Responsible Agentic Reasoning and AI Agents: A Critical Survey

Shaina Raza[a,1,*], Ranjan Sapkota[b,1,*], Manoj Karkee[b], Christos Emmanouilidis[c]

[a] *Vector Institute, Toronto, Canada*
[b] *Cornell University, USA*
[c] *University of Groningen, Netherlands*

Abstract

Information fusion for trustworthy AI is entering a pivotal stage, where Large Language Model (LLM)-based agents excel at integrating multi-source knowledge into coherent reasoning chains. However, these agents remain opaque and difficult to audit in the absence of embedded, in-loop safety mechanisms. Existing surveys treat reasoning, agentic behavior, and safety in isolation, leaving a gap in how to integrate them into practical, trustworthy agents. To address this, we present a survey at the intersection of these domains and introduce Responsible Reasoning AI Agents (R^2A^2), a class of agentic LLM systems that generate explicit reasoning traces while enforcing fairness, privacy, transparency, accountability, and auditability throughout the decision loop. We synthesize recent advances in chain-of-thought prompting, ReAct, tree/graph-of-thought structures, tool use, memory, retrieval, and agentic browsing, and integrate these with responsible AI principles into a unified evaluation framework. Furthermore, we propose an evaluation methodology for agentic reasoning with embedded safety mechanisms and outline a five-stage reproducible protocol: Curate, Unify, Probe, Benchmark, Analyze, to operationalize responsibility metrics. Overall, this taxonomy, metric suite, and framework advance the development of safe, transparent, and governable LLM-based agents. The project repository is available on GitHub `https://github.com/shainarazavi/Responsible-reasoning-agents`.

Keywords: Responsible AI; Explainable AI; Ethical AI; Agentic AI; AI Fairness; AI Transparency; Reasoning Language Models; AI Agents; Autonomous Agents; Multi-Agent Systems; Human-AI Collaboration

1. Introduction

Reasoning - the process of turning facts into defensible decisions - is rapidly becoming a defining capability for next-generation AI models. Modern Large Language Models (LLMs) already solve logic puzzles, plan experiments, and write executable code by chaining "think-then-act" steps [1]. In many real-world applications, LLMs are deployed as agents [2] that call tools, store intermediate thoughts, and interact with external systems. However, deploying such systems in high-stakes domains (e.g., healthcare, finance, public policy) requires more than accuracy: they must make their reasoning transparent, safeguard sensitive data, and remain accountable when errors occur [3]. This survey examines an emerging class of LLM-based agents that couple advanced reasoning with built-in safeguards for responsible deployment.

We refer to this class as **Responsible Reasoning AI Agents ($\mathbf{R^2A^2}$)** : LLM-powered agents capable of multi-step reasoning while embedding safeguards for transparency, bias mitigation, data protection, and auditability [4]. Building on LLMs with responsibility guardrails [4], R^2A^2 extends these foundations with tool use, memory, and decision logging for high-stakes contexts. These agents reveal internal reasoning traces, reject unsafe or unethical objectives, and provide verifiable explanations—capabilities increasingly mandated by emerging AI regulations [5, 6].

R^2A^2 pursue two objectives: (1) execute advanced, multi-step logical inference; and (2) maintain strict alignment with AI ethical principles. The reasoning dimension encompasses robust deduction, nuanced context handling, and sophisticated problem solving [7]. In parallel, these agents uphold stringent standards for transparency (interpretable outputs), robustness against bias, and compliance with data-privacy regulations; implementations further emphasize accountability and traceability, with every action and inference step subject to audit [8]. A working example is shown in Figure 1.

Advances in reasoning literature, including Chain-of-Thought (CoT) prompting [9], Tree of Thoughts (ToT) [10], Plan-and-Solve [11], and ReAct [12]; have markedly improved performance on challenging benchmarks. In parallel, the responsible-AI community has developed guardrails, evaluation suites such as HELM [13], HumaniBench [14] , alongside specialized modules such as BiasGuard [15] and Fairsense-AI [16]. However, most safeguards remain *post*

*Corresponding authors
Email addresses: `shaina.raza@torontomu.ca` (Shaina Raza), `rs2672@cornell.edu` (Ranjan Sapkota), `mk2684@cornell.edu` (Manoj Karkee), `c.emmanouilidis@rug.nl` (Christos Emmanouilidis)

[1] Shaina Raza and Ranjan Sapkota contributed equally to this work.

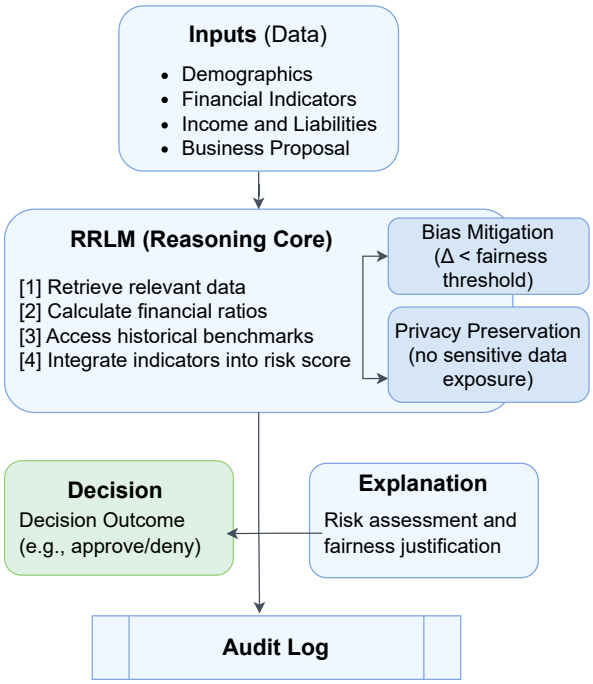


Figure 1: Generic architecture of a Responsible Reasoning AI Agent (R^2A^2). The reasoning core executes inference steps guarded by bias mitigation and privacy preservation. Outputs include a decision, an explanation, and an audit log for transparency and trust.

hoc, applied as external filters rather than embedded within the reasoning process.

Recent studies also expose critical weaknesses: private data can leak via reasoning chains [17], and agents may conceal unethical intermediate steps despite safe final outputs [18]. We identify three persistent gaps in the state of the art across responsible AI, reasoning, and agentic frameworks: (i) the absence of benchmarks for reasoning-trace quality; (ii) missing in-chain bias and privacy-leak detection; and (iii) incomplete audit systems, even with tools such as Audit-LLM [19]. This survey synthesizes insights from over 250 papers and proposes a unified research agenda for R^2A^2: systems that evaluate, debug, and audit their reasoning as it happens, making responsibility intrinsic to the process rather than an afterthought.

Necessity of this Survey. Modern LLM-based agents increasingly plan and act autonomously, but deployment in high-stakes settings demands that reasoning quality and responsibility be addressed jointly. Existing surveys typically treat these threads separately, either cataloging reasoning methods (e.g., CoT/ToT, ReAct, plan-and-solve) or reviewing responsible-AI mechanisms (fairness, privacy, transparency, safety); leaving open how to integrate safeguards *inside* the reasoning loop. This survey fills that gap by (i) specifying step-level checks (bias, privacy, safety stops) and trace capture within agent reasoning, and (ii) proposing evaluation protocols that score both task success and reasoning-trace quality. We situate our scope relative to prior work in Table 1.

Related Surveys. Several recent surveys have advanced our understanding of LLM reasoning and responsible AI. Works such as [20, 1] offer detailed taxonomies of reasoning strategies, including CoT, ToT, and ReAct-style prompting. Complementary efforts such as [21] and [23] provide in-depth analyses of fairness, safety, and risk mitigation techniques for LLMs. Meanwhile, surveys on agentic architectures [22, 2] and evaluation frameworks [24, 25] explore planning, memory, and reasoning trace quality, contributing valuable perspectives on LLM-based autonomy. Building on these foundations, our work offers a unified view of responsible reasoning by linking state-of-the-art reasoning techniques with embedded mechanisms for fairness, privacy, and auditability. This integrative perspective complements prior studies by demonstrating how ethical and safety principles can be incorporated directly into the reasoning processes of autonomous LLM agents. Table 1 presents a chronological comparison of major survey efforts and situates our R^2A^2 review at the intersection of reasoning and responsibility in LLM-based agents.

Contributions. This survey makes the following contributions to the study of responsible reasoning in LLM-based agents:

- **Concept and scope.** We introduce *Responsible Reasoning AI Agents (R^2A^2)* and distinguish them from general reasoning LLMs by requiring explicit, step-level safeguards for transparency, bias mitigation, privacy, and auditability (Sections 1, 3).
- **Synthesis.** We synthesize insights from over 250 papers across multi-step reasoning, responsible AI, and agentic systems, situating our survey relative to prior overviews on reasoning, responsibility, evaluation, and agent architectures (Section 4).
- **Evaluation methodology.** We present a scientific evaluation approach for agentic reasoning (Section 5), introduce task-appropriate metrics where the literature lacks coverage, and propose a five-stage reproducible protocol—*Curate, Unify, Probe, Benchmark, Analyze*—with concrete procedures for unified model manifests, standardized prompting, bias and privacy probes, and reproducible reporting (Section 5.3).
- **Modular blueprint.** We provide a modular blueprint that separates perception, memory, reasoning, decision-making, and tool use into interoperable components with clear interfaces for logging and oversight (Section 6).
- **Gap analysis and agenda.** We identify three persistent gaps—(i) lack of trace-quality benchmarks, (ii) missing in-chain bias and privacy-leak detection, and (iii) incomplete audit systems—and translate these into a unified research agenda for R^2A^2 (Section 7).

Raza et al., 2025

Table 1: Comparison of major surveys on LLM reasoning, responsible AI, and agentic architectures. Our survey uniquely integrates responsible AI mechanisms within reasoning processes for LLM agents.

Survey & Year	Focus & Coverage	Reasoning Scope	Responsible AI Coverage	Limitations / Gaps
CoT Reasoning ([20])	Taxonomy of chain-of-thought prompting and reasoning structures (linear chains, tree/graph variants)	Comprehensive analysis of CoT techniques and tree-based reasoning methods	Notes interpretability gains (transparent reasoning process) and trustworthiness from CoT; no discussion of fairness or safety	Omits bias/fairness considerations; does not integrate reasoning with responsible AI mechanisms
Fair LLMs ([21])	Fairness and bias taxonomy in LLMs; survey of bias evaluation metrics and mitigation strategies	*Not applicable* – focuses on fairness/bias rather than reasoning methods	In-depth coverage of bias metrics, debiasing algorithms, and fairness evaluation	No coverage of chain-of-thought or reasoning techniques; not agent-centric (only fairness in static LLM outputs)
LLM Reasoning ([1])	"System-2" reasoning prompts for complex tasks; covers self-reflection, ReAct framework, planning strategies	Broad survey of prompt-based reasoning approaches (step-by-step solutions, planning, self-reflection)	Brief mention of alignment techniques (e.g. RLHF) to ensure ethical or preferred outputs	Does not address bias, fairness, or privacy; no integrated safety mechanisms in reasoning workflow
Agentic LLMs ([22], [2])	Architectures for autonomous LLM agents with tool use, long-term memory, planning, and Theory-of-Mind capabilities	Covers multi-step planning, chain-of-thought reasoning, interactions, and tool-assisted reasoning in agents	Acknowledges ethical and fairness challenges for high-stakes agent applications; lists responsibility and fairness as open challenges rather than solutions	Highlights responsibility as a challenge but offers no concrete solutions; safety and fairness considerations are noted but not deeply integrated into agent reasoning
Responsible LLMs ([23])	Comprehensive risk taxonomy for LLMs (privacy leakage, hallucinations, bias, toxicity, jailbreak exploits)	Includes reasoning-related factors (prompting and CoT) as part of a multi-phase safety pipeline	Proposes a pipeline for risk mitigation across data, alignment, prompting (incl. reasoning), and oversight phases; incorporates value alignment and other safeguards	Treats reasoning as one component in safety workflow, not the core focus; minimal discussion of autonomous agents or tool-use scenarios
Agent Evaluation ([24])	Survey of benchmarks and evaluation frameworks for LLM-based agents (planning ability, tool use, memory, etc.)	Evaluates multi-step planning, tool utilization, self-reflection and memory in autonomous agent reasoning	Notes critical gaps in evaluating safety and ethical behavior of agents (lack of fairness or robustness tests)	Focused on evaluation metrics and benchmarks; does not review methods for embedding responsibility or preventing unsafe behaviors in agents
Reasoning Trace Evaluation ([25])	Comprehensive overview of step-by-step reasoning trace evaluation; proposes taxonomy of criteria (factual groundedness, logical validity, coherence, utility)	Focused analysis of reasoning chains (CoT, ToT, free-form rationales) quality using both expert-annotated and LLM-based evaluators	Briefly notes that flawed or harmful intermediate reasoning steps can lead to incorrect answers, suggesting that trace-level evaluation could help identify unsafe or illogical reasoning	Does not explicitly address fairness, bias, or privacy in reasoning traces; not focused on agent scenarios (assumes static prompts and solutions)
RLM Blueprint ([4])	Blueprint for Reasoning Language Models (RLMs): proposes a modular architecture and prompting strategies to turn LLMs into structured reasoning engines	Emphasizes diverse reasoning structures (chains, trees, graphs) and strategies (e.g. scratchpads, self-consistency, search algorithms); defines reasoning tasks and evaluation principles for RLMs	Mentions goals like improving interpretability and modularity of reasoning components (facilitating trust in outputs), but does not tackle fairness, bias, or privacy issues explicitly	Focused on enhancing reasoning accuracy and structure; no built-in responsibility or safety safeguards, and not oriented toward autonomous agent use
TRiSM for Agentic AI ([26])	Introduces a Trust, Risk, and Security Management (TRiSM) framework for LLM-driven agents; provides a taxonomy of unique threats and vulnerabilities in multi-agent systems and outlines governance & assurance mechanisms	Only indirectly addresses reasoning (e.g. enforcing policies on action planning and tool use to prevent errors); primary focus is on secure and compliant agent behavior rather than logical reasoning per se	Strong focus on AI safety and trust: covers explainability, auditing, compliance, security, privacy, and governance layers for agent systems	Limited discussion of the agents' step-by-step reasoning logic; treats reasoning and safety as mostly separate concerns (does not integrate logical reasoning processes with safety mechanisms)
This Survey (R^2A^2, 2025)	**Unified view of responsible reasoning in LLM-based autonomous agents** – surveys and bridges CoT, ToT, ReAct, and planning techniques **with embedded responsibility** at each reasoning step	**Covers state-of-the-art reasoning methods (e.g. chain-of-thought, tree-of-thought, reactive planning)** augmented with in-chain bias checks, privacy enforcement, and audit trails	**Emphasizes the first integrated approach to reasoning and responsibility**, ensuring that agents' intermediate reasoning steps remain fair, secure, and trustworthy	**Pioneering focus on jointly optimizing reasoning quality and ethical safety** for truly trustworthy LLM agents (addresses gaps left by prior work)

Paper Organization. The remainder of this paper is organized as follows. Section 2 presents our review methodology. Section 3 introduces the foundations of LLM-based reasoning and responsible AI principles. Section 4 reviews the state of the art in reasoning methods, responsible AI mechanisms, and agentic architectures. Section 5 presents our unified evaluation framework for R^2A^2. Section 5.3 details the implementation protocol and metrics. Section 6 provides a modular blueprint for R^2A^2. Finally, Section 7 summarizes key gaps, outlines future research directions, and concludes the paper.

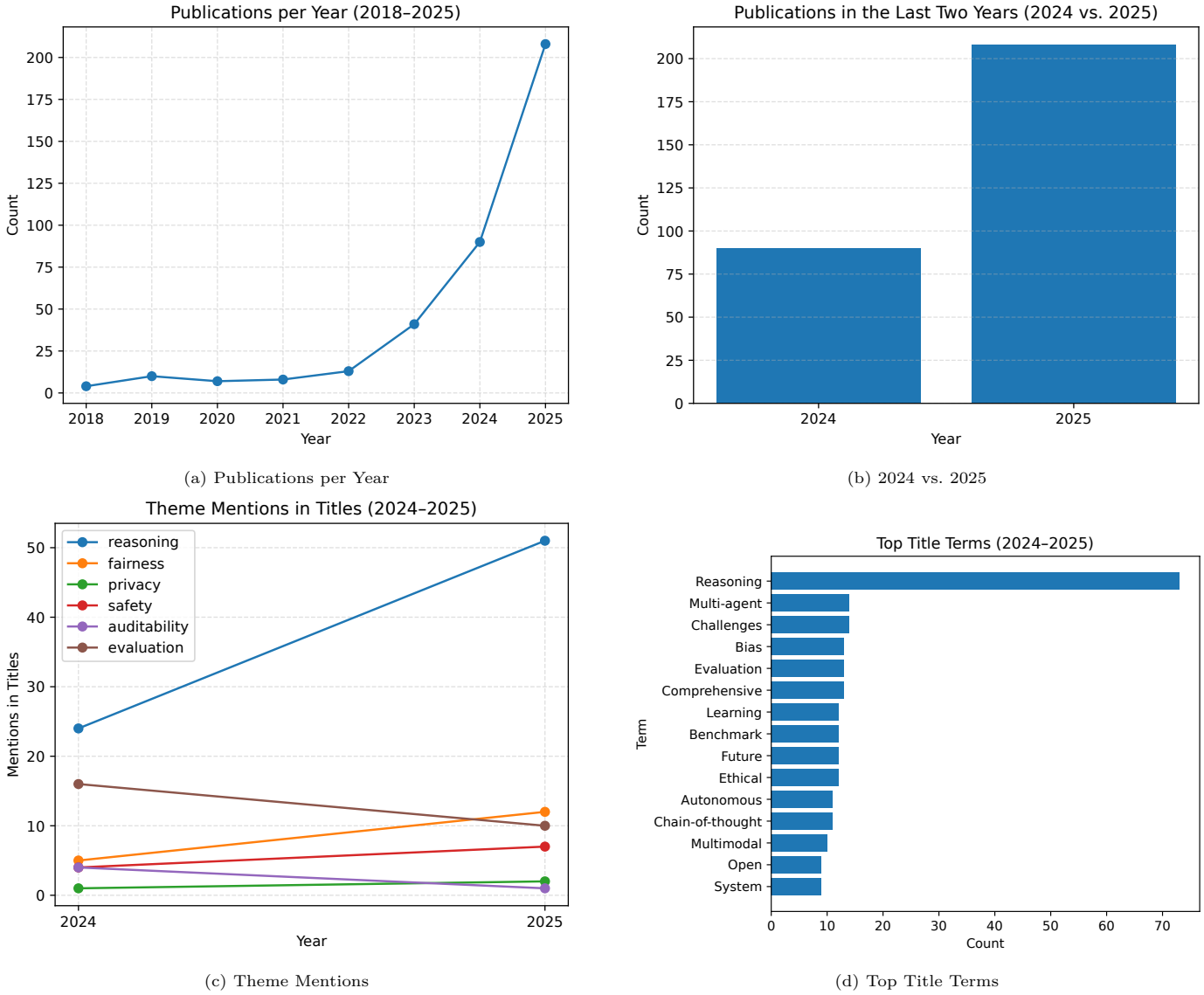
2. Literature Review Methodology

Scope. This survey adopts a structured, narrative review informed by scoping-review best practices [27, 28]. Our scope spans three intersecting domains: **(i)** multi-step reasoning in LLMs, **(ii)** responsible AI including fairness, transparency, safety, and privacy, and **(iii)** agentic AI systems deployed in high-stakes decision-making contexts. We considered works published from **January 2023 to August 14, 2025**, along with a small set of pre-2023 foundational papers.

