Human-Centered AI Agents for Healthcare and Education: A Systematic Literature Review

Lifu Gao, Joshua Sherwood, Nawwaf Aleisa, Andrews Damoah, Yingzhou Lu, and Xiaodong Qu[0000−0001−7610−6475]

The George Washington University

Abstract. Artificial Intelligence (AI) agents have evolved from narrow, rule-based programs to versatile, learning-driven systems capable of perceiving, reasoning, and acting in complex, dynamic environments. Advances in deep learning, reinforcement learning, large language models, and multi-agent coordination have enabled increasingly integrated architectures that combine cognition, planning, memory, and interaction. However, deploying these systems in human-centered domains such as healthcare and education requires more than technical sophistication; usability, trust, interpretability, and ethical alignment are equally critical.

This paper presents a Systematic Literature Review (SLR) conducted in accordance with PRISMA guidelines, covering peer-reviewed research from January 2018 to March 2025. Guided by three research questions, we (**RQ1**) map the core architectural components, design patterns, and enabling technologies that define contemporary AI agents; (**RQ2**) examine applications in healthcare and education to identify domain-specific HCI considerations; and (**RQ3**) synthesize key challenges and future research opportunities for building robust, adaptable, and trustworthy human-centered agents. Our analysis integrates insights from cognitive science-inspired models, hierarchical reinforcement learning, hybrid symbolic–subsymbolic approaches, and large language model-based reasoning. By combining a technical synthesis with human-centered perspectives, this review provides a roadmap for advancing AI agents from experimental prototypes to reliable partners in real-world, user-facing contexts.

Keywords: AI agents · human-centered AI · systematic literature review · PRISMA · cognitive architectures · large language models · reinforcement learning · healthcare applications · educational technology · trust and interpretability

1 Introduction

Artificial intelligence (AI) agents—autonomous systems capable of perceiving their surroundings, reasoning about possible courses of action, and executing decisions—have undergone rapid evolution over the past seven decades. In this paper, we define an *AI agent* as an integrated system that combines perception,

reasoning, action, and interaction modules, operating in dynamic environments with varying levels of autonomy. Figure 1 illustrates a typical modular architecture, including memory, planning, tool use, and interaction capabilities, that underpins modern AI agents.

Figure 1 illustrates a representative architecture, in which the agent integrates memory, planning, tool usage, and action execution. This modular organization supports adaptability, multi-step reasoning, and interactive capabilities—key traits for human-centered AI systems.

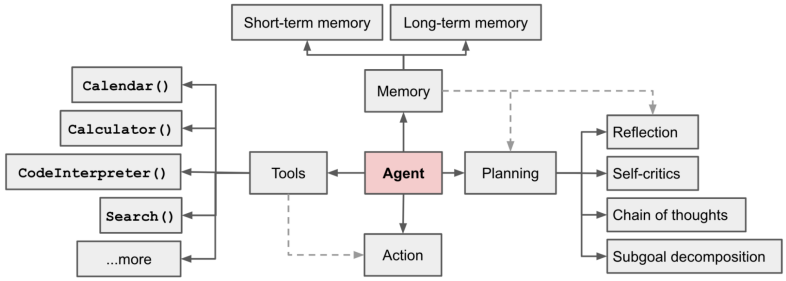


Fig. 1. A high-level modular architecture of a modern AI agent, showing key components such as memory (short- and long-term), planning, tool use, and action execution. These components interact to enable perception, reasoning, and decision-making in dynamic environments.

Early AI agents, rooted in symbolic reasoning systems of the 1950s and 1960s, relied on hand-crafted rules and logic-based methods, excelling in constrained domains but struggling with adaptability and uncertainty [70, 12]. The introduction of statistical learning and probabilistic reasoning in the 1980s and 1990s enhanced reliability, while the emergence of reinforcement learning (RL) enabled agents to learn policies through trial-and-error interactions [16, 20, 27, 73]. The integration of deep neural networks with RL (DeepRL) led to milestones such as superhuman performance in Atari games and Go [67, 68]. More recently, advances in perception, natural language understanding, and cognitive-inspired principles—combined with large language models (LLMs)—have enabled agents to adapt, collaborate, and mirror aspects of human reasoning in dynamic, open-ended environments [4, 35, 47].

Today, AI agents are increasingly deployed in high-stakes, human-facing contexts: self-driving cars navigating congested urban environments [74, 54], autonomous laboratories accelerating scientific discovery [30, 92], virtual assistants managing complex user queries [75], and automated trading systems operating in financial markets [8]. These deployments are enabled by advances in deep learning for perception [32, 23, 12], RL for decision-making [50, 49], LLMs for communication and reasoning [9, 4], and multi-agent coordination frameworks [85].

Despite these advances, developing *unified, human-centered AI agents* remains a grand challenge. Such systems must integrate perception, abstract reasoning, hierarchical planning, and flexible communication while ensuring safety, interpretability, and alignment with human values. In human–computer interaction (HCI) contexts, these requirements extend to usability, trust calibration, ethical compliance, and user experience design. Failures in these areas—whether due to distributional shift, opaque decision-making, or misaligned objectives—can erode user trust and limit adoption.

This paper presents a Systematic Literature Review (SLR), conducted in accordance with PRISMA guidelines, of AI agent architectures, methodologies, and applications from January 2018 to March 2025. We synthesize findings from peer-reviewed literature across multiple databases, focusing on foundational frameworks, key enabling technologies, and application patterns in domains such as healthcare and education—two areas where human-centered AI agents are both urgently needed and underrepresented in prior surveys. We also identify critical challenges—robustness, explainability, resource efficiency, and ethical alignment—and propose a structured design framework to guide newcomers and practitioners in developing trustworthy, human-centered AI agents.

