Connection to Relevant Surveys

Table 2 summarizes surveys related to specific categories within our paradigms. As shown, most existing surveys remain fragmented across subfields and do not explicitly address the directionality between humans and LLMs—that is, whether the model primarily serves, collaborates with, or emulates humans. In addition to the task-specific surveys listed in the table, several recent papers have begun to explore broader frameworks for LLM–human interaction (Zou et al., 2025; Wang et al., 2024a). For example, Huang et al. (2025b) provide an overview of human–model cooperation and propose a new taxonomy offering a unified lens to summarize existing approaches. Compared with such works, our paradigm focuses more explicitly on the role orientation of LLMs, providing a conceptual structure to situate benchmarks, applications, and evaluation methods under the *LLMs for, with, as Humans* framework. Likewise, Shen et al. (2024) investigate how human–AI alignment is defined and operationalized; in contrast, our work emphasizes the underlying interaction direction and its implications for modeling, evaluation, and system design.

C Selection Criteria

This survey focuses on research that directly examines interactions between humans and large language models. To ensure comprehensive and balanced coverage across paradigms, studies were selected according to four main criteria: relevance to human–LLM interaction, methodological contribution, representativeness within each paradigm, and recency of publication.

Search strategy. Relevant literature was identified through search on Google Scholar. The search used combinations of keywords such as *human–LLM interaction, human–AI collaboration, alignment, feedback learning, AI assistants, LLMs as judges, multi-agent LLMs, and simulation.* Backward and forward citation tracking was also employed to capture influential works not directly retrieved by keyword search.

Relevance. Included studies explicitly examine how LLMs assist, collaborate with, or emulate humans within assistive, collaborative, or simulative settings. This includes empirical studies, frameworks, evaluation methodologies, and theoretical or position papers that analyze human roles in training, supervision, or interaction. Papers focused solely on model architecture, scaling, or data efficiency without a human-centered component were excluded.

Methodological contribution. Selected works contribute methodological insight into the design, evaluation, or alignment of human–LLM systems. This includes techniques for preference modeling, feedback integration, interpretability, and interactive evaluation. Works that only apply LLMs to downstream tasks without addressing the human–model relationship were not considered central to the survey.

Representativeness and recency. Within each paradigm, representative studies were chosen to capture diversity in application domains and methodological approaches. When multiple papers addressed similar topics, preference was given to those providing broader insight or stronger empirical validation. The survey emphasizes research published after 2020, following the widespread adoption of general-purpose LLMs, while referencing earlier foundational studies when relevant.

This selection strategy balances breadth and analytical depth, providing a transparent and repro-

Paradigm	Subcategories	Representative surveys / Focus sreas
LLMs-for-Humans	Coordinating, Problem-solving, Consulting, Tool Learning, and Personalizing Assistance	– Surveys on task assistance (Qu et al., 2024) and domain-specific (e.g., education, healthcare, code generation) (Chen et al., 2024b; Xu et al., 2024; Lai et al., 2023a; Liu et al., 2025a; Bedi et al., 2024; Ramos et al., 2024; Wen et al., 2024; Qu et al., 2024) – Reviews of human–AI collaboration (Li et al., 2024c), planning (Cao et al., 2025a; Aghzal et al., 2025), agents (Huang et al., 2024; Plaat et al., 2025; Luo et al., 2025a) and personalization (Huang et al., 2025a; Zhang et al., 2024d; Wu et al., 2024; Li et al., 2024f)
LLMs-with-Humans	Human Data Shaping, Interactive Alignment, and Collaborative Co-evolution	– Surveys on alignment methods (RLHF, DPO, human feedback loops) (Shen et al., 2023a; Ji et al., 2023; Kaufmann et al., 2023; Wang et al., 2024e; Shen et al., 2024; Cao et al., 2025b) – Reviews on prompt engineering, interactive learning, and co-adaptive frameworks (Wu et al., 2022; Fragiadakis et al., 2024; Puerta-Beldarrain et al., 2025)
LLMs-as-Humans	Human-like Reasoning, Persona Simulation, and Social Simulation	– Surveys on cognitive modeling (Wang et al., 2024a), LLM-as-a-Judge (Li et al., 2024b; Chang et al., 2024; Gu et al., 2024), and Theory-of-Mind capabilities (Chen et al., 2025b; Saritas et al., 2025) – Reviews on persona grounding, emotion modeling, and multi-agent social dynamics (Sun et al., 2025c; Tran et al., 2025; Gao et al., 2023; Li et al., 2024e; Mou et al., 2024; Wang et al., 2024a)

Table 2: Survey coverage across three paradigms of Human–LLM interaction. Each paradigm summarizes its key subsections and representative areas reviewed in related survey literature.

ducible basis for the taxonomy and synthesis presented in this survey.

D Detailed Literature Review

D.1 LLMs-for-Humans Paradigm

LLMs-for-Humans paradigm refer to systems that act as interactive partners, supporting users in accomplishing tasks, making decisions, or enhancing productivity. Specifically, we break the assistance type into four main types: *coordinating*, *problem-solving*, *consulting*, and *personalizing*. Coordinating assistance involves managing workflows, agents, or tools to achieve complex goals efficiently. Problem-solving assistance supports reasoning, planning, and task execution, such as debugging code or synthesizing information. Consulting assistance offers expert advice or recommendations based on domain knowledge, while personalizing assistance adapts outputs to individual user preferences, behavior, and context. Together, these capabilities enable LLMs to integrate seamlessly into human workflows and augment creative, technical, and decision-making processes.