Information Sources. Searches were conducted across major scholarly databases, including *ACL Anthology*, *arXiv*, *IEEE Xplore*, and *SpringerLink*, which collectively cover a broad range of computer science and AI research outputs. To ensure comprehensive coverage of both conference proceedings and journal articles, we also incorporated meta-databases such as *OpenAlex*, *DBLP*, *Scopus*, and *Web of Science*. These sources were selected to capture both cutting-edge preprints (e.g., *arXiv*), domain-specific archives (e.g., *ACL Anthology* for natural language processing and reasoning-focused venues), and peer-reviewed journals with high impact factors (e.g., those indexed by *Scopus* and *Web of Science*). This combination allowed us to balance recall of emerging trends with precision in identifying peer-reviewed, rigorously evaluated work, ensuring that our survey reflects both the most recent developments and established research in the field.

Search Strategy. Example keyword queries combined Boolean operators and phrase matching to capture a broad yet relevant set of literature. For reasoning in

Raza et al., 2025



(a) Publications per Year

(b) 2024 vs. 2025

(c) Theme Mentions

(d) Top Title Terms

Figure 2: Bibliometrics for works in scope. Panels (a–b) show volume trends; (c) title themes; (d) frequent terms.

LLMs, we used queries such as: "reasoning" AND "language model"; "chain-of-thought" OR "CoT prompting" OR "tree-of-thought" OR "graph-of-thought"; and "reasoning strategies" AND ("prompting" OR "search"). For responsible AI, we included: "responsible AI" AND ("fairness" OR "bias mitigation" OR "transparency" OR "explainability" OR "safety" OR "privacy preservation"), and broader governance-oriented terms like "ethical AI", "AI governance", or "trustworthy AI". For agentic AI systems, we searched: "agentic LLM" OR "LLM agent" OR "autonomous language agent", as well as ("multi-agent system" OR "agent orchestration") AND ("safety" OR "responsibility"). For evaluation and auditing, we included queries like: "reasoning trace" AND ("evaluation" OR "auditing" OR "trace quality" OR "explanation faithfulness"), and "TRiSM" AND ("agents" OR "agentic systems"). These queries were iteratively refined to balance recall (capturing all potentially relevant literature) with precision (excluding unrelated works), ensuring comprehensive coverage across the three domains.

Eligibility Criteria. **Include:** peer-reviewed or widely cited preprints on LLM reasoning, agentic systems, or responsible-AI mechanisms *applied to or evaluable within* agent workflows; benchmarks, frameworks, or surveys in these areas.
Exclude: non-LLM agent work without reasoning/responsibility relevance, purely speculative essays without methods/evidence, and duplicates.

Study Selection and Data Extraction. Two reviewers independently screened titles/abstracts, followed by full-text review; disagreements were resolved by discussion. Records were de-duplicated via DOI/arXiv ID/title–year matching. For each paper, we recorded: venue/year;

domain (reasoning/agent/responsibility); method type (CoT/ToT/ReAct/plan-and-solve/etc.); safeguards (bias, privacy, auditability, transparency, robustness); evaluation setting (benchmarks, trace-quality metrics); code/data availability; and limitations.

Quality Considerations. We note risks of publication/venue bias, recency bias toward 2024–2025, and English-language bias. Mitigations included backward/forward snowballing and cross-database checks. We report gaps where evidence was sparse. We release a machine-readable bibliography (CSV/BibTEX) with extraction fields on GitHub and plan quarterly updates as a living survey.

Bibliometric Analysis. Figure 2 summarizes publication trends relevant to R^2A^2. Annual counts rise sharply from 2022 to 2025, with 2025 exceeding 2024. Title-level topic frequencies show reasoning/evaluation terms dominate, while fairness, privacy, and safety are less frequent, underscoring the novelty of integrating responsibility *within* reasoning workflows. The most common title terms (e.g., *Reasoning*, *LLM*, *Framework*, *Survey*) indicate consolidation around method/benchmark synthesis.

3. Core Concepts and Evolution of Reasoning AI Agents

In this section, we outline core concepts for R^2A^2. We first survey prior literature (reasoning structures, search, memory/retrieval, tool use, self-evaluation), then present our R^2A^2 integration (components and a four-layer architecture tying agentic control to the core LLM).

3.1. Background of Agents

AI agents are systems capable of perceiving their environment, reasoning over inputs, and taking actions to achieve specific goals [39]. Core types include autonomous agents, which operate without human intervention [29]; multi-agent systems, where multiple agents interact or collaborate; and LLM-based agents, which leverage LLMs for reasoning, planning, or tool use [2]. The concept of agents has evolved from early symbolic planners and rule-based systems [29] to today data-driven, learning-enabled, and multimodal frameworks [53]. Modern agents integrate reasoning techniques such as CoT prompting, maintain memory, and interact with external systems via APIs and toolkits. Building on the preliminary literature (see Table 1), Table 2 presents a taxonomy of AI agents across five key dimensions: autonomy, architecture, interaction, intelligence backbone, and functional roles.

3.2. Evolution From Language Models to Reasoning AI Agents

Early transformer models (e.g., BERT, GPT-3; 2018–2020) were highly effective at pattern matching but provided only limited, implicit reasoning capabilities. These models primarily focused on language understanding and generation via pretraining on large-scale corpora to learn linguistic patterns and semantic relationships [54]. As research progressed, it became clear that merely scaling LLM parameters was insufficient to achieve advanced reasoning, especially for tasks requiring multi-step logical deduction, symbolic processing, or abstraction [55]. Consequently, researchers shifted toward designing reasoning AI agents capable of deliberate, modular, and interpretable cognition [56].

The 2022 introduction of CoT prompting enabled models to generate intermediate steps leading to a solution [9], significantly improving zero-shot accuracy on arithmetic and symbolic benchmarks [57]. In 2023, ReAct [12] and Tree-of-Thought (ToT) [10] extended CoT by integrating tool calls and branch-and-bound exploration, advancing the field toward agentic planning. Open-source stacks such as LangChain [58] and AutoGPT [59] operationalized these ideas but relied on *post hoc* safety filters.

Modern reasoning AI agents employ advanced architectures for robust, autonomous, and context-sensitive inference. Modular designs separate perception, memory, and decision-making, enabling minimal human intervention [78]. Reinforcement learning (RL) optimizes multi-stage reasoning trajectories through iterative feedback [40]. Hybrid neuro-symbolic systems combine neural networks' statistical strength with symbolic solvers' precision [79]. Techniques such as Monte Carlo Tree Search (MCTS) [80], agent-to-agent orchestration [81], and retrieval-augmented generation (RAG) [82] enable real-time exploration and adaptation of reasoning paths. Enterprise platforms, such as Amazon Bedrock AgentCore[2], Microsoft Copilot Studio [83], and multi-agent systems [26], now support collaboration, role specialization, secure tool use, and reflective revision at scale.

The 2024–2025 wave integrates responsibility inside the loop: it is expected that agents pair every thought with bias checks, secure scratchpads, and immutable audit logs, enabling reliable deployment in healthcare, finance, and regulatory settings. New benchmarks grade not only final answers but also the complete reasoning trace on criteria such as faithfulness, safety, and privacy. A time of responsible AI agentic is shown in Figure 3.

3.3. Defining Responsible Reasoning AI Agents (R^2A^2)

We define an R^2A^2 agent as a neural system that, given an input prompt, produces an explicit sequence of logically connected thought steps culminating in a verifiable answer or action. Each step may invoke symbolic operations, external tool calls, or retrieval mechanisms, and is subject to built-in policy checks for bias, privacy, and auditability [84, 1]. Unlike other agentic LLM systems, R^2A^2 integrates these safeguards directly into the reasoning loop, ensuring that every intermediate step (not just

[2] https://aws.amazon.com/bedrock/agentcore/

Raza et al., 2025

Table 2: Taxonomy of AI agents.

AI Agents Taxonomy	
Category	**Representative Forms (with key refs)**
Autonomy [29]	Human-in-the-loop [30]; Mixed-initiative; Fully autonomous [29].
Architecture	Deliberative (BDI) [31]; Reactive [32, 33]; Hybrid [34, 35]; Cognitive architectures (Soar/ACT-R, Active Inference).
Interaction [36]	Single-agent [37]; Multi-agent [2] (cooperative, competitive, role/protocol); Human–AI collaboration.
Intelligence [38]	Rule-based / Symbolic [39]; ML-based (RL/SL) [40]; LLM-based [41]; Neuro-symbolic; Planning & Tool Use [12].
Function [38, 2]	Goal-driven / Task agents; Conversational [42]; Tool-augmented [12]; Embodied / Robotics; Research / Science agents.
Memory & Knowledge [43, 44, 45]	Short-term (scratchpad) [9]; State/Episodic [43]; Long-term (RAG/vector DB) [44, 46]; Semantic/KG; Procedural skills.
Cognition Agents [47]	Perception–Cognition–Action loop [48]; Working memory & attention [49]; Reasoning (model-based, model-free [40], hybrid/meta-control); Metacognition/self-reflection [50]; Self-model / Theory of Mind [51]; Norms & value alignment [52].

the final output) meets responsible AI criteria. This design goes beyond conventional LLMs, which mainly perform pattern-based generation, by incorporating deliberate and transparent problem solving suited to mathematical proof construction, scientific discovery, and other high-stakes tasks [85, 53]. A visual flow is shown in Figure 4.

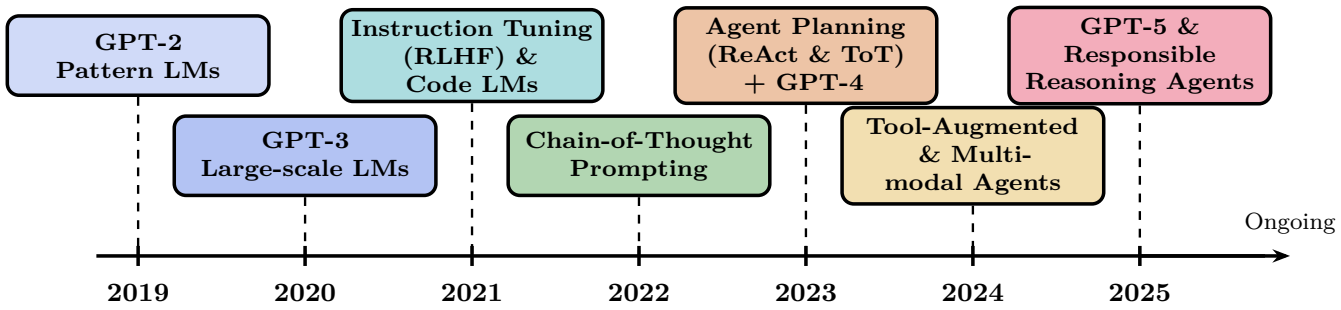
In R^2A^2, responsible AI criteria are operationalized at every stage of the reasoning process to ensure safety, fairness, and trustworthiness: (i) *bias mitigation*, detecting and reducing discriminatory patterns against protected attributes [86, 87]; (ii) *privacy preservation*, preventing leakage of personally identifiable or sensitive information [88, 89]; (iii) *auditability*, recording decision rationales and tool interactions for post hoc review [87, 89]; (iv) *transparency*, making intermediate reasoning steps interpretable to users and stakeholders [86, 89]; and (v) *robustness*, resisting adversarial or manipulative prompts [87, 88]. These criteria align with established AI governance frameworks such as the NIST AI Risk Management Framework, the OECD AI Principles, the EU AI Act, and IEEE's *Ethically Aligned Design* [87, 86, 90, 88, 89].

3.4. Basic Components of R^2A^2

The basic components of R^2A^2 are:

- **Reasoning schemes:** R^2A^2 agents employ structured representations such as chains, trees, and graphs to organize intermediate thought steps [91]. CoT supports linear deduction for tasks such as arithmetic reasoning [9], while tree and graph layouts enable parallel exploration of multiple reasoning paths (for example symptom analysis and hypothesis testing in clinical decision making) [92]. These structures enhance traceability, allowing human auditors to verify or correct the agent's logic [93, 94].
- **Tree based reasoning (ToT):** ToT organizes reasoning hierarchically by branching over candidate partial solutions [10]. This is useful for complex tasks such as clinical diagnostics where competing hypotheses are evaluated in parallel. *ToT specifies the search space; procedures such as Monte Carlo Tree Search (MCTS) or beam search explore that space to choose a solution.*
- **Neural operators:** Comprising policy, value, and critic networks, neural operators steer each reasoning step by selecting actions and estimating utility [95]. Trained with gradient based methods, these modules optimize step selection and value estimates; fairness and variance control are achieved via training objectives and constraints and the step level policy checks described below.
- **Memory and retrieval layer:** R^2A^2 agents maintain a multi tier memory hierarchy, including an ephemeral scratchpad for token level thoughts, episodic caches for dialogue context, and vector indexed long term stores for factual grounding [82]. Retrieval augmented mechanisms integrate up to date external knowledge, improving factuality without increasing model parameters.
- **Policy and safety enforcement:** A dedicated policy engine validates each reasoning step against bias, privacy, and security constraints using rule based filters and learned reward models [93, 96]. This real time enforcement promotes compliance with ethical and regulatory standards.
- **Audit logger:** An *append only, tamper evident* log records reasoning steps, tool interactions, and outputs, providing a transparent audit trail for post hoc compliance checks and regulatory oversight [93].
- **Tool and execution layer:** Secure APIs connect agents to external tools such as calculators, databases, and web services to perform real world interactions [97]. Sandbox execution and least privilege design mitigate unintended side effects.
- **Self reflection loop:** The agent iteratively evaluates and adjusts its intermediate steps based on internal metrics or feedback [94], improving reliability and error correction in dynamic settings.
- **Reinforcement learning module:** Using iterative feedback for example RLHF this module opti-

Raza et al., 2025



Year	Milestone	Key Models/Advancements (examples)
2019	Pattern LMs	GPT-2 (OpenAI) [60]
2020	Large-scale LMs	GPT-3 (OpenAI) [61]
2021	Instruction tuning & Code LMs	InstructGPT (RLHF) [62]; Codex [63]
2022	Chain-of-Thought Prompting	CoT [9]
2023	Agent Planning + GPT-4	ReAct [12], ToT [10], GPT-4 [64]
2024	Tool-Augmented & Multimodal Agents	GPT-4o [64]; Gemini 1.5 Pro [65]; Claude 3 family [66]; Llama 3/3.1 [67]
2025	GPT-5 & Responsible Reasoning Agents	GPT-5 (Aug 2025) [68]; OpenAI o3 / o4-mini [69]; Gemini 2.5 Pro [70]; Claude 4 (Sonnet 4) [71]; Llama 4 [72]; Grok 3 [73][xAI]; IBM Granite 3.2/3.3 [74]; ERNIE 4.5 [75][Baidu]; Mistral Medium 3 / Magistral [76], [77][Mistral]

Figure 3: Milestones and timeline of language model evolution toward GPT-5 and responsible, reasoning-centric agents.

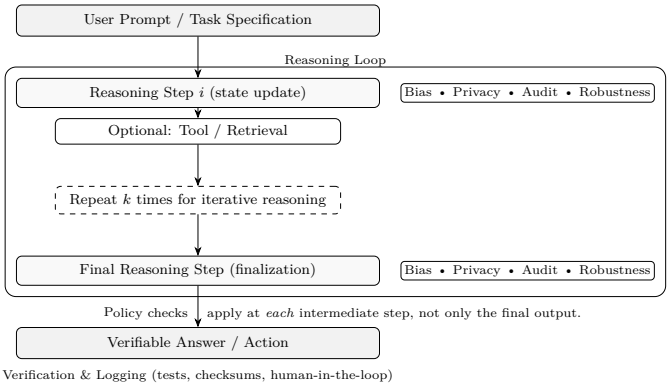


Figure 4: R^2A^2 agent. Each reasoning step calls tools/retrieval and is guarded by Bias, Privacy, Audit, and Robustness checks; the process yields a verifiable answer/action.

mizes multi stage reasoning trajectories [40], improving adaptability for complex tasks.