Research Questions

To guide our review and ensure alignment with our human-centered focus, we address the following research questions:

- **RQ1:** What are the core architectural components, design patterns, and enabling technologies that define contemporary AI agents?
- **RQ2:** How are AI agents currently applied in healthcare and education, and what HCI-specific considerations emerge from these domains?
- **RQ3:** What are the key challenges, limitations, and research opportunities for developing robust, adaptable, and trustworthy human-centered AI agents?

By combining a historical perspective with a contemporary synthesis of research, this work aims to bridge the gap between ambitious visions for AI agents and the practical realities of designing, evaluating, and deploying them in real-world, human-facing contexts.

2 Related Work

The rapid evolution of AI agents in both research and industry has led to a surge in literature consolidating their historical development, architectural principles, and practical applications. Existing works fall broadly into two categories: (i) *systematic or domain-specific reviews* synthesizing trends and challenges, and (ii) *technical frameworks* that have directly shaped agent design and evaluation, often with implications for human–AI interaction (HCI).

2.1 Survey-Based Contributions

Several comprehensive surveys have examined the AI agent landscape from different perspectives. Wang et al. [76] provide a holistic review of large language model (LLM)-based agents, covering foundational design principles, application domains, and evaluation strategies. Their work emphasizes the architectural building blocks—planning, memory, and tool use—but remains broad in scope, offering limited guidance for human-centered design.

Guo et al. [21] focus on multi-agent systems for simulation-based research, contrasting single-agent and multi-agent paradigms in terms of profiling, communication, and decision-making. While insightful for collaborative and competitive agent behaviors, their emphasis is primarily on simulation rather than real-world, user-facing applications.

Xi et al. [81] explore LLMs as foundational models for AI agents, with applications in agent-to-agent, human-agent, and multi-agent interactions. They highlight autonomy and adaptability as key properties, yet their review does not deeply address usability or trust in human-facing domains.

Xie et al. [84] examine multimodal agents, analyzing how textual, visual, and auditory capabilities influence design frameworks and evaluation. This perspective is crucial for HCI, as multimodality enables richer user experiences, but their review is technology-centric and omits structured design guidance for domain-specific adoption.

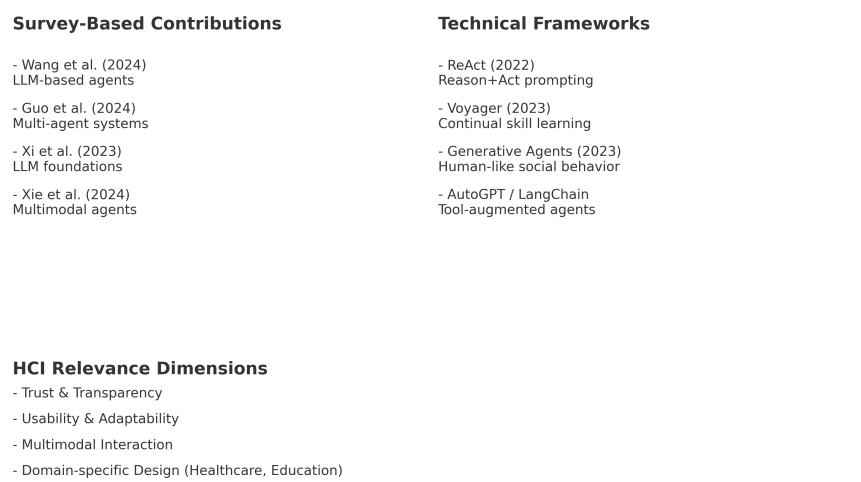


Fig. 2. Taxonomy of prior work on AI agents. Survey-based contributions consolidate literature on architectures, components, and domains, while technical frameworks demonstrate new methods and interaction paradigms. The HCI relevance dimensions highlight how these works contribute to trust, usability, multimodal interaction, and domain-specific design.

2.2 Influential Technical Frameworks

Alongside surveys, several seminal technical works have directly influenced HCI-relevant AI agent design.

ReAct (Reason + Act) [86] introduced a prompting paradigm that interleaves reasoning traces with action steps, enabling interpretable, multi-step decision-making. By allowing human observers to follow an agent's thought process, ReAct contributes to transparency and trust—critical factors for user acceptance.

Voyager [78] demonstrated continual, open-ended skill acquisition in the Minecraft environment through autonomous exploration, curriculum generation, and skill reuse. Although not explicitly HCI-focused, its mechanisms for skill composition and adaptation have strong implications for agents operating in long-term, user-facing contexts.

Generative Agents [55] presented a sandbox environment populated with agents exhibiting believable, human-like social behaviors. This work offers a blueprint for designing agents that can maintain coherent identities and social dynamics, directly aligning with HCI concerns in domains such as education, healthcare, and collaborative work.

Other open-source frameworks, such as AutoGPT and LangChain Agents, have lowered the barrier for developing tool-augmented LLM agents, further enabling rapid prototyping and integration in user-centric applications. However, these frameworks often lack rigorous evaluation for safety, interpretability, and usability.

Figure 2 provides a visual taxonomy summarizing how prior works cluster into survey-based contributions and technical frameworks, and how each relates to key HCI relevance dimensions.

2.3 Synthesis and Research Gap

Table 1 synthesizes prior surveys and technical frameworks, comparing their scope, contributions, HCI relevance, and limitations. While existing surveys provide valuable overviews, and technical works push the boundaries of agent capabilities, few combine architectural synthesis with explicit HCI design principles or offer domain-specific guidance for healthcare and education. This gap motivates our work: a systematic literature review integrating architectural trends, human-centered design considerations, and practical case studies.

3 Methodology

We conducted a Systematic Literature Review (SLR) in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [51] to ensure transparency, reproducibility, and methodological rigor. The review targeted peer-reviewed studies on AI agents published between **January 2018 and March 2025**, with an emphasis on works relevant to both

Table 1. Summary of key prior works on AI agents, with scope, contributions, HCI relevance, and gaps addressed by this SLR.