D.1.1 Coordinating Assistance

Task management. In this stream, LLMs act as workflow coordinators that plan, manage, and execute complex tasks (Huang et al., 2024). Human users typically provide high-level strategic goals, while the model decomposes these goals into executable plans, prioritizes subtasks, and monitors progress across different applications and environments (Prasad et al., 2023). Early approaches explored human-instructed planning, where users provide stepwise instructions that guide the model's decomposition and scheduling (Prasad et al., 2023; Shen et al., 2023b). Datasets such as Mind2Web (Deng et al., 2023) enable LLMs to interact with digital environments, including websites, APIs to achieve complex objectives and assess task completion.

Multi-Agent collaboration. This stream leverages multiple LLMs working together to coordinate subtasks across specialized roles or domains. Systems like MetaGPT (Hong et al., 2023) and ChatDev (Qian et al., 2023) exemplify frameworks where LLM agents communicate, negotiate, and assign responsibilities to efficiently accomplish shared goals. These multi-agent workflows are

particularly useful for software engineering, multi-step project planning, and game-based simulations, enabling each agent to focus on domain-specific subtasks while the ensemble collectively executes the overall plan (Tran et al., 2025; Agashe et al., 2023). It allows LLMs to act as dynamic coordinators, orchestrating distributed workflows that adapt to changing requirements and user feedback.

D.1.2 Problem-Solving Assistance

Technical problem-solving. LLMs function as embedded creative and technical partners, seamlessly integrated into diverse workflows such as coding (Zhang et al., 2024a; Rozière et al., 2023), writing (Wasi et al., 2024; Yuan et al., 2022), solving mathematical problems (Cobbe et al., 2021; He-Yueya et al., 2023), or executing domain-specific tasks (Schick et al., 2023; Zhuang et al., 2023). Their role extends beyond passive response generation: interaction with LLMs is inherently iterative and proactive. For instance, within a code editor, an AI does not merely follow user commands; it can propose solutions, suggest optimizations, refactor existing code, and even identify and debug errors in real time. This dynamic transforms traditionally solitary creative or technical processes into a human-AI duet, where the model not only amplifies human capability but also reshapes the workflow.

Tool learning. An important topic under the category is *tool learning*, where LLMs are equipped with the ability to invoke external tools, APIs, or knowledge sources to enhance reasoning, retrieval, and task execution (Qu et al., 2024; Qin et al., 2023). For example, GPT4Tools (Yang et al., 2023) enhances the open-source LLMs in terms of visual problems with LoRA optimization. AppWorld (Trivedi et al., 2024) provides an execution environment composed of everyday applications, enabling LLM-based agents to perform interactive code generation through APIs. Qin et al. (2023) enable LLMs to perform external actions, such as setting reminders or booking flights.

D.1.3 Consulting Assistance

Professional consulting. LLMs can also act as professional in-domain expertise, making suggestions and helping users explore complex domains like research (Du et al., 2025a), law (Lai et al., 2023a; Chen et al., 2024b), healthcare (Jin et al., 2023; Bedi et al., 2024), etc. Instead of discrete queries, humans engage in dynamic, multi-turn dialogues to synthesize information, understand nuances, and receive explanations, much like consulting a domain expert who understands the full history of the conversation. For example, in biology, GeneGPT (Jin et al., 2023) is augmented with Web APIs to answer genomics-related questions. In chemistry, relevant domain knowledge is retrieved to support reasoning and problem-solving (Inaba et al., 2023; Ramos et al., 2024; Bran et al., 2023). Similar approaches have been applied in other domains, including physics (Cherian et al., 2024), economics (Zhang et al., 2024b), and medicine (Jin et al., 2024), where LLMs are integrated with external knowledge sources or tools to enhance their task-specific capabilities.

Decision support. LLMs can also function by aggregating relevant information, generating structured analyses, surfacing trade-offs, and articulating multiple perspectives to support complex professional or strategic decisions (Eigner and Händler, 2024; Li et al., 2025b). Unlike traditional automated systems, LLMs in consulting are interactive and dialogic: they engage in iterative clarification, contextual adaptation, and explanation generation, enabling consultants and clients to explore alternatives dynamically. The value of LLM decision support lies not only in speed or scale but in its ability to foster reflective reasoning and augment human expertise, while maintaining transparency, accountability, and alignment with organizational goals.

D.1.4 Personalizing Assistance

LLMs can be integrated into workflows such as scheduling, summarization, and recommendations, providing personalized user experiences (Zhang et al., 2024d; Liu et al., 2025a). Beyond task-specific assistance, personalization can be achieved through adapting to user preferences, behavior, and expertise, enabling more context-aware and relevant outputs (Feng et al., 2024; Magister et al., 2024). Approaches to personalization include prompt-level customization (Wang et al., 2023a; Mysore et al., 2023), memory-augmented models that retain user context (Madaan et al., 2022; Zhang et al., 2023b), and fine-tuning techniques such as LoRA or RLHF to align the model with individual users' needs (Wo'zniak et al., 2024; Wu et al., 2023; Tan et al., 2024). Furthermore, multi-modal personalization allows LLMs to incorporate signals from text, images, or other user-generated content, enhancing the model's understanding of the user's goals and context (Wu et al., 2024; Wei et al.,

2024).

D.2 LLMs-with-Humans Paradigm

The development paradigm for LLMs is shifting from a technology-centric model to one of Human-AI Co-Evolution. This paradigm re-envisions LLMs not as static artifacts to be merely used, but as dynamic partners engaged in a symbiotic process with humans. This co-evolutionary journey unfolds across three distinct phases: first shaping the model's foundational capabilities from static human data, then Aligning its behavior through interactive feedback, and finally Co-evolving with it in a long-term, adaptive collaboration.

D.2.1 Human Data Shaping

This initial phase is dedicated to shaping the model's foundational knowledge and capabilities. The process relies on vast, static corpora of human-generated data—such as text from the internet and curated datasets—to mold the model's understanding of language, reasoning, and world knowledge. This knowledge transfer is unidirectional and performed offline, establishing a powerful but unrefined baseline. The model is effectively a raw intellectual artifact, shaped by the patterns in past human data but lacking the ability to dynamically adapt its behavior to live human interaction.

Shaping LLMs via pre-training. The first and most extensive form of shaping occurs during pre-training, where models implicitly absorb world knowledge and linguistic competence from massive, static text corpora. This large-scale knowledge transfer was enabled by the Transformer architecture, whose attention mechanism proved highly parallelizable and allowed for unprecedented scaling (Vaswani et al., 2017). While early approaches established a "pre-training for fine-tuning" paradigm with models like BERT (Devlin et al., 2019), a pivotal discovery was that scaling itself unlocked new capabilities. GPT-2, for example, demonstrated that a sufficiently large model could perform diverse tasks in a zero-shot setting (Radford et al., 2019). This trajectory culminated in massively scaled models like GPT-3, which revealed powerful in-context learning abilities, enabling it to perform competitively on many tasks directly from prompts without any gradient updates (Brown et al., 2020). This process results in a model with immense latent capabilities, establishing a potent but fundamentally undirected foundation.

Shaping LLMs via fine-tuning. Following pre-training, Supervised Fine-Tuning (SFT), or instruction tuning, introduces a more explicit and targeted layer of shaping. This process refines the model's raw potential by fine-tuning it on a diverse collection of curated instruction-response pairs, teaching it to align with human intent and follow commands (Wei et al., 2021; Sanh et al., 2021). The goal is to transform the model from a probabilistic text completer into a reliable instruction-following agent, effectively unlocking its zero-shot generalization to unseen tasks (Wei et al., 2021; Sanh et al., 2021). The effectiveness of SFT is such that smaller, instruction-tuned models can significantly outperform larger, raw counterparts (Ouyang et al., 2022; Wei et al., 2021; Lyu et al., 2024; Zhang et al., 2024c). The composition of this data is critical for sculpting specific skills; for instance, incorporating chain-of-thought (CoT) examples is crucial for enhancing complex reasoning abilities (Chung et al., 2024). While SFT significantly improves usability, it still relies on a static dataset, serving as the final step in creating a more controllable LLM, now prepared for the subsequent, dynamic phase of Alignment through interactive feedback (Ouyang et al., 2022).

D.2.2 Interactive Alignment

Following the initial shaping, the critical phase of aligning begins. This marks the transition from learning from static data to refining the model through dynamic, interactive feedback. The primary goal is to align the model's behavior with nuanced human preferences, values, and conversational norms, which cannot be fully captured in static datasets. Through techniques like Reinforcement Learning from Human Feedback (RLHF), the model's responses are fine-tuned based on explicit human judgments within short-term, session-based interactions. This phase calibrates the pre-shaped model, making it a more helpful, harmless, and collaborative conversationalist.

Aligning LLMs via prompt engineering. Model performance is enhanced within an interaction through structured prompting methods that create dynamic, inference-time feedback loops to guide the model's reasoning process. The foundational Chain-of-Thought (CoT) approach prompts the model to generate intermediate steps that form an internal feedback loop, enabling it to solve complex problems by sequentially reasoning over

its own generated context (Wei et al., 2022). The ReAct framework establishes an external feedback loop by synergizing reasoning and acting through an interleaved "Thought-Action-Observation" cycle (Yao et al., 2023b). Feedback from external tools grounds the model's internal reasoning, making it more reliable and factually accurate (Yao et al., 2023b). More advanced methods enable deliberative self-correction by generalizing reasoning into a strategic search. Frameworks such as Tree of Thoughts (ToT) (Yao et al., 2023a) and Graph of Thoughts (GoT) (Besta et al., 2024) guide the model to explore multiple reasoning paths and, crucially, perform self-evaluation to critique and select the most promising direction. This paradigm also extends to learning across trials, where frameworks such as Reflexion use linguistic self-reflections on past failures as explicit feedback to guide subsequent attempts (Shinn et al., 2023). Through these techniques, the model is aligned at inference time, using dynamic feedback to refine its behavior without updating its parameters.

Aligning LLMs via preference learning. Aligning LLMs is predominantly achieved through preference learning, historically pioneered by Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). This multi-stage paradigm first trains a reward model (RM) on human-ranked outputs and then uses this RM as a reward function to fine-tune the language model policy with reinforcement learning, typically PPO (Ouyang et al., 2022; Schulman et al., 2017; Lyu et al.). To improve the dynamic nature of this process, some work has explored iterated online training, continuously updating the model with new feedback (Bai et al., 2022). A significant paradigm shift came with Direct Preference Optimization (DPO), which obviates the need for explicit reward modeling and complex RL training by reframing the alignment objective as a direct classification loss on preference pairs (Rafailov et al., 2023). This innovation spurred a host of subsequent methods aimed at enhancing stability, efficiency, and performance. These include more robust direct optimization algorithms like IPO (Azar et al., 2024) and DPOP (Pal et al., 2024) that address theoretical and practical failure modes of DPO; simpler online RL alternatives to PPO like ReMax (Li et al., 2023c); and monolithic, reference-model-free architectures such as ORPO (Hong et al., 2024) and SimPO (Meng et al., 2024) that merge supervised fine-tuning with preference alignment into a single, efficient stage. This trajectory reflects a clear trend toward more direct and computationally tractable alignment methods that effectively distill human preferences into model behavior.

D.2.3 Collaborative Co-evolution

This phase represents the frontier of the "LLMs-with-Humans" paradigm, aiming for true co-evolution. It moves beyond session-based alignment to foster a long-term, symbiotic partnership where adaptation is continuous and bidirectional. In this stage, the model leverages long-term memory and reinforcement learning to progressively personalize its support to a specific user's goals, context, and feedback history. Concurrently, the human partner adapts their workflow to capitalize on the AI's evolving, deeply integrated capabilities. The objective is to create a unified system where both humans and AI continuously learn from each other and evolve together, becoming more effective over time.

Co-evolving LLMs via memory evolution. To facilitate long-term symbiotic collaboration, agentic LLMs are integrating dynamic memory systems that actively manage and evolve experiential knowledge. This process transcends mere archival by distilling raw trajectories into abstract, reusable components like "insights" (Zhao et al., 2024) and "workflows" (Wang et al., 2024f). The memory architecture itself is an evolvable component, with systems like Mem0 introducing structured operations (e.g., add, update, delete) and graph-based representations to maintain a coherent, dynamic knowledge base (Chhikara et al., 2025). Furthermore, these systems incorporate cognitive principles; for instance, frameworks like SAGE and MemoryBank use mechanisms inspired by the Ebbinghaus forgetting curve to manage information lifecycle (Liang et al., 2024; Zhong et al., 2024), while others autonomously augment raw data with structured semantic attributes to improve retrieval (Salama et al., 2025). More advanced paradigms drive memory evolution directly from performance feedback, using reinforcement learning to update the value of stored experiences (Zhang et al., 2023a) or employing self-play to generate new knowledge without human data (Guan et al., 2024). By evolving memory in its content, structure, and value, agents build an increasingly sophisticated model of their tasks, fostering a

deeply adaptive partnership.

Co-evolving LLMs via reinforcement learning. Reinforcement Learning (RL) is the critical mechanism for the continuous and bidirectional adaptation required for human-AI co-evolution. To support this dynamic alignment, efficient algorithms are necessary, such as Group Relative Policy Optimization (GRPO), which enhances resource efficiency by replacing the memory-intensive critic model with a group average reward baseline (Shao et al., 2024). Furthermore, advanced frameworks such as Cooper address the integrity of the partnership by co-optimizing the policy and the reward model to mitigate Reward Hacking (Hong et al., 2025a), while DAPO ensures stable convergence in live interaction streams (Yu et al., 2025). The ultimate form of this co-evolution is the agent's capacity for autonomous self-improvement, achieved through unified architectures like URPO (Lu et al., 2025) and Self-Rewarding Models (Yuan et al., 2024), which integrate and iteratively refine the policy (generator) and the critic (reward model) within a single system. Essential for a trustworthy partnership, meta-cognition is fostered through paradigms like Post-Completion Learning (PCL), which embeds internal reflection (Fei et al., 2025), and PAG (Policy as Generative Verifier), which enables intelligent, selective self-correction (Jiang et al., 2025). Finally, methods like Critique-GRPO ensure transparent collaboration by leveraging both numerical rewards and natural language critiques to train the model on the process of reasoning, thereby achieving more nuanced and interpretable alignment (Zhang et al., 2025b).

D.3 LLMs-as-Humans Paradigm

The "LLMs-as-Humans" paradigm examines settings where language models are positioned as *stand-ins for human actors*. In contrast to work that measures reasoning ability in isolation, research in this paradigm focuses on cases where models are explicitly evaluated in place of humans, such as judges, role-players, social participants, or domain experts. This framing emphasizes substitution in human roles rather than abstract cognitive capability, raising questions about how convincingly and responsibly models can simulate human-like behavior. To capture the breadth of work in this space, we organize our review around four core human dimensions: reasoning, identity and expression, social interaction, and knowledge and skill. These dimensions correspond to four categories of simulation: *Human judgement Simulation*, *Persona Simulation*, *Social Simulation*, and *Expertise Simulation*. Each reflects a distinct way in which LLMs are studied as human analogs, and together they provide a comprehensive view of how models are evaluated when treated as proxies for people.

In the subsections that follow, we highlight representative works under each category, illustrating how researchers have positioned LLMs as human substitutes across cognitive, social, and professional dimensions. We also discuss the challenges and implications that arise when models are evaluated not only as tools but as actors imitating human roles.

D.3.1 Human-like Reasoning Simulation

This category corresponds to the human dimension of *reasoning*, encompassing understanding, inference, and judgment. In this line of work, LLMs are examined as substitutes for human reasoning roles, for example, when serving as evaluators in quality assessment tasks, simulating human responses in cognitive or psychological experiments, or acting as moral and social decision-makers.

Judging system outputs. A central line of work uses LLMs as evaluators for open-ended NLP tasks, eliciting direct-assessment scores and pairwise preferences that would otherwise require human annotators. Early machine translation studies prompted GPT-based evaluators and reported correlation with human-referenced MQM metrics at scale (Kocmi and Federmann, 2023). Subsequent methods augmented judging with explicit rubrics and short rationales to boost agreement in summarization and dialogue and to probe prompt sensitivity and bias toward LLM-generated text (Liu et al., 2023c; Zheng et al., 2023; Wang et al., 2024b). Community benchmarks and platforms such as MT-Bench and Chatbot Arena improved this setting at scale by aggregating pairwise LLM judgments and comparing them to human preferences (Zheng et al., 2023; Chiang et al., 2024). Beyond generic prompting, studies also explore calibration strategies (Liu et al., 2024; Huang et al., 2023a), reference-aware versus reference-free judging (Badshah and Sajjad, 2024), and mixtures of pairwise and absolute scoring to better mirror established human evaluation protocols (Gilardi et al., 2023; Sun et al., 2025a).

LLMs as proxies in cognitive and psychological experiments. Another direction uses LLMs as ar-

tificial participants in tasks originally designed for humans, to compare model outputs to established behavioral patterns. Representative studies probe theory of mind with false-belief and perspective-taking tasks, where models sometimes produce human-like response patterns on canonical items (Chen et al., 2025b; Kosinski, 2024). Follow-up work varies problem wording and structure, showing that performance can drop under perturbations that remain easy for people, which motivates more controlled stimulus design and attention to generalization (Street et al., 2024; Ullman, 2023; Hu et al., 2025a). Other studies place LLMs in logical and causal reasoning settings, using human experimental baselines to characterize when model responses align with human inferences and when they diverge (Gandhi et al., 2023b).

LLMs in moral and social judgment. A third direction examines whether LLMs can stand in for people when making value-sensitive judgments. Benchmarks such as ETHICS target normative judgments across justice, deontology, virtue, and commonsense morality (Hendrycks et al., 2021). Systems like Delphi elicit everyday moral judgments to study descriptive norms in free-form scenarios (Jiang et al., 2022). Follow-up analyses compare model responses to human survey data across cultures and framings (AlKhamissi et al., 2024; Li et al., 2024a) and investigate prompt formats that elicit more consistent judgments. This work maps where LLMs approximate aggregate human judgments and where their responses track dataset regularities rather than underlying normative reasoning (Abdulhai et al., 2024; Ji et al., 2025b).

Building dedicated evaluator models. To improve reproducibility, researchers have trained dedicated evaluator models: Prometheus-style approaches fine-tune open models for direct assessment and pairwise ranking, including extensions to multimodal and multilingual settings (Kim et al., 2024a,b); JudgeLM assembles outputs with reference judgments to fine-tune scalable judges and analyzes biases introduced during training (Zhu et al., 2025b).

D.3.2 Persona Simulation

This category corresponds to the human dimension of *identity and expression*, including self-presentation, style, and emotional expression. Here, LLMs are studied as substitutes for human communicators who adopt specific personas or conversational identities. Research on persona simulation with large language models (LLMs) has converged on four themes: data-driven conditioning, long-term identity control, affect and style, and structured environments for evaluation.

Persona construction and control. Early work encoded persistent speaker traits to improve consistency in open-domain dialogue (Li et al., 2016). PERSONA-CHAT framed chit-chat as conditioning on self and partner profiles to elicit engaging, identity-consistent responses (Zhang et al., 2018a). Controllable generation introduced lightweight mechanisms to steer style and persona without retraining (Dathathri et al., 2020). More recently, data-driven pipelines ground personas in external evidence: survey- or corpus-based synthesis for task-relevant profiles (Jung et al., 2025), population-aligned sampling that matches psychometric distributions (Hu et al., 2025b), and audits of fully synthetic personas that document fidelity and bias trade-offs (Li et al., 2025a). Compact feature controls and conditioning have also been shown to modulate emergent behavioral patterns (Wang et al., 2025c). In parallel, large-scale pretraining and augmentation on persona-rich data aim further to stabilize identity expression (Hong et al., 2025b).

Maintaining persona consistency. A central objective is to sustain an assigned identity over time. Dialogue NLI operationalizes consistency as entailment with respect to profile statements and shows that NLI supervision improves self-consistency (Welleck et al., 2019). Unlikelihood training reduces contradictions and repetition (Welleck et al., 2020; Li et al., 2020), and later variants introduce contrastive or constraint-oriented objectives (Li et al., 2022). Beyond surface facts, systems simulate deeper linguistic and cognitive traits and track stability across topics and sessions (Wang et al., 2025d). Quantitative analyses measure how demographic and social variables shape the *persona effect* on outputs and outcomes (Hu and Collier, 2024).

Affect and stylistic expression. Persona is expressed through affective stance and style. EMPATHETICDIALOGUES introduces situation-conditioned conversations to evaluate perceived empathy (Rashkin et al., 2019). ESCONV defines emotional support dialogue with strategy annotations to model supportive helper identities (Liu

et al., 2021); follow-up work augments supporters with persona and strategy modeling to match intended identities better (Cheng et al., 2023). Toolkits and design frameworks instantiate controlled persona settings for interactive study and reproducible ablations. TINYTROUPE supports multi-agent simulations where distinct personas coordinate on tasks, enabling analysis of persona-driven behavior at scale (Salem et al., 2025). PERSONAFLOW offers patterns for constructing expert-oriented personas and evaluating decision support under explicit identity constraints (Liu et al., 2025b).

D.3.3 Social Simulation

This subsection examines how LLM agents *simulate human social interaction* in cooperative work, negotiation, and group-level dynamics. The focus is on whether multi-agent systems reproduce characteristic human behaviors and outcomes.

Cooperation. Role-based setups test whether agents coordinate like human teammates. CAMEL pairs complementary roles with inception prompting, yielding stable division of labor and shared routines (Li et al., 2023a). Open-ended environments such as SOTOPIA evaluate social goal pursuit through interactive role-play and reveal systematic differences across models (Zhou et al., 2024a); SOTOPIA-π adds interaction-derived learning (behavior cloning + self-reinforcement) to improve goal completion (Wang et al., 2024c). Broader benchmarks increase comparability: MULTIAGENTBENCH standardizes collaboration/competition scenarios and metrics, while BATTLEAGENTBENCH highlights persistent failures on harder cooperative tasks (Zhu et al., 2025a; Wang et al., 2024d). Domain testbeds adapt human team problems to agent societies, e.g., cooperative planning with partial observability in COLLAB-OVERCOOKED, and frameworks that emphasize coordinated task planning and communication (Sun et al., 2025b). Complementary studies examine consensus seeking via iterative protocols, mixed cooperative/competitive patterns in multi-agent systems, and survey evaluation practices for cooperation in dynamic environments (Chen et al., 2025a; Tran et al., 2025; Gao et al., 2023).

Negotiation. Negotiation benchmarks probe perspective taking, preference reconciliation, and commitment strategies. CICERO combines long-horizon planning with natural-language bargaining to reach human-level performance in Diplomacy, establishing a reference point for mixed human–agent play (Bakhtin et al., 2022). Interactive multi-issue testbeds provide tunable roles and payoffs with metrics for deal quality, fairness, and role adherence (Abdelnabi et al., 2024). Recent analyses broaden coverage: assistive role-play agents add remediation for norm compliance in business-style negotiations (Hua et al., 2024b); a 35-task zero-shot suite decomposes competence into comprehension, partner modeling, and generation, noting strong but brittle strategic responses (Kwon et al., 2024); and a large-scale autonomous competition links linguistic "warmth" to agreement rates and value creation (Vaccaro et al., 2025). Personality-conditioned negotiators and game-theoretic workflows study how traits and strategy scaffolds shift outcomes toward more principled bargains (Noh and Chang, 2024; Hua et al., 2024a). Persuasion-oriented evaluations complement bargaining metrics with human-facing influence measures (Breum et al., 2024).

Group dynamics. Population studies ask whether agent societies exhibit human-like macro-patterns. GENERATIVE AGENTS shows that memory/plan/reflect loops can produce routines, coordination, and community events reminiscent of small-town life (Park et al., 2023). Purpose-built architectures (e.g., CRSEC) elicit norm creation, spread, and compliance under controlled conditions (Ren et al., 2024b). Controlled population experiments report the emergence of shared linguistic conventions and collective bias dynamics in decentralized groups (Ashery et al., 2025). To mirror human heterogeneity, population-aligned sampling matches psychometric distributions, and mixture-of-personas conditioning elicits controlled diversity in group responses, linking micro-level agent design to macro-level outcomes (Bui et al., 2025a).

E LLMs-against-Humans: Paradigm Specific and Cross-Cutting Risks

Although the three paradigms—*LLMs-for-humans*, *LLMs-with-humans*, and *LLMs-as-humans*—describe cooperative interaction, deployments increasingly reveal cases where model behavior diverges from human goals or produces unintended effects. We refer to these as **LLMs-against-humans**, encompassing both paradigm-specific and cross-paradigm risks and harms. This

Benchmark	for-Humans	with-Humans	as-Humans	Primary Domain
MATH / GSM8K (Hendrycks et al., 2020; Cobbe et al., 2021)	✓ Math assistance	✓ Tutoring feedback	—	Mathematical reasoning
ToolBench / Gorilla / AppWorld (Xu et al., 2023; Trivedi et al., 2024)	✓ Tool orchestration	✓ Collaborative agents	—	API / Tool usage
Mind2Web / WebArena / ALFWorld (Deng et al., 2023; Zhou et al., 2023b; Shridhar et al., 2020)	✓ Web automation	✓ Shared environment	—	Web navigation / Embodied tasks
HumanEval / MBPP (Chen et al., 2021; Austin et al., 2021)	✓ Code generation	✓ Pair programming	—	Software development
BigBench Hard (Srivastava et al., 2022)	✓ Tool-aided reasoning	—	✓ Reasoning simulation	Multi-step reasoning
Persona-Chat (Zhang et al., 2018b)	✓ Personalized assistance	—	✓ Persona simulation	Persona / Identity
SocialIQa / MMLU (Sap et al., 2019; Hendrycks et al., 2021)	✓ Expert consultation	—	✓ Knowledge / social reasoning proxy	World knowledge / Social reasoning
CaSiNo / Deal or No Deal Negotiator (Chawla et al., 2021; Lewis et al., 2017)	—	✓ Strategic interaction	✓ Negotation Simulation	Negotiation / Bargaining
MindAgent / GameNGen / MultiAgentBench / Agents (Gong et al., 2024; Zhu et al., 2025a; Zhou et al., 2023c)	—	✓ Multi-agent interaction	✓ Behavior / agent simulation	Multi-agent collaboration / Competition
SWE-bench / AgentBench / BOLAA / BFCL / WebShop / OSWorld (Jimenez et al., 2023; Liu et al., 2023b,d; Patil et al.; Yao et al., 2022; Xie et al., 2024)	✓ Task / Issue resolution	✓ Agentic alignment	✓ Behavior simulation	Multi-domain agentic tasks
AlpacaEval / MT-Bench (Li et al., 2023b; Zheng et al., 2023)	✓ Conversation assistance	✓ Preference learning	✓ Judge simulation	Multi-turn dialogue evaluation

Table 3: Optimized benchmark table spanning multiple human–LLM interaction paradigms. Checkmarks indicate primary applicability; dashes indicate limited relevance. Benchmarks are merged where possible for clarity.

perspective provides a unified lens for analyzing when systems intended to operate *for*, *with*, or *as* humans generate outcomes that conflict with or misrepresent human intent. Table 5 summarizes these risks.

E.1 LLMs-for-Humans: Manipulative or Unreliable Assistance

In assistive settings, misalignment occurs when models intended to help users produce outputs that mislead, over-persuade, or diminish user agency. Fluent but factually incorrect content remains a core limitation, as hallucinations persist across domains and degrade trust in high-stakes decision-making (Bang et al., 2025; Ji et al., 2024; Huang et al., 2025c; Sahoo et al., 2024).

Studies show that sustained exposure to automated assistance fosters cognitive offloading, where users defer reasoning and evaluation to models, weakening analytical and creative engagement (Zhai et al., 2024; Jose et al., 2025; Doshi et al., 2024; Spatharioti et al., 2025; Kumar et al., 2025). These effects highlight a tension between usability and autonomy in human–AI interaction.

Bias and misinformation remain pervasive. Models reflect and amplify inequities embedded in their training data and are vulnerable to adversarial manipulation, raising concerns about fairness and robustness (Lin et al., 2025b; Peng et al., 2024). Privacy and security risks further complicate deployment, with evidence of prompt injection, inference attacks, and unintentional data exposure (Zeng et al., 2025; Pasch and Cha, 2025).

Finally, when connected to external tools or APIs, LLMs often mismanage or misapply them, compounding reasoning errors and reducing overall reliability (Cui et al., 2025; Zhang et al., 2025c,a).

E.2 LLMs-with-Humans: Feedback Corruption and Misalignment

In collaborative and feedback-driven settings, misalignment arises when adaptive mechanisms amplify bias or distort user intent. Human feedback may be incomplete, inconsistent, or adversarial, causing models to overfit to narrow preferences or exploit reward signals in unintended ways

Evaluation Method	LLMs-for-Humans	LLMs-with-Humans	LLMs-as-Humans
Automatic	Task success , BLEU/ROUGE, F1, accuracy , pass rate (coding), hallucination rate, **reward alignment** (Qin et al., 2023; Jin et al., 2020)	**Agreement with human feedbac** (Cohen's κ, Krippendorff's α, Spearman's ρ), LLM-as-a-judge (Zheng et al., 2023; Li et al., 2023b), reward model score (Lambert et al., 2025), reasoning accuracy (Cobbe et al., 2021), task success rate (Liu et al., 2023b)	**Reliability and agreement** (κ, α, ρ), Accuracy (Kosinski, 2024; Strachan et al., 2024), moral calibration (Hendrycks et al., 2021; Ji et al., 2025b) Persona consistency (Zhang et al., 2018b), response diversity (Jafari et al., 2025), negotiation win rates (Lewis et al., 2017; Chan et al., 2024), group task success (Park et al., 2023)
Human-centered	Helpfulness, harmlessness , accessibility, faithfulness (Nakano et al., 2021), satisfaction, **trust** , workload (NASA-TLX) (Miah and Zhu, 2024)	Collaboration quality, preference satisfaction, **trust calibration** , pairwise comparison and Elo rating (Zheng et al., 2023; Li et al., 2024d), helpfulness and harmlessness (Bai et al., 2022), instruction following (Zhou et al., 2023a)	**Plausibility and fairness** , believability (Sclar et al., 2023), coherence (Dillion et al., 2025), identity authenticity (Dillion et al., 2025), persuasiveness (Wang et al., 2019), perceived fairness (Jiang et al., 2022), social plausibility (Park et al., 2023)
Hybrid	Task success + user metrics, user improvement vs. model accuracy, **synergy gain** (Miah and Zhu, 2024), consistency under feedback	Automatic RL metrics (rewards) + human preference assessments, **Complementarity gain** , adaptation curves, **synergy improvement** (Zhu et al., 2024; Ouyang et al., 2022), reward–preference integration	Automatic metrics + subjective human assessments (fluency, naturalness), **Fidelity of human emulation** , perception alignment (Bui et al., 2025b)
Adversarial	**Robustness** to misleading instructions, over-reliance detection, noise perturbations (Yu et al., 2023), adversarial test cases , prompt injections, red-teaming	**Feedback-loop vulnerabilities** , reward model exploitation, red-teaming, adversarial prompts , bias under conflicting goals (Oikarinen et al., 2020; Zhu et al., 2024)	**Robustness and bias sensitivity** , manipulation resistance, Persona modulation, prompt injection, reward exploitation, CoT hijacking (Robey et al., 2023)

Table 4: Evaluation methods and representative metrics across the three human–LLM paradigms. Color highlighting denotes shared or overlapping measures supporting *cross-paradigm* comparison: alignment and reliability , trust and experience , complementarity and synergy , and robustness and bias evaluation .

(Dahlgren Lindström et al., 2025; Amodei et al., 2016). Alignment elasticity—where models revert to pre-training behavior—reduces reliability, while alignment faking produces superficially compliant but deceptive responses that undermine trust (Ji et al., 2025a; Greenblatt et al., 2024).

Safety vulnerabilities remain a persistent concern, including prompt manipulation, fine-tuning circumvention, and reward hacking, each of which can generate biased or harmful outputs (Zhou et al., 2024b; Dahlgren Lindström et al., 2025). Feedback loops further reinforce narrow or polarized content, weakening model expressivity and communicative diversity over time (Shen et al., 2023a; Ji et al., 2025a). Cultural and linguistic biases also limit generalizability, as models often reflect English- and Western-centric norms that reduce fairness and inclusivity (Shen et al., 2023a; Yao et al., 2024; Ryan et al., 2024; Liang et al., 2023).

Aligned systems risk anthropomorphic and persuasive misalignment: users may overtrust, over-personalize, or develop dependency on conversational models (Cohn et al., 2024; Strachan et al., 2024; Breum et al., 2024; Street, 2024). Sycophantic responses further degrade epistemic reliability by echoing user biases and reinforcing misinformation (Rrv et al., 2024; Sharma et al., 2023b; Malmqvist, 2024).

E.3 LLMs-as-Humans: Deceptive Simulation and Identity Distortion

When models are positioned as substitutes for human judgment or behavior, misalignment arises through inaccurate, biased, or misleading representations of human roles. LLMs acting as evaluators, participants, or personas may produce outputs that appear reliable but misrepresent human reasoning, diversity, or expertise, reducing the validity of evaluation and collaboration.

As judges or annotators, models risk overestimating alignment with human assessments, where surface correlations are mistaken for genuine understanding, resulting in inconsistent or biased evaluations (Li et al., 2024b; AlKhamissi et al., 2024; Adilazuarda et al., 2024; Chang et al., 2024; Chen et al., 2024a; Thakur et al., 2025). Moral or social reasoning tasks often replicate patterns from training data rather than normative reasoning, reinforcing stereotypes and cultural bias (Kosinski, 2024). Replacing human annotators may also conceal labor inequities and reduce perspective diversity.

Persona simulation introduces risks of identity distortion, where synthetic profiles simplify or misrepresent social groups and may reproduce sensitive identities without consent (Tseng et al., 2024; Gao et al., 2025; Wang et al., 2025a; Wu et al., 2025). In social simulations, multi-agent systems can display coordinated behaviors that appear human-like but reflect training artifacts, leading to biased or persuasive outcomes (Park et al., 2023; Fisher et al., 2025; Hua et al., 2024b; Kwon et al., 2024; Moëll and Sand Aronsson, 2025).

E.4 Cross-Paradigm: Adversarial and Competitive Dynamics

Certain forms of misalignment extend across all paradigms, linking assistive, collaborative, and simulative settings through shared vulnerabilities and feedback effects. Errors in factuality, bias mitigation, or feedback handling can propagate from one paradigm to another, creating system-level interactions that are difficult to isolate. For instance, synthetic annotations produced by models acting *as* humans may bias the preference data that shapes systems trained *with* humans, while reward objectives optimized in collaborative alignment can reinforce persuasive or manipulative behaviors when models act *for* users. These cross-paradigm dynamics show that failures of calibration, transparency, and accountability are not confined to specific interaction modes but emerge from interconnected development and deployment pipelines (Amodei et al., 2016; Zhou et al., 2024b). At scale, competition between human and model-generated content introduces further risks to epistemic reliability. Automated data generation can reduce the visibility of human contributions, affect data diversity, and amplify feedback loops in which models increasingly learn from their own outputs, leading to distributional drift and bias.

Paradigm	Human-LLM relationship	"LLMs-against-Humans": misalignment manifestations
LLMs-for-Humans	Assistive systems where models support user tasks and decision-making.	*Manipulative or unreliable assistance:* hallucinated or biased outputs mislead users; over-reliance reduces human agency and critical thinking; persuasive phrasing steers user choices; privacy risks through overexposed data.
LLMs-with-Humans	Collaborative or feedback-driven co-adaptation between humans and models.	*Corrupted feedback and misalignment:* adversarial users exploit feedback loops; reward hacking distorts objectives; preference data encode population bias; iterative co-adaptation drifts from intended values.
LLMs-as-Humans	Simulation of human roles such as evaluators, teammates, or participants.	*Deceptive substitution and identity distortion:* models impersonate human roles without disclosure; synthetic judgments mistaken for human consensus; anthropomorphic framing erodes accountability; biased or culturally narrow simulations misrepresent human behavior.
Cross-cutting (All)	Applies across all three paradigms.	*Adversarial and competitive dynamics:* emergent model behaviors undermine trust, spread misinformation, or compete for user attention and authority; malicious use amplifies disinformation or manipulates social systems.

Table 5: *LLMs-against-humans* dimension across the three human–LLM interaction paradigms. This paradigm, "against" represents a cross-cutting lens for analyzing when model behavior diverges from or conflicts with human objectives.