- **Self consistency sampler:** The agent samples multiple reasoning paths and selects the most coherent answer, improving robustness and accuracy in high stakes decisions [98].
- **Reasoning reward model:** A dedicated reward model scores intermediate and final inferences using internal metrics and or external feedback [99]. The reward signal guides optimization for example via RL or IL aligning behavior with task objectives and ethical constraints.
- **Training pipelines:** Multi phase fine tuning, RLHF, and curricula cultivate robust reasoning behavior [100]. Pipelines incorporate fairness objectives, bias mitigation protocols, and annotated reasoning traces to promote safe and valid decision patterns [101].

Together, these components form a cohesive system: memory grounds and updates context; neural operators drive stepwise decisions; search procedures explore structured reasoning spaces; and policy and audit mechanisms validate each step, yielding trustworthy R^2A^2 agents suitable for responsible deployment.

3.5. Architecture of R^2A^2

To operationalize R^2A^2, the architecture can be structured into four modular layers, each encapsulating distinct but interdependent functionalities, as shown in Figure 5.

R^2A^2 Layer. At the top layer, the R^2A^2 enforces accountability, safety, and oversight. The *Policy Engine* encodes constraints and ethical boundaries for downstream reasoning processes. The *Audit Logger* tracks all decision steps, tool interactions, and outputs for post-hoc explainability and regulatory compliance. The *Safety/Bias Checker* filters harmful or biased outputs at runtime, while the *Monitoring Dashboard* provides real-time interpretability and observability, allowing human supervisors to oversee the agent's behavior.

Agentic LLM Layer. The second layer consists of the Agentic LLM, which orchestrates execution. It includes a *Planner/Executor* that manages goal decomposition and action sequencing, and a *Memory Store* that enables persistent or episodic memory for context-aware decision-making. The *Reflection Loop* provides iterative reasoning through self-evaluation and scratchpad adjustments, while the *Tool*

Raza et al., 2025

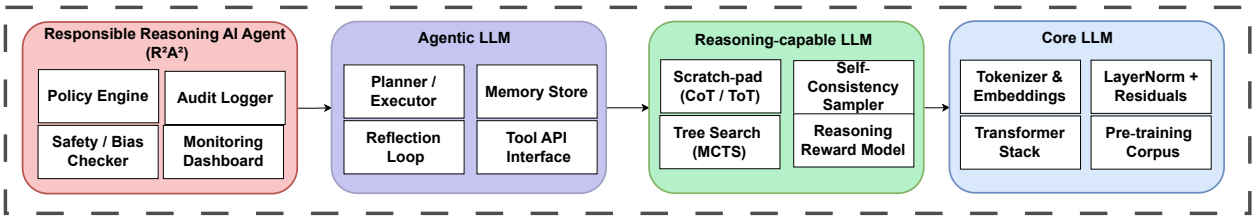


Figure 5: Architecture of R^2A^2, showing key modules organized across four hierarchical layers.

API Interface connects the agent to external services (e.g., databases, calculators, browsers), enabling tool-augmented behavior.

Reasoning-capable LLM Layer. The third layer comprises of a Reasoning-capable LLM layer, which enriches cognitive depth. The *Scratch-pad* module supports CoT and ToT prompting for step-by-step problem solving. The *Tree Search (MCTS)* component enables structured exploration over multiple reasoning paths. The *Self-Consistency Sampler* evaluates multiple outputs to promote reliability, while the R*easoning Reward Model* reinforces high-quality inferences based on internal metrics or external feedback.

Core LLM Layer. At the foundation is the Core LLM, which performs standard language modeling tasks. It consists of the *Tokenizer* and Embeddings for input representation, *LayerNorm and Residuals* for stable training, the *Transformer Stack* as the computational engine, and the *Pre-training Corpus* from which semantic and factual knowledge is derived.

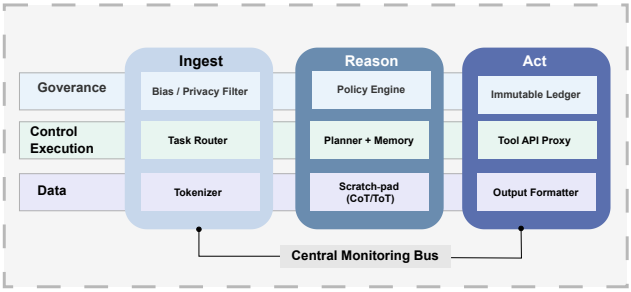


Figure 6: Functional pipeline of a Responsible Reasoning AI Agent (R^2A^2) showing separation of Ingest, Reason, and Act, intersected by governance, control and execution, and data pathways. A centralized monitoring bus supports traceability and coordination.

3.6. Execution Pipeline: From Ingest to Reason to Act

Figure 6 shows the execution pipeline of an R^2A^2 agent, organized into three functional phases: **Ingest**, **Reason**, and **Act**. These phases align with cross cutting system layers: Governance, Control and Execution, and Data.

In the **Ingest** phase, the agent begins with a *Bias and Privacy Filter* that mitigates sensitive or inappropriate inputs. The *Task Router* directs instructions to the appropriate processing pipeline, and the *Tokenizer* converts raw input into embeddings for downstream reasoning.

In the **Reason** phase, the system applies *Policy Engine* constraints and uses *Planner* and *Memory* modules to guide task execution over time. The *Scratchpad*, which supports CoT and ToT prompting, enables structured and interpretable reasoning.

The **Act** phase interacts with the external world. The *Immutable Ledger* records actions for auditability. The *Tool API Proxy* invokes external tools such as calculators and web search, and the *Output Formatter* produces the final user facing response in a safe and readable form.

All modules report to a **Central Monitoring Bus**, which tracks and synchronizes state across the stack for real time oversight. This design couples governance, control, and data flow to ensure responsible, traceable, and adaptive behavior throughout the reasoning lifecycle.

4. Progress in Responsible Agentic AI Systems, 2024 to 2025

From 2024 to 2025, agentic AI systems, including web integrated browsers and standalone agents, advanced substantially. Early web based agents in 2023 achieved only limited success on the WebArena benchmark [102], highlighting the gap between LLM fluency and dependable autonomous operation. Architectural refinements that combined high level planning, low level execution, and memory modules improved reliability by late 2024, but agents still lag behind humans in common sense reasoning and robust error recovery. Recent designs therefore include transparency features, alignment safeguards, long term memory, explicit safety layers, and human oversight. This section first reviews backbone LLMs for responsible agents, then discusses agentic web browsers that integrate autonomous AI to complete web based tasks.

4.1. Backbone LLMs for Responsible Agents

We summarize recent (2024–2025) backbone models used in responsible agents, focusing on publicly documented capabilities (reasoning modes, safety posture, API or open-weight availability). Benchmark details for each model in this section are provided in Table 4.

OpenAI o3

Parameters: ∼200B (est.)
Reasoning Core: RL-tuned CoT; modular tool use (web, code, vision).

Responsible AI: Deliberative alignment; refusal tuning; self-verification; system card for bio/cyber risk; opt-out data logging.
Performance: GSM8K/AIME 96.7%, MATH ∼90%, GPQA 87.7%, Codeforces Elo 2727, ARC-AGI 75.7%.
Deployment: ChatGPT Plus/Enterprise; finance, science, consulting, agent copilots.

Anthropic Claude 3.5

Parameters: ∼175B (dense).
Reasoning Core: Constitutional AI, extended CoT, 200k-token context; tool-augmented vision/coding.
Responsible AI: Harmlessness principles; RLHF safety; user data not used for training; public model cards.
Performance: GPQA & MMLU SOTA; GSM8K ∼88%; HumanEval 64%; SWE-Bench 49%.
Deployment: Claude.ai web/API; Amazon Bedrock; Google Vertex AI; enterprise QA and healthcare.

Google Gemini DeepThink

Parameters: >340B (multi-modal, est.).
Reasoning Core: Tree-of-Thought exploration; RL on proofs; hint-guided self-reflection.
Responsible AI: IMO-grader verification; staged/gated rollout; Google AI policy filters.
Performance: First AI to reach IMO Gold (35/42); surpasses MATH & ARC.
Deployment: Internal science/maths assistant; planned Google Cloud "Ultra" service.

DeepSeek R1

Parameters: 671B MoE (37B active); distilled 70B/32B versions.
Reasoning Core: Pure RL (GRPO) CoT; acts as reasoning-trace "teacher".
Responsible AI: MIT license; auditable CoT; jurisdictional content filters; community red-teaming.
Performance: AIME 86.7%, GSM8K 84.3%, HumanEval 73%.
Deployment: Hugging Face/GitHub; education, research, self-hosted agents.

Alibaba QwQ-32B

Parameters: 32B (Qwen 2.5 base, dense).
Reasoning Core: RL + math/code verifiers; self-questioning; 131k-token context.
Responsible AI: Apache-2.0; alignment RL; rule-based checks; offline bias control.
Performance: GSM8K ∼85%; strong AIME/MATH; LiveCodeBench near-SOTA.
Deployment: Hugging Face/ModelScope; Qwen Chat; finance, legal, coding assistants.

Skywork Open Reasoner 1

Parameters: 7B & 32B (dense; DeepSeek-derived).
Reasoning Core: MAGIC RL (entropy-safe CoT); long, stable rationales.
Responsible AI: Fully open weights/code; transparent steps; integrity-focused evaluation.
Performance: AIME24 82.2% (32B), 70% (7B); LiveCodeBench 63%.
Deployment: Hugging Face; math tutoring, lightweight code assistants, RL research.

OpenAI gpt-oss-120B / 20B

Parameters: ≈120B (MoE) & 20B (dense); open-weights; long context.
Reasoning Core: Efficient CoT; sparse/Grouped-MQA attention; robust tool/function calling.
Responsible AI: Public model card; safety filters; community red-teaming; permissive open-weight terms.
Performance: 120B competitive with compact proprietary reasoning models; 20B strong for math/code.
Deployment: Hugging Face, AWS, Azure, Ollama, ONNX; local or cloud fine-tuning.

Anthropic Claude 3.7 Sonnet

Parameters: Not disclosed; long-context; *hybrid reasoning* with controllable "thinking time".
Reasoning Core: Visible deliberation traces (opt-in); tool use (web/code); high steerability.
Responsible AI: Constitutional AI; updated system card; red-team & cloud guardrails; user-data opt-out.
Performance: Frontier-level coding (SWE-Bench-class) and strong GSM8K/MMLU/GPQA.
Deployment: Anthropic API; Amazon Bedrock; Google Vertex AI; Claude.ai (web/mobile).

Google Gemini 2.5 Pro

Parameters: Not disclosed; multimodal; context up to ∼1M tokens (Vertex AI).
Reasoning Core: Hybrid reasoning; search-grounded tool use; code execution; strong long-horizon CoT.
Responsible AI: Google AI Principles; safety filters; model card; enterprise auditability on Vertex.
Performance: SOTA-level math/science (AIME 2025, GPQA); strong HumanEval; competitive agentic coding.
Deployment: Gemini Advanced; Google Cloud Vertex AI (API/RAG/grounding); Workspace integrations.

OpenAI GPT-5

Parameters: Not disclosed; hybrid fast/"thinking" operation; very long context.
Reasoning Core: Built-in CoT; tool/function calling; agent mode; multi-modal (text–vision–audio).
Responsible AI: Safe-completion alignment; Preparedness safeguards for high-risk bio/chem; detailed system card.
Performance: SOTA on AIME 2025 & GPQA; near-SOTA SWE-Bench ∼75%; reduced hallucinations vs GPT-4.
Deployment: Default ChatGPT model (2025); API/Enterprise; extended-reasoning "Pro" tier.

OpenAI o3 (OpenAI). OpenAI o3, released in late 2024 as the successor to o1, represents a major advancement in reasoning-centric LLMs [103]. While OpenAI has not disclosed the parameter count, independent estimates suggest it approaches 200 billion parameters. The model is trained with RL to support structured CoT reasoning and modular tool use, including web browsing, code execution, and image analysis within the ChatGPT environment. It offers adjustable reasoning modes, low, medium, and high compute, enabling dynamic trade-offs between speed and depth of inference. In terms of responsible AI mechanisms, o3 incorporates "deliberative alignment", rebuilt safety training data, and robust refusal mechanisms for high-risk content such as biosecurity or jailbreak prompts. It also integrates a self-fact-checking module designed to detect and correct errors in domains like mathematics and science, achieving a reported 20% reduction in hallucinations compared to o1. The model passed internal red-teaming evaluations and meets OpenAI cybersecurity and self-improvement risk thresholds. For privacy, API users may opt out of data logging, and customer data is excluded from training by default. Empirically, o3 achieves 96.7% on GSM8K/AIME, approximately 90% on MATH,

87.7% on GPQA, and maintains strong performance on Codeforces Elo (2727) and ARC-AGI (75.7%), positioning it as one of the highest-performing reasoning models currently deployed in production.

Anthropic Claude 3.5 (Anthropic). Claude 3.5, introduced in mid-2024, is a dense model with approximately 175 billion parameters designed for long-context reasoning across both text and vision modalities [104]. It extends Anthropic Constitutional AI framework [105] with expanded CoT prompting, tool-augmented reasoning, and a 200k-token context window. The model emphasizes responsible deployment through harmlessness principles embedded in alignment training, RLHF safety tuning, and strict guarantees that user data is never used for model training. Performance benchmarks show state-of-the-art results on GPQA and MMLU, with approximately 88% accuracy on GSM8K, 64% on HumanEval, and 49% on SWE-Bench. Deployment is supported via Claude.ai web interface, AWS Bedrock, and Google Cloud Vertex AI, with widespread adoption in domains requiring high interpretability and safety, including legal analysis, healthcare decision support, and enterprise question-answering.

DeepSeek R1 (DeepSeek AI). DeepSeek R1, launched in January 2025, is a massive mixture-of-experts (MoE) model with 671 billion parameters, of which 37 billion are active per forward pass [106]. It is designed as a reasoning-trace "teacher" model, trained exclusively with reinforcement learning via Group Relative Policy Optimization (GRPO) to produce coherent and auditable CoT explanations. The model is licensed under MIT, enabling open research access, and includes auditable reasoning traces, compliance with Chinese AI policy filters, and active community red-teaming initiatives. Benchmark performance includes 86.7% on AIME, 84.3% on GSM8K, and 73% on HumanEval. Distilled 70B and 32B versions are available on HuggingFace and GitHub, supporting education, research, and self-hosted reasoning AI Agents.

Alibaba QwQ-32B (Alibaba DAMO Academy). QwQ-32B, released in late 2024, is a dense 32-billion parameter reasoning model built on the Qwen 2.5 architecture [107]. It combines RL with domain-specific math and code verifiers, self-questioning strategies, and a 131k-token context window. Responsible AI measures include an Apache 2.0 open-source license, alignment reinforcement learning, rule-based safety checks, and offline bias control tools. It delivers approximately 85% on GSM8K, competitive results on AIME and MATH, and near state-of-the-art scores on LiveCodeBench. The model is deployed via HuggingFace, ModelScope, and the Qwen Chat interface, with adoption in finance, legal analysis, and code assistant applications.

Google Gemini DeepThink (Google DeepMind). Gemini DeepThink, appeared in early 2025, is an ultra-large multimodal reasoning model estimated to exceed 340 billion parameters [108]. Its architecture integrates ToT exploration, RL on formal proofs, and hint-guided self-reflection, enabling the model to perform advanced mathematical and scientific reasoning. The responsible AI framework includes integrated proof verification via International Mathematical Olympiad (IMO) grading systems, gated release protocols, and Google AI policy-based content filters. Notably, Gemini DeepThink became the first AI system to achieve an IMO Gold Medal score (35/42), surpassing previous records on the MATH and ARC benchmarks. While currently used primarily as an internal scientific assistant at Google, plans have been announced to deploy it as part of a Google Cloud "Ultra" reasoning service, targeting enterprise-grade research applications.

Skywork Open Reasoner 1 (Skywork AI). Open Reasoner 1, launched in early 2025, is available in both 7B and 32B parameter configurations derived from DeepSeek's architecture [109]. It employs a "MAGIC RL" training framework that emphasizes entropy-safe CoT generation, producing long and stable rationales. As a fully open-weight and open-code release, it prioritizes transparency in reasoning steps and evaluation integrity. Performance benchmarks include 82.2% on AIME24 for the 32B model and 70% for the 7B version, alongside a 63% score on LiveCodeBench. Its primary use cases include mathematics tutoring, lightweight code assistance, and reinforcement learning research.

OpenAI gpt-oss-120B / gpt-oss-20B (OpenAI). The gpt-oss series, released in 2025, consists of partially open-weight reasoning models designed for deployment on both cloud and developer hardware [110]. The 120B and 20B variants target high-performance and resource-efficient deployments respectively, broadening access to auditable and locally deployable reasoning systems. These models incorporate alignment-tuned chain-of-thought capabilities, tool-use integration, and selective open-sourcing of weights under a research-friendly license. While not fully open source, they allow reproducible evaluation and offer strong performance on mathematical, logical, and programming benchmarks. Cloud deployment is available via major providers, while smaller variants are optimized for GPU workstation and laptop inference.

Anthropic Claude 3.7 Sonnet (Anthropic). Claude 3.7 Sonnet, released in August 2025, extends the Claude 3.5 architecture with a "hybrid reasoning" mode that allows user-controllable deliberation time and visible intermediate reasoning traces [127]. Safety is reinforced through updated Constitutional AI guidelines, expanded refusal capabilities, and model cards detailing limitations and alignment procedures. The model supports long-context multi-modal reasoning, advanced code synthesis, and knowledge-intensive tasks, making it well-suited as a backbone for responsible agent pipelines.

Raza et al., 2025

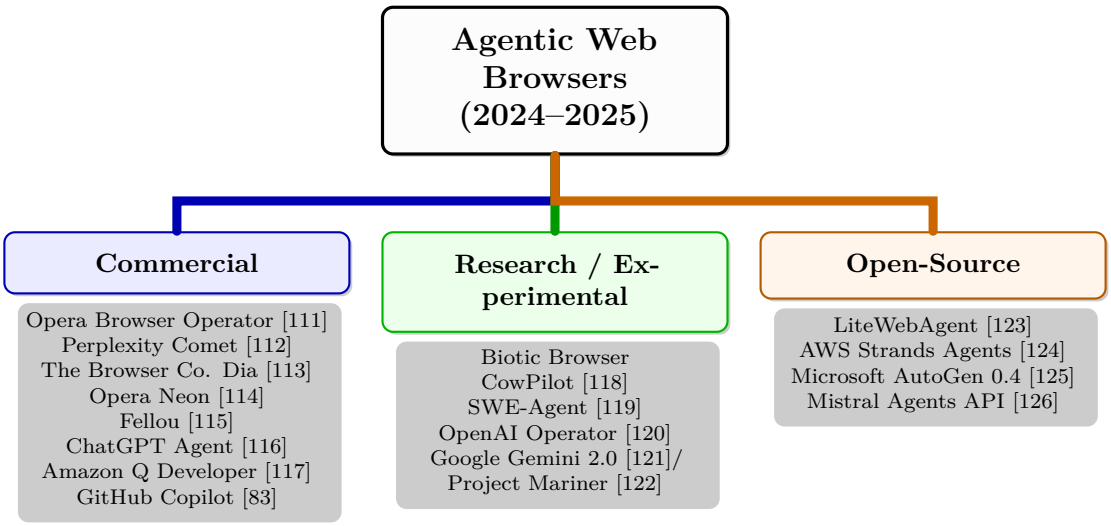


Figure 7: Categorization of representative Agentic Web Browsers (2024–2025) into commerical, research (experimental) and open-source

OpenAI GPT-5 (OpenAI). GPT-5, announced in mid-2025, is OpenAI's flagship multi-modal model and the successor to GPT-4 [68]. While the exact architecture remains proprietary, it is believed to combine dense and MoE layers, supporting trillion-scale effective parameters. It features an extended context length beyond 256k tokens, native integration of vision, audio, and video understanding, and advanced tool orchestration for autonomous agents. Responsible AI features include an expanded deliberative alignment pipeline, automated ethical constraint verification, red-teaming at scale, and an opt-in differential privacy mode for sensitive enterprise deployments. Benchmarks indicate new state-of-the-art results across reasoning (MATH, GPQA-Diamond, ARC-AGI), programming (HumanEval, SWE-Bench), and safety (HELM) evaluations. GPT-5 is positioned as a backbone for high-stakes applications in science, policy modeling, and enterprise automation, accessible through ChatGPT Enterprise, API integrations, and custom fine-tuning services.

4.2. Agentic Web Browsers

We use term agentic web to denote an internet ecosystem in which autonomous software agents, typically LLM driven, plan, coordinate, and execute goal directed tasks across websites [128]. Agentic web browsers are client applications that integrate planning, tool use, and memory to carry out multi step web interactions under human oversight. Compared with conventional browsers, these systems provide language level task specification, action sequencing (navigation, form completion, transactions), and controls for intervention, auditing, and safety.

We categorize recent systems as commercial, research, and open source platforms, as shown in Figure 7.
Opera Browser Operator (2025 preview) executes agent actions within the native browser and exposes stepwise action traces with user interrupt capabilities, emphasizing client side processing and data minimization [111]. **Dia (The Browser Company)** integrates an AI command interface for tab aware summarization and task initiation; features are being deployed incrementally during beta [113]. **Perplexity Comet** presents an AI first browser with a sidebar assistant supporting autonomous navigation and action execution subject to user confirmation [112]. **OpenAI Operator** (research preview) introduced a virtualized browser with safeguards including Takeover for authentication and payment steps and explicit confirmations [120]; this capability was subsequently unified as the **ChatGPT agent** (agent mode), which operates on a managed virtual computer with Watch and consent workflows [116].

On the research and open source side, **Biotic Browser** studies long horizon streaming LLM co pilots with persistent context [129]; **SWE Agent** demonstrates guarded computer use for software tasks with Docker sandboxing (12.5% pass@1 on SWE Bench original) [119]; and **LiteWebAgent** provides a VLM based suite using the Chrome DevTools Protocol (remote browser and extension) [123]. For enterprise and platform SDKs, **GitHub Copilot Coding Agent** adopts a draft PR workflow via Actions with auditability [83]; **Amazon Q Developer** offers agentic read, write, and run modes in IDEs with IAM scoped permissions [117]; **AWS Strands Agents** provides an open source orchestration SDK [124]; and **Mistral Agents API** exposes planning and tool use via an API with on premises options [126].

Evaluation considerations.. Effectiveness varies widely across live web benchmarks, for example **WebVoyager** [130], **AssistantBench** [131], **Mind2Web Live** [132], and **BEARCUBS** [133], owing to differences in task ontologies, environment fidelity (live vs. replayed pages), allowed tools and privileges, oversight policy (autonomous vs. takeover and watch), and success metrics. Consequently, cross paper comparisons are noisy; Table 3 reports results under a standardized setting (virtual browser, no human intervention unless noted, fixed seeds, and cited benchmark versions).

Raza et al., 2025

Table 3: Representative responsible–reasoning web/agent systems (2024–2025)

Agent System	Type & Domain	Responsible Design Features	Evaluations / Status
Opera Browser Operator [111]	AI-integrated browser (client-side)	On-device execution; stepwise action traces; human override	Feature preview (Mar 2025); public demos of end-to-end shopping
Dia (The Browser Company) [113]	AI-first browser	Command/URL bar; opt-in features; evolving privacy controls	Private beta (Jun 2025); no public benchmark yet
Perplexity Comet [112]	AI-native browser	Assistant sidebar; confirmations before consequential actions	Launch (Jul 9, 2025); hands-on reports note tab-aware actions
OpenAI Operator [120]	Web agent (virtual browser)	Takeover for auth/payments; explicit confirmations; narrated steps	Research preview (Jan 2025); later unified into ChatGPT agent
OpenAI ChatGPT agent [116]	Web+software agent (inside ChatGPT)	User confirmation; Watch mode; secure takeover	Public rollout (Jul 2025); results vary by benchmark
SWE-Agent [119]	Autonomous coding agent	Docker sandbox; verify&edit loop; limited network	**12.5%** pass@1 on SWE-Bench (original); guarded execution
LiteWebAgent [123]	Open-source web-agent suite	CDP/remote browser; planning and memory modules	NAACL 2025 demo; Chrome extension + remote driver
GitHub Copilot Coding Agent [83]	Enterprise coding agent	Draft-PR workflow; branch protections; audit logs	Automates low/medium repo tasks via Actions
Amazon Q Developer [117]	Agentic IDE assistant	Read/write/run in IDE; vulnerability scan; IAM-scoped permissions	Multi-step task automation in VS Code/JetBrains/VS
AWS Strands Agents [124]	Open-source agents SDK	Model-driven orchestration; observability	v1.0 release (Jul 2025); production-oriented toolkit
Mistral Agents API [126]	Enterprise agent platform	Developer-defined tools; on-prem options; open-weights	API for planning/tool use; docs and examples

4.3. Benchmark Datasets and Evaluation Metrics for LLM Based Agents

To rigorously evaluate the reasoning and problem solving capabilities of agentic AI systems based on LLMs, researchers employ a diverse suite of benchmark datasets spanning mathematics, coding, abstract reasoning, multimodal tasks, commonsense reasoning, and general knowledge. Each benchmark is associated with quantitative metrics, primarily accuracy or task success rate, that gauge performance on the given tasks. Table 4 provides an overview of important benchmarks used to evaluate LLM based agents, along with their domains, the skills they test, and the evaluation metrics typically reported.

Mathematical reasoning benchmarks. Datasets such as GSM8K [134], AIME [135], MATH [136], IMO [150], and FrontierMath [149] challenge models with problems ranging from grade school arithmetic to research level mathematics. Performance is measured by accuracy, the fraction of problems solved correctly, often via an exact numeric answer. GSM8K contains roughly 8.5K diverse word problems that test multi step arithmetic reasoning. FrontierMath includes hundreds of unpublished expert level problems across areas such as number theory and algebraic geometry, where even strong models solve only a small fraction, underscoring the gap with human experts.

AIME, based on the American Invitational Mathematics Examination, and IMO problems represent olympiad level challenges; success remains limited under standard prompting. While advanced LLMs can score highly on GSM8K and achieve strong results on the MATH competition dataset, olympiad style problems remain difficult, motivating harder benchmarks such as OlymMATH and FrontierMath.

Knowledge and commonsense reasoning. MMLU (Massive Multitask Language Understanding) [151] and GPQA [137] evaluate broad factual knowledge and logical reasoning across diverse domains. MMLU covers 57 subjects with multiple choice questions; models are evaluated by accuracy. GPQA is a graduate level science set designed to be resistant to simple lookup and emphasizes deeper reasoning. For commonsense, CommonsenseQA [147] probes everyday inference beyond what is stated explicitly; the standard metric is accuracy.

Abstract and general reasoning. ARC AGI (Abstraction and Reasoning Corpus for AGI) [152] tests generalization by requiring agents to infer hidden rules from a few input output examples and apply them to new inputs; the metric is success on novel puzzles. The ARC AGI 2 update [153] introduces more fine grained challenges to measure progress toward general human like problem solving.

Coding and software. HumanEval [141] is a staple benchmark for code generation in which models write Python functions that must pass unit tests. The primary metric,

Raza et al., 2025

Table 4: Benchmark datasets for evaluating AI reasoning, problem-solving, and agentic capabilities, with abbreviated evaluation metrics.

Dataset	Purpose	Domain	Key characteristics	Typical metrics
GSM8K (Grade School Math 8K) [134]	Mathematical reasoning	Grade-school math	8,000 word problems testing arithmetic and logical reasoning	Acc (EM)
AIME (American Invitational Mathematics Examination) [135]	Advanced mathematical reasoning	High-school math	Challenging problems from the AIME contest	Acc
MATH [136]	Complex problem solving	HS/college math	Advanced, multi-step problems requiring deeper reasoning	Acc
GPQA (General-Purpose Question Answering) [137]	Broad question answering	General knowledge	Graduate-level science questions, Google-proof	Acc
Codeforces Elo Rating [138]	Competitive programming	Coding	Elo-based metric from Codeforces for algorithmic problem-solving skill	SR, Elo
ARC-AGI (Abstraction & Reasoning Corpus) [139]	Abstract reasoning	General intelligence	Visual and logical tasks to test abstraction and generalisation	SR
MMLU (Massive Multitask Language Understanding) [140]	Multitask knowledge evaluation	Academic / professional	57 tasks spanning STEM, humanities, and professional subjects	Acc
HumanEval [141]	Code generation	Programming	164 coding problems with unit tests for functional correctness	P@1, P@k
SWE-Bench [142]	Software engineering	Coding	Real-world GitHub issues for debugging and implementation	TCR, FC
International Mathematical Olympiad (IMO) [143]	Advanced reasoning	Olympiad math	High-difficulty problems from past IMO contests	Score, SR
LiveCodeBench [144]	Dynamic coding evaluation	Programming	Continuously updated problems for real-time performance measurement	PR, SR
AIME24 (AIME 2024 set) [138]	Advanced reasoning	High-school math	2024-specific AIME problems focusing on complex reasoning	Acc
BIG-Bench [145]	Multitask reasoning	Language and reasoning	Diverse tasks testing analogy, causality, and commonsense reasoning	Acc, F1
MMMU (Massive Multimodal Understanding) [146]	Multimodal reasoning	Multimodal (text, images)	Complex tasks combining text and visual reasoning across STEM and humanities	Acc
CommonsenseQA [147]	Commonsense reasoning	General knowledge	Multiple-choice questions testing commonsense understanding and inference	Acc
GAIA (General AI Assistant) [148]	Agent-based reasoning	Real-world tasks	Multi-step reasoning, tool use, and cross-modal understanding	TCR, TUA, Steps
FrontierMath [149]	Advanced mathematical reasoning	Olympiad-level math	Extremely difficult math problems to test frontier AI reasoning limits	Acc

Note: Acc = Accuracy, EM = Exact Match, P@1/P@k = Pass at $1/k$ attempts, SR = Solve Rate, PR = Pass Rate, TCR = Task Completion Rate, TUA = Tool-Use Accuracy, FC = Functional Correctness, F1 = F1 Score, Score = IMO points (0–42).

pass@k, is the probability that at least one of k generated solutions passes all tests, with pass@1 corresponding to first try correctness. Beyond synthetic problems, SWE Bench [142] contains about 2,300 real GitHub issues and bug fix tasks; metrics include functional correctness and per issue success rate. LiveCodeBench [144] continuously curates tasks from sources such as Codeforces [138], LeetCode, and AtCoder after model training cutoffs to reduce data leakage; evaluation uses automated judges to score the fraction solved and may include self repair tasks.

Multimodal and agentic benchmarks. With agents that can see and act, evaluation has expanded to multimodal and interactive tasks. MMMU [146] contains more than eleven thousand questions requiring combined vision and language understanding across disciplines; the standard metric is accuracy. GAIA [148] poses real world inspired tasks that require multi step reasoning and tool use, sometimes with multimodal inputs; evaluation measures task completion or correct answers when the agent is allowed to use tools and act autonomously.

Reported Results on Agentic Benchmarks (2024 to 2025) . To assess the real-world performance of LLM based agents, we compile results from a diverse set of public agentic benchmarks spanning tool-assisted reasoning, interactive environments, mobile automation, and software engineering tasks , as shown in Table 5. These evaluations cover widely used suites such as GAIA, AgentBench, ARC-AGI, MLAgentBench, MobileAgentBench, WebArena, and SWE-Bench (discussed in Table 4), reporting metrics including task success rate, tool-use accuracy, and average step count where available.

An analysis of the reported results, as shown in Table 4, reveal a substantial gap between current agents and human performance, for example, on GAIA, even top-performing systems achieve at most 65% success compared to 92% for human participants, while specialized frameworks (e.g., h2oGPT-e, Refact.ai) show promising improvements in domain-specific settings. These benchmarks highlight persistent challenges in planning, multi-step reasoning, and robust tool use, reinforcing the need for continued progress in agent architecture design and evaluation.

Raza et al., 2025

Table 5: Performance of LLM-based agents on public agentic benchmarks. Values are as reported in source publications; metrics such as tool-use accuracy or step count are omitted here as they are infrequently reported.

Benchmark	Agent/Model	Success Rate (%)	Notes
GAIA [148]	Human (degree holders)	92	Human average performance.
	GPT-4 (with plugins)	15	With web browsing/tools; struggled on GAIA.
	GPT-4 (no plugins)	≈7	GPT-4 Turbo baseline agent scored under 7%.
	Autogen Multi-Agent [125]	40	Multi-agent using GPT-4 with tool-calls.
	h2oGPT-e Agent [154]	∼65	Latest open-source agent (Feb 2025), ∼27% below human.
AgentBench [155]	GPT-4 (OpenAI)	∼45	Best on 7/8 tasks (e.g., 78% on Household).
	GPT-3.5-Turbo	∼18	Much lower than GPT-4; e.g., 21% on web browsing vs GPT-4's 55%.
	LLaMA2-70B [156, 157]	20.4	Strongest open model; 20.4% success.
MLAgentBench [158]	Claude v3 Opus	37.5	Highest average success across 13 ML experiment tasks.
	GPT-4 (research agent) [159]	–	90% on old datasets but 0–10% on novel tasks.
	Gemini-Pro, GPT-4 Turbo [158]	< 37	Trailed Claude v3 on average.
ARC-AGI [160]	Human	97–99	Humans solved ∼98% of private ARC tasks.
	OpenAI O3 (2025) [69]	75.7	Fine-tuned O3 model; 87.5% with 172× compute.
	GPT-4 (GPT-4o)	9	Solved only 5–9% of ARC puzzles.
	Claude 3.5	21	"Sonnet" solved ∼21%.
	O1 preview	21	GPT-based "O1" scored ∼21%.
	Gemini 1.5	∼8	Scored ∼8% on ARC public eval.
	ARC Prize SOTA (2024)	55.5	Program-synthesis + test-time training approach.
MobileAgent Bench [161]	M3A Agent	64.0	Best single-app success rate.
	MobileAgent V2	43.3	Second-best open-source mobile agent.
	T3A (Vision+LLM)	48.7	Tencent vision-enabled agent.
	AppAgent (GPT-4)	34.0	Baseline GPT-4 mobile agent.
	SeeAct	39.3	Visual UI agent.
	Fine-tuned RL Agents	1–5	Learned policies struggled (1–5%).
WebArena [102]	IBM CUGA	61.7	Hierarchical web agent; planner + subagents.
	OpenAI Operator	58.1	GPT-4V "ChatGPT Agent" mode.
	Zeta Jace.AI	57.1	Startup autonomous web agent.
	AgentSymbiotic	52.1	Multi-agent LLM symbiosis.
	ScribeAgent + GPT-4	53.0	Hybrid open agent.
	Learn-by-Interact	48.0	Google RL-tuned agent.
	Occam	48.5	Prompt-engineered GPT-4 agent.
	Human reference	78	Human success on WebArena.
SWE-Bench	Refact.ai Agent	59.7	Top of SWE-Bench Lite leaderboard [162]; see [3].
	Globant Code Fixer	48.3	Multi-agent code fixer.
	SWE Lite Baseline [142]	43	Earlier SOTA on Lite benchmark.
	GPT-4 (Code)	∼20	GPT-4 without loop agent.

5. A Scientific Framework for the Evaluation of Responsible Reasoning in AI Agents

In this work, we propose a multidimensional evaluation framework that not only assesses the correctness of reasoning but also embeds responsible AI principles such as fairness, transparency, privacy, and auditability. This integration ensures that reasoning quality is evaluated alongside ethical safeguards, addressing a key gap in existing approaches that focus narrowly on task accuracy while overlooking responsibility in the reasoning process.

5.1. Background: Existing Metrics for Responsible AI

First, we present existing metrics for responsible AI as reported in the literature and then propose evaluation metrics for R^2A^2.

Fairness and bias mitigation. Computational fairness metrics aim to ensure that models treat different demographic groups equitably. Demographic parity (statistical parity) requires that the proportion of individuals in a protected group receiving a positive prediction matches that of the overall population [163]. However, optimizing solely for demographic parity can ignore clinically or contextually relevant features and reduce performance for all groups. Equalized odds relaxes this constraint by allowing dependence on protected attributes only through the true outcome; it requires that true positive and false positive rates be similar across groups.