Paper	Scope	Key Contributions	HCI Relevance	Gaps
Wang et al. (2024)	LLM agent survey	Components, applications, metrics	Core modules; limited user focus	No design guidance
Guo et al. (2024)	Multi-agent sims	Single vs. multi-agent, domains	Collab. behavior insights	No real-world HCI
Xi et al. (2023)	LLM agent foundations	Autonomy, adaptability, interaction	Highlights adaptability	Limited trust/usability
Xie et al. (2024)	Multimodal agents	Text, vision, audio integration	Richer interactions	No domain HCI guidance
ReAct (2022)	Reason+Act prompting	Interleaves reasoning/actions	Improves interpretability	Prompt-based only
Voyager (2023)	Continual skill learning	Open-ended exploration, skills	Adaptable to users	Game-only context
Generative Agents (2023)	Social sandbox agents	Coherent identities, social dynamics	Social HCI potential	No applied eval.

foundational architectures and human–AI interaction (HCI) considerations. This process was designed to directly address the three research questions outlined in Section Introduction, by:

- Mapping the **core architectural components and enabling technologies** of AI agents (**RQ1**);
- Identifying and analyzing applications in **healthcare and education** with explicit HCI considerations (**RQ2**);
- Synthesizing **challenges and opportunities** to inform future development of robust, adaptable, and trustworthy human-centered agents (**RQ3**).

3.1 Search Strategy

To address the limitations of prior surveys and strengthen methodological rigor, we expanded the search beyond Google Scholar to include six major scholarly databases: *Google Scholar*, *ACM Digital Library*, *IEEE Xplore*, *Scopus*, *Web of Science*, and *PubMed*. This multi-database approach ensured coverage of both computer science and application-specific research in healthcare, education, and other domains, directly supporting the breadth needed for **RQ1** and **RQ2**.

We used a combination of general and domain-specific search strings, including:

- `"AI Agent"` OR `"Autonomous Agent"` OR `"Intelligent Agent"`
- `"Reinforcement Learning Agent"` OR `"Multi-Agent System"`
- `"Large Language Model Agent"` OR `"LLM-based Agent"`
- Domain-specific combinations such as `"AI Agent" AND Healthcare`, `"AI Agent" AND Education`.

The full list of queries by topic is provided in Appendix A.
The review process followed three stages:

1. **Identification:** Database queries retrieved 310 records.
2. **Screening:** After removing duplicates, 310 unique records were screened by title and abstract, resulting in 66 candidate papers.
3. **Eligibility:** Full-text review of these 66 papers led to the inclusion of 29 studies in the final synthesis.

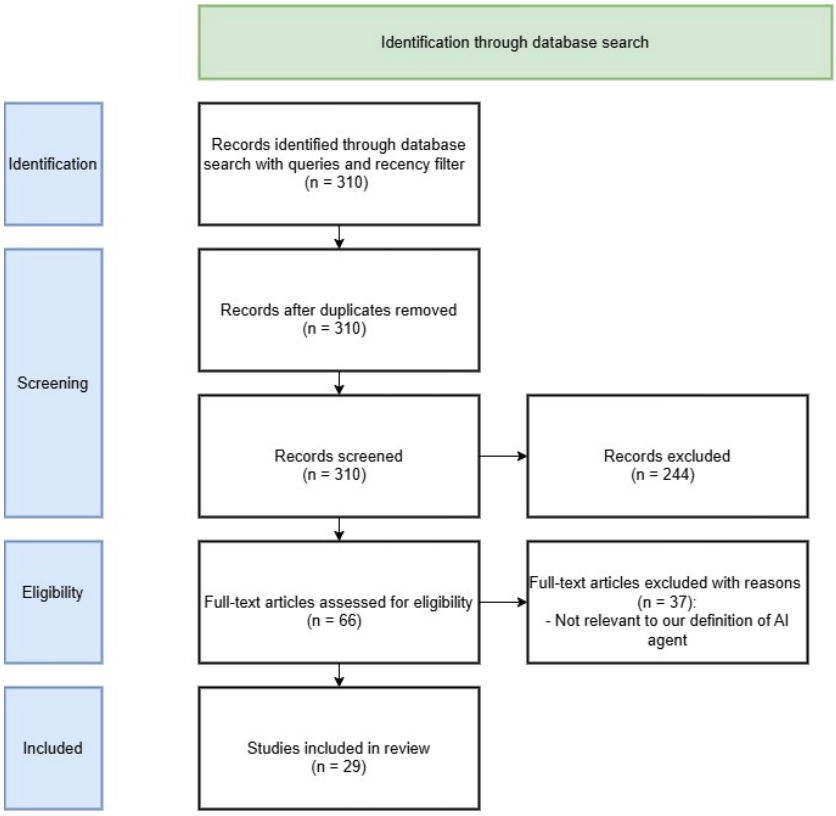


Fig. 3. PRISMA flow diagram summarizing the literature search and selection process. Records identified (n=310), after duplicates removed (n=310), after title/abstract screening (n=66), and final included studies (n=29).

This multi-database approach ensured coverage of both computer science and application-specific research in healthcare, education, and other domains [65, 52].

3.2 Inclusion and Exclusion Criteria

Studies were included if they:

- Presented a peer-reviewed research paper or reputable preprint available in full text.
- Focused directly on AI agent architectures, frameworks, or applications (**RQ1**, **RQ2**).
- Were published between January 2018 and March 2025 in English.

Studies were excluded if they:

- Were inaccessible in full text (e.g., abstract-only or behind a paywall without institutional access).
- Mentioned AI agents only tangentially without substantive technical or HCI contributions (**RQ1**–**RQ3**).
- Were non-English, to maintain analysis consistency and avoid misinterpretation.

3.3 Data Extraction and Analysis

For each included study, we extracted:

- **Bibliographic metadata:** authors, year, venue.
- **Technical focus:** architecture, reasoning method, perception modules, interaction model (**RQ1**).
- **HCI considerations:** usability, transparency, trust, user evaluation (**RQ2**).
- **Domain relevance:** e.g., healthcare, education, business, entertainment (**RQ2**).
- **Key contributions and limitations** (**RQ3**).

We then applied thematic coding to group studies into three major categories: *Core Components* (e.g., memory, planning, tool use), *Applications* (healthcare, education, business, entertainment), and *Paradigm-Shifting Designs* (e.g., symbolic–neural integration, multimodal reasoning). This categorization ensured that the synthesis directly addressed **RQ1** by mapping technical foundations, **RQ2** by evaluating domain-specific use cases, and **RQ3** by identifying persistent challenges and emerging research opportunities.

4 Results

This section presents the results of our systematic review, organized into three main areas: (i) core architectures and components of modern AI agents, (ii) application domains, and (iii) paradigm-shifting designs. The goal is to provide a structured synthesis of technical advances without interpretation; human-centered implications are discussed in Section Discusssion. Findings are presented in relation to the research questions (RQs) outlined in Section Introduction.

4.1 Core Architectures and Components (RQ1)

Modern AI agents integrate perception, reasoning, decision-making, and interaction into unified architectures. Figure 1 illustrates a representative modular design, highlighting memory, planning, tool use, and interaction subsystems. These components, supported by advances in large language models (LLMs), reinforcement learning (RL), and symbolic reasoning, enable agents to operate adaptively in dynamic environments.

Memory Memory modules allow agents to store and retrieve relevant information over varying time scales [80, 25]. Short-term memory is constrained by the model's context window, while long-term memory can be parametric (encoded within model weights) or external (e.g., vector databases). Declarative memory stores explicit facts, whereas procedural memory encodes learned skills. Figure 4 summarizes memory sources, forms, and operations. Memory-enhanced agents improve task continuity and personalization but face challenges in scalability and integration with external knowledge sources [72, 90].

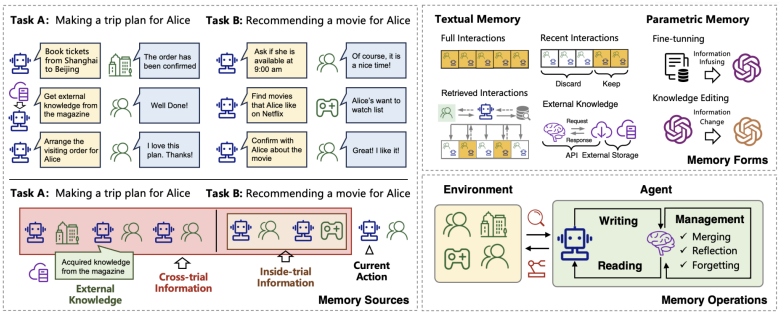
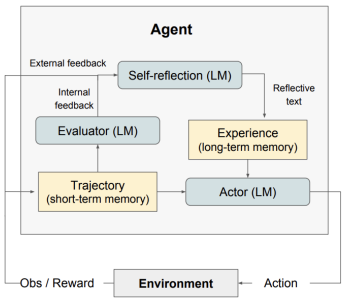


Fig. 4. Sources, forms, and operations of memory in LLM-based agents [90].

Planning Planning enables agents to sequence actions toward long-term goals. Modern approaches combine symbolic planning [20] with subsymbolic methods, including hierarchical RL [33] and model-based RL. Techniques such as Chain-of-Thought prompting [79] improve reasoning transparency, while frameworks like Reflexion [66] and Chain-of-Hindsight [44] incorporate self-reflection and historical feedback to refine decisions without extensive fine-tuning. Planning modules benefit from meta-learning [17] and continual learning [7] for cross-task generalization, refine decisions without extensive fine-tuning [13, 59].

Tools Tool integration extends an agent's native capabilities [48, 39]. Tools can include search engines, APIs, code interpreters, and robotics control systems.



Algorithm 1 Reinforcement via self-reflection

Initialize Actor, Evaluator, Self-Reflection: M_a, M_e, M_{sr}
Initialize policy $\pi_\theta(a_i|s_i)$, $\theta = \{M_a, mem\}$
Generate initial trajectory using π_θ
Evaluate τ_0 using M_e
Generate initial self-reflection sr_0 using M_{sr}
Set $mem \leftarrow [sr_0]$
Set $t = 0$
while M_e not pass or $t <$ max trials **do**
  Generate $\tau_t = [a_0, o_0, \dots a_i, o_i]$ using π_θ
  Evaluate τ_t using M_e
  Generate self-reflection sr_t using M_{sr}
  Append sr_t to mem
  Increment t
end while
return

Fig. 5. The Reflexion framework integrates heuristic functions and linguistic feedback for self-refinement [66].

LLM-based agents dynamically select and invoke tools to perform specialized tasks. Key challenges include managing hallucinations, planning complexity, and error propagation, which impact reliability in real-world applications [82].

Perception Perception modules transform sensory input (text, images, audio, LiDAR) into structured representations. Vision relies on convolutional neural networks [32, 23] and vision transformers [12], while speech recognition and natural language understanding leverage transformer-based LLMs [9, 4]. Sensor fusion combines modalities for more complete environmental understanding.Perception modules transform sensory input (text, images, audio, LiDAR) into structured representations [41, 29].

Representation and Abstraction Representation layers encode high-dimensional data into compact latent vectors using self-supervised and contrastive learning [6], enabling flexible reasoning and generation [87]. In LLM-based agents, embeddings capture semantic and syntactic relationships, enabling flexible reasoning and generation [83].

Interaction and Communication Interaction modules enable communication with humans and other agents. Methods include natural language interfaces [9], emergent communication protocols in multi-agent systems [37], and grounded language acquisition [3]. Effective interaction design supports coordination and collaboration.

4.2 Applications of AI Agents (RQ2)

AI agents are deployed in a variety of domains, leveraging the above components to deliver adaptive, context-aware solutions. The domains reviewed include both

technical and human-centered applications, with healthcare and education as primary foci.

Healthcare. Healthcare applications leverage AI agents for diagnostic, assistive, and decision-support roles [60, 57]. Applications include diagnostic and decision-support systems [28], patient-facing virtual assistants [14], and AI-assisted robotic surgery [19]. Advances in data integration and context awareness have improved clinical decision-making and workflow efficiency [29].

Education. In education, AI agents support personalized learning, adaptive feedback, and cognitive impact assessment [62, 58, 56, 63, 61]. AI agents support personalized learning, adaptive feedback, and team teaching scenarios [36, 77]. Systems like Mentigo [89] adjust strategies based on student state and problem-solving stage, enhancing engagement and learning outcomes.

Business and Industry. Enterprise uses span customer service chatbots [64], supply chain optimization [15], and financial decision-making [43]. AI agents improve operational efficiency, reduce costs, and enable real-time decision-making.

Science and Research. Automated laboratories [34] and AI research assistants [45] accelerate discovery in fields from biology to materials science, supporting experiment design, data analysis, and hypothesis generation.

Public Services and Urban Planning. Applications include urban resource management [40], collaborative decision-making in public administration [2], and transportation optimization [31].

Entertainment and Creativity. AI agents contribute to game AI, interactive storytelling [69], and content creation [10, 88], enhancing user engagement through personalization and adaptive behaviors.

4.3 Paradigm-Shifting Designs (RQ3)

Beyond incremental improvements to components, several architectural innovations redefine agent capabilities.

Cognitive-Inspired Architectures Hybrid models integrate symbolic reasoning with neural networks [35, 47], combining the interpretability and compositionality of symbolic approaches with the adaptability of deep learning. These designs are well-suited for domains requiring precise reasoning and abstraction.

Hierarchical and Modular Approaches Hierarchical control decomposes complex tasks into subtasks, enabling long-horizon planning [33]. Modular architectures assign specialized roles to subcomponents, improving scalability, reusability, and fault isolation [24, 42]. Integrating these approaches with meta-learning and transfer learning enhances adaptability across domains.

5 Discussion

This section synthesizes the findings from our systematic review, highlighting their implications for human–AI interaction (HCI), offering design recommendations for new researchers entering the field, and outlining a consolidated set of challenges and future research opportunities. While our Results section provided a technical overview, the following discussion focuses on interpretive insights and human-centered considerations, aligning with HCII's emphasis on usability, trust, and real-world adoption. Each subsection below ties back to the research questions (RQs) defined in Section Introduction.

5.1 Synthesis of Findings and HCI Implications (RQ1, RQ2)

Our review revealed that modern AI agents integrate increasingly sophisticated components—memory, planning, tool use, perception, and interaction modules—into modular architectures capable of operating in dynamic, open-ended environments (**RQ1**). While these advancements have significantly improved adaptability and task performance, their deployment in human-facing domains such as healthcare and education (**RQ2**) raises distinct design and adoption considerations.

Trust and Transparency. Techniques like Chain-of-Thought reasoning [79] and Reflexion [66] enhance interpretability by making reasoning steps observable. In healthcare, this transparency is critical for clinician trust; in education, it helps instructors validate AI-generated guidance.

Usability and Accessibility. Memory and planning modules support context retention and goal decomposition, enabling smoother interactions. However, user experience still suffers from unpredictable behavior and complex configuration requirements, which can hinder adoption by non-experts.

Domain-Specific Constraints. In high-stakes applications, domain alignment is essential. For example, healthcare agents must integrate with clinical workflows and comply with privacy regulations, while educational agents must adapt to diverse learning styles and remain inclusive for neurodiverse populations and accessible to non-experts [53].

5.2 Design Recommendations for Newcomers (RQ3)

For researchers and practitioners entering AI agent development, we propose the following recommendations, distilled from the literature and our synthesis of best practices. These recommendations are intended to help bridge the gap between the current technical state (**RQ1**), domain-specific requirements (**RQ2**), and identified research gaps (**RQ3**):

1. **Establish a Strong Theoretical Foundation.** Begin with core concepts in reinforcement learning, planning, multi-agent coordination, and decision theory. Use accessible resources such as *Reinforcement Learning: An Introduction* [73] and introductory reviews [80] to build conceptual grounding.

2. **Start with Controlled, Measurable Projects.** Use simulation environments like OpenAI Gym, CARLA [5], or PettingZoo [18] for initial projects, ensuring that evaluation metrics (e.g., task completion rate, convergence speed, interpretability) are defined from the outset.
3. **Leverage Toolkits and Frameworks.** Adopt open-source frameworks (e.g., LangChain, AutoGPT) to rapidly prototype and experiment with agent capabilities such as tool integration and memory augmentation. Engage with the open-source community to accelerate learning and maintain awareness of emerging practices.
4. **Iterate with Feedback and Reproducibility in Mind.** Implement iterative testing, leveraging user or peer feedback to refine design choices. Share code and datasets to facilitate reproducibility and community validation.
5. **Identify and Address Research Gaps Early.** Focus on underexplored intersections such as hybrid symbolic–subsymbolic architectures for improved interpretability, or continual learning frameworks for long-term adaptability.

5.3 Challenges, Opportunities, and Research Directions (RQ3)

While AI agents have achieved notable milestones, their widespread adoption in human-centered domains depends on overcoming both technical and social barriers. The challenges and opportunities identified here directly inform **RQ3**, providing a roadmap for future research.

Technical Barriers Robustness and Safety. Agents remain sensitive to distributional shifts and adversarial perturbations [1], limiting their reliability in open-world settings. **Generalization and Transfer.** Cross-domain adaptation remains limited [91], necessitating advances in meta-learning and domain adaptation. **Scalability and Efficiency.** High compute and energy demands [71] restrict accessibility, underscoring the need for model compression and efficient architectures.

HCI-Related Barriers Interpretability. Opaque decision-making processes hinder trust [11], particularly in regulated domains. **Ethical and Social Impacts.** Bias, privacy risks, and accountability gaps [26, 22] require transparent evaluation and regulatory frameworks. **Adoption Challenges.** Complex interfaces and lack of integration with existing workflows slow real-world uptake.

Emerging Research Opportunities Neuroscience-Inspired Mechanisms. Incorporating predictive coding, synaptic plasticity, and dendritic computation [46] could yield more stable and interpretable learning. **Continual and Interactive Learning.** Architectures that learn from ongoing interaction while retaining prior knowledge [7] are key to real-world adaptability. **Hybrid Symbolic–Subsymbolic Models.** Blending structured reasoning with deep learning [35, 47] may improve transparency without sacrificing performance. **Multi-Agent Governance and Coordination.** Protocols for negotiation, resource

sharing, and conflict resolution [85, 38] will be critical for large-scale, cooperative agent ecosystems.

5.4 Summary

In sum, the next stage of AI agent research must balance technical innovation with human-centered design. Robustness, interpretability, and domain alignment are essential for deployment in sensitive contexts (**RQ2**), while open-source collaboration and reproducibility will ensure that the field progresses in a transparent and inclusive manner (**RQ3**).

6 Conclusion

AI agents have evolved from narrow, rule-based programs into autonomous systems capable of perceiving, reasoning, acting, and collaborating across diverse domains. This review addressed **RQ1** by synthesizing developments in core architectural components, enabling technologies, and paradigm-shifting designs, mapping how advances in reinforcement learning, large language models, planning, and multimodal perception have expanded agent capabilities.

We addressed **RQ2** by examining the deployment of AI agents in human-facing domains—particularly healthcare, education, and public services—highlighting domain-specific constraints such as workflow integration, regulatory compliance, and accessibility. These analyses emphasized that real-world adoption depends equally on meeting human-centered requirements, including trust calibration, interpretability, and usability.

Finally, **RQ3** was addressed by identifying persistent challenges and outlining future research opportunities, including robustness under distributional shift, ethical alignment, hybrid symbolic–subsymbolic integration, continual learning, and multi-agent governance. We emphasized that agents must be interpretable to their human collaborators, adaptable to diverse user needs, and aligned with societal values. These priorities are especially critical in sensitive applications, where usability, safety, and integration with existing workflows determine long-term impact.

Through interdisciplinary collaboration and sustained attention to human-centered design, AI agents can progress from experimental prototypes to reliable partners in research, education, healthcare, and beyond—serving not only as intelligent tools but as trustworthy collaborators in advancing shared human goals.

References

1. Amodei, D.e.a.: Concrete problems in ai safety. arXiv:1606.06565 (2016)
2. Bauer, M., Sanchez, L., Song, J.: Iot-enabled smart cities: Evolution and outlook. Sensors **21**(13), 4511 (2021)
3. Bisk, Y.e.a.: Experience grounds language. In: EMNLP. pp. 8718–8735 (2020)

4. Brown, T.B.e.a.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
5. CARLA: Carla (autonomous driving simulator). https://github.com/carla-simulator/carla (2025)
6. Chen, T.e.a.: A simple framework for contrastive learning of visual representations. ICML pp. 1597–1607 (2020)
7. De Lange, M.e.a.: A continual learning survey: Defying forgetting in classification tasks. IEEE TPAMI **44**(7), 3366–3385 (2022)
8. Deng, Y., Bao, F., Kong, Y., Ren, Z., Dai, Q.: Deep direct reinforcement learning for financial signal representation and trading. IEEE transactions on neural networks and learning systems **28**(3), 653–664 (2017)
9. Devlin, J.e.a.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
10. Ding, S., Chen, X., Fang, Y., Liu, W., Qiu, Y., Chai, C.: DesignGPT: Multi-agent collaboration in design. In: 2023 16th International Symposium on Computational Intelligence and Design (ISCID). pp. 204–208. https://doi.org/10.1109/ISCID59865.2023.00056, https://ieeexplore.ieee.org/abstract/document/10494260?casa_token=pYK0QW_jLRkAAAAA:mShLDj7Q52Zxm3_gKla5JEsTSIGaqhwsN64NWBBvlKaErH5ib0XHFl0Evm4KLUhxxSDCBXvblg, ISSN: 2473-3547
11. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017)
12. Dosovitskiy, A.e.a.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
13. Dou, G., Zhou, Z., Qu, X.: Time majority voting, a pc-based eeg classifier for non-expert users. In: International Conference on Human-Computer Interaction. pp. 415–428. Springer (2022)
14. Du, D., Bhardwaj, S., Parker, S.J., Cheng, Z., Zhang, Z., Van Eyk, J.E., Yu, G., Clarke, R., Herrington, D.M., et al.: Abds: tool suite for analyzing biologically diverse samples. bioRxiv (2023)
15. Du, D., Bhardwaj, S., Wang, Y., Parker, S.J., Zhang, Z., Van Eyk, J.E., Yu, G., Clarke, R., Herrington, D.M., et al.: Embracing the informative missingness and silent gene in analyzing biologically diverse samples. Scientific Reports **14**(1), 28265 (2024)
16. Feigenbaum, E.A., McCorduck, P.: The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World. Addison-Wesley (1983)
17. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. pp. 1126–1135 (2017)
18. Foundation, F.: Pettingzoo (multi-agent reinforcement learning environments). https://github.com/Farama-Foundation/PettingZoo (2024)
19. Fu, Y., Lu, Y., Wang, Y., Zhang, B., Zhang, Z., Yu, G., Liu, C., Clarke, R., Herrington, D.M., Wang, Y.: Ddn3. 0: Determining significant rewiring of biological network structure with differential dependency networks. Bioinformatics p. btae376 (2024)
20. Ghallab, M., Nau, D., Traverso, P.: Automated Planning: Theory and Practice. Elsevier (2004)
21. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large Language Model based Multi-Agents: A Survey of Progress and Challenges (Apr 2024). https://doi.org/10.48550/arXiv.2402.01680, http://arxiv.org/abs/2402.01680, arXiv:2402.01680 [cs]

22. Hagendorff, T.: The ethics of ai ethics: An evaluation of guidelines. Minds and Machines **30**(1), 99–120 (2020)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
24. Huang, T.e.a.: Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: ICLR (2023)
25. Jiang, Y.H., Li, R., Zhou, Y., Qi, C., Hu, H., Wei, Y., Jiang, B., Wu, Y.: Ai agent for education: Von neumann multi-agent system framework. In: Proceedings of the 28th Global Chinese Conference on Computers in Education (GCCCE 2024). pp. 77–84 (2024)
26. Jobin, A., Ienca, M., Vayena, E.: The global landscape of ai ethics guidelines. Nature Machine Intelligence **1**, 389–399 (2019)
27. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of artificial intelligence research **4**, 237–285 (1996)
28. Katz, D., Jin, C., Smith, G., Clark, N., Iyer, G., Keshishian, H., Hart, P., Sanford, J., Zhang, Z., Ge, Y., et al.: The multi-omic, multi-tissue response to acute endurance and resistance exercise: Results from the molecular transducers of physical activity consortium. Circulation **150**(Suppl_1), A4143199–A4143199 (2024)
29. Key, M.L., Mehtiyev, T., Qu, X.: Advancing eeg-based gaze prediction using depthwise separable convolution and enhanced pre-processing. In: International Conference on Human-Computer Interaction. pp. 3–17. Springer (2024)
30. Kitson, P.J., Marie, G., Fossey, J.S., Jesus, P., Cronin, L.: Configurable robotic platform for automated synthesis. Nature **549**, 70–75 (2017)
31. Kouziokas, G.N.: The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment. Transportation research procedia **24**, 467–473 (2017)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
33. Kulkarni, T.D.e.a.: Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. NeurIPS pp. 3675–3683 (2016)
34. Kusne, A.G., McDannald, A.: Scalable multi-agent lab framework for lab optimization. Matter **6**(6), 1880–1893 (Jun 2023). https://doi.org/10.1016/j.matt.2023.03.022, http://dx.doi.org/10.1016/j.matt.2023.03.022
35. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Sciences **40**, e253 (2017)
36. Lan, Y.J., Chen, N.S.: Teachers' agency in the era of llm and generative ai. Educational Technology & Society **27**(1), I–XVIII (2024)
37. Lazaridou, A., Baroni, M.: Emergent multi-agent communication in the deep learning era. Neural Computation **31**(5), 753–763 (2019)
38. Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J.A., Shen, B., et al.: Political-llm: Large language models in political science. arXiv preprint arXiv:2412.06864 (2024)
39. Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., Li, Y.: Api-bank: A comprehensive benchmark for tool-augmented llms. arXiv preprint arXiv:2304.08244 (2023)
40. Li, P., Yang, Q., Geng, X., Zhou, W., Ding, Z., Nian, Y.: Exploring diverse methods in visual question answering. In: 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). pp. 681–685. IEEE (2024)

41. Li, W., Zhou, N., Qu, X.: Enhancing eye-tracking performance through multi-task learning transformer. In: International Conference on Human-Computer Interaction. pp. 31–46. Springer (2024)
42. Li, Y., Ma, T., Zhu, S.: Learning knowledge graphs with language models: A survey. arXiv:2301.01037 (2023)
43. Liu, F., Guo, S., Xing, Q., Sha, X., Chen, Y., Jin, Y., Zheng, Q., Yu, C.: Application of an ann and lstm-based ensemble model for stock market prediction. In: 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE). pp. 390–395. IEEE (2024)
44. Liu, H., Sferrazza, C., Abbeel, P.: Chain of hindsight aligns language models with feedback. arXiv preprint arXiv:2302.02676 (2023)
45. Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D.: The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery (Sep 2024). https://doi.org/10.48550/arXiv.2408.06292, http://arxiv.org/abs/2408.06292, arXiv:2408.06292 [cs]
46. Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Frontiers in Computational Neuroscience **10**, 94 (2016)
47. Marcus, G.: The next decade in ai: Four steps towards robust artificial intelligence. Medium (2020)
48. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: Augmented language models: a survey. arXiv preprint arXiv:2302.07842 (2023)
49. Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Playing atari with deep reinforcement learning. In: NIPS Deep Learning Workshop (2013)
50. Mnih, V.e.a.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
51. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. Bmj **339** (2009)
52. Murungi, N.K., Pham, M.V., Dai, X., Qu, X.: Trends in machine learning and electroencephalogram (eeg): a review for undergraduate researchers. In: International Conference on Human-Computer Interaction. pp. 426–443. Springer (2023)
53. Murungi, N.K., Pham, M.V., Dai, X.C., Qu, X.: Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. In: The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2023)
54. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on Intelligent Vehicles **1**(1), 33–55 (2016)
55. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)
56. Qu, X., Hall, M., Sun, Y., Sekuler, R., Hickey, T.J.: A personalized reading coach using wearable eeg sensors (2019)
57. Qu, X., Hickey, T.J.: Eeg4home: A human-in-the-loop machine learning model for eeg-based bci. In: International Conference on Human-Computer Interaction. pp. 162–172. Springer (2022)
58. Qu, X., Key, M., Luo, E., Qiu, C.: Integrating hci datasets in project-based machine learning courses: a college-level review and case study. In: International Conference on Human-Computer Interaction. pp. 124–143. Springer (2024)

59. Qu, X., Liu, P., Li, Z., Hickey, T.: Multi-class time continuity voting for eeg classification. In: International Conference on Brain Function Assessment in Learning. pp. 24–33. Springer (2020)
60. Qu, X., Liukasemsarn, S., Tu, J., Higgins, A., Hickey, T.J., Hall, M.H.: Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on eeg patterns. Frontiers in psychiatry **11**, 541659 (2020)
61. Qu, X., Mei, Q., Liu, P., Hickey, T.: Using eeg to distinguish between writing and typing for the same cognitive task. In: International Conference on Brain Function Assessment in Learning. pp. 66–74. Springer (2020)
62. Qu, X., Sherwood, J., Liu, P., Aleisa, N.: Generative ai tools in higher education: A meta-analysis of cognitive impact. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–9 (2025)
63. Qu, X., Sun, Y., Sekuler, R., Hickey, T.: Eeg markers of stem learning. In: 2018 IEEE Frontiers in Education Conference (FIE). pp. 1–9. IEEE (2018)
64. Satheesh, M., Nagaraj, S.: Applications of artificial intelligence on customer experience and service quality of the banking sector. International Management Review **17**(1), 9–86 (2021)
65. Saunders, T., Aleisa, N., Wield, J., Sherwood, J., Qu, X.: Optimizing the literature review process: Evaluating generative ai models on sum-marizing undergraduate data science research papers. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024)
66. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems **36** (2024)
67. Silver, D.e.a.: Mastering the game of go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
68. Silver, D.e.a.: Mastering the game of go without human knowledge. Nature **550**, 354–359 (2017)
69. Stefnisson, I., Thue, D.: Mimisbrunnur: AI-assisted authoring for interactive storytelling **14**(1), 236–242. https://doi.org/10.1609/aiide.v14i1.13046, https://ojs.aaai.org/index.php/AIIDE/article/view/13046
70. Stone, P., Veloso, M.: Multiagent systems: A survey from a machine learning perspective. Autonomous Robots **8**(3), 345–383 (2000)
71. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in nlp. ACL pp. 3645–3650 (2019)
72. Su, Z., Zhang, J., Qu, X., Zhu, T., Li, Y., Sun, J., Li, J., Zhang, M., Cheng, Y.: Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm (2024), https://arxiv.org/abs/2408.12076
73. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT press (2018)
74. Tang, J., Lin, C., Zhao, Z., Wei, S., Wu, B., Liu, Q., Feng, H., Li, Y., Wang, S., Liao, L., et al.: Textsquare: Scaling up text-centric visual instruction tuning. arXiv preprint arXiv:2404.12803 (2024)
75. Tür, G., De Mori, R.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley (2011)
76. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (Mar 2024). https://doi.org/10.1007/s11704-024-40231-1, https://doi.org/10.1007/s11704-024-40231-1

77. Wang, T., Wu, T., Liu, H., Brown, C., Chen, Y.: Generative co-learners: Enhancing cognitive and social presence of students in asynchronous learning with generative ai. arXiv preprint arXiv:2410.04365 (2024)
78. Wang, X., Xie, T., et al.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
79. Wei, J.e.a.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS pp. 24824–24837 (2022)
80. Weng, L.: Llm-powered autonomous agents. lilianweng.github.io (Jun 2023), https://lilianweng.github.io/posts/2023-06-23-agent/
81. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., Gui, T.: The Rise and Potential of Large Language Model Based Agents: A Survey (Sep 2023). https://doi.org/10.48550/arXiv.2309.07864, http://arxiv.org/abs/2309.07864, arXiv:2309.07864 [cs]
82. Xiao, Z., Mai, Z., Cui, Y., Xu, Z., Li, J.: Short interest trend prediction with large language models. In: Proceedings of the 2024 International Conference on Innovation in Artificial Intelligence. p. 1. ICIAI '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3655497.3655500, https://doi.org/10.1145/3655497.3655500
83. Xiao, Z., Mai, Z., Xu, Z., Kwon, Y., Li, J.: Short interest trend prediction. In: 2024 6th International Conference on Natural Language Processing (ICNLP). pp. 352–356 (2024). https://doi.org/10.1109/ICNLP60986.2024.10692439
84. Xie, J., Chen, Z., Zhang, R., Wan, X., Li, G.: Large Multimodal Agents: A Survey (Feb 2024). https://doi.org/10.48550/arXiv.2402.15116, http://arxiv.org/abs/2402.15116, arXiv:2402.15116 [cs]
85. Yang, Y.e.a.: Mean field multi-agent reinforcement learning. In: International Conference on Machine Learning. pp. 5571–5580 (2018)
86. Yao, S., Shinn, N., Gopinath, A., Narasimhan, K.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
87. Yi, L., Qu, X.: Attention-based cnn capturing eeg recording's average voltage and local change. In: International Conference on Human-Computer Interaction. pp. 448–459. Springer (2022)
88. Yunoki, I., Berreby, G., D'Andrea, N., Lu, Y., Qu, X.: Exploring ai music generation: A review of deep learning algorithms and datasets for undergraduate researchers. In: International Conference on Human-Computer Interaction. pp. 102–116. Springer (2023)
89. Zha, S., Liu, Y., Zheng, C., XU, J., Yu, F., Gong, J., XU, Y.: Mentigo: An intelligent agent for mentoring students in the creative problem solving process. arXiv preprint arXiv:2409.14228 (2024)
90. Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z., Wen, J.R.: A survey on the memory mechanism of large language model based agents. arXiv preprint arXiv:2404.13501 (2024)
91. Zhao, Y.e.a.: Towards domain-agnostic and open-domain generalization of vision and language agents. In: NeurIPS. pp. 19567–19578 (2021)
92. Zhao, Z., Tang, J., Wu, B., Lin, C., Wei, S., Liu, H., Tan, X., Zhang, Z., Huang, C., Xie, Y.: Harmonizing visual text comprehension and generation. arXiv preprint arXiv:2407.16364 (2024)