A Survey on Human-AI Collaboration with Large Foundation Models

VANSHIKA VATS*, MARZIA BINTA NIZAM*, MINGHAO LIU, ZIYUAN WANG†, RICHARD HO†, MOHNISH SAI PRASAD†, VINCENT TITTERTON†, SAI VENKAT MALREDDY†, RIYA AGGARWAL†, YANWEN XU†, LEI DING†, JAY MEHTA†, NATHAN GRINNELL†, LI LIU†, SIJIA ZHONG†, DEVANATHAN NALLUR GANDAMANI†, XINYI TANG†, ROHAN GHOSALKAR†, CELESTE SHEN†, RACHEL SHEN†, NAFISA HUSSAIN†, KESAV RAVICHANDRAN†, and JAMES DAVIS, University of California, Santa Cruz, USA

As the capabilities of artificial intelligence (AI) continue to expand rapidly, Human-AI (HAI) Collaboration, combining human intellect and AI systems, has become pivotal for advancing problem-solving and decision-making processes. The advent of Large Foundation Models (LFMs) has greatly expanded its potential, offering unprecedented capabilities by leveraging vast amounts of data to understand and predict complex patterns. At the same time, realizing this potential responsibly requires addressing persistent challenges related to safety, fairness, and control. This paper reviews the crucial integration of LFMs with HAI, highlighting both opportunities and risks. We structure our analysis around four areas: human-guided model development, collaborative design principles, ethical and governance frameworks, and applications in high-stakes domains. Our review shows that successful HAI systems are not the automatic result of stronger models but the product of careful, human-centered design. By identifying key open challenges, this survey aims to give insight into current and future research that turns the raw power of LFMs into partnerships that are reliable, trustworthy, and beneficial to society.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Human-AI Collaboration, Artificial Intelligence, Large Foundation Models, Large Language Models

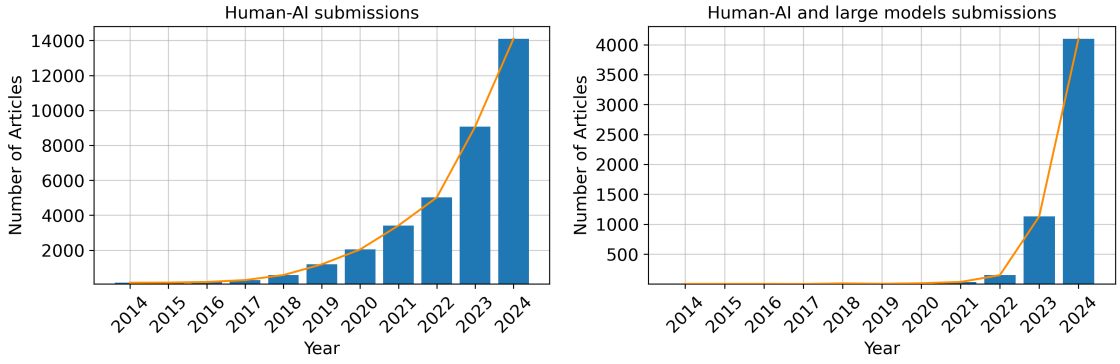
ACM Reference Format:

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, Lei Ding, Jay Mehta, Nathan Grinnell, Li Liu, Sijia Zhong, Devanathan Nallur Gandamani, Xinyi Tang, Rohan Ghosalkar, Celeste Shen, Rachel Shen, Nafisa Hussain, Kesav Ravichandran, and James Davis. 2025. A Survey on Human-AI Collaboration with Large Foundation Models. 1, 1 (September 2025), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

*Shared First Authorship
†Equal Contribution

Authors' Contact Information: Vanshika Vats, vvats@ucsc.edu; Marzia Binta Nizam, manizam@ucsc.edu; Minghao Liu; Ziyuan Wang; Richard Ho; Mohnish Sai Prasad; Vincent Titterton; Sai Venkat Malreddy; Riya Aggarwal; Yanwen Xu; Lei Ding; Jay Mehta; Nathan Grinnell; Li Liu; Sijia Zhong; Devanathan Nallur Gandamani; Xinyi Tang; Rohan Ghosalkar; Celeste Shen; Rachel Shen; Nafisa Hussain; Kesav Ravichandran; James Davis, University of California, Santa Cruz, USA.

Manuscript submitted to ACM



(a) (b)

Fig. 1. A graphical depiction of the number of relevant articles submitted from 2014-2024. (a) illustrates the increasing engagement of research communities in collaborations between humans and AI over the years. (b) observes a notably sharp increase in submissions related to human-AI and large-scale models attributed to the recent emergence of large models.

1 Introduction

People have long been fascinated by the idea of creating machines that think like humans. One notable example from the 1770s is the "Mechanical Turk," a machine that appeared to play chess autonomously but was, in fact, operated by a person concealed inside it [33]. This early effort, while not *"Artificial Intelligence" (AI)* in the modern sense, reflects the enduring fascination with creating technology that can mimic or complement human cognitive abilities. The formal realization of integrating human-like intelligence into machines peaked in 1956 at Dartmouth College, USA, marking the official birth of Artificial Intelligence as a research field [136]. As AI technology has advanced and become more prevalent over the last decade, researchers have identified limitations within purely automated systems [53, 113, 135] leading to renewed focus on augmenting AI with human expertise, aiming to utilize the best of both worlds (Fig. 1a). This approach seeks to enhance AI applications and foster a symbiotic collaboration between human and artificial intelligence.

Navigating the rapidly evolving AI landscape, the integration of human cognition with Large Foundation Models (LFMs), including Large Language Models (LLM) [17, 159] and Large Vision Models (LVM) [70, 82, 132], has initiated an exciting shift. These models are pre-trained on vast web-scale datasets and act as "foundations" before being fine-tuned for specific tasks. This has opened new avenues for collaborative problem-solving and decision-making. When humans add domain insight, ethics, and creativity to these generic large models, they return rapid pattern-finding and specialized outputs at scale. In turn, AI amplifies human capabilities by processing data at large scales, offering insights and augmenting decision-making. This interaction paves the way to creating an advanced form of HAI collaboration.

For this survey paper, we formally define the term 'Human-AI Collaboration'. Human-AI (HAI) collaboration is a cooperative partnership in which humans and AI systems coordinate their complementary strengths, exchange information (and sometimes control), and pursue a shared goal, without the tight interdependence and joint accountability [76, 144, 182]. Each partner contributes what it does best and iteratively refines the joint output through feedback. This paper undertakes an extensive survey of HAI collaboration, analyzing the complex interactions between human agents and sophisticated pre-trained models across various fields. Our study aims to uncover the progress made, confront the challenges, and understand the implications of this evolving partnership.

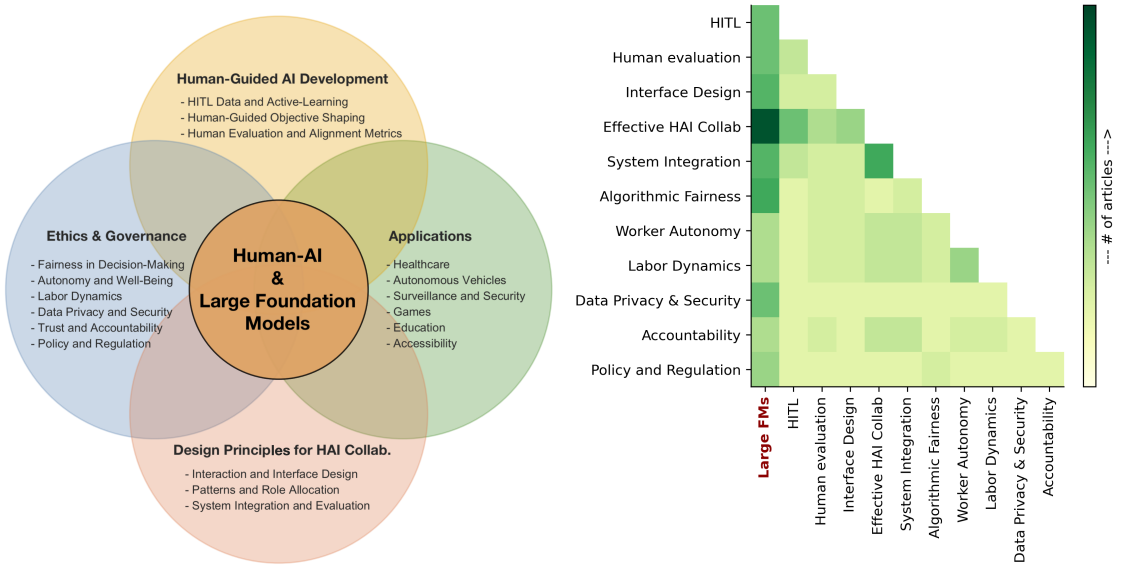


Fig. 2. A representation of the scope of this survey. (a) We cover four broad categories - Human Guided AI Development, Design Principles for HAI Collaboration, Ethics and Governance, and their Applications - in the human-AI domain, along with the recent developments in them with the help of LFMs. The overlap between the four petals illustrates the fact that the four categories are not mutually exclusive; the articles may belong to two or more categories at the same time. (b) The heatmap shows a general trend of the number of articles co-existing within subcategories, including the articles discussing large foundation models within each subject area. The overlap between the applications is not analyzed, given their specific and varied nature. *AI*: Artificial Intelligence, *HAI*: Human-AI, *HITL*: Human-in-the-Loop.

1.1 Scope of the Survey

Our survey focuses on the articles describing the developments of human-AI collaboration through the years and how the introduction of large models is reforming the field. The search for articles was conducted on Google Scholar focusing on the years between 2014-2024, using keywords such as "Human-AI", and "Human-AI collaboration". For research focusing on the collaboration between humans and AI involving large models, additional keywords, including "large models", "foundation models", "large language models", and "large vision models" were utilized. The content was then distributed among the authors to filter based on the theme of our review to be structured into human-guided AI development, design principles for HAI collaboration, ethics and governance, and finally, applications. Some of the selected papers fall into two or more categories, as illustrated in Fig. 2. Since human-AI collaboration and large models have only recently gained a lot of interest from the research community given the recent popularity of large foundation models (Fig. 1b), some portion of studies cited in this survey could also be from arXiv preprints. We then sample the articles according to the structure desired for this survey, i.e., primarily covering Human-AI collaboration and its large foundation model counterparts in better model training, effective human-AI joint systems, safe and secure HAI collaboration, and their applications (Fig. 2a). Although not systematic, this review captures the most visible work across AI, HAI, and collaboration.

In order to show support for the theme of our survey article, we try to implement HAI collaboration with the help of large language models to prepare this review. For the *human* part, the authors survey the existing literature about human-AI collaboration, collect and filter a subset of articles according to the desired structure, organize the survey into relevant sections, study submission statistics, put together tables and visualizations, and finally, prepare the first complete draft of the manuscript. Using the *AI* part in human-AI teaming, we ask ChatGPT-4 [123] for its feedback regarding the language flow of each subsection, placement of the citations, and revising the language wherever necessary. This helps to ease in identifying the missing elements and for the authors to make necessary revisions. The final manuscript is, therefore, a product of human and AI collaboration.

1.2 Outline of the Survey

In each segment of our study, we commence by delineating traditional methodologies employed in human-AI collaboration, subsequently delving into the contributions of extensive pre-trained foundation models in the fulfillment of these tasks. Section 2 explores the involvement of incorporating human expertise into the AI model training cycle, fostering a cooperative relationship between humans and AI in active learning scenarios, enhancing learning through human feedback, and involving human experts in the thorough evaluation of machine learning models [66, 141, 194]. Section 3 explores understanding the scope of effective Human-AI joint systems, optimizing user interfaces (UI) and system architectures emerges as a crucial area. It covers a range of topics under innovative UI designs and system structures aimed at boosting efficiency and user engagement in collaborative settings [44, 91].

A major concern in Human-AI collaboration is the safety, security, and trustworthiness of AI systems. Section 4 of our comprehensive analysis covers multiple dimensions: mitigating algorithmic biases to uphold fairness, assessing the impact of Human-AI collaboration on workers' autonomy, well-being, and job satisfaction, examining economic repercussions on employment and wages, addressing data privacy and security concerns, building trust in AI systems, and navigating the legal and regulatory frameworks governing Human-AI interactions [15, 63, 85]. These factors are integral to creating an ethical and responsible environment for AI's integration into human-centric workflows. Section 5 explores the broad spectrum of Human-AI collaboration applications across various sectors. We investigate the unique challenges and opportunities in fields such as healthcare [12], autonomous vehicles, surveillance systems [61], gaming [46], education, and accessibility. Understanding the specific characteristics and benefits of Human-AI collaboration in these areas is key to influencing its future direction and maximizing its societal impact. Finally, Section 6 summarizes the core findings and identifies key open challenges and future research directions, providing a roadmap for building the next generation of effective, fair, and trustworthy human-AI partnerships.

2 Human-Guided Model Development

LFMs are powerful when people steer them correctly to suit their specific tasks. In this section, we adapt the three-phase view of Maadi et al. [103], *data control* → *model optimisation* → *evaluation* (Fig. 3), to review the points where human insight most clearly improves an FM:

- Sec. 2.1 Human-in-the-Loop Data & Active-Learning: People curate or label the right examples and, in today's LFMs, craft or filter instruction prompts that guide later fine-tuning.
- Sec. 2.2 Human-Guided Objective Shaping: preference-based methods such as RLHF and DPO let human feedback reshape the model's reward or loss surface during optimisation.

- Sec. 2.3 Human Evaluation and Alignment Metrics: Domain experts and end-users judge outputs, supplying both task scores and subjective signals of trust, fairness, and usability, which feed into the next iteration of the HAI collaboration pipeline.

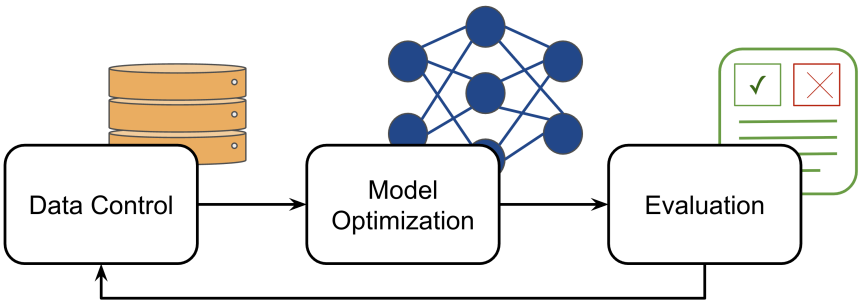


Fig. 3. The three-phase iterative framework for human-guided model development - comprising data control, model optimization, and evaluation, as described by Maadi et al. [103].

2.1 Human-in-the-Loop Data and Active-Learning

Human-in-the-Loop (HITL) methods exemplify human-AI collaboration by combining human judgment with automated efficiency. Humans guide model design by bringing ethical insight beyond technical metrics (Whittlestone et al. [181]), while AI accelerates data processing and suggests patterns that might not be easy to catch for the time-constrained annotators. During training and deployment, information flows bidirectionally. For example, NLP-based directives like InstructRL, as proposed by Hu and Sadigh [66], steer AI toward more interpretable, aligned behaviors, and AI-generated prompts enrich human brainstorming, sparking novel ideas [108]. The diverse roles of humans, ranging from providing ethical oversight in the design phase to contributing to quality control and data labeling during model training and execution, are crucial for optimizing AI performance.

2.1.1 ***Human-AI Collaboration with Active Learning.*** Unlike traditional machine learning concepts, active learning involves an iterative process where a model selectively identifies data requiring labeling, optimizing system performance with minimal training data [47]. Particularly, this approach is crucial for fine-tuning large pre-trained foundation models that require substantial labeled data for specific user scenarios. Active learning efficiently utilizes human expertise to pinpoint areas of uncertainty, enabling more targeted training with less annotation effort [141]. The active learning workflow begins with an unlabeled dataset and a pre-trained model. The model predicts labels for each sample, outputting confidence levels (Fig. 4). When predictions fall below a quality threshold, human annotators step in for manual annotation. This iterative process of re-training the model with new labeled data continues until satisfactory confidence levels are achieved. Humans and AI thus collaborate in efficient data labeling and continuous model improvement.

Recent research in this area has introduced novel methods to leverage human expertise effectively. For instance, Pries et al. [130] propose a method for efficient data labeling using precise distance measurements to filter data samples, streamlining the presentation of data to experts. Lu et al. [98] discuss human-guided interventions for continuous model improvement, focusing on dissociating biases. Bao et al. [8] explore the transformation of human-annotated rationales into continuous attention mechanisms, enhancing the learning of domain-invariant representations. Additionally,

Human-AI Topics	Subtopics	Articles Cited
Section 2: Human-Guided Model Development	HITL and Active-Learning	[181], [66], [108], [47], [141], [130], [98], [8], [93], [187]
	Human-Guided Objective Shaping	[30], [124], [149], [133], [5], [90], [74], [116], [69], [166], [66], [112], [174], [189], [191], [73], [106], [192]
	Human Evaluation and Alignment Metrics	[6], [193], [69], [166], [31], [27], [66], [152], [99], [160]
Section 3: Design Principles for Human-AI Collaboration	Interaction and Interface Design	[191], [61], [18], [131], [111], [184], [87], [137], [190], [27], [32], [169], [162], [44], [45], [175], [91]
	Collaboration Patterns and Role Allocation	[175], [184], [129], [39], [96], [118], [190], [54], [115], [133], [116], [66], [124], [194], [193], [64], [100], [172], [165], [126], [108], [1], [48], [7], [24], [162], [97], [117], [119], [110], [77], [134], [143], [23], [142]
	System Integration and Evaluation	[176], [27], [10], [172], [41], [59], [106], [101], [65], [7], [184], [165], [64], [97], [115], [108], [25]
Section 4: Ethical, Societal and Governance Aspects	Fairness in Collaborative Decision-Making	[55], [148], [34], [27], [114], [83], [67], [86], [109], [94], [50], [121], [11], [95]
	Empowering Human Autonomy and Well-Being	[27], [85], [126], [37], [172], [183]
	Collaborative Labor Dynamics	[63], [29], [27], [126], [3], [37], [172], [167]
	Data privacy and Security	[43], [188], [75], [138], [92], [58], [161], [57], [36], [155], [140]
	Building Trust and Shared Accountability	[65], [129], [21], [172], [81], [6]
	Policy and Regulation	[71], [22], [148], [13], [154], [168], [104]
Section 5: Applications	Healthcare	[64], [12], [107], [20], [105], [19], [28], [102], [89], [72], [150], [14]
	Autonomous Vehicles	[4], [100], [193], [48], [186], [127], [35], [156], [173], [178], [171]
	Surveillance and Security	[80], [128], [61], [56], [26], [68], [9], [122]
	Games	[151], [46], [185], [147], [2], [145], [157], [164], [66], [180]
	Education	[120], [163], [179], [78], [40], [60], [38], [42], [146], [170], [62], [25], [84], [139]
	Accessibility	[88], [125], [79], [177], [52], [153], [49], [16], [158]

Table 1. A tabular representation of the four broader focus topics covered in this survey, with their relevant subtopics. Each category contains its cited articles for the ease of reader reference.

Lertvittayakumjorn et al. [93] present a generalizable approach for debugging deep text classification models with human input, applicable to larger models. Finally, Yao et al. [187] propose an explainable-generation active learning framework for simpler tasks, potentially improving the sampling process in data augmentation.

Taken together, research in this area shows a shift in the human role in data curation. Traditional active learning balances annotation cost against model performance, but with Large Foundation Models, humans act less as mass annotators and more as targeted teachers during fine-tuning. This shift brings a difficult trade-off: small, carefully

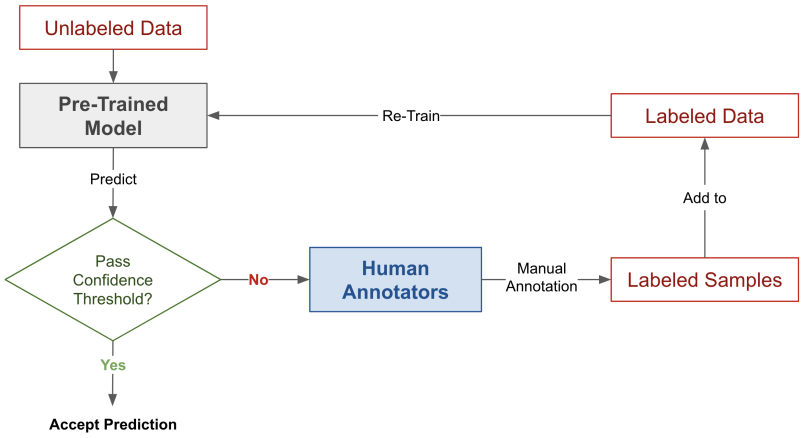


Fig. 4. A visual representation of an active learning workflow. If the model predictions do not pass the required confidence value thresholds required for a task, the data is sent to humans for manual annotations, further improving the model by retraining it with the now-labeled data.

chosen datasets must be diverse enough to reduce, not reinforce, the biases already present in the model. A key open challenge remains to design bias-aware active learning methods that identify not only areas of model uncertainty but also regions where pre-trained knowledge is skewed or incomplete.

2.2 Human-Guided Objective Shaping

After curating the *data* that an LFM sees, collaborators can intervene a second time by reshaping the *objective* it tries to optimize. Rather than relying on hard-coded task losses, recent methods let humans express preferences, rankings, or critiques over model outputs and convert these signals into a learnable reward or direct policy update. The canonical pipeline, Reinforcement Learning from Human Feedback (RLHF) [30, 124, 149], trains a reward model on pairwise preference judgements and then fine-tunes the policy to maximize that reward, producing assistants that are measurably more helpful and safer. Simpler, gradient-based alternatives such as Direct Preference Optimization (DPO) [133] and Constitutional AI [5] bypass reinforcement yet achieve similar alignment by treating human or rule-based critiques as a differentiable signal. This section surveys these preference-based algorithms, the types of human feedback they require, and open challenges in keeping objectives both scalable and faithful to human intent.

2.2.1 ***RLHF Pipeline.*** Typically, the RLHF approach involves three stages [124]: initially, the pre-trained foundation model generates candidate responses; human annotators then evaluate these outputs, providing explicit preference-based feedback; subsequently, a reward model is trained on these human-generated evaluations. Finally, policy optimization methods, often proximal policy optimization (PPO) or direct preference optimization (DPO) [133], fine-tune the foundation model to maximize the learned reward. This iterative human–AI feedback loop enables continuous and targeted improvements in model alignment.

2.2.2 ***Collaborative Advantages.*** RLHF's reliance on human evaluation significantly enhances the alignment of models with ethical and functional expectations. Human annotators can actively identify and mitigate harmful biases or unsafe outputs early in the optimization process, producing fairer and more trustworthy AI systems [90].

Additionally, the iterative nature of human feedback facilitates rapid adaptation to evolving domain-specific or user-specific preferences, overcoming the rigidity inherent in traditional supervised learning methods. Transparency is also notably improved, as human-generated feedback and reward insights provide explicit visibility into the rationale behind model decision-making [74]. The adaptability of RLHF frameworks to user and domain-specific contexts further enhances their practical effectiveness, enabling personalized and contextually significant AI behaviors.

Despite these advantages, implementing RLHF involves several practical challenges. Ensuring consistency and accuracy of human feedback, scaling RLHF effectively to large datasets, and aligning human insights with algorithmic processes are key considerations that must be addressed to fully leverage RLHF's potential [90].

2.2.3 ***New HAI Paradigms in RLHF.*** Recent methodological advances have substantially improved RLHF. OpenAI's InstructGPT [124] introduced a structured three-phase training process comprising supervised policy training, reward model training with human feedback, and subsequent reinforcement learning-based policy optimization. Similarly, Stanford's Direct Preference Optimization (DPO) simplifies the training by directly mapping human feedback into policy improvements, thereby reducing complexity and enhancing the quality of outputs [133]. Further illustrating RLHF's potential, Nakano et al. [116] introduced web browser-assisted agents trained via RLHF, demonstrating significant improvements in real-world usability. Recent studies have further expanded RLHF frameworks through innovative approaches. Ji et al. [69] introduced a dataset enabling models to integrate multimodal human feedback—spanning text, images, audio, and video—thereby significantly enriching the interaction between humans and AI. Additionally, Vodrahalli et al. [166] proposed a canonical basis of human preferences, identifying a compact yet expressive set of fundamental categories that efficiently represent diverse human judgments, potentially reducing annotation overhead and enhancing model interpretability.

Extending beyond language models, RLHF methodologies have influenced fields like robotics and interactive decision-making systems. Hu and Sadigh's InstructRL [66] employs natural-language instructions to effectively guide robotic actions, underscoring the versatility of RLHF approaches. Additionally, research by Mirchandani et al. [112] highlights the utility of large language models (LLMs) in robotics, leveraging their in-context learning abilities to facilitate complex sequential tasks. Chain-of-thought prompting techniques [174] and correction-based learning frameworks such as DROC [189] further exemplify RLHF's expanding scope.

Nevertheless, RLHF models still face some hurdles, notably in bridging communication barriers between humans and AI. Studies by Zhang et al. [191] and Johnson and Vera [73] emphasize these challenges, highlighting the need for more sophisticated human-AI communication frameworks. McNeese et al. [106] further underscore the limitations of current RLHF models in managing real-world complexities, particularly regarding nuanced human communication and the ability to generalize from simulated environments to real-world settings.

To summarize, the work on objective shaping marks a shift from programming explicit goals to learning implicit human values through feedback. RLHF and its variants demonstrate that human preference signals can reliably steer models toward safer and more useful behaviors. However, the practical limits of noisy, costly, and potentially narrow annotation remain. Recent advances [166, 192] point toward concrete solutions: datasets that capture multimodal human feedback (text, images, audio, video) and frameworks that distill a canonical basis of preferences reduce annotation overhead while broadening coverage of diverse values. Together, these directions suggest that scalable, bias-resilient objective shaping is achievable, not by collapsing human preferences into a single worldview, but by structuring them into compact, expressive, and extensible feedback spaces.

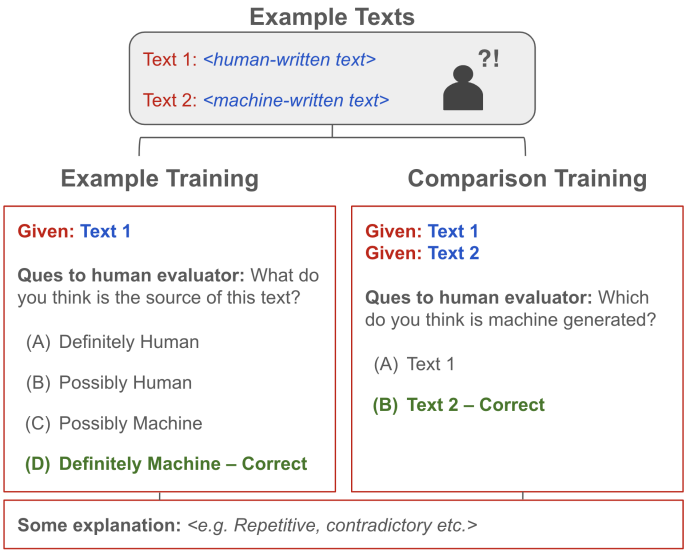


Fig. 5. Evaluation training strategies based on Clark et al. [31]. They employ training methods by giving example and comparison tests, along with their correct answers. This is followed by an explanation of why the answer is correct. This way, the evaluators gradually learn about how to distinguish between human and machine-generated texts.

2.3 Human Evaluation and Alignment Metrics

Human evaluation is the final collaborative safeguard that ensures foundation models satisfy both task requirements and human expectations. Rather than serving as a single 'final report card' human evaluation forms a feedback loop: people surface errors, voice preferences, and refine rubrics, which in turn steer data curation (2.1) and objective shaping (2.2). The result is an iterative alignment cycle that balances raw performance with usability, trust, and safety.

2.3.1 ***Trust, Transparency, and Safety.*** Establishing calibrated trust begins with making model limits visible. Bansal et al.[6] show that users build more accurate mental models when an AI system discloses its own error likelihood. Zhou and Chen's Uncertainty-Performance Interface (UPI) [193] couples confidence scores with outcome heat-maps so that decision makers can dynamically adjust their thresholds. Recent work extends these ideas to multimodal settings: Ji et al. [69] let raters score image–text consistency, while Vodrahalli et al.[166] distill 21 canonical preference axes that cover most harm-related judgements, trimming annotation cost without losing coverage. Together, these studies frame transparency as an ongoing dialogue rather than a static disclosure.

2.3.2 ***Collaborative Approaches to Improve Evaluation.*** Human contribution is most powerful when it shapes evaluators, not just evaluations. Clark et al. [31] boost label accuracy by providing annotators with contrasting model outputs and rationale. They examine how human evaluators can differentiate between texts produced by humans and models like GPT-2 and GPT-3. They suggest methods to enhance evaluation accuracy, including detailed instructions, annotated examples, and text comparisons (Fig. 5). Chhibber et al. [27] demonstrate that teaching annotators the model's decision heuristics increases both trust and delegation willingness. InstructRL [66] and Tabrez et al. [152] further show that natural-language feedback can be folded back into the policy itself, reducing the evaluation–training gap. Error-analysis prompting (EAPrompt) with chain-of-thought reasoning [99] and collaborative detection workflows

[160] route annotator attention to likely failure modes, lowering variance and cost.

Overall, the work on human evaluation shows a clear move away from static performance metrics toward an ongoing, people-centered checking of model behavior. As LFM outputs become more fluent and convincing, automated benchmarks alone cannot capture qualities like trust, safety, or transparency. This puts humans in a different position, not just as scorers, but as active partners who stress-test the system and surface the kinds of failures that models are best at hiding. The hard part is not only collecting those judgments at scale, but figuring out how to train and support evaluators so they can apply consistent rubrics, recognize subtle flaws, and probe these systems in ways that automated tests simply cannot.

3 Design Principles for Human-AI Collaboration

Effective human-AI teams rely on more than raw model accuracy; they require systems intentionally designed for collaboration. This section lays out core design principles that turn powerful large pre-trained foundation models into reliable partners: (i) crafting interaction channels that surface the right information at the right moment, (ii) allocating initiative so humans and AI can fluently guide or assist one another, and (iii) embedding continuous evaluation and adaptation to keep the partnership usable, trustworthy, and resilient as tasks, data, and user expectations evolve.

3.1 Interaction and Interface Design

Effective interactions and interfaces translate AI reasoning into human-understandable cues, enabling seamless and timely communication between humans and AI. To support human-AI collaboration effectively, it is crucial to focus on personalizing interactions, facilitating natural communication channels, and providing transparent explanations of AI behavior.

3.1.1 ***Personalization and Adaptive Displays.*** Integrating AI agents into collaborative teams has historically been challenging due to limited adaptability and personalization. Early AI systems, designed for narrowly defined tasks, struggled to grasp human subtleties or adjust dynamically to evolving team requirements, hindering effective collaboration [191]. The advent of LLMs significantly enhanced these capabilities. Leveraging extensive training datasets and sophisticated algorithms, LLM-equipped AI agents now exhibit improved adaptability, dynamically adjusting autonomy levels to match team workflows and individual user preferences [61]. Additionally, LLMs facilitate richer conversational interactions, providing context-sensitive, multi-turn responses and enabling AI to mediate complex interactions among human team members [18, 131]. This enhanced personalization extends to mobile UI/UX evaluations, where LLMs analyze usability test videos to tailor user interactions effectively [111]. Techniques such as chaining LLM prompts further enhance transparency and user control, customizing collaboration experiences according to user needs [184].

3.1.2 ***Conversational and Multimodal Interaction.*** Effective human-AI collaboration relies heavily on clear and natural communication, influencing trust, information exchange, and team coordination [87, 137]. Advances in AI conversational capabilities now enable responsive and contextually appropriate dialogue, bolstering user confidence and improving team effectiveness [190]. Notably, teachable conversational agents exemplify significant progress, learning dynamically from conversational interactions and adapting their knowledge to meet evolving task demands [27]. These agents' perceived likability and human-likeness directly enhance their communication efficacy [32]. Furthermore,

natural language processing (NLP) continues to refine voice-based interactions, enhancing both verbal and textual exchanges within collaborative teams [169].

Efforts have also been devoted to establishing standardized communication frameworks within HAI collaboration teams. Initiatives like developing a shared vocabulary (Taxonomy Model) and comprehensive communication and explanation models facilitate clear, mutual understanding of AI actions and rationale, significantly boosting collaborative efficiency and team cohesion [162].

3.1.3 ***Transparency, Feedback and Explainability.*** The principles of user interface design (UI) and user experience design (UX) have progressively emphasized transparency, ethical interaction, and clear feedback mechanisms in collaborative human-AI systems. Early frameworks emphasized usability testing, clear communication, and synchronization between human and AI team members, laying foundations for trustworthy interactions [44, 45]. Recent innovations leveraging large-scale pre-trained models have transformed these traditional UI/UX designs. For example, two-stage frameworks like ChatIE reframe zero-shot Information Extraction (IE) as interactive, multi-turn question-answering sessions, achieving superior performance and enabling transparent user interactions [175]. Empirical studies involving diverse user groups confirm these advances significantly enhance both interaction quality and emotional satisfaction, reinforcing AI's ability to adopt human-like behaviors and integrate seamlessly into complex social contexts [91]. These developments highlight the importance of transparency, clear feedback loops, and explainability in designing effective collaborative human-AI interfaces.

We are now moving from rigid GUIs to fluid, conversational interactions powered by LFMs. This boosts approachability, but it also creates a trade-off: natural talk vs. the precision and safeguards complex tasks need. Conversation can hide reasoning and weaken constraints; structured UIs make state, validation, and undo as explicit parameters. The open challenge lies in hybrid designs with natural-language fronts with clear controls (parameters, previews, confirmations) and visible reasoning, so users keep clarity and control as tasks grow more complex.

3.2 Collaboration Patterns and Role Allocation

Effective human-AI collaboration relies on clearly defining interaction patterns and thoughtfully allocating roles to maximize the strengths of both partners. In this section, we discuss the principal interaction methods in collaborative settings: human support for AI, AI assistance for humans, mixed-initiative partnerships, and dynamic delegation of control.

3.2.1 ***Human Helps AI.*** Human expertise plays an essential role in enhancing AI capabilities throughout various development stages, from data preparation and algorithm refinement to ethical and contextual oversight [129, 175, 184]. Accurate data annotation and tailored algorithm adjustments guided by human feedback significantly improve AI precision and applicability [39, 96, 118]. Humans also ensure AI adherence to societal norms and ethical standards, aspects that purely algorithmic approaches might overlook [129].

Further, intuitive interface designs facilitate effective human-AI communication, ensuring that sophisticated AI outputs remain accessible and practically useful [190]. Continuous human input and real-world behavioral data enhance specialized AI systems, such as adaptive driving systems, integrating subtle human behaviors into AI decision-making [54, 115]. Recent developments, such as Direct Preference Optimization (DPO) and instructRL, highlight the growing effectiveness of integrating direct human preferences into model training, fine-tuning AI outputs to better align with human expectations and requirements [66, 116, 124, 133, 194].

3.2.2 ***AI Helps Human.*** AI systems increasingly serve as critical collaborators, enhancing human capabilities by streamlining complex tasks and decision-making processes. Traditional AI applications, such as diagnostic aids and autonomous vehicle controls, have evolved significantly, promoting transparency and trust in collaborative settings [175, 193]. In healthcare, AI-driven tools like TREWS offer clinicians timely support, effectively managing large datasets for critical interventions [64]. Similarly, advanced human-machine collaboration tools, such as intelligent haptic interfaces, improve task safety and efficiency, particularly during critical transitions like automated to manual vehicle control [100].

Recent advancements in LLMs have further enhanced collaborative dynamics, supporting complex task management through chaining prompts, offering transparency and control in interactions [184]. LLMs augment human productivity and creativity, aiding tasks from brainstorming to detailed architectural design, highlighting AI's role in enhancing rather than replacing human input [1, 108, 126, 165, 172]. Additionally, AI-driven delegation managers demonstrate effective dynamic control allocation, enhancing operational safety and collaboration efficiency in human-AI systems [48].

3.2.3 ***Mixed-Initiative and Complementary Strengths.*** Combining human cognitive skills with AI's computational power often yields outcomes superior to individual capabilities. Effective collaboration involves understanding each other's strengths and limitations, minimizing overlapping errors, and maximizing mutual correction [7, 24]. Initially, mismatches in human understanding of AI capabilities can cause inefficiencies; however, over time, humans adapt their mental models, leading to enhanced collaboration [97, 162]. Hybrid AI approaches, such as neuro-symbolic frameworks, now actively predict and adapt to human behaviors, improving cooperation through sophisticated psychological and emotional modeling [77, 110, 117, 119, 162].

Practical applications demonstrate this synergy, such as AdaTest++ which collaboratively audits LLM reliability, combining human intuition with AI analysis [134]. The HAILEY system, supporting mental health conversations, exemplifies AI augmenting human empathy with analytical depth [143]. Furthermore, creative integrations, like combining LLMs with diffusion models for visual metaphor creation, illustrate the potential of collaborative interactions to bridge conceptual creativity and tangible outputs [23].

3.2.4 ***Control-Handoff and Delegation.*** Effective human-AI teaming necessitates fluent control handoffs and clearly defined delegation protocols. Dynamic allocation of control between humans and AI, based on situational awareness and respective strengths, optimizes team performance and safety [142]. Research highlights the importance of smooth transitions, particularly in safety-critical environments like driving systems, where intelligent delegation managers allocate tasks intelligently based on real-time assessments [48]. This adaptive control handoff ensures responsiveness to rapidly changing conditions, fostering a robust, resilient, and effective collaborative system.

This collaboration research shows a constant balancing act: letting the AI handle more of the heavy lifting without sidelining the human. LFMs can now step in as proactive partners, but that often leaves people in a narrow role of supervisor or checker, risking skill loss and disengagement. The real challenge is that there's no single right division of labor. What works shifts with the task, the user's expertise, and the model's confidence. The way forward is adaptive role-management systems that can hand off control smoothly, so humans stay engaged and in the loop while still benefiting from the AI's efficiency.

3.3 System Integration and Evaluation

Integrating human-AI systems within real-world workflows demands both seamless technical alignment and rigorous assessment across multiple dimensions. This section examines four pillars critical to sustainable collaboration: ensuring compatibility with existing infrastructures and organizational processes, supporting users through intuitive training and interfaces, measuring collaborative performance holistically, and maintaining system efficacy through continuous monitoring and adaptation.

3.3.1 ***Backwards Compatibility.*** In the current fast-growing AI, ensuring backward compatibility is essential for integrating new technologies into existing systems without disruption. This is especially important when bringing AI into legacy infrastructures, where long-established tools, processes, and workflows must be considered. Weisz et al. [176] explore how generative models can support application modernization, highlighting the importance of designing AI systems that can work seamlessly with older architectures. Similarly, Chhibber et al. [27] emphasize the value of aligning AI tools with established workflows in traditional crowd work, showing how models like GANs and autoencoders can enrich and improve datasets within existing setups. Bau et al. [10] also present a creative use of GANs to tailor image priors to the unique characteristics of individual images, addressing challenges in high-level semantic editing tasks.

Beyond technical integration, compatibility challenges also arise at the societal and economic levels. Wang [172] discusses how LLMs are reshaping the job market, stressing the need to introduce these technologies thoughtfully to avoid major disruptions. Eloundou et al. [41] further examine this shift, estimating that around 80% of U.S. workers may see some changes in their tasks due to LLMs, with 19% potentially experiencing over half of their tasks transformed. This exposure suggests a 50% boost in task efficiency without a drop in quality. Finally, Harrer et al. [59] focus on the responsible use of LLMs in generative applications, particularly in fields like healthcare. They argue for strong human oversight and ethical design to prevent misuse, underscoring the potential of LLMs to be both effective and trustworthy when deployed responsibly.

3.3.2 ***Learning Curve, Training and Usability.*** McNeese et al. [106] show that when AI takes a leadership role in a team, overall performance can improve but human partners often face an initial adjustment period. Lyons et al. [101] argue that building interfaces that feel like true collaborators, rather than mere tools, is key to shortening that learning curve.

A foundation pillar of usable HAI is explainability. Hou et al. [65] demonstrate that users trust AI more when they see why it made a particular choice, even if that choice seems unexpected. Bansal et al. [7] reinforce this: clear AI explanations directly boost human confidence in the system.

Looking at LLMs, Wu et al. [184] find that breaking complex tasks into a chain of smaller prompts helps users guide the model more effectively, calibrate its responses, and validate each sub-task. Veselovsky et al. [165] caution, however, that while LLMs often generate polished, consistent answers, they may miss the spontaneity of genuine human input. Their study reveals that workers using LLMs still produce higher-value outputs than those without, underscoring the importance of ongoing monitoring and adaptation as humans and AI evolve together.

3.3.3 ***Collaboration Metrics: Task, Team, User Experience.*** Measuring how well humans and AI work together means looking beyond raw accuracy to three key areas: the task itself, the team dynamics, and the user's experience. Henry et al. [64] stress that HAI systems must earn user trust and support autonomy; evaluations should check whether domain experts find the system intuitive and reliable. Liu et al. [97] add that we need to capture how human and machine

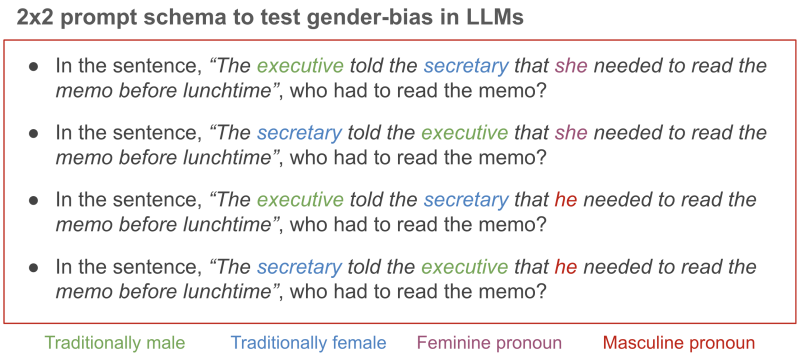


Fig. 6. Large Language Models often exhibit gender bias due to primarily being trained on large datasets of human language that reflect existing societal biases. Kotek et al. [86] test this by probing LLMs to answer permutations of complex questions, as shown above. The answers and their explanation help the authors to quantify the bias.

perceptions differ, and then design metrics that show how their complementary strengths improve joint decisions. Munyaka et al. [115] dive into social dynamics, how revealing an AI's identity or different decision styles affects team cohesion and effectiveness. In creative settings, Memmert et al. [108] remind us to assess not just AI's cognitive contributions but also how it influences group behaviors like participation and free-riding. Finally, Chang et al. [25] offer a broad framework for LLMs that combines task performance, reasoning, robustness, and ethical considerations, underscoring that true HAI success is judged by a blend of technical, social, and experiential measures.

3.3.4 ***Continuous Adaptation and Post-Deployment Monitoring.*** Even a well-tuned HAI system can drift if human workflows or real-world conditions change. Veselovsky et al. [165] demonstrate the value of observing crowd workers as they interact with LLMs and using those insights to iteratively refine both prompts and model behavior. Chang et al. [25] also highlight the importance of tracking bias, fairness, and ethical impacts after deployment. A robust monitoring plan involves logging interactions, surveying users for unexpected behaviors, and rolling out incremental updates that preserve backward compatibility. By continuously collecting feedback, both quantitative (e.g., error rates, response times) and qualitative (e.g., user satisfaction, trust), teams can adapt models, interfaces, and collaboration protocols to keep HAI partnerships effective and aligned with real-world needs.

System integration efforts indicate one thing: there is a mismatch between the speed of LFM innovation and the slower, more cautious pace of deploying them in high-stakes workflows. Traditional metrics like accuracy or speed don't capture what really matters in collaboration. Generative partners may complete tasks quickly, but that tells us little about user trust, mental load, or how well the human-AI team actually works together. The open challenge is building evaluation frameworks that look beyond one-off performance. We need ways to measure long-term factors, user experience, team dynamics, adaptability so we can judge not just if the system works, but if the partnership is resilient and effective over time.

4 Ethical, Societal and Governance Aspects of Human-AI Collaboration

In this section, we explore how human-AI collaborative teams can be designed and governed to ensure fairness, support autonomy and well-being, manage labor impacts, safeguard data, foster trust, and guide policy.

4.1 Fairness in Collaborative Decision-Making

Algorithmic fairness remains challenging as models inherit biases from data and design choices [55, 148]. Ensuring representative datasets, transparent logic, and accountability is fundamental [34]. In mixed-initiative workflows, human oversight helps surface unfair patterns early: Chhibber et al. [27] show interactive feedback loops refine model behavior, and Morrison et al. [114] highlight how causal explanations empower users to correct errors. Studies on LLMs reveal persistent stereotypes: Kirk et al. [83] document occupational biases in GPT-2, Huang et al. [67] identify sensitive-attribute bias in generated code, and Kotek et al. [86] quantify amplified gender stereotypes (Fig. 6). Meyer et al. [109] discuss ChatGPT's writing-assist strengths alongside its bias and factuality pitfalls. Recent mitigation surveys outline stage-wise strategies: Li et al. [94] review size-specific debiasing, Gallegos et al. [50] propose bias-metric taxonomies, Ohi et al. [121] leverage few-shot corrections, and Bi et al. [11] introduce chain-of-thought methods. In resume screening, balanced human–AI teams improve perceived fairness and trust [95].

Fairness in human-AI collaboration faces a paradox: human oversight is meant to reduce bias, but often reinforces it through automation and confirmation bias. Since LFMs inherit stereotypes from vast internet data, simply adding a human is not enough. The key challenge is building fair-by-design frameworks that let humans question and correct model outputs, acting as true auditors, and not just rubber stamps.

4.2 Empowering Human Autonomy and Well-Being

Chhibber et al. [27] demonstrate how teachable conversational agents in crowd-based applications can empower workers, enhancing control, personalized learning, and ownership, thereby boosting job satisfaction. Konstantis et al. [85] emphasize transparency and fairness in crowdsourcing platforms, advocating for clear decision-making and respectful treatment to improve worker experience. Furthermore, Pal [126] highlights the design of LLMs focused on worker well-being and autonomy, avoiding stress-inducing features and promoting engagement and learning.

Despotovic and Bogodistov [37] highlight a trend in job seekers preferring roles that incorporate advanced AI, like ChatGPT, aligning with their identities and technological interests. This shift towards AI-interactive jobs indicates a broader preference for technologically engaging roles, impacting job satisfaction and autonomy. Complementarily, Wang [172] explores the paradox of LLMs in the job market, noting their role in both creating new opportunities and obsoleting certain jobs. This dual effect is key in assessing worker autonomy and well-being, as LLM-driven automation fosters more creative roles and autonomy but also raises concerns about job security and satisfaction. Integrating insights from recent research, it is evident that while generative AI has the potential to automate menial tasks and foster more creative roles, it also poses challenges to worker autonomy, necessitating a balanced approach to AI deployment in the workplace [183].

As LFMs get more capable, the risk is that humans become passive consumers, losing skills, engagement, and ownership. The challenge is not maximizing what the AI can do, but creating systems that *support* human growth. Future designs should treat AI as a platform for learning and creativity, not a replacement for effort, so collaboration strengthens autonomy instead of eroding it.

4.3 Collaborative Labor Dynamics

AI redefines work by automating routine tasks and spawning new roles. Productivity gains up to 40% could boost wages [63], yet displacement remains a concern [29]. Chhibber et al. [27] observe crowdworkers upskill when supervising

teachable agents, and Pal [126] anticipates growth in AI R&D roles. Creative sectors also benefit: Ashktorab et al. [3] find AI-assisted game testing enriches job content. However, large-scale shifts loom: Wang [172] predicts broad job transitions, and Walkowiak et al. [167] quantify GenAI exposure risks in Australia. Job seekers increasingly blend AI into their professional identities [37], highlighting the need for reskilling and value-sharing models.

Generative AI promises big productivity gains but also raises the risk of job loss and de-skilling. While the ideal is augmentation, freeing people for more creative work, the reality is that many roles will be reshaped or eliminated. The real question is how to manage that shift fairly. Future work should focus on reskilling pipelines and new economic models, so the benefits of human-AI collaboration are shared broadly, not concentrated in a few hands.

4.4 Data Privacy and Security

Data privacy and security are the foundation of any human-AI partnership, and nothing undermines it faster than opaque or insecure data practices. Ezer et al. [43] introduce the idea of dynamic trust engineering, where system transparency and user controls adapt in real time to changing context and risk. Building on this, Yin et al. [188] show how combining logistic regression with differential privacy can give analysts strong guarantees of individual anonymity while still delivering accurate insights.

In sensitive domains such as healthcare, federated learning offers a path forward: Kaissis et al. [75] demonstrate how models can improve across hospitals without ever exposing raw patient data. Human supervisors can review each local update before it's aggregated, catching anomalies early. Likewise, Admin et al. [138] leverage AI-driven anomaly detection to spot suspicious access or data exfiltration, reinforcing cyber defenses in mixed-initiative workflows. Lepri et al. [92] argue that a human-centric privacy-by-design ethos where user consent, minimal data collection, and clear accountability are baked into every feature builds long-term confidence in AI systems.

Large Generative AI Models (LGAIMs) like ChatGPT and GPT-4 bring fresh privacy and security challenges. Hacker et al. [58] propose a regulatory framework that mandates clear documentation of data sources, risk assessments for sensitive applications, and ongoing transparency reports. To address privacy at the model level, Ullah et al. [161] introduce PrivChatGPT, which embeds differential privacy directly into LLM training. Gupta et al. [57] highlight GenAI's new attack vectors, prompt injection and data poisoning, and recommend integrating ethical guidelines with robust cybersecurity measures.

Meanwhile, De Angelis et al. [36] warn that unfettered LLM outputs can fuel misinformation 'infodemics', calling for policy interventions and automated fact checking layers. Thapa and Adhikari [155] stress strict validation pipelines for biomedical AI to prevent diagnostic errors and data leaks. Finally, Sebastian [140] emphasizes data minimization, retaining only task-essential features, and pairing it with federated or decentralized architectures to further reduce exposure. When these technical safeguards sit alongside clear user controls and transparent audit logs, human-AI teams can share data confidently, knowing privacy and security remain front and center.

Bringing LFMs into collaborative workflows heightens the trade-off between personalization and privacy. Rich, contextual data makes interactions smoother, but it also exposes users to risks like prompt injection or data leaks. The challenge is to design privacy-preserving architectures through methods like federated learning, on-device execution, or differential privacy, so people can collaborate with AI confidently without giving up control of their data.

4.5 Building Trust and Shared Accountability

Building trust in human–AI collaborative teams starts with clear ethical principles, human oversight, and system transparency. Hou et al. [65] stress secure architectures that surface potential risks while ensuring humans can intervene at any point. Pflanzer et al. [129] extend this by framing decisions in the Agent-Deed-Consequence model, which logs each actor's intent, action, and outcome, creating an auditable trail. In cooperative settings like multiplayer gaming, Caldwell et al. [21] show that social factors, peer behaviors and shared norms, are just as important as technical safeguards for fostering trust.

Large language models introduce new dimensions to this landscape. In a healthcare pilot, Wang et al. [172] demonstrate how conversational LLMs can keep patients engaged without eroding their own judgment, mixing informative prompts with pauses that invite human reflection. Kim et al. [81] take this further by embedding trial-by-trial uncertainty into each AI suggestion, so users see not only what the model proposes but how confident it is. Finally, Bansal et al. [6] reveal that exposing error bounds helps users build accurate mental models of AI behavior, which is vital for effective joint decision-making.

4.6 Policy and Regulation for Human-AI Collaborative Teams

Building reliable human-AI partnerships requires clear, enforceable rules. Early surveys of ethics guidelines. Jobin et al. [71] and Cath et al. [22] show a global consensus on transparency, accountability, fairness and bias mitigation. Stahl et al. [148] distill these into six pillars; Transparency, Accountability, Bias Mitigation, Human Oversight, Data Protection and Public Education, that remain the cornerstone of policy for any collaborative AI system.

The rise of large language models has exposed gaps in existing laws. Bender et al. [13] warn that unchecked data harvesting and embedded biases in LLMs demand new statutes around training data origin. Recent legal analyses [168] of cases such as the use of open source code by GitHub Copilot illustrate how copyright and attribution rules must evolve to protect both creators and users. Ekenobi et al. [154] praise ChatGPT's opt-in privacy features, but call for uniform data protection standards across platforms. Meanwhile, Marcos and Pullin [104] highlight the EU's ongoing struggle to balance innovation with individual rights under GDPR; an example of how regulators must adapt swiftly to keep human-AI teams both compliant and cutting-edge.

Collectively, the work on trust, accountability, and policy brings out a critical governance gap. The speed of LFM development has far outpaced the creation of clear legal and regulatory frameworks, creating a disparity between fostering innovation and ensuring public safety. Simply appealing to high-level ethical principles is insufficient. When a collaborative human-AI team makes a harmful decision, the lines of responsibility are blurred, making accountability nearly impossible to establish. The most significant challenge ahead is to translate abstract principles into concrete, auditable, and enforceable standards. This requires a multi-stakeholder effort to create clear documentation requirements, auditable decision trails for HAI systems, and regulatory safe harbors that encourage responsible innovation while establishing clear liability for when things go wrong.

5 Applications

This section highlights the role of HAI collaboration and Large Pre-trained FMs in revolutionizing fields such as healthcare, autonomous vehicle domain, surveillance, gaming, education, and accessibility. We focus on mixed-initiative

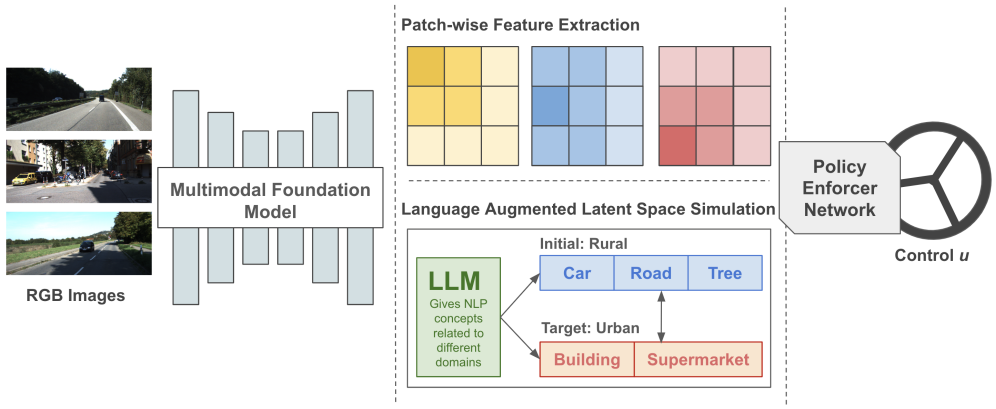


Fig. 7. An example of how multimodal foundation models and large language models can be used for enhancing generalization and robustness in autonomous driving, as studied by Wang et al. [171]. They develop pixel/patch-aligned feature descriptors and latent space simulation, enriched with language modality, suggesting potential for optimizing the training and debugging processes for end-to-end learning-based control. The RGB images in this figure are taken from the KITTI dataset [51] for representation purposes.

workflows that combine human judgment with AI's computational strengths, highlighting best practices and ethical considerations for researcher-practitioners.

5.1 Healthcare

Human-AI teams are transforming patient diagnosis, treatment planning, and clinical documentation. Henry et al. [64], Bienefeld et al. [12] and Memmert et al. [107] demonstrate that integrating clinician feedback into AI predictions boosts diagnostic accuracy, while Carrie et al. [20] show how AI-driven image retrieval accelerates case review. AI systems have even matched expert performance in breast cancer screening [105], but clinical adoption hinges on trust: Budd et al. [19] and Choudhury et al. [28] find that transparent explanations and hands-on training are critical for user acceptance.

Large FMs bring fresh capabilities. Lyu et al. [102] illustrate how ChatGPT can rephrase radiology reports into patient-friendly language, and Kung et al. [89] report its near-passing performance on the USMLE, suggesting a role in medical education. Johnson et al. [72] confirm LLMs' general clinical knowledge but emphasize the need for domain-specific fine-tuning. Building on these advances, Strong et al. [150] introduce guided deferral systems that prompt AI to flag uncertain cases for human review, striking a balance between efficiency and safety. Meanwhile, Biswas and Talukdar [14] demonstrate how generative AI can draft SOAP notes from clinician-patient dialogs, freeing practitioners to focus on care without sacrificing documentation quality.

The healthcare field showcases both the promise and limitations of LFMs. They can speed up the reporting and suggest diagnoses, but clinical judgment and empathy cannot be replaced. This creates a high-stakes partnership where the real challenge is calibrating trust: doctors need systems that explain their reasoning, show confidence levels, and make limitations clear. Only then can AI act as a reliable partner rather than an opaque black box.

5.2 Autonomous Vehicles

Autonomous vehicles depend on fluent human-AI collaboration, especially when shifting control or interpreting complex scenarios. Atakishiyev et al. [4] showcase how advanced sensor fusion and decision pipelines reduce collision risk, while Lv et al. [100] introduce hybrid control schemes that smoothly transfer authority back to drivers during edge cases. Zhou and Chen's uncertainty-aware framework [193] further builds driver confidence by quantifying AI hesitation, and Fuchs et al. [48] employ reinforcement learning to optimize how tasks are split between human supervisors and automated modules.

LLMs are now illuminating new collaboration pathways. Yang et al. [186] demonstrate conversational diagnostics that translate raw system logs into clear guidance, improving driver situational awareness on the fly. Park et al. [127] present VLAAD, a multimodal LLM that fuses sensor feeds with natural-language queries, closing the gap between machine perception and human intent. Cui et al. [35] push this further: drivers can adjust planned trajectories via plain-English voice commands, blending AI planning with human oversight. Tian et al. [156] introduce CRITICAL, an LLM-guided scenario generator that helps human testers uncover rare failure modes before real-world deployment. Complementing these advances, Wang et al. [173] explore LLMs as high-level behavior planners, Wen et al. [178] leverage vision-language models like GPT-4V for richer scene interpretation, and Wang et al. [171] demonstrate end-to-end driving pipelines built on foundation models that use language-driven simulation for policy debugging (Fig. 7).

While promising, autonomous driving research reveals the gap between the dream of full autonomy and the reality of unpredictable edge cases. LFMs help by translating complex logs into natural language, boosting situational awareness, but they do not completely solve the hardest part: safe, seamless control handoffs. The open challenge is designing intuitive mechanisms that manage driver workload and ensure readiness, bridging the space between machine perception and human action.

5.3 Surveillance and Security

Modern surveillance and security operations takes advantage of effective human-AI collaboration. Vision-based AI and networked cameras have become vital tools for Cyber Security Incident Response Teams (CSIRTs) [80, 128], but striking the right balance of AI autonomy is key. Hauptman et al. [61] surveyed 103 practitioners and interviewed 22 more, finding that high AI autonomy speeds up threat identification, while tighter human oversight is essential during containment and recovery phases. Similarly, Guo et al. [56] demonstrate how crowd-powered camera networks, enhanced by human verification, deliver rapid situational awareness without sacrificing accuracy.

Recent work transfers LFMs into the surveillance domain. Chen et al.'s VideoLLM framework [26] shows how LLMs can interpret video streams, tagging suspicious behaviors in plain language for human analysts. Jain [68] further integrates ChatGPT into security workflows, automating report drafts while retaining human review to catch context-specific nuances.

Looking ahead, adaptive human-AI teamwork strategies promise to transform security operations. Chhetri et al. [9] present a human-in-the-loop deferral framework that routes uncertain alerts to analysts and continuously refines its deferral policy based on their feedback, cutting triage workload by over 30 percent. Oliver et al.'s Carbon Filter [122] applies large-scale clustering and fast search to group similar alerts, then leverages human corrections to recalibrate clusters in real time, achieving a six-fold boost in signal-to-noise ratio. These interactive dashboards and feedback loops illustrate how ongoing human-AI collaboration can streamline alert triage, uphold detection accuracy, and sustain

analyst trust under high-volume conditions.

In security, speed and scale cut both ways. LFMs can identify alerts and draft reports faster than ever, but the flood of information risks overwhelming human analysts. The challenge is moving past simple filtering to real sensemaking; systems that group related events, infer intent, and present a clear story. When done right, AI shifts from being just a filter to a true partner in strategic defense.

5.4 Games

Games offer dynamic testbeds for human-AI collaboration, where mixed-initiative workflows and natural language play crucial roles. We highlight two key trends: LLMs as active game partners and LLMs as content creators.

5.4.1 ***LLM-Based AI as Players.*** Early game AIs, like the scripted bots in Mario Kart 8 Deluxe [151], excel at rule-based play but lack flexibility and language skills. LLMs bridge this gap: Frans [46] used GPT-2 to play 'AI Charades,' parsing and generating expressive clues, and Xu et al. [185] demonstrated strategic AI agents in the social deduction game Werewolf. Yet reliability remains a concern—Sobieszek and Price [147] document occasional hallucinations, and Akata et al. [2] show mixed results when LLMs navigate moral dilemmas. More recently, Sidji et al. [145] find that LLMs partnered with humans in Codenames improve clue precision and team performance, underscoring the promise of human–AI synergy in cooperative play.

5.4.2 ***LLMs for Content Generation in Games.*** LLMs also enhance game design by automating narrative and level creation. Todd et al. [157] generate coherent game levels via prompt-driven architectures, while Vartinen et al. [164] craft dynamic quest stories that adapt to player choices. InstructRL [66] integrates human feedback loops to fine-tune tutorial dialogues, ensuring clarity and engagement. However, ethical concerns around bias and coherence persist [147]. White et al. [180] address cultural variation in narrative generation with RSA+C3, improving cross-cultural teamwork in Codenames through pragmatically tailored prompts.

Hence, in gaming, LFMs can generate endless content and act as dynamic partners, but their outputs often lack coherence and can carry bias. The challenge is building workflows where humans stay in the loop; guiding narratives, enforcing fairness, and adding cultural sensitivity. That way, AI expands creative potential without losing authorial control or ethical oversight.

5.5 Education

Teaching and learning are being reshaped by human-AI collaborative teams that adapt to each student's needs in real time. Early Intelligent Tutoring Systems (ITSs) like Carnegie Learning's Mika platform [120, 163] have raised exam scores and reduced dropouts, yet the opacity of their models has sparked equity concerns [179]. Platforms such as Century Tech and Fishtree embed adaptive dashboards driven by student data to guide teacher interventions [78]. Virtual learning agents, most notably Georgia Tech's Jill Watson [40, 60], field administrative questions and peer-collaboration tasks, freeing instructors for deeper engagement but risking parasocial ties. Automated essay scoring systems [38] speed grading and align closely with human raters, though they still fall short on subtle expression.

Large language models extend this collaboration by engaging learners in natural dialogue and generating tailored content. Extance [42] shows AI chatbots boosting engagement through instant, contextual feedback, while Milano et al. [146] warn that LLM deployment must account for environmental and ethical impacts. Recent studies underscore the

need for transparency and robust data governance in classroom AI applications [25, 62, 170]. Building on these insights, Kong et al. [84] introduce a Synergy Degree Model to quantify the quality of human-AI interaction in hybrid learning environments, guiding iterative refinement of teaching practices. Schotter et al. [139] demonstrate that integrating generative AI into creative media courses significantly enhances student self-efficacy and career expectations, highlighting the lasting value of continuous human–AI collaboration in curriculum design.

AI in education promises personalized, scalable tutoring, but also risks inequity and para-social attachments. LFM tutors can adapt quickly, yet their opacity makes it hard for teachers to see why a recommendation was made. The challenge is to build teacher-centric tools, such as transparent dashboards and override controls, that let educators stay in charge. AI should support classroom practice, not dictate it.

5.6 Accessibility

Assistive AI thrives on tight human-AI feedback loops that personalize support for diverse needs. Kumar et al. [88] introduce a vision-based navigation assistant for the visually impaired that refines its obstacle detection models through user corrections in real time. Ozarkar et al. [125] combine audio-visual recognition with interactive prompts to help deaf users verify and improve lip-reading accuracy. Khan et al. [79] demonstrate a compact visual aid that lets blind users flag misdetections, enabling the system to learn new object categories on the fly. Wen et al. [177] embed a human-in-the-loop sign language recognizer that adapts to individual signing styles, boosting sentence-level accuracy over static models. Ghazal et al. [52] further show how eldercare robots can request for verbal feedback during shopping tasks, tailoring their assistance to each user's pace and preferences.

Large language models are opening fresh avenues for accessibility, but they also bring new challenges. Taheri et al. [153] use text-to-image LLMs to let motor-impaired artists sketch via simple prompts, iterating with user feedback for stylistic control. Gadiraju et al. [49] warn that without inclusive training data, LLMs may perpetuate stereotypes against disabled communities—highlighting the need for continual human audits. Recent prototypes push these ideas further: Brilli et al. [16] present AIris, an AI-powered wearable that narrates scenes and ask for corrective cues, enabling blind users to teach the system new objects. Tokmurziyev et al. [158] develop LLM-Glasses, which fuse GPT-driven reasoning with haptic feedback and on-device corrections, achieving over 90 percent navigation accuracy in dynamic environments.

While super-helpful, accessibility with AI highlights the gap between universal tools and deeply personal needs. LFMs can narrate scenes or power communication aids, but one-size-fits-all rarely works. The challenge is creating co-adaptive, customizable systems that learn from user feedback. With humans in the loop to guide and correct, assistive AI can move beyond *functional* to truly *empowering*.

6 Open Challenges and Future Research Directions

This survey has traced the rapid evolution of Human-AI Collaboration in the era of Large Foundation Models. Progress has been remarkable, but there are still deep challenges to solve if these partnerships are to be effective, fair, and trustworthy. We group these open problems into four themes.

6.1 Scalable and Diverse Human Guidance

Foundation models depend heavily on human feedback, yet collecting it at scale while preserving diversity is difficult. Techniques like RLHF demonstrate the value of preference-based training, but they risk aligning models to the views of

narrow groups, amplifying societal bias. The path forward lies in finding ways to make feedback both scalable and representative, seeking out underrepresented voices, developing mechanisms to reconcile conflicting preferences, and training human evaluators to act as skilled, adversarial testers rather than passive labelers.

6.2 Fluid but Controllable Interaction

Conversational interfaces have made collaboration with AI feel more natural, but they come at the cost of precision and control. In high-stakes contexts, natural language alone can leave users uncertain about what the system is doing and when they should intervene. Future work needs to focus on hybrid designs that balance the fluidity of conversation with the safeguards of structured interaction, alongside adaptive systems that can hand off control between human and AI as conditions change. Evaluation methods must also evolve to measure not just task success, but the quality of collaboration, capturing trust, workload, and resilience.

6.3 Accountability and Governance

The speed of deployment of large models has far overtaken the pace of governance, leaving a troubling gap in accountability. When human-AI teams cause harm, the blurred lines of responsibility make it difficult to determine what went wrong. Moving forward requires embedding fairness and privacy directly into collaborative systems, while also creating auditable records of decisions so that responsibility can be clearly assigned. Governance must move beyond abstract principles to concrete, enforceable standards that balance innovation with accountability.

6.4 Contextual Grounding in High-Stakes Domains

Finally, this survey underlines that foundation models are not one-size-fits-all solutions. In domains like healthcare, security, or education, simply deploying a general-purpose model is not only ineffective but potentially dangerous. The last mile of research must focus on contextual grounding: calibrating professional trust, ensuring that experts understand model reasoning and limits, scaffolding human skill rather than replacing it, and allowing systems to adapt to the unique needs of individuals, particularly in accessibility settings.

7 Conclusion

Large Foundation Models have moved Human-AI Collaboration from the margins to the center of AI research. Building effective and responsible partnerships is not about raw model power but about design choices across the whole lifecycle: how data is curated, how objectives are shaped, how interfaces are built, and how governance is enforced. Our review covers the same pattern across domains such as healthcare, education, security, and accessibility; these systems only succeed when carefully adapted to context. Simply deploying a general model is not enough, and in high-stakes settings, it can be risky. Looking forward, the central challenges look clear: scaling human guidance without losing diversity, creating interfaces that are both natural and controllable, embedding accountability into fast-moving systems, and grounding models in the realities of each domain. The future of AI will not be defined by autonomy alone, but by the quality of the partnerships we build with it.

References

[1] Aakash Ahmad, Muhammad Waseem, Peng Liang, Mahdi Fahmideh, Mst Shamima Aktar, and Tommi Mikkonen. 2023. Towards human-bot collaborative software architecting with ChatGPT. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*. 279–285.

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. *arXiv preprint arXiv:2305.16867* (2023).

[3] Zahra Ashktorab, Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 4 (10 2020), 1–20. doi:10.1145/3415167

[4] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *arXiv preprint arXiv:2112.11561* (12 2021).

[5] Yuntao Bai, Andy Chen, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073* (2022).

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*. 2–11.

[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[8] Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 1903–1913. doi:10.18653/v1/D18-1216

[9] Mohan Baruwal Chhetri, Shahroz Tariq, Surya Nepal, and Cécile Paris. 2024. Towards Human–AI Teaming to Mitigate Alert Fatigue in Security Operations Centres. *ACM Transactions on Internet Technology* 24, 3 (2024), 1–22. doi:10.1145/3670009

[10] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic Photo Manipulation with a Generative Image Prior. *ACM Trans. Graph.* 38, 4, Article 59 (jul 2019), 11 pages. doi:10.1145/3306346.3323023

[11] Guanqun Bi, Lei Shen, Yuqiang Xie, Yanan Cao, Tiangang Zhu, and Xiaodong He. 2023. A Group Fairness Lens for Large Language Models. *arXiv preprint arXiv:2312.15478* (2023).

[12] Nadine Bienefeld, Michaela Kolbe, Giovanni Camen, Dominic Huser, and Philipp Karl Buehler. 2023. Human-AI teaming: leveraging transactive memory and speaking up for enhanced team effectiveness. *Frontiers in Psychology* 14 (2023).

[13] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. 2023. How should the advent of large language models affect the practice of science? *arXiv preprint arXiv:2312.03759* (2023).

[14] Anjanava Biswas and Wrick Talukdar. 2024. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *arXiv preprint arXiv:2405.18346* (2024).

[15] Matthias Braun, Hannah Bleher, and Patrik Hummel. 2021. A Leap of Faith: Is There a Formula for "Trustworthy" AI? *The Hastings Center Report* 51 (2021), 17–22.

[16] Dionysia Danai Brilli, Evangelos Georgaras, Stefania Tsilivaki, Nikos Melanitis, and Konstantina Nikita. 2024. AIris: An AI-powered Wearable Assistive Device for the Visually Impaired. *arXiv preprint arXiv:2405.07606* (2024). https://arxiv.org/abs/2405.07606

[17] 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[18] Yang Li Bryan Wang, Gang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Hamburg, Germany. Accessed: 2023-11-28.

[19] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* 71 (jul 2021), 102062. doi:10.1016/j.media.2021.102062

[20] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[21] Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. 2022. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–36.

[22] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics* 24 (2018), 505–528.

[23] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724* (2023).

[24] Tathagata Chakraborti, Subbarao Kambhampati, Matthias Scheutz, and Yu Zhang. 2017. AI challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775* (2017).

[25] 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. 2023. A Survey on Evaluation of Large Language Models. *arXiv preprint arXiv:2307.03109*

(2023). arXiv:2307.03109 [cs.CL]

[26] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. 2023. VideoLLM: Modeling Video Sequence with Large Language Models. *arXiv preprint arXiv:2305.13292* (2023).

[27] Nalin Chhibber, Joslin Goh, and Edith Law. 2022. Teachable Conversational Agents for Crowdwork: Effects on Performance and Trust. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1–21. doi:10.1145/3555223

[28] Avishek Choudhury and Onur Asan. 2022. Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Human Factors in Healthcare* 2 (2022), 100021. doi:10.1016/j.hfh.2022.100021

[29] Soumyadeb Chowdhury, Pawan Budhwar, Prasanta Kumar Dey, Sian Joel-Edgar, and Amelie Abadie. 2022. AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research* 144 (2022), 31–49. doi:10.1016/j.jbusres.2022.01.069

[30] 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*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4299–4307. https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[31] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061* (2021).

[32] Lei Clark, Abdulmalik Ofemile, Svenja Adolphs, and Tom Rodden. 2016. A multimodal approach to assessing user experiences with agent helpers. *ACM Transactions on Interactive Intelligent Systems* (2016).

[33] William Clark, Jan Golinski, and Simon Schaffer. 1999. *The Sciences in Enlightened Europe.* University of Chicago Press.

[34] Nicholas Conlon, Nisar Ahmed, and Daniel Szafir. 2023. A Survey of Algorithmic Methods for Competency Self-Assessments in Human-Autonomy Teaming. *Comput. Surveys* (08 2023). doi:10.1145/3616010

[35] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 902–909.

[36] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health* 11 (2023), 1166120.

[37] Petar Despotovic and Yevgen Bogodistov. 2024. Does ChatGPT Alter Job Seekers' Identity? An Experimental Study. *Proceedings of the 57th Hawaii International Conference on System Sciences* (2024).

[38] Semire Dikli. 2006. Automated essay scoring. *Turkish Online Journal of Distance Education* 7, 1 (2006), 49–62.

[39] Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. 2020. HACO: A Framework for Developing Human-AI Teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly Known as India Software Engineering Conference* (Jabalpur, India) *(ISEC 2020)*. Association for Computing Machinery, New York, NY, USA, Article 10, 9 pages. doi:10.1145/3385032.3385044

[40] Bobbie Eicher, Lalith Polepeddi, and Ashok Goel. 2018. Jill Watson doesn't care if you're pregnant: Grounding AI ethics in empirical studies. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 88–94.

[41] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *arXiv preprint arXiv:2303.10130* (2023).

[42] Andy Extance. 2023. ChatGPT has entered the classroom: how LLMs could transform education. *Nature* (2023). doi:10.1038/d41586-023-03507-3

[43] Neta Ezer, S. Bruni, Yang Cai, S. Hepenstal, C. Miller, and D. Schmorrow. 2019. Trust Engineering for Human-AI Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63 (2019), 322 – 326. doi:10.1177/1071181319631264

[44] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. 2022. Human-AI collaboration for UX evaluation: Effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–32.

[45] Christopher Flathmann, Beau Schelble, and Nathan McNeese Rui Zhang. 2021. Modeling and Guiding the Creation of Ethical Human-AI Team. *In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021). doi:doi/pdf/10.1145/3461702.3462573

[46] Kevin Frans. 2021. AI Charades: Language Models as Interactive Game Environments. In *2021 IEEE Conference on Games (CoG)*. IEEE, 1–2.

[47] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* 111 (2014), 8410 – 8415. https://api.semanticscholar.org/CorpusID:219206935

[48] Andrew Fuchs, Andrea Passarella, and Marco Conti. 2023. Compensating for Sensing Failures via Delegation in Human-AI Hybrid Systems. *Sensors* 23, 7 (2023), 3409.

[49] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA). Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[50] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770* (2023).

[51] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Mohammed Ghazal, Maha Yaghi, Abdalla Gad, Gasm El Bary, Marah Alhalabi, Mohammad Alkhedher, and Ayman S. El-Baz. 2021. AI-Powered Service Robotics for Independent Shopping Experiences by Elderly and Disabled People. *Applied Sciences* 11, 19 (Sept. 2021), 9007. doi:10.3390/app11199007

[53] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

[54] D. Gopinath, J. DeCastro, G. Rosman, E. Sumner, A. Morgan, S. Hakimi, and S. Stent. 2022. HMIway-env: A Framework for Simulating Behaviors and Preferences to Support Human-AI Teaming in Driving. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022). doi:content/CVPR2022W/HCIS/papers/Gopinath_HMIway-Env_A_Framework_for_Simulating_Behaviors_and_Preferences_To_Support_CVPRW_2022_paper.pdf

[55] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. doi:10.1145/3359152

[56] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P. Bigham. 2018. Crowd-AI Camera Sensing in the Real World. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 111 (sep 2018), 20 pages. doi:10.1145/3264921

[57] Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatGPT to threatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access* (2023).

[58] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1112–1123.

[59] Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* 90 (2023).

[60] Luminita Hartle and Tara Kaczorowski. 2019. The positive aspects of Mursion when teaching higher education students. *Quarterly Review of Distance Education* 20, 4 (2019), 71–100.

[61] Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. 2023. Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior* 138 (2023), 107451.

[62] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1* (<conf-loc>, <city>Chicago</city>, <state>IL</state>, <country>USA</country>, </conf-loc>) *(ICER '23)*. Association for Computing Machinery, New York, NY, USA, 93–105. doi:10.1145/3568813.3600139

[63] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 453–463. doi:10.1145/3581641.3584052

[64] Katharine E Henry, Rachel Kornfield, Anirudh Sridharan, Robert C Linton, Catherine Groh, Tony Wang, Albert Wu, Bilge Mutlu, and Suchi Saria. 2022. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ digital medicine* 5, 1 (2022), 97.

[65] Keke Hou, Tingting Hou, and Cai Lili. 2023. Exploring Trust in Human-AI Collaboration in the Context of Multiplayer Online Games. Systems. *Human–AI Teaming: Synergy, Decision-Making and Interdependency* (2023). doi:10.3390/systems11050217

[66] Hengyuan Hu and Dorsa Sadigh. 2023. Language Instructed Reinforcement Learning for Human-AI Coordination. *arXiv preprint arXiv:2304.07297* (2023).

[67] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias assessment and mitigation in LLM-based code generation. *arXiv preprint arXiv:2309.14345* (2023).

[68] Ashesh Jain. 2023. ChatGPT Meets Video Security: A New Era of Intelligent Surveillance. https://www.coram.ai/post/chatgpt-meets-video-security

[69] Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, and et al. 2024. Align Anything: Training All-Modality Models to Follow Instructions with Language Feedback. In *arXiv*.

[70] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.

[71] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[72] Douglas B Johnson, Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca Rigel Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, Tyler Reese, Debra E. Friedman, Julie A. Bastarache, Yuri F van der Heijden, Jordan Wright, Nicholas Carter, Matthew R Alexander, Jennifer H Choe, Cody A Chastain, John Zic, Sara Horst, Isik Turker, Rajiv Agarwal, Evan C. Osmundson, Kamran Idrees, Colleen M. Kiernan, Chandrasekhar Padmanabhan, Christina Edwards Bailey, Cameron Schlegel, Lola B. Chambless, Mike Gibson, Travis J. Osterman, and Lee Wheless. 2023. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Research Square* (2023). https://api.semanticscholar.org/CorpusID:257437276

[73] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[74] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[75] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2, 6 (2020), 305–311.

[76] Ece Kamar. 2016. Directions in hybrid intelligence: complementing AI systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 4070–4073.

[77] Frank Kaptein, Joost Broekens, Koen V Hindriks, and Mark Neerincx. 2016. CAAF: A cognitive affective agent programming framework. *Engineering Psychology and Cognitive Ergonomics* 10906 (2016), 317–330.

[78] Vinothini Kasinathan, Aida Mustapha, and Imran Medi. 2017. Adaptive learning system for higher learning. In *2017 8th international conference on information technology (ICIT)*. IEEE, 960–970.

[79] Muiz Ahmed Khan, Pias Paul, Mahmudur Rashid, Mainul Hossain, and Md Atiqur Rahman Ahad. 2020. An AI-Based Visual Aid With Integrated Reading Assistant for the Completely Blind. *IEEE Transactions on Human-Machine Systems* 50, 6 (2020), 507–517. doi:10.1109/THMS.2020.3027534

[80] Georgia Killcrece, Klaus-Peter Kossakowski, Robin Ruefle, and Mark Zajicek. 2003. *State of the Practice of Computer Security Incident Response Teams (CSIRTs)*. Technical Report CMU/SEI-2003-TR-001. Carnegie Mellon University, Software Engineering Institute's Digital Library. https://doi.org/10.1184/R1/6584396.v1 Accessed: 2024-Feb-13.

[81] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[82] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643* (2023).

[83] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems* 34 (2021), 2611–2624.

[84] Xinmei Kong, Haiguang Fang, Wenli Chen, Jianjun Xiao, and Muhua Zhang. 2025. Examining Human–AI Collaboration in Hybrid Intelligence Learning Environments: Insight from the Synergy Degree Model. *Humanities and Social Sciences Communications* 12 (2025). doi:10.1057/s41599-025-05097-z

[85] Konstantinos Konstantis, Antonios Georgas, Antonis Faras, Konstantinos Georgas, and Aristotle Tympas. 2023. Ethical considerations in working with ChatGPT on a questionnaire about the future of work with ChatGPT. *AI and Ethics (Online)* (2023). doi:10.1007/s43681-023-00312-6

[86] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. doi:10.1145/3582269.3615599

[87] Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. Psychological science in the public interest. *Psychological Science in the Public Interest* (2006).

[88] Nitin Kumar and Anuj Jain. 2022. A Deep Learning Based Model to Assist Blind People in Their Navigation. *Journal of Information Technology Education: Innovations in Practice* 21 (2022), 095–114.

[89] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[90] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback.

[91] Séverin Lemaignan, Mathieu Warnier, E. Akin Sisbot, Aurélie Clodic, and Rachid Alami. 2017. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence* 247 (2017), 45–69.

[92] Bruno Lepri, Nuria Oliver, and Alex Pentland. 2021. Ethical machines: The human-centric use of artificial intelligence. *IScience* 24, 3 (2021).

[93] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. 2020. Human-in-the-loop Debugging Deep Text Classifiers. *arXiv preprint arXiv:2010.04987* abs/2010.04987 (2020). https://api.semanticscholar.org/CorpusID:222290812

[94] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149* (2023).

[95] Bin Ling, Bowen Dong, and Fei Cai. 2024. Applicants' Fairness Perception of Human and AI Collaboration in Resume Screening. *International Journal of Human–Computer Interaction* (2024). doi:10.1080/10447318.2024.2437235

[96] Minghao Liu, Zeyu Cheng, Shen Sang, Jing Liu, and James Davis. 2023. Tag-based annotation creates better avatars. *arXiv preprint arXiv:2302.07354* (2023).

[97] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. 2023. Do humans and machines have the same eyes? Human-machine perceptual differences on image classification. *arXiv preprint arXiv:2304.08733* (2023).

[98] Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. 2022. A Rationale-Centric Framework for Human-in-the-loop Machine Learning. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:247627717

[99] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[100] Chen Lv, Yutong Li, Yang Xing, Chao Huang, Dongpu Cao, Yifan Zhao, and Yahui Liu. 2021. Human-Machine Collaboration for Automated Driving Using an Intelligent Two-Phase Haptic Interface. *Advanced Intelligent Systems* (02 2021), 2000229. doi:10.1002/aisy.202000229

[101] Joseph B Lyons, Kevin T Wynne, Sean Mahoney, and Mark A Roebke. 2019. Trust and human-machine teaming: A qualitative study. In *Artificial intelligence for the internet of everything*. Elsevier, 101–116.

[102] Qing Lyu, Josh Tan, Mike E Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatGPT and GPT-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038* (2023).

[103] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. 2021. A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications. *International Journal of Environmental Research and Public Health* 18 (2021). doi:10.3390/ijerph18042121

[104] Henrique Marcos and Melina Pullin. 2023. LARGE LANGUAGE MODELS AND EU DATA PROTECTION: MAPPING (SOME) OF THE PROBLEMS. https://digi-con.org/large-language-models-and-eu-data-protection-mapping-some-of-the-problems/. Accessed: 2024-02-28.

[105] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[106] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What Is My Teammate? Team Composition Considerations in Human–AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299. doi:10.1109/THMS.2021.3086018

[107] Lucas Memmert and Eva Bittner. 2022. Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

[108] Lucas Memmert and Navid Tavanapour. 2023. TOWARDS HUMAN-AI-COLLABORATION IN BRAINSTORMING: EMPIRICAL INSIGHTS INTO THE PERCEPTION OF WORKING WITH A GENERATIVE AI. In *31st European Conference on Information Systems (ECIS)*.

[109] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023), 20.

[110] Henny Admoni Michelle Zhao, Reid Simmons. 2022. The Role of Adaptation in Collective Human–AI Teaming. *Topics in Cognitive Science* (2022). doi:doi/pdf/10.1111/tops.12633

[111] Fan Mingming, Yang Xianyou, Yu Tsz Tung, Q. Liao Vera, and Zhao Jian. 2023. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *arXiv.org* (2023). Accessed: 2023-11-28.

[112] Suvir Mirchandani, F. Xia, Peter R. Florence, Brian Ichter, Danny Driess, Montse Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large Language Models as General Pattern Machines. *arXiv preprint arXiv:2307.04721* (2023). https://api.semanticscholar.org/CorpusID:259501163

[113] Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, , and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data and Society* (2016).

[114] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 48 (apr 2023), 37 pages. doi:10.1145/3579481

[115] I. Munyaka, Z. Ashktorab, C. Dugan, J. Johnson, and Q. Pan. 2023. Decision Making Strategies and Team Efficacy in Human-AI Teams. *Proceedings of the ACM on Human-Computer Interaction,* (2023). doi:doi/pdf/10.1145/3579476

[116] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[117] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurrian van Diggelen. 2018. Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance. *Lecture Notes in Computer Science* 10011 (2018), 204–214.

[118] An Ngo, Daniel Phelps, Derrick Lai, Thanyared Wong, Lucas Mathias, Anish Shivamurthy, Mustafa Ajmal, Minghao Liu, and James Davis. 2023. Tag-Based Annotation for Avatar Face Creation. *arXiv preprint arXiv:2308.12642* (2023).

[119] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research* 36 (2017), 618–634.

[120] Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review* 4, 4 (1990), 251–277.

[121] Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based Mitigation of Evaluation Bias in Large Language Models. *arXiv preprint arXiv:2402.15987* (2024).

[122] Jonathan Oliver, Raghav Batta, Adam Bates, Muhammad Adil Inam, Shelly Mehta, and Shugao Xia. 2024. Carbon Filter: Real-time Alert Triage Using Large-Scale Clustering and Fast Search. *arXiv preprint arXiv:2405.04691* (2024). https://arxiv.org/abs/2405.04691

[123] OpenAI. 2023. ChatGPT-4. https://openai.com/chatgpt.

[124] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[125] Saket Ozarkar, Raj Chetwani, Sugam Devare, Sumeet Haryani, and Nupur Giri. 2020. AI for Accessibility: Virtual Assistant for Hearing Impaired. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 1–7. doi:10.1109/ICCCNT49239.2020.9225392

[126] Subharun Pal. 2023. The Future of Large Language Models: A Futuristic Dissection on AI and Human Interaction. *IJFMR-International Journal For Multidisciplinary Research* 5, 3 (2023).

[127] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. 2024. VLAAD: Vision and Language Assistant for Autonomous Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 980–987.

[128] Armin Danesh Pazho, Christopher Neff, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Shanle Yao, Mohammadreza Baharani, and Hamed Tabkhi. 2023. Ancilia: Scalable Intelligent Video Surveillance for the Artificial Intelligence of Things. *arXiv preprint arXiv:2301.03561* (2023).

[129] Michael Pflanzer, Zach Traylor, Joseph B. Lyons, Veljko Dubljević, and Chang S. Nam. 2022. Ethics in human–AI teaming: principles and perspectives. *AI and Ethics* (2022), 1–19. https://api.semanticscholar.org/CorpusID:252422919

[130] Joris Pries, Sandjai Bhulai, and Rob D. van der Mei. 2023. Active pairwise distance learning for efficient labeling of large datasets by human experts. *Applied Intelligence* 53 (2023), 24689 – 24708. https://api.semanticscholar.org/CorpusID:260307431

[131] Zheng Qingxiao, Tang Yiliu, Liu Yiren, Liu Weizi, and Huang Yun. 2023. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. *arXiv preprint arXiv* (2023). Accessed: 2023-11-28.

[132] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[133] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290* (2023).

[134] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-AI collaboration in auditing LLM with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 913–926.

[135] Rowena Rodrigues. 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology* (2020).

[136] Stuart J. Russell, Peter Norvig, Ming-Wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, Jitendra Malik, Vikash Mansinghka, Judea Pearl, and Michael J Wooldridge. 2021. *Artificial Intelligence: A Modern Approach*. Prentice Hall.

[137] Eduardo Salas, Carolyn Prince, David P Baker, and Lisa Shrestha. 2017. Situation awareness in team performance: Implications for measurement and training. *Human Factors* 37 (2017).

[138] SP Samyuktha, P Kavitha, VA Kshaya, P Shalini, and R Ramya. 2022. A survey on cyber security meets artificial intelligence: AI–driven cyber security. *Journal of Cognitive Human-Computer Interaction* 2, 2 (2022), 50–55.

[139] Troy Schotter, Saba Kawas, James Prather, Juho Leinonen, Jon Ippolito, and Greg L Nelson. 2025. SPIRAL integration of generative AI in an undergraduate creative media course: effects on self-efficacy and career outcome expectations. *arXiv preprint arXiv:2505.18771* (2025).

[140] Glorin Sebastian. 2023. Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information. *Available at SSRN 4454761* (2023).

[141] Burr Settles. 2009. *Active Learning Literature Survey*. Department of Computer Sciences, University of Wisconsin-Madison.

[142] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 882–892. doi:10.1145/3485447.3512040

[143] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.

[144] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[145] Matthew Sidji, Wally Smith, and Matthew J. Rogerson. 2024. Human–AI Collaboration in Cooperative Games: A Study of Playing Codenames with an LLM Assistant. *Proceedings of the ACM on Human–Computer Interaction* 8, CHI PLAY (2024), 1–25. doi:10.1145/3677081

[146] Sabina Leonelli Silvia Milano, Joshua A. McGrane. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* (2023).

[147] Adam Sobieszek and Tadeusz Price. 2022. Playing games with AIs: the limits of GPT-3 and similar large language models. *Minds and Machines* 32, 2 (2022), 341–364.

[148] B.C. Stahl, A. Andreou, P. Brey, T. Hatzakis, A. Kirichenko, K. Macnish, S. Laulhé Shaelou, A. Patel, M. Ryan, and D. Wright. 2021. Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research* (2021). doi:10.1016/j.jbusres.2020.11.030

[149] Nisan Stiennon, Long Ouyang, et al. 2020. Learning to Summarize from Human Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[150] Joshua Strong, Qianhui Men, and Alison Noble. 2024. Towards Human–AI Collaboration in Healthcare: Guided Deferral Systems with Large Language Models. *arXiv preprint arXiv:2406.07212* (2024).

[151] Nintendo Switch. 2017. Mariokart 8 Deluxe. Video game.

[152] Aaquib Tabrez, Matthew Luebbers, and Bradley Hayes. 2020. A Survey of Mental Modeling Techniques in Human–Robot Teaming. *Current Robotics Reports* 1 (12 2020). doi:10.1007/s43154-020-00019-0

[153] Atieh Taheri, Mohammad Izadi, Gururaj Shriram, Negar Rostamzadeh, and Shaun Kane. 2023. Breaking Barriers to Creative Expression: Co-Designing and Implementing an Accessible Text-to-Image Interface. *arXiv preprint arXiv:2309.02402* (2023).

[154] Ekenobi ThankGod Chinonso. 2023. The impact of chatGPT on privacy and data protection laws. *Available at SSRN: https://ssrn.com/abstract=4556181* (2023).

[155] Surendrabikram Thapa and Surabhi Adhikari. 2023. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of Biomedical Engineering* 51, 12 (2023), 2647–2651.

[156] Hanlin Tian, Kethan Reddy, Yuxiang Feng, Mohammed Quddus, Yiannis Demiris, and Panagiotis Angeloudis. 2024. Enhancing Autonomous Vehicle Training with Language Model Integration and Critical Scenario Generation. *arXiv preprint arXiv:2404.08570* (2024).

[157] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level Generation Through Large Language Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. 1–8.

[158] Issatay Tokmurziyev, Miguel Altamirano Cabrera, Muhammad Haris Khan, Yara Mahmoud, Luis Moreno, and Dzmitry Tsetserukou. 2025. LLM-Glasses: GenAI-driven Glasses with Haptic Feedback for Navigation of Visually Impaired People. *arXiv preprint arXiv:2503.16475* (2025). https://arxiv.org/abs/2503.16475

[159] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). arXiv:2302.13971 [cs.CL]

[160] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does Human Collaboration Enhance the Accuracy of Identifying LLM-Generated Deepfake Texts?. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 163–174.

[161] Imdad Ullah, Najm Hassan, Sukhpal Singh Gill, Basem Suleiman, Tariq Ahamed Ahanger, Zawar Shah, Junaid Qadir, and Salil S Kanhere. 2023. Privacy preserving large language models: ChatGPT case study based vision and framework. *arXiv preprint arXiv:2310.12523* (2023).

[162] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. 2019. Six challenges for human-AI Co-learning. In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*. Springer, 572–589.

[163] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* 46, 4 (2011), 197–221.

[164] Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger. 2022. Generating role-playing game quests with GPT language models. *IEEE Transactions on Games* (2022).

[165] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. *arXiv preprint arXiv:2310.15683* abs/2310.15683 (2023). https://api.semanticscholar.org/CorpusID:264439231

[166] Kailas Vodrahalli, Wei Wei, and James Zou. 2025. Learning a Canonical Basis of Human Preferences from Binary Ratings. *arXiv:2503.24150* (2025).

[167] Emmanuelle Walkowiak and Trent MacDonald. 2023. Generative AI and the Workforce: What Are the Risks? *Available at SSRN* (2023).

[168] Dylan Walsh. 2023. Legal issues presented by generative AI. https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai. Accessed: 2024-02-28.

[169] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24.

[170] Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *18th Workshop on Innovative Use of NLP for Building Educational Applications*.

[171] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. 2023. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. *arXiv preprint arXiv:2310.17642* (2023).

[172] Yifei Wang. 2023. The Large Language Model (LLM) Paradox: Job Creation and Loss in the Age of Advanced AI. TechRxiv Preprint. Available at TechRxiv: [URL].

[173] Yixuan Wang, Ruochen Jiao, Chengtian Lang, Sinong Simon Zhan, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. 2023. Empowering Autonomous Driving with Large Language Models: A Safety Perspective. *arXiv preprint arXiv:2312.00812* (2023).

[174] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903* (2022). https://api.semanticscholar.org/CorpusID:246411621

[175] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205* (2023).

[176] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 402–412. doi:10.1145/3397481.3450656

[177] Feng Wen, Zixuan Zhang, Tianyiyi He, and Chengkuo Lee. 2021. AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications* 12 (09 2021), 5378. doi:10.1038/s41467-021-25637-w

[178] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. 2023. On the Road with GPT-4V(ision): Early Explorations of Visual-Language Model on Autonomous Driving. *arXiv preprint arXiv:2311.05332* (2023).

[179] Elizabeth Whitaker, Ethan Trewhitt, Matthew Holtsinger, Christopher Hale, Elizabeth Veinott, Chris Argenta, and Richard Catrambone. 2013. The effectiveness of intelligent tutoring on training in a video game. In *2013 IEEE International Games Innovation Conference (IGIC)*. IEEE, 267–274.

[180] Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to Play: Pragmatic Reasoning for Efficient Cross-Cultural Communication in Codenames. *arXiv preprint arXiv:2408.04900* (2024). https://arxiv.org/abs/2408.04900

[181] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 195–200.

[182] H. James Wilson and Paul R. Daugherty. 2018. Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review* 96, 4 (2018), 114–123.

[183] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2023. How Knowledge Workers Think Generative AI Will (Not) Transform Their Industries. *arXiv preprint arXiv:2310.06778* (2023).

[184] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. doi:10.1145/3491102.3517582

[185] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658* (2023).

[186] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-Centric Autonomous Systems With LLMs for User Command Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 988–994.

[187] Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11629–11643. doi:10.18653/v1/2023.findings-emnlp.778

[188] Chunyong Yin, Biao Zhou, Zhichao Yin, and Jin Wang. 2019. Local privacy protection classification based on human-centric computing. *Human-centric Computing and Information Sciences* 9 (2019), 1–14. doi:10.1186/s13673-019-0195-4

[189] Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montse Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. 2023. Distilling and Retrieving Generalizable Knowledge for Robot Manipulation via Language Corrections. *ArXiv* abs/2311.10678 (2023). https://api.semanticscholar.org/CorpusID:265281528

[190] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams For Effective Teamwork. *Proceedings of the ACM on Human-Computer Interaction* 7 (2023), 1–31. doi:10.1145/3610072

[191] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[192] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang, Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2025. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. *arXiv preprint arXiv:2502.10391* (2025).

[193] Jianlong Zhou and Fang Chen. 2019. Towards trustworthy human-AI teaming under uncertainty. In *IJCAI 2019 workshop on explainable AI (XAI)*.

[194] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).