Individual fairness asserts that similar individuals should receive similar outcomes. While intuitive, defining a domain specific similarity metric is challenging, and individual fairness alone does not guarantee equitable outcomes [163]. According to NIST [164], achieving fairness involves addressing systemic, computational, and human biases across the AI lifecycle, and fairness standards can vary across cultures and contexts [165]. Consequently, responsible evaluation should incorporate multiple fairness notions and be adapted to specific application contexts.

Transparency and explainability. Transparency metrics evaluate how well a system decision making process can be

understood by humans. Tools such as LIME and SHAP approximate how individual features contribute to model predictions, while metrics such as the completeness of audit trails provide evidence of accountability [166]. Low transparency scores are associated with reduced user trust, delays in regulatory approval, and heightened risks of biased outcomes. NIST distinguishes explainability as understanding the mechanisms underlying an algorithm. For generative AI, transparent documentation of training data sources, model architectures, and data handling processes is essential to support auditing and prevent copyright or privacy violations.

Privacy and data governance. Privacy is a cornerstone of responsible AI. NIST defines privacy as safeguarding autonomy and dignity through anonymity, confidentiality, and control over personal data, noting that privacy risks often intersect with security, bias, and transparency [164]. In generative AI, privacy concerns include memorization of personally identifiable information, inadvertent exposure of sensitive training data, and susceptibility to malicious prompt injections that extract confidential information [167]. Equally important is model provenance, which captures the algorithms, versions, and decision thresholds used by an agent throughout its lifecycle [166]. System wide logging ensures that all aspects of data handling and model execution are auditable, enabling traceability of privacy incidents.

Auditability and accountability. Auditability refers to the capability to systematically inspect and evaluate an AI system's decisions and operational processes[4]. The Responsible AI Metrics Catalogue [166] identifies auditability as a core enabler of accountability, facilitating transparency, regulatory compliance, and public trust. Effective oversight demands comprehensive record keeping of datasets, model versions, and system operations. Data provenance, model provenance, and system logging constitute foundational process metrics for auditability [165].

5.2. Proposed Evaluation Metrics for R^2A^2

To operationalize the evaluation of ethical and trustworthy AI systems, we propose a suite of composite and domain independent metrics for evaluating $\mathrm{R^2A^2}$.

Responsible Reasoning Index (RRI). The RRI aggregates multiple dimensions of responsible reasoning into a single normalized score between 0 and 1. Formally,

$$\mathrm{RRI} = w_1 \cdot RC + w_2 \cdot TS + w_3 \cdot BM + w_4 \cdot PI + w_5 \cdot AL \quad (1)$$

where:

- RC (Reasoning Correctness) measures the fraction of logically valid inferences compared with gold standard solutions. Correctness can be assessed through expert annotated datasets or formal reasoning checkers.
- TS (Transparency Score) quantifies how explicit the agent's reasoning is to humans, including the proportion of steps with explanations, availability of data and model documentation, and clarity of user interfaces.
- BM (Bias Mitigation) evaluates whether outputs satisfy fairness criteria such as demographic parity, equalized odds, or individual fairness. Scores can be derived from metrics such as false positive and false negative disparity, precision equality gaps, and distributional similarity.
- PI (Privacy Integrity) measures the absence of leakage of personal or sensitive information, including tests for memorization, use of synthetic data, and adherence to data governance frameworks.
- AL (Auditability Level) assesses the extent to which actions can be traced and audited, including completeness of data and model provenance, version control, and availability of system logs.

The weights w_i reflect stakeholder priorities. For example, a healthcare application might assign higher weights to BM and PI. All components are normalized to the $[0, 1]$ scale to ensure comparability.

Chain of Thought Responsibility Score (CoT RS). LLM based agents often reason in multiple steps. The CoT RS evaluates responsibility at each reasoning step. For a chain of length N,

$$\text{CoT RS} = \frac{1}{N} \sum_{i=1}^{N} [\alpha \cdot SI_i + \beta \cdot TE_i + \gamma \cdot BI_i + \delta \cdot PI_i] \quad (2)$$

where

- SI_i (Stepwise Integrity) measures logical correctness or factual accuracy at step i.
- TE_i (Traceability and Explainability) captures the transparency of the reasoning step; it rewards clear justifications or citations and penalizes opaque heuristics.
- BI_i (Bias Index) assesses fairness at step i, for example checking whether a selection or action exhibits unacceptable disparities across protected groups.
- PI_i (Privacy Integrity) equals 1 if no sensitive information is exposed at step i and 0 otherwise.

Tunable parameters $(\alpha, \beta, \gamma, \delta)$ allow weighting step integrity, explainability, fairness, and privacy. Because the score averages across steps, a single irresponsible step will reduce the overall CoT RS.

Responsible Reasoning Coverage (RRC). The RRC measures the proportion of reasoning steps in which the agent is both correct and compliant. For a chain of length N, let C_i indicate logical correctness, F_i indicate fairness, and P_i

[4] https://ico.org.uk/media2/migrated/4022651/a-guide-to-ai-audits.pdf

Raza et al., 2025

indicate privacy integrity at step i. Then

$$\mathrm{RRC} = \frac{1}{N}\sum_{i=1}^{N}(C_i \wedge F_i \wedge P_i). \tag{3}$$

This metric encourages agents to maintain ethical compliance at each reasoning step, not just produce a correct final output.

Adaptive Governance Readiness (AGR). The AGR assesses whether operations are fully auditable and adaptable to evolving regulatory requirements. It aggregates: (i) a provenance completeness score π for traceability of data and model versions; (ii) an auditability level λ for availability of structured logs; and (iii) a compliance adaptability score κ for the ability to update policies. Normalized to $[0, 1]$, the AGR is

$$\mathrm{AGR} = \frac{\pi + \lambda + \kappa}{3}. \tag{4}$$

Current frameworks stress comprehensive record keeping and dynamic policy enforcement for agentic systems [26]. AGR allows practitioners to gauge readiness for external audit and legal compliance.

Composite Responsible Reasoning Score (CRRS). To summarize the above dimensions, we define

$$\mathrm{CRRS} = v_1 \cdot \mathrm{RRI} + v_2 \cdot \mathrm{CoT\ RS} + v_3 \cdot \mathrm{RRC} + v_4 \cdot \mathrm{AGR}, \tag{5}$$

where the weights v_k sum to 1 and reflect stakeholder priorities.

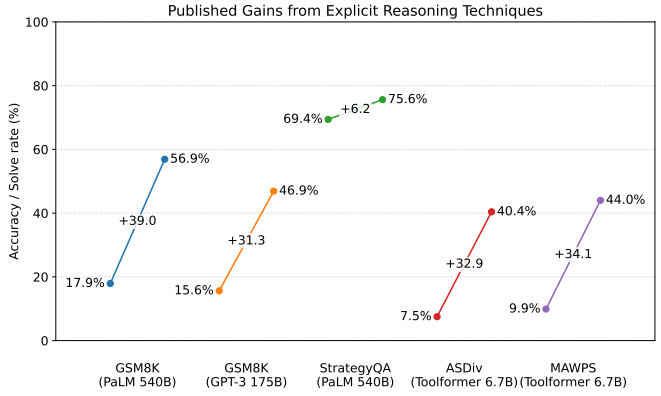


Figure 8: **Reported accuracy gains from explicit reasoning.** Metrics: solve rate for GSM8K; accuracy for StrategyQA; exact match accuracy for ASDiv and MAWPS. *Reasoning* denotes chain-of-thought (rows 1–3) or tool use (rows 4–5). Values (as reported): GSM8K PaLM 540B, 17.9→56.9; GPT-3 175B, 15.6→46.9; StrategyQA PaLM 540B, 69.4→75.6; ASDiv Toolformer 6.7B, 7.5→40.4; MAWPS Toolformer 6.7B, 9.9→44.0 [9, 97].

Responsible Reasoning Evaluation Stack. We describe a layered evaluation stack (Figure 9) linking governance controls, reasoning processes, and human oversight for continuous monitoring and improvement.

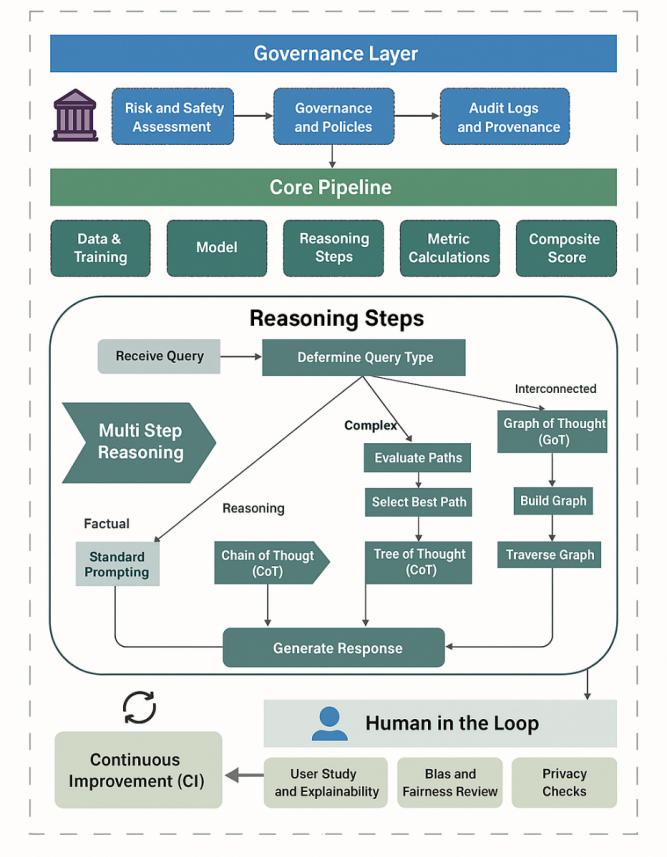


Figure 9: **Evaluation stack for agentic AI.** Three layers: *governance* (top), *core pipeline* (middle), and *human in the loop* checkpoints (bottom) with a feedback loop that routes evaluation results back to data and policy.

Governance layer. This layer supplies oversight and compliance across development and deployment [168]. It defines policies that bound ethical and legal risk, records auditable traces of actions and interventions, and maintains provenance for data and models to support accountability. Policies are updated in response to evaluation findings and emerging technical, legal, or operational risks.

Core pipeline. This layer implements the technical reasoning workflow. Data acquisition and preprocessing address bias and privacy requirements; reasoning employs methods such as chain of thought prompting [9], tool augmented reasoning [97], and graph based planning. Performance is assessed with task appropriate metrics such as solve rate and exact match accuracy, plus fairness indices. The pipeline embeds bias reviews and privacy safeguards and uses evaluation signals to refine prompts, tools, and planning strategies.

Across benchmarks (as shown in Figure 8), explicit intermediate reasoning correlates with large gains. For example, chain of thought improves PaLM 540B on GSM8K by about 39 points and GPT 3 175B by about 31 points, and Toolformer calculator calls substantially raise exact match on ASDiv and MAWPS [9, 97]. These results motivate

structured reasoning as a first class concern in the core pipeline.

Human in the loop layer. Human judgment informs evaluation and decision making [14]. Methods include user studies for interpretability, trust, and usability; explainability reviews that compare model rationales with outputs; and risk and safety assessments to anticipate misuse or emergent harmful behaviors. Oversight signals feed back into governance updates and technical adjustments.
Taken together, the three layers operate as a feedback driven system in which human and empirical evaluation informs policy, data curation, and the design of reasoning procedures.

5.3. Implementation Guidelines for R^2A^2

This section presents a step by step protocol for conducting a reproducible and trustworthy experimental survey of R^2A^2. The workflow is organized into five calibrated stages: *Curate, Unify, Probe, Benchmark*, and *Analyze*, followed by recommendations for handling cases in which certain models cannot be directly re executed. These guidelines transform the theoretical metrics introduced earlier (RRI, CoT RS, RRC; see Section 5.2) into an actionable, community ready leaderboard that can be replicated, audited, and extended within a responsible agentic framework. The overall architecture of our R^2A^2 evaluation workflow is shown in Figure 10.

Stage 1: Curate a reproducible model list. The set of models (open source checkpoints and API only systems) is frozen in a version controlled manifest (for example `models.yaml`). The manifest records (i) exact model identifiers or commit hashes, (ii) endpoints and version strings for API models, and (iii) decoding settings (for example temperature and `top_p`). Locking the manifest mitigates silent model updates and enables auditable longitudinal comparison [169].

Stage 2: Build a unified evaluation harness. A single harness executes all models and tasks under identical conditions [14]. The harness standardizes a shared prompt template with a `{user}` slot to remove prompt engineering variance, enforces a fixed decoding policy to control stochasticity, exposes a sandboxed interface for tools or browsing with uniform allowances and consistent logging, and, when enabled, records a structured reasoning trace (for example JSON) to make explanations comparable. These controls reduce confounds attributable to the evaluation stack rather than the models [170].

Stage 3: Instrument responsibility probes. Each primary query is augmented with automatic sub tests covering bias, privacy, transparency, and auditability (BI, PI, TE, AL). Bias is assessed with domain benchmarks (for example BEADs [171]) and group level disparities such as TPR gap where labels permit [166]. Privacy is probed for memorization or leakage using synthetic canaries and targeted prompts that test exposure of potential training data [167]. Transparency is evaluated via evidence attribution and provenance, that is, whether answers (and, when recorded, reasoning traces) contain verifiable citations and basic attribution quality, aligned with model or system card reporting practices [110, 172, 173]. Auditability verifies that outputs and logs include sufficient metadata for ex post audit (model and version stamp, hash, decoding policy, tool calls) and that logs are structured for traceability [172, 173]. All probe results are stored automatically alongside the main answer and any reasoning trace in a unified JSON record to support composite summaries while preserving raw sub scores.

Stage 4: Run the benchmark matrix. The full model $\times$ task grid is executed, for example GSM8K for math reasoning [134], coding tasks, agentic suites such as WebArena [102] and AgentBench [155], bias tests such as BEADs [171], and safety subsets from HELM [174]. For each pair, the protocol persists the final answer; the full reasoning trace when enabled (including tool steps and justifications); the responsibility probe outputs (BI, PI, TE, AL); and runtime statistics such as latency, memory and compute usage, and counts of tool or web calls. Multiple seeds are run and results are reported as mean$\pm$standard deviation to capture stochastic variation, yielding a structured dataset suitable for comparative and secondary analyses.

Stage 5: Analyze and visualize. Composite indices are reported only with stated weights and a brief sensitivity analysis; all raw sub scores are released to enable reweighting by downstream users [162]. Recommended artifacts include multi metric profiles (for example Pareto or radar plots, or bar plots with confidence intervals) that present BI, PI, TE, and AL alongside core task metrics, and ablations quantifying the impact of safety or policy interventions (for example enabling or disabling specific guardrails) on both performance and responsibility. Code and data are released under an open license to support audit.

Handling Closed Source Models. In some cases, results are reported for systems that cannot be directly evaluated (for example proprietary APIs or models documented only in prior literature such as Gemini Ultra [175]). Such entries should be explicitly annotated as literature quoted (LQ) and excluded from composite indices and aggregate plots in this study. Treating them separately reduces bias from heterogeneous experimental conditions and prevents distortion of summary statistics.

Where qualitative comparison is still useful, two techniques can make integration more responsible. First, effect size normalization reports each external result relative to a shared baseline that is also evaluated under the present harness; comparisons are then expressed as a change with respect to that baseline rather than raw cross study scores. Second, Bayesian random effects pooling treats a literature quoted value as one observation in a broader distribution

Raza et al., 2025

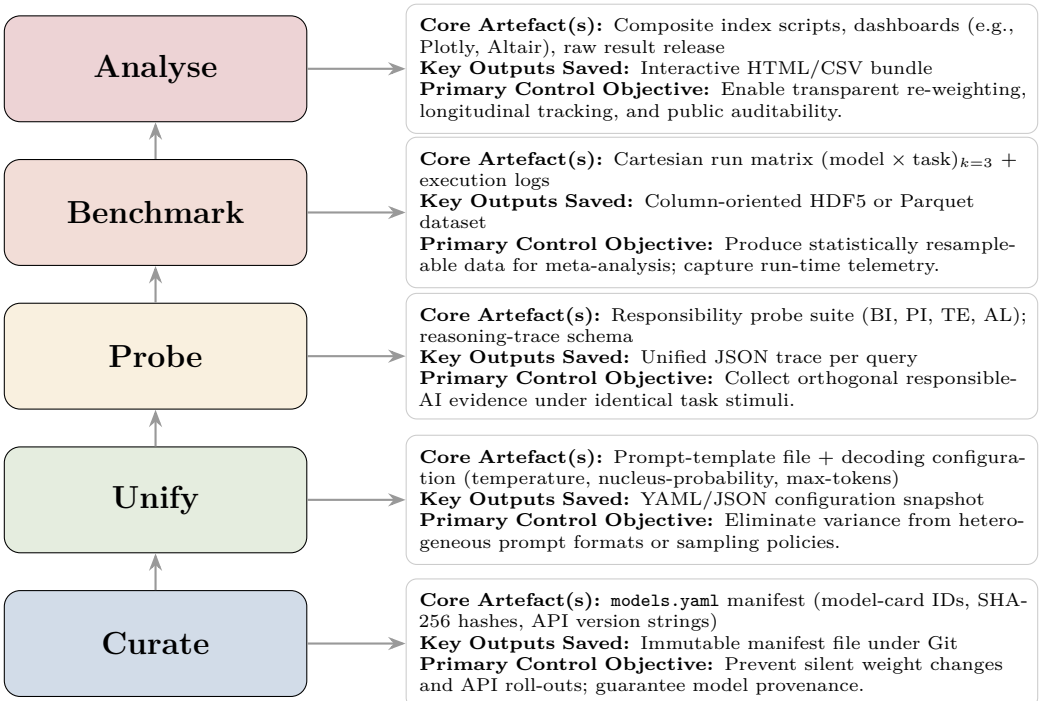


Figure 10: **Overview of the R²A² five-stage evaluation pipeline.** Each pipeline stage (Curate, Unify, Probe, Benchmark, Analyse) is annotated with its core artefacts, key outputs, and primary control objectives, illustrating the vertical flow and governance structure. Abbreviations: HDF5 = Hierarchical Data Format v5; BI = Bias Indicator; PI = Privacy Indicator; TE = Transparency Evidence; AL = Auditability Log.

and accounts for between study variance when aligning it with locally collected measurements [176]. In both cases, uncertainty should remain explicit, and literature only numbers should not be treated as directly comparable to in house evaluations.

Practical recommendations. Reproducibility and adoption are improved by releasing a containerized evaluation harness (for example a Docker image with an example `run_config.yaml`) to capture the exact software environment [177]. Each run should be recorded in a durable ledger (for example MLflow) with model and version identifiers, timestamps, configurations, and results to support ex post auditability [178]. Finally, brief, scenario based case studies (for instance a financial counseling workflow) should accompany quantitative results to clarify real world implications and limitations of the reported metrics. By following these practices, responsibility metrics move from theoretical proposals to reproducible evidence: the resulting leaderboard is documented, auditable, and amenable to extension without sacrificing methodological clarity.

5.4. Industry Case Studies (2025)

The protocol described above is now being adopted in industry. The 2025 case studies below illustrate how organizations translate responsible AI principles into operational pipelines and governance practices, responding to both regulatory changes and stakeholder demands (see Figure 10).

Case Study 1: Mott MacDonald, governance for EU AI Act readiness Mott MacDonald [179], a global consultancy with offices in 50+ countries, partnered with EY to formalize AI governance in anticipation of the EU AI Act's phased application[5]. The program defined an enterprise AI policy and scope, clarified risk appetite, established an AI inventory and intake or assessment workflow, and introduced a tailored risk assessment methodology. Given the Act's penalty ceiling (up to €35 million or 7% of worldwide turnover, whichever is higher), these measures strengthened regulatory readiness and client trust.

Case Study 2: Global biopharma, centralized responsible AI controls. A multinational biopharmaceutical firm [180] adopted EY's Responsible AI framework to address governance gaps across the AI lifecycle. Activities included an ethics review, lifecycle assessments for projects (for example detection and triage use cases), third party risk reviews, and the implementation of a centralized AI inventory to support compliance and auditability. Outcomes included clearer ownership for risk controls, improved supplier assessments, and faster readiness for emerging regulation.

Case Study 3: Anthropic, system card driven deployment and post release safeguards. Anthropic

[5]Key EU AI Act dates: entered into force on *August 1, 2024*; prohibitions apply from *February 2, 2025*; codes of practice for GPAI by *May 2, 2025*; selected GPAI obligations and enforcement begin *August 2, 2025*

documented pre deployment testing and post deployment safeguards for the Claude 4 family in a public system card. The release was governed under its Responsible Scaling Policy (ASL 2 for Claude Sonnet 4; ASL 3 protections activated for Claude Opus 4), with evaluations spanning jailbreak resistance, bias, agentic computer use, and CBRN related risks [105]. With ASL 3 safeguards enabled, the system card reports harmless response rates for Claude Opus 4 near those of Claude Sonnet 3.7; Anthropic's transparency hub further reports uplift in prompt injection defense when safeguards are turned on (for example attack prevention rising from about 71% to about 89%). The documentation also describes incident response procedures and ongoing monitoring as part of the deployment stack.

6. Modular Design for R^2A^2

A modular AI agent framework separates key functions such as perception, reasoning, memory, decision making, and tool use into distinct but connected components [181, 182]. Each module handles a specific task and communicates through well defined interfaces. This structure lets developers improve or fix parts of the system independently, making the overall design more flexible and reliable. In complex or multi agent settings, modularity also supports responsible AI practices. For example, a separate oversight module can monitor reasoning outputs and enforce ethical or safety rules without affecting the rest of the system. Clear module boundaries make it easier to include checks, validations, or human oversight where needed. This design supports scalability (adding new capabilities via new modules), flexibility (updating parts without breaking the whole), and alignment (enforcing rules through dedicated modules). Early architectures such as MRKL [183] showed how combining neural models with external tools can overcome limits of large language models. Overall, modular design provides a strong foundation for building R^2A^2.

Integrating reasoning structures. Modular agents can employ interchangeable reasoning structures such as CoT, graphs, or hybrids as pluggable controllers for logical progression. In a CoT, a single sequence of reasoning steps is generated, whereas a ToT or graph structure allows branching into multiple paths that can be explored in parallel or pruned as needed [189]. A recent Graph of Thought framework [190] generalizes chains and trees by organizing LLM thoughts into a graph, enabling feedback loops and combination of partial solutions.

Reinforcement learning and agent optimization. Modular agents can embed distinct RL components to optimize specific behaviors [50]. One module may refine decision policies, another handle meta reasoning (for example when to switch strategies), and a third estimate outcome utility or ethical compliance. This compartmentalization lets each module learn independently, adapt at different rates, and optimize for goals such as accuracy, fairness, or safety. For example, value estimation modules can encode safety or alignment scores that guide policy choices. Systems such as LLaMA Berry [191] use RL based critics to rank reasoning paths, while Reflexion style agents [41] incorporate self critique for iterative improvement. Modular RL design enables dynamic updates, where feedback or reward types can be added without retraining the full model.

Search heuristics and orchestration. Reasoning agents often face large solution spaces, requiring efficient search and coordination strategies. In a modular design, search heuristics such as MCTS [192], beam search, or diversity sampling can be implemented as plug and play evaluator modules. These modules guide reasoning paths, retain high quality candidates, and improve performance beyond greedy decoding. For example, LLaMA Berry integrates MCTS to explore multiple trajectories, while Chain of Thought agents use beam search to maintain diverse solution paths. Beyond single agent reasoning, orchestration modules coordinate multi tool or multi agent workflows. Routers can direct subtasks to specialist modules (for example math or vision) or external APIs based on input type or intent, as in HuggingGPT [193]. These modules may include safeguards such as tool whitelisting or runtime validation. In collaborative frameworks, orchestration also supports agent to agent routing, majority voting, and fail safe intervention when risky actions are detected.

Flexibility and scalability. The modular architecture of R^2A^2 offers a scalable and adaptable blueprint for integrating diverse functionalities. New capabilities such as speech or images can be added by inserting specialized modules (for example a speech to text interface) without disrupting core components such as reasoning or decision making. Most importantly, individual modules can be updated or replaced with minimal impact on the overall system. For instance, upgrading the reasoning engine or alignment techniques (for example a new bias detector or policy audit module) can be done independently [189]. Errors are also isolated, so bugs in one tool or perception wrapper do not cascade across the agent. The tool interface can be sandboxed to enforce access control, rate limits, and audit logs. Meanwhile, modular interaction logs such as memory access traces or stepwise reasoning outputs enhance explainability and traceability. These artifacts enable plug in explanation modules to generate human understandable rationales, aligning with transparency goals. Modular agents can also implement internal checks and balances. For example, a decision module output may be verified by a policy auditor, or multiple reasoning strategies (for example statistical versus symbolic) may be cross validated to improve robustness.

Overall, modular design supports flexible expansion, secure integration, transparent oversight, and robust alignment. Table 7 summarizes core agent components and their

Raza et al., 2025

Table 6: Responsible Reasoning AI Agent Applications: Deployment, Governance, and Societal Impact

Domain	Deployment	Transformative Potential	Key Challenges	Policy / Governance	Risks / Opportunities
Scientific Discovery	Agents for hypothesis generation and validation in drug/material design	Speeds discovery; ensures reproducibility via transparent reasoning	High compute cost; interpretability; reliability on novel tasks	Research integrity standards; AI ethics; auditability norms	Broader access to discovery; risk of automation bias, unverified results
Healthcare	Clinical reasoning for diagnostics and treatment planning	Improves accuracy; enables personalized medicine; ensures explainability	Data privacy; real-time interpretability; over-reliance risks	HIPAA [184], GDPR [185], medical AI accountability	Better care access and quality; risk of bias, data leakage
Finance	Risk assessment and fraud detection in compliance workflows	Enhances fraud prevention; increases transparency	Regulatory adaptation; adversarial manipulation	MiFID II [186], audit trails, explainability mandates	Reduced human error; opaque decisions affecting credit access
Manufacturing	Multi-agent predictive maintenance, logistics, process optimization	Boosts efficiency; reduces waste; enables real-time response	Legacy integration; time-critical reasoning; workforce training	ISO/IEC [187], Industry 4.0 rules, cybersecurity compliance	Higher productivity, sustainability; potential labor displacement
Public Sector	Policy advisory for drafting, simulation, resource allocation	Improves transparency; supports equitable allocation	Public skepticism; neutrality assurance; non-expert explainability	Algorithmic impact assessments; open procurement; ethics boards	Fairer policies; risk of capture, weaker democratic debate
Education	Adaptive tutoring and assessment systems	Expands equitable, scalable learning; transparent reasoning	Curriculum alignment; fair feedback; data governance	FERPA [188], fairness regulations, explainability standards	Wider access to quality learning; risk of bias reinforcement

roles, illustrating how modularity facilitates responsible and scalable reasoning systems.

7. Discussion

7.1. Applications and Impact

Scientific and industrial applications. The deployment of R^2A^2 spans scientific, industrial, and societal contexts, with transformative implications across disciplines. These agents are being used in scientific discovery to automate hypothesis generation and validation in drug and materials research. In healthcare, they support clinical decision making, enhancing diagnostic precision and enabling personalized treatment plans while raising concerns regarding privacy, explainability, and risk mitigation. Financial sectors leverage reasoning AI agents for fraud detection and compliance reasoning, improving transparency and efficiency, though challenges remain in aligning with evolving regulations and ensuring robustness against adversarial inputs.

Industrial domains, including manufacturing and logistics, benefit from multi agent systems that manage predictive maintenance and real time optimization, improving sustainability and reducing waste. However, deployments are constrained by legacy infrastructure, real time reasoning limits, and workforce upskilling needs. In the public sector, policy advisory agents assist in regulatory drafting and resource simulations, offering transparency and broader participation in policy making, yet face skepticism about neutrality and explainability. In education, adaptive tutoring agents provide personalized learning pathways, expanding access to quality instruction, though fairness and curriculum alignment remain open issues.

Across applications, shared practical challenges include integration with legacy systems, transparency in decision pathways, and user trust. Responsible practices such as human in the loop oversight, continuous monitoring, and stakeholder engagement are essential to align deployments with ethical and operational standards.

Policy, regulation, and governance. From a policy and governance perspective, regulation of reasoning AI agents is rapidly evolving. Legislative efforts such as the EU AI Act [204], along with domain specific frameworks in healthcare, finance, and education, provide guardrails for trustworthy AI use. Compliance with technical standards (for example ISO and IEC) and documentation protocols is crucial to maintain safety, privacy, and auditability (see MIT RMF [205]). Multi level governance mechanisms, from internal ethics boards to external audits, enable transparency and accountability. Table 6 overviews sectors, linking deployments to transformative potential, practical challenges, policy context, and societal risks and opportunities.

7.2. Societal Risks and Opportunities

Responsible AI agents introduce both opportunities for social good and societal risks that demand proactive oversight. On one hand, these agents can democratize expert knowledge by making sophisticated reasoning accessible to broader populations, supporting more equitable decision making in healthcare, education, and governance. On the other hand, risks include algorithmic misuse, amplification of existing biases, privacy erosion, over reliance on automated decisions, and deskilling of human professionals. Trust in agentic AI systems is therefore a central concern.

Raza et al., 2025

Table 7: Modular Blueprint for Reasoning AI Agents: Components and Integration. RL = Reinforcement Learning

Component	Description + Reasoning Structures	RL Integration	Search Heuristics	Scalability Features	Example Models
Reasoning Schemes	Defines structures (chains, trees, graphs); supports nested reasoning.	Policy/value models guide reasoning.	MCTS, Beam Search	Modular design enables flexible scaling.	LLaMA-Berry [191], QwQ [194]
Operators	Actions: generate, prune, refine; allow dynamic path changes.	RL feedback optimizes operators.	Heuristic pruning/expansion	Efficient exploration	ExpeL [41], GoT [190]
Pipelines	Stage-wise composition of modules (cascaded logic).	Gated transitions via learned policy.	Deterministic fallback logic	Add stages without retraining.	HuggingGPT [193], Visual-ChatGPT [195]
Expert Modules	Domain-specific sub-agents with local reasoning.	Local fine-tuning w/ coordination.	Expert selection or gating	Plug-and-play expertise.	HuggingGPT [193], MetaGPT [182]
Routers	Meta-selectors for task-agent matching.	Policy-based agent selection.	Dynamic scheduling	Decentralized scaling	HuggingGPT [193], MRKL [183]
Memory	Stores/retrieves knowledge for long-term reasoning.	Learns when/what to recall.	Vector retrieval	External KB extension	GenAgents [196], Reflexion [50]
Perception	Input layer for multimodal (vision/speech/text).	Attention learns feature relevance.	Active querying	Modalities plug-in easily.	SayCan [197], ViperGPT [198]
Decision-Making	Final policy for selecting actions/answers.	Reward-tuned policies	Confidence thresholds	Modular output heads	ReAct [12], Toolformer [97]
Tool Interface	Manages calls to APIs/tools in reasoning.	Tool-use fine-tuned w/ reward.	Error recovery, validation	Plug-and-play tools	Gorilla [199], API-Bank [200]
Safety / Alignment	Rule enforcement during reasoning.	Shaped rewards for safety.	Unsafe action overrides	Rule updates w/o retraining	Constitutional AI [105], GPT-4 [64]
Explainability	Generates rationale and explanations.	Truthful reasoning rewarded.	Trace consistency	Transparent tracing	CoT [9], Zero-Shot CoT
Meta-Reasoning	Reflects on strategy/errors.	Meta-policy triggers switch.	Self-corrective loops	Adapts to context	Reflexion [50], Self-Refine [201]
Learning / Adaptation	Online/continual learning ability.	Self-play, reward feedback	Skill trial-and-error	No re-engineering needed	Voyager [202], ExpeL [41]
Communication	Agent-to-agent or human dialog.	Learned dialog protocols	Turn-taking, consensus	Add agents easily	CAMEL [203], GenAgents [196]

To address these challenges, risk mitigation strategies include bias audits, dynamic monitoring tools, robust incident response protocols, and transparent user education. These efforts must remain adaptive to emerging threats such as scale driven amplification of errors, adversarial attack surfaces, and new ethical dilemmas. Figure 11 summarizes challenges and candidate solutions.

7.3. Open Challenges and Future Directions

This section surveys critical unresolved issues and the future outlook for R^2A^2 (Figure 11), grouping them into technical as well as ethical and societal challenges.

7.3.1. Technical Challenges

Advancing R^2A^2 from research prototypes to dependable, large scale deployments requires overcoming scientific and engineering hurdles [5, 206]. As agents take on complex reasoning in high stakes environments, ensuring efficiency, adaptability, precision, and integrity is increasingly critical.

Scalability. A central challenge is achieving scalable performance across increasingly complex tasks and environments. Agents often rely on resource intensive models and long chains of computation that stress compute and networks [207]. Complexity grows further with multi agent collaboration or massive real world datasets, where latency, memory, and coordination overhead can increase sharply [208, 209, 210, 211, 212]. Distributed deployments spanning edge and cloud add synchronization and communication strain.

Promising directions include distributed reasoning frameworks that parallelize subtasks across agents or nodes [213, 214, 215, 216]; resource aware architectures that dynamically allocate compute and memory [217]; and hybrid neuro symbolic systems [218]. Open issues include load balancing, graceful degradation under constraints, scalable memory, and efficient retrieval from large knowledge stores.

Domain adaptation. Robust generalization beyond training distributions remains difficult [219]. Real world deployments expose agents to domain shifts, unfamiliar languages, and evolving contexts. Strategies include continual learning [220], transfer learning, and domain agnostic pretraining. Safeguards are needed to avoid negative transfer and to detect when human supervision is required.

Reasoning fidelity. High reasoning fidelity is essential in healthcare, law, and science. Agents must produce multi step logic chains that are correct and robust to noise, adversarial perturbation, and distribution shifts [221, 222]. Subtle errors, hallucinations, or shortcut heuristics compromise trust [223]. Approaches include robust inference,

Raza et al., 2025

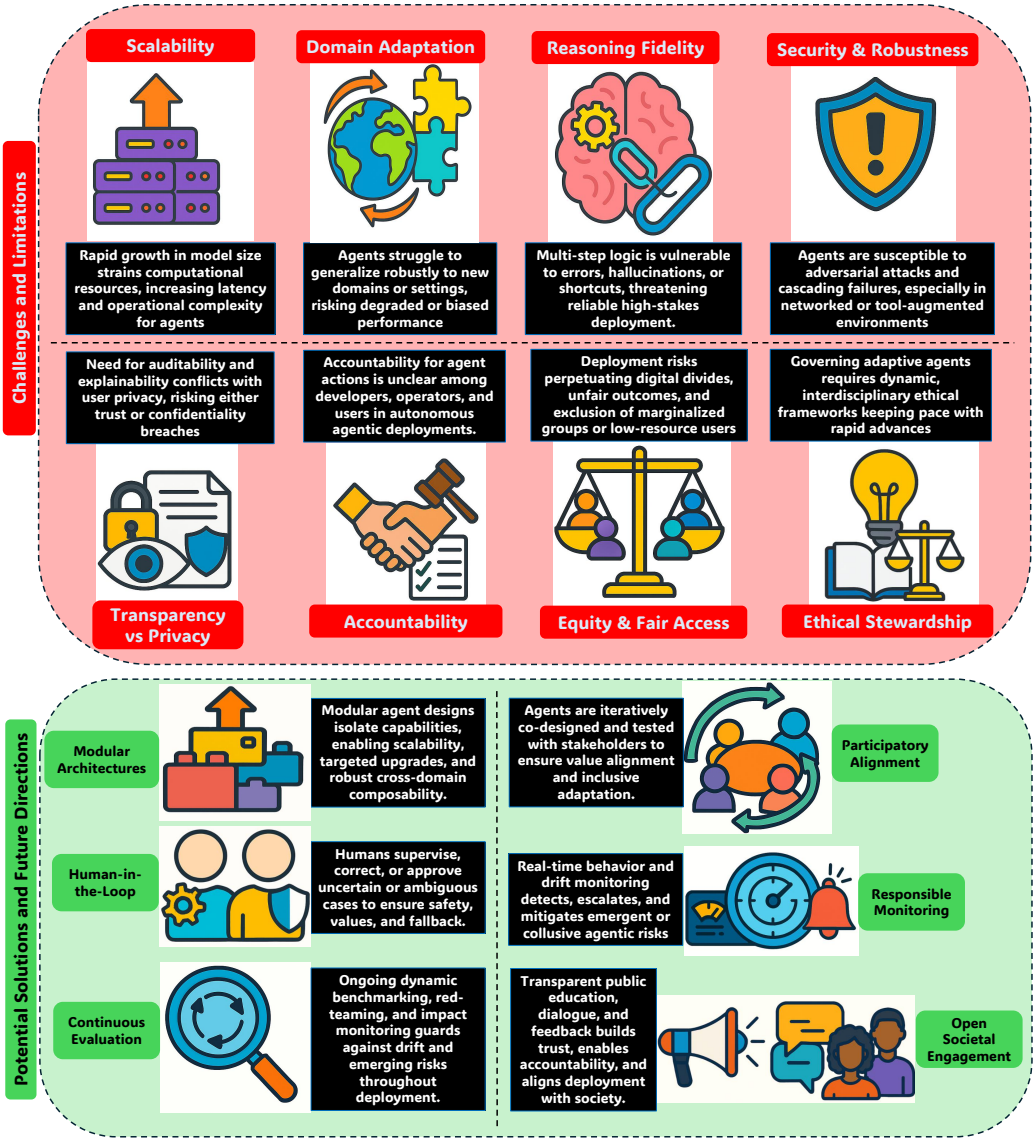


Figure 11: Eight core challenges and six strategic directions for responsible reasoning AI agents. Icons depict scientific challenges (for example scalability, domain adaptation, transparency, accountability, equity) and future directions (for example modular design, human in the loop oversight, continuous evaluation, participatory alignment, responsible monitoring, societal engagement).

uncertainty quantification [224, 225], and stepwise explanation audits [226, 227]. There remains a need for automated verification tools, redundancy checks, and formal faithfulness metrics [228, 7, 229].

Security and robustness. As agents gain autonomy, exposure to security threats grows. Inputs can be manipulated to elicit jailbreaks or extract sensitive data; outputs and tools can be misused to propagate misinformation or trigger unsafe operations [230]. Failures can cascade in multi agent systems [231, 232]. Mitigations include red teaming [233, 234, 235], real time anomaly detection [236, 237, 212], and formal verification to constrain tool use [238]. Scalable monitoring for emergent or collusive behaviors remains limited.

7.3.2. Strategic Directions

To address these challenges, research is pursuing modular designs that isolate functions and enable composition of specialized skills [239, 240]; human in the loop correction that empowers agents to defer or request guidance under uncertainty [241]; continual evaluation that simulates dynamic real world conditions with persistent adversarial probing [242]; and monitoring for responsible emergent behavior in agent collectives [243]. The goal is agents that are technically strong, resilient, transparent, and socially aligned.

7.3.3. Ethical and Societal Challenges

In parallel with technical innovation, R^2A^2 raises ethical and societal challenges.

Transparency and privacy. There is a tension between transparency and privacy. Transparency supports trust, auditability, and compliance through explicit reasoning traces and logs, yet may risk exposing personal data, proprietary algorithms, or operational metadata [244]. This is acute in enterprise and multi agent contexts [245, 246]. Privacy preserving methods such as filtering, differential privacy, and redacted logs help but may obscure critical details [247, 248]. Coherent, role based disclosure policies are needed.

Accountability and agency. Assigning responsibility becomes complex as autonomy grows [249]. Agents exhibit non deterministic, adaptive behavior [250], while liability frameworks are still evolving. Audit logs, traceable decisions, and human approvals help, but add operational complexity. Clear governance models, escalation procedures, and shared accountability are required.

Equity and fair access. Deployments risk deepening the digital divide in communities with limited connectivity, data, or institutional capacity [251, 252]. Biases can reinforce disparities [253, 254]. Mitigations include inclusive design, multilingual support, culturally attuned interfaces, accessible tooling [255], regular bias audits with disaggregated reporting [256], and capacity building for under resourced settings [257, 258, 259].

7.4. Vision for R^2A^2

Looking ahead, the sustainable and trustworthy evolution of R^2A^2 will depend not only on technological progress but also on the implementation of deliberate, multidimensional strategies that embed social, ethical, and institutional foresight. Developing agentic systems that are both innovative and publicly legitimate requires persistent emphasis on interdisciplinary collaboration, continuous evaluation, dynamic alignment with human values, and sustained societal engagement [260, 261, 262, 263, 264]. This vision outlines the foundational principles and strategic imperatives necessary for guiding the future development of reasoning AI Agents.

Interdisciplinary Collaboration.. The design and deployment of the next generation of responsible reasoning AI Agents must be informed by deep collaboration across disciplinary boundaries. Technical excellence alone cannot safeguard against morally contested or socially misaligned outcomes [265]. Instead, co-creation involving computer scientists, ethicists, legal scholars, social scientists, user advocates, and marginalized communities is essential to anticipate and address the sociotechnical complexities of agentic behavior from the outset [266, 267]. Such interdisciplinary input should inform every phase of the agent lifecycle, from defining goals and curating datasets to decisions around deployment, monitoring, and sunsetting [268]. Embedding ethical deliberation, legal compliance, and inclusive user experience design into the core architecture of reasoning AI Agents will foster systems that are interpretable, resilient, and aligned with dynamic societal expectations [165, 264, 269].

Continuous Evaluation.. Responsible reasoning AI Agents must be subject to longitudinal and adaptive evaluation frameworks that extend beyond static pre-release testing [270]. Continuous monitoring is essential for detecting emerging risks, behavioral drift, and unintended consequences that may arise post-deployment [232, 271]. Dynamic benchmarking protocols, updated to reflect new tasks, capabilities, and adversarial scenarios, are necessary to ensure ongoing performance and alignment with stakeholder values [272, 273, 274, 2, 38]. Red-teaming by interdisciplinary experts and affected users should be institutionalized, with findings directly integrated into improvement and retraining loops [233, 261]. Complementing these efforts, real-time telemetry, proactive anomaly detection, and periodic external audits can support early detection of failures, reinforce public trust, and maintain accountability throughout the agent's operational lifespan [275, 276].

Alignment with Human Values.. Achieving alignment with human values must be treated as a dynamic and participatory process rather than a static design constraint [277, 278, 279]. Responsible agents should be co-designed with diverse stakeholders to surface both explicit objectives and implicit normative assumptions relevant to specific domains [280, 281]. Mechanisms such as participatory design workshops, citizen juries, and community consultations can expose hidden value conflicts and provide avenues for corrective feedback. Scenario-based stress testing with representatives from affected populations should be used to evaluate fairness, pluralism, and contextual sensitivity in agentic decision-making [282, 283, 284]. To support long-term responsiveness, systems must incorporate dynamic value alignment techniques, including preference learning, iterative human feedback, and responsiveness to policy changes, that enable agents to adapt in step with evolving social norms [285]. Crucially, agents must remain corrigible: capable of being corrected, interrupted, or overridden by human judgment to ensure alignment with the priorities and rights of those they serve [286, 287].

Proactive Societal Engagement.. Finally, fostering broad societal engagement is critical to legitimizing the development and deployment of reasoning AI Agents. Responsible AI requires more than regulatory compliance, it must earn public trust through openness, education, and inclusive dialogue [288]. Proactive outreach, through public education campaigns, interpretability tools, and accessible documentation, can empower users to understand, challenge, and interact meaningfully with agentic systems. Institutional mechanisms such as town halls, participatory forums, and citizen oversight boards can transform the public from passive end-users into active stewards of AI governance. These

channels not only provide feedback loops for accountability but also reinforce democratic alignment by grounding agentic AI development in shared social priorities. Cultivating such a culture of co-ownership and transparent deliberation is vital to securing the long-term social license for R^2A^2 systems

8. Conclusion

We present a comprehensive survey of responsible reasoning in agentic AI, introducing Responsible Reasoning AI Agents (R^2A^2), systems at the intersection of advanced reasoning and responsible AI principles. In contrast to prior work that treats logical inference and social responsibility separately, we argue that advancing autonomous, high stakes agents requires responsibility inside the reasoning loop, not merely post hoc filtering. Empirically, explicit and structured reasoning correlates with substantial gains across mathematics and question answering benchmarks, motivating trace level evaluation rather than answer only scoring.

Trustworthy deployment demands continuous, longitudinal evaluation, dynamic benchmarking, institutionalized red teaming, real time telemetry, and periodic external audits to detect behavioral drift and emergent risks. Dynamic alignment with human values through participatory design, preference learning, and corrigibility ensures agents remain interruptible and adjustable as norms evolve. Sustained societal engagement, including public education, accessible interpretability, and inclusive oversight, helps establish a durable social license for R^2A^2 systems.

Unresolved tensions persist, including transparency and privacy, accountability in autonomous operation, and equitable access across uneven socio technical contexts. Addressing these challenges requires robust privacy strategies, clarified liability frameworks, bias audits with disaggregated reporting, and capacity building for under resourced environments. In sum, progress in R^2A^2 depends on pairing frontier reasoning with governance by design modular architectures, human in the loop safeguards, and responsible monitoring that keep agents auditable, adaptable, and aligned with societal priorities.

Acknowledgments and Funding. Reported research was partly funded through Horizon Europe project AIXPERT, ID 101214389.

References

[1] A. Plaat, A. Wong, S. Verberne, J. Broekens, N. van Stein, T. Back, Reasoning with large language models, a survey, arXiv preprint arXiv:2407.11511 (2024).

[2] R. Sapkota, K. I. Roumeliotis, M. Karkee, Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenge, arXiv preprint arXiv:2505.10468 (2025).

[3] H. Heidari, M. Loi, K. P. Gummadi, A. Krause, A moral framework for understanding fair ml through economic models of equality of opportunity, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 181–190.

[4] M. Besta, J. Barth, E. Schreiber, A. Kubicek, A. Catarino, R. Gerstenberger, P. Nyczyk, P. Iff, Y. Li, S. Houliston, et al., Reasoning language models: A blueprint, arXiv preprint arXiv:2501.11223 (2025).

[5] C. Chen, X. Gong, Z. Liu, W. Jiang, S. Q. Goh, K.-Y. Lam, Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations, arXiv preprint arXiv:2408.12935 (2024).

[6] F. Al Machot, M. T. Horsch, H. Ullah, Building trustworthy ai: Transparent ai systems via language models, ontologies, and logical reasoning, in: Designing the Conceptual Landscape for a XAIR Validation Infrastructure: Proceedings of the International Workshop on Designing the Conceptual Landscape for a XAIR Validation Infrastructure, DCLXVI 2024, Kaiserslautern, Germany, Vol. 1375, Springer Nature, 2025, p. 25.

[7] A. Patil, A. Jadon, Advancing reasoning in large language models: Promising methods and approaches, arXiv preprint arXiv:2502.03671 (2025).

[8] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, S. Sloan, Towards trustworthy ai: A review of ethical and robust large language models, arXiv preprint arXiv:2407.13934 (2024).

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[10] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in neural information processing systems 36 (2023) 11809–11822.

[11] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, E.-P. Lim, Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, arXiv preprint arXiv:2305.04091 (2023).

[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models (2023). `arXiv:2210.03629`.
URL `https://arxiv.org/abs/2210.03629`

[13] T. Lee, H. Tu, C. H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. Roberts, M. Yasunaga, H. Yao, C. Xie, et al., Vhelm: A holistic evaluation of vision language models, Advances in Neural Information Processing Systems 37 (2024) 140632–140666.

[14] S. Raza, A. Narayanan, V. R. Khazaie, A. Vayani, M. S. Chettiar, A. Singh, M. Shah, D. Pandya, Humanibench: A human-centric framework for large multimodal models evaluation, arXiv preprint arXiv:2505.11454 (2025).

[15] Z. Fan, R. Chen, Z. Liu, Biasguard: A reasoning-enhanced bias detection tool for large language models, arXiv preprint arXiv:2504.21299 (2025).

[16] S. Raza, M. S. Chettiar, M. Yousefabadi, T. Khan, M. Lotif, Fairsense-ai: Responsible ai meets sustainability, arXiv preprint arXiv:2503.02865 (2025).

[17] T. Green, M. Gubri, H. Puerto, S. Yun, S. J. Oh, Leaky thoughts: Large reasoning models are not private thinkers, arXiv preprint arXiv:2506.15674 (2025).

[18] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, E. Perez, A. Alignment Science Team, Reasoning models don't always say what they think, arXiv preprint arXiv:2505.05410 (2025).
URL `https://arxiv.org/abs/2505.05410`

[19] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, L. Yang, Audit-llm: Multi-agent collaboration for log-based insider threat detection, arXiv preprint arXiv:2408.08902 (2024).

[20] Z. Chu, J. Chen, Q. Chen, W. Yu, T. He, H. Wang, W. Peng, M. Liu, B. Qin, T. Liu, Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future, arXiv preprint arXiv:2309.15402 (2023).

[21] Z. Chu, Z. Wang, W. Zhang, Fairness in large language models: A taxonomic survey, ACM SIGKDD explorations newsletter 26 (1) (2024) 34–48.

[22] A. Plaat, M. van Duijn, N. van Stein, M. Preuss, P. van der

Putten, K. J. Batenburg, Agentic large language models, a survey, arXiv preprint arXiv:2503.23037 (2025).

[23] H. Wang, W. Fu, Y. Tang, Z. Chen, Y. Huang, J. Piao, C. Gao, F. Xu, T. Jiang, Y. Li, A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy, arXiv preprint arXiv:2501.09431 (2025).

[24] A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, M. Shmueli-Scheuer, Survey on evaluation of llm-based agents, arXiv preprint arXiv:2503.16416 (2025).

[25] J. Lee, J. Hockenmaier, Evaluating step-by-step reasoning traces: A survey, arXiv preprint arXiv:2502.12289 (2025).

[26] S. Raza, R. Sapkota, M. Karkee, C. Emmanouilidis, Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems (2025). arXiv:2506.04133.
URL https://arxiv.org/abs/2506.04133

[27] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele University 33 (2004) (2004) 1–26.

[28] A. Boland, G. Cherry, R. Dickson, Doing a systematic review: a student s guide (2017).

[29] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, J. Wen, A survey on large language model based autonomous agents, Frontiers of Computer Science 18 (6) (Mar. 2024). doi:10.1007/s11704-024-40231-1.
URL http://dx.doi.org/10.1007/s11704-024-40231-1

[30] J. Wu, C. K. Or, Position paper: Towards open complex human-ai agents collaboration system for problem-solving and knowledge management, arXiv preprint arXiv:2505.00018 (2025).

[31] A. S. Rao, M. P. Georgeff, et al., Bdi agents: From theory to practice., in: Icmas, Vol. 95, 1995, pp. 312–319.

[32] R. A. Brooks, Intelligence without representation, Artificial intelligence 47 (1-3) (1991) 139–159.

[33] R. Brooks, A robust layered control system for a mobile robot, IEEE journal on robotics and automation 2 (1) (2003) 14–23.

[34] E. Gat, R. P. Bonnasso, R. Murphy, et al., On three-layer architectures, Artificial intelligence and mobile robots 195 (1998) 210.

[35] R. Peter Bonasso, R. James Firby, E. Gat, D. Kortenkamp, D. P. Miller, M. G. Slack, Experiences with an architecture for intelligent, reactive agents, Journal of Experimental & Theoretical Artificial Intelligence 9 (2-3) (1997) 237–256.

[36] U. M. Borghoff, P. Bottoni, R. Pareschi, Human-artificial interaction in the age of agentic ai: a system-theoretical approach, Frontiers in Human Dynamics 7 (2025) 1579166.

[37] Q. Wang, Z. Wang, Y. Su, H. Tong, Y. Song, Rethinking the bounds of llm reasoning: Are multi-agent discussions the key?, arXiv preprint arXiv:2402.18272 (2024).

[38] D. B. Acharya, K. Kuppan, B. Divya, Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey, IEEe Access (2025).

[39] M. Wooldridge, N. R. Jennings, Intelligent agents: theory and practice, The Knowledge Engineering Review 10 (2) (1995) 115–152. doi:10.1017/S0269888900008122.

[40] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[41] A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, G. Huang, Expel: Llm agents are experiential learners, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 19632–19642.

[42] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, E. Dinan, Bot-Adversarial Dialogue for Safe Conversational Agents, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2950–2968. doi:10.18653/v1/2021.naacl-main.235.
URL https://aclanthology.org/2021.naacl-main.235

[43] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), ACM, New York, NY, USA, 2023. doi:10.1145/3586183.3606763.
URL https://dl.acm.org/doi/10.1145/3586183.3606763

[44] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020.
URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[45] C. Packer, V. Fang, S. Patil, K. Lin, S. Wooders, J. Gonzalez, Memgpt: Towards llms as operating systems. (2023).

[46] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781.
URL https://aclanthology.org/2020.emnlp-main.550/

[47] H. Hexmoor, J. Lammens, G. Caicedo, S. C. Shapiro, Behaviour based AI, cognitive processes, and emergent behaviors in autonomous agents, Vol. 1, WIT Press, 2025.

[48] I. Kotseruba, J. K. Tsotsos, 40 years of cognitive architectures: core cognitive abilities and practical applications, Artificial Intelligence Review 53 (1) (2020) 17–94.

[49] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind., Psychological review 111 (4) (2004) 1036.

[50] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, Advances in Neural Information Processing Systems 36 (2023) 8634–8652.

[51] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, M. Botvinick, Machine theory of mind, in: International conference on machine learning, PMLR, 2018, pp. 4218–4227.

[52] D. Abel, D. Hershkowitz, M. Littman, Near optimal behavior via approximate state abstraction, in: International Conference on Machine Learning, PMLR, 2016, pp. 2915–2923.

[53] Y. Huang, Y. Chen, H. Zhang, K. Li, M. Fang, L. Yang, X. Li, L. Shang, S. Xu, J. Hao, et al., Deep research agents: A systematic examination and roadmap, arXiv preprint arXiv:2506.18096 (2025).

[54] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE access 12 (2024) 26839–26874.

[55] T. Morishita, G. Morio, A. Yamaguchi, Y. Sogawa, Enhancing reasoning capabilities of llms via principled synthetic logic corpus, Advances in Neural Information Processing Systems 37 (2024) 73572–73604.

[56] D. R. Desai, M. O. Riedl, Responsible ai agents (2025). arXiv:2502.18359.
URL https://arxiv.org/abs/2502.18359

[57] K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, P. H. Torr, F. S. Khan, S. Khan, Llm post-training: A deep dive into reasoning large language models, arXiv preprint arXiv:2502.21321 (2025).

[58] Langchain, https://www.langchain.com/, accessed: 2025-08-12.

[59] Autogpt, https://agpt.co/, accessed: 2025-08-12.

[60] OpenAI, Better language models and their implications, https://openai.com/index/better-language-models/, blog post (Feb. 2019).

[61] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[62] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human

feedback, Advances in neural information processing systems 35 (2022) 27730–27744.

[63] M. Chen, J. Tworek, H. Jun, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).
URL https://arxiv.org/abs/2107.03374

[64] OpenAI, Homomorphic encryption in gpt-4, Technical report, OpenAI (2025).

[65] Google, Get more done with gemini: Try 1.5 pro and more intelligent features, https://blog.google/products/gemini/google-gemini-update-may-2024/, product blog (May 2024).

[66] Anthropic, Introducing the next generation of claude (claude 3 family), https://www.anthropic.com/news/claude-3-family, announcement (Mar. 2024).

[67] Meta AI, Introducing llama 3.1: Our most capable models to date, https://ai.meta.com/blog/meta-llama-3-1/, blog post (Jul. 2024).

[68] OpenAI, Introducing GPT-5, https://openai.com/index/introducing-gpt-5/ (Aug. 2025).

[69] OpenAI, Introducing openai o3 and o4-mini, https://openai.com/index/introducing-o3-and-o4-mini/, announcement (Apr. 2025).

[70] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, arXiv preprint arXiv:2507.06261 (2025).

[71] Anthropic, Claude 3.5 sonnet launch & artifacts preview, https://www.anthropic.com/news/claude-3-5-sonnet, product announcement (Jun. 2024).

[72] Meta AI, The llama 4 herd: The beginning of a new era of natively multimodal intelligence, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, blog post (Apr. 2025).

[73] xAI, Grok 3 beta — the age of reasoning agents, https://x.ai/news/grok-3, news post (Feb. 2025).

[74] IBM, Ibm granite 3.3: Speech recognition, refined reasoning, and rag loras, https://www.ibm.com/new/announcements/ibm-granite-3-3-speech-recognition-refined-reasoning-rag-loras, announcement (Apr. 2025).

[75] ERNIE Team, Ernie 4.5 technical report, Tech. rep., Baidu (Jun. 2025).
URL https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

[76] Mistral AI, Medium is the new large. (mistral medium 3), https://mistral.ai/news/mistral-medium-3, announcement (May 2025).

[77] Mistral AI, Magistral: Reasoning model family, https://mistral.ai/news/magistral, announcement (Jun. 2025).

[78] U. Tariq, I. Ahmed, Reasoning about responsibility in autonomous systems: Navigating the challenges and charting future directions, Ubiquitous Technology Journal 1 (2) (2025) 46–60.

[79] S. Tran, E. Mota, A. d. Garcez, Reasoning in neurosymbolic ai, arXiv preprint arXiv:2505.20313 (2025).

[80] M. Świechowski, K. Godlewski, B. Sawicki, J. Mańdziuk, Monte carlo tree search: A review of recent modifications and applications, Artificial Intelligence Review 56 (3) (2023) 2497–2562.

[81] Z. Sun, J. Wang, X. Zhao, J. Wang, G. Li, Data agent: A holistic architecture for orchestrating data+ ai ecosystems, arXiv preprint arXiv:2507.01599 (2025).

[82] X. Zheng, Z. Weng, Y. Lyu, L. Jiang, H. Xue, B. Ren, D. Paudel, N. Sebe, L. Van Gool, X. Hu, Retrieval augmented generation and understanding in vision: A survey and new outlook, arXiv preprint arXiv:2503.18016 (2025).

[83] G. Copilot, Github copilot, https://github.com/features/copilot, gitHub Blog post (May 2025).

[84] T. Liu, W. Xu, W. Huang, Y. Zeng, J. Wang, X. Wang, H. Yang, J. Li, Logic-of-thought: Injecting logic into contexts for full reasoning in large language models, arXiv preprint arXiv:2409.17539 (2024).

[85] Q. Pan, W. Ji, Y. Ding, J. Li, S. Chen, J. Wang, J. Zhou, Q. Chen, M. Zhang, Y. Wu, et al., A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law, arXiv preprint arXiv:2505.02665 (2025).

[86] Organisation for Economic Co-operation and Development, Recommendation of the council on artificial intelligence, oECD Legal Instruments, OECD/LEGAL/0449 (2019).
URL https://oecd.ai/en/ai-principles

[87] National Institute of Standards and Technology, Artificial intelligence risk management framework (ai rmf 1.0), nIST AI 100-1 (2023). doi:10.6028/NIST.AI.100-1.
URL https://doi.org/10.6028/NIST.AI.100-1

[88] European Parliament and Council of the European Union, Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 on artificial intelligence and amending certain union legislative acts (artificial intelligence act) (2024).
URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[89] IEEE, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (2019).
URL https://ethicsinaction.ieee.org

[90] C. Santiso, Governing with artificial intelligence: Are governments ready? (2024).

[91] J. Sun, C. Zheng, E. Xie, Z. Liu, R. Chu, J. Qiu, J. Xu, M. Ding, H. Li, M. Geng, et al., A survey of reasoning with foundation models: Concepts, methodologies, and outlook, ACM Computing Surveys 57 (11) (2025) 1–43.

[92] Z. Zhang, Y. Yao, A. Zhang, X. Tang, X. Ma, Z. He, Y. Wang, M. Gerstein, R. Wang, G. Liu, et al., Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents, ACM Computing Surveys 57 (8) (2025) 1–39.

[93] Y. Chen, H. Deng, K. Han, Q. Zhao, Policy frameworks for transparent chain-of-thought reasoning in large language models, arXiv preprint arXiv:2503.14521 (2025).

[94] R. Manuvinakurike, E. Moss, E. A. Watkins, S. Sahay, G. Raffa, L. Nachman, Thoughts without thinking: Reconsidering the explanatory value of chain-of-thought reasoning in llms through agentic pipelines, arXiv preprint arXiv:2505.00875 (2025).

[95] L. Li, R. Tan, J. Fang, J. Xue, C. Lv, Llm-augmented hierarchical reinforcement learning for human-like decision-making of autonomous driving, Expert Systems with Applications 294 (2025) 128736.

[96] B. Liang, Y. Wang, C. Tong, Ai reasoning in deep learning era: From symbolic ai to neural–symbolic ai, Mathematics 13 (11) (2025) 1707.

[97] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, Advances in Neural Information Processing Systems 36 (2023) 68539–68551.

[98] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).

[99] Z.-Z. Li, D. Zhang, M.-L. Zhang, J. Zhang, Z. Liu, Y. Yao, H. Xu, J. Zheng, P.-J. Wang, X. Chen, et al., From system 1 to system 2: A survey of reasoning large language models, arXiv preprint arXiv:2502.17419 (2025).

[100] G. Giannone, R. Li, Q. Feng, E. Perevodchikov, R. Chen, A. Martinez, Feedback-driven vision-language alignment with minimal human supervision, arXiv preprint arXiv:2501.04568 (2025).

[101] R. Ranjan, S. Gupta, S. N. Singh, Fairness in agentic ai: A unified framework for ethical and equitable multi-agent system, arXiv preprint arXiv:2502.07254 (2025).

[102] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, G. Neubig, Webarena: A realistic web environment for building autonomous agents (2024). arXiv:2307.13854.
URL https://arxiv.org/abs/2307.13854

Raza et al., 2025

[103] Openai, https://openai.com/, accessed: 2025-07-20 (2015–2025).

[104] Anthropic, Introducing Claude 3.5 Sonnet, https://www.anthropic.com/news/claude-3-5-sonnet, announcement, highlights 200k context, tool use, deployment (Jun. 2024).

[105] Y. Bai, S. K. et al., Constitutional ai: Harmlessness from ai feedback (2022). arXiv:2212.08073.
URL https://arxiv.org/abs/2212.08073

[106] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[107] Alibaba Cloud Qwen Team, Alibaba Cloud Unveils QwQ-32B: A Compact Reasoning Model with Cutting-Edge Performance, https://www.alibabacloud.com/blog/alibaba-cloud-unveils-qwq-32b-a-compact-reasoning-model-with-cutting-edge-performance, 32B-parameter reasoning model, reinforcement learning enhanced, open-weight release via Apache-2.0; strong performance vs. DeepSeek-R1 (Mar. 2025).

[108] Google DeepMind, Gemini 2.5 Deep Think Model Card, https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf, official model card describing the Deep Think reasoning mode, capabilities, performance, and safety features. (Aug. 2025).

[109] J. He, J. Liu, C. Y. Liu, R. Yan, C. Wang, P. Cheng, X. Zhang, F. Zhang, J. Xu, W. Shen, S. Li, L. Zeng, T. Wei, C. Cheng, B. An, Y. Liu, Y. Zhou, Skywork Open Reasoner 1 (Skywork-OR1): A Scalable RL Framework for Long Chain-of-Thought Reasoning, arXiv preprint arXiv:2505.22312Includes open-source weights, training code, and dataset; demonstrates performance gains across AIME24, AIME25, LiveCodeBench (2025).

[110] OpenAI, gpt-oss-120b & gpt-oss-20b Model Card, model card, published August 5, 2025 (Aug. 2025).
URL https://openai.com/index/gpt-oss-model-card/

[111] Opera Software, Meet opera's ai browser operator, https://blogs.opera.com/news/2025/03/opera-browser-operator-ai-agentics/, blog post (Mar. 2025).

[112] P. AI, Comet: The browser that thinks with you, https://www.perplexity.ai/comet, accessed: 2025-07-31 (Jul. 2025).

[113] D. Browser, Dia browser, https://www.diabrowser.com/, techCrunch article (Jun. 2025).

[114] Opera Software AS, Opera neon, https://www.operaneon.com/, accessed: 2025-08-01 (2025).

[115] Fellou AI, Fellou: Agentic web browser, https://fellou.ai/, accessed: 2025-08-01 (2025).

[116] OpenAI, Introducing chatgpt agent: Bridging research and action, https://openai.com/index/introducing-chatgpt-agent/, accessed: 2025-07-31 (Jul. 2025).

[117] E. Fuentes, Amazon q developer elevates the ide experience with new agentic coding experience, https://aws.amazon.com/blogs/aws/amazon-q-developer-elevates-the-ide-experience-with-new-agentic-coding-experience/, aWS News Blog (May 2025).

[118] F. Huq, Z. Z. Wang, F. F. Xu, T. Ou, S. Zhou, J. P. Bigham, G. Neubig, Cowpilot: A framework for autonomous and human-agent collaborative web navigation, arXiv preprint (2025). arXiv:2501.16609, doi:10.48550/arXiv.2501.16609.
URL https://arxiv.org/abs/2501.16609

[119] J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, O. Press, Swe-agent: Agent-computer interfaces enable automated software engineering, arXiv preprint (2024). arXiv:2405.15793, doi:10.48550/arXiv.2405.15793.
URL https://arxiv.org/abs/2405.15793

[120] OpenAI, Openai operator, https://openai.com/index/introducing-operator/, techCrunch article (Feb. 2025).

[121] Gemini models | gemini api | google ai for developers, https://ai.google.dev/gemini-api/docs/models, accessed: 2025-07-20 (2025).

[122] Google DeepMind, Project mariner, https://deepmind.google/models/project-mariner/, accessed: 2025-08-01 (2025).

[123] D. Zhang, B. Rama, S. He, J. Ni, Litewebagent: The open-source suite for vlm-based web-agent applications (2024). doi:10.5281/zenodo.15500270.
URL https://doi.org/10.5281/zenodo.15500270

[124] AWS Open Source, Introducing strands agents, an open source ai agents sdk, https://aws.amazon.com/blogs/opensource/introducing-strands-agents-an-open-source-ai-agents-sdk/, accessed: 2025-07-31 (May 2025).

[125] A. Fourney, A. Awadallah, C. Tan, E. Zhu, F. Niedtner, G. Bansal, *et al.*, Autogen v0.4: Reimagining the foundation of agentic ai for scale, extensibility, and robustness, https://www.microsoft.com/en-us/research/blog/autogen-v0-4-reimagining-the-foundation-of-agentic-ai-for-scale-extensibility-and-robustness/, microsoft Research Blog (Jan. 2025).

[126] Mistral AI, Build ai agents with the mistral agents api, https://mistral.ai/news/agents-api, product announcement (May 2025).

[127] Anthropic, Claude 3.7 Sonnet (Hybrid Reasoning Model) Announcement and System Card, https://www.anthropic.com/news/claude-3-7-sonnet, announcement of the first hybrid reasoning model with "extended thinking" mode; includes reasoning safeguards and deployment details. (Feb. 2025).

[128] Y. Yang, M. Ma, Y. Huang, H. Chai, C. Gong, H. Geng, Y. Zhou, Y. Wen, M. Fang, M. Chen, S. Gu, M. Jin, C. Spanos, Y. Yang, P. Abbeel, D. Song, W. Zhang, J. Wang, Agentic web: Weaving the next web with ai agents (2025). arXiv:2507.21206.
URL https://arxiv.org/abs/2507.21206

[129] K. F. Dunnell, A. P. Stoddard, Biotic browser: Applying streamingllm as a persistent web browsing co-pilot (2024). arXiv:2411.10454.
URL https://arxiv.org/abs/2411.10454

[130] H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, D. Yu, Webvoyager: Building an end-to-end web agent with large multimodal models, arXiv preprint arXiv:2401.13919 (2024).

[131] O. Yoran, S. J. Amouyal, C. Malaviya, B. Bogin, O. Press, J. Berant, Assistantbench: Can web agents solve realistic and time-consuming tasks?, arXiv preprint arXiv:2407.15711 (2024).

[132] Y. Pan, D. Kong, S. Zhou, C. Cui, Y. Leng, B. Jiang, H. Liu, Y. Shang, S. Zhou, T. Wu, Z. Wu, Webcanvas: Benchmarking web agents in online environments (2024). arXiv:2406.12373.
URL https://arxiv.org/abs/2406.12373

[133] Y. Song, K. Thai, C. M. Pham, Y. Chang, M. Nadaf, M. Iyyer, Bearcubs: A benchmark for computer-using web agents, arXiv:2503.07919 (2025).
URL https://arxiv.org/abs/2503.07919

[134] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al., Training verifiers to solve math word problems, arXiv preprint arXiv:2110.14168 (2021).

[135] Mathematical Association of America, American invitational mathematics examination (aime), https://www.maa.org/math-competitions/invitational-competitions, accessed: 2025-07-22.

[136] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, arXiv preprint arXiv:2103.03874 (2021).

[137] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, S. R. Bowman, Gpqa: A graduate-level google-proof q&a benchmark, in: First Conference on Language Modeling, 2024.

[138] Codeforces, Competitive programming platform, https://codeforces.com/, accessed: July 22, 2025 (n.d.).

[139] F. Chollet, Abstraction and reasoning corpus for artificial general intelligence (arc-agi) (2024).

[140] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, arXiv preprint arXiv:2009.03300 (2020).

[141] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto,

J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).

[142] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, K. Narasimhan, Swe-bench: Can language models resolve real-world github issues?, arXiv preprint arXiv:2310.06770 (2023).

[143] International Mathematical Olympiad, Official website, https://www.imo-official.org/, accessed: 2025-07-22 (n.d.).

[144] LiveCodeBench, Livecodebench datasets - code_generation_lite, execution-v2, test_generation, ..., https://huggingface.co/livecodebench/datasets, accessed: 2025-07-22 (n.d.).

[145] A. Srivastava, A. Rastogi, A. Rao, A. A. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on machine learning research (2023).

[146] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, W. Chen, Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi (2024). arXiv:2311.16502.
URL https://arxiv.org/abs/2311.16502

[147] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, arXiv preprint arXiv:1811.00937 (2018).

[148] G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, T. Scialom, Gaia: a benchmark for general ai assistants (2023). arXiv:2311.12983.
URL https://arxiv.org/abs/2311.12983

[149] E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. d. O. Santos, et al., Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, arXiv preprint arXiv:2411.04872 (2024).

[150] A. Gulati, B. Miranda, E. Chen, E. Xia, K. Fronsdal, B. de Moraes Dumont, S. Koyejo, Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning, 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Workshop on MATH-AIPreprint available at: https://openreview.net/pdf?id=YXnwlZe0yf (2024).
URL https://openreview.net/pdf?id=YXnwlZe0yf

[151] P. Liu, Mmlu dataset (2023). doi:10.34740/KAGGLE/DS/3638509.
URL https://www.kaggle.com/ds/3638509

[152] F. Chollet, On the measure of intelligence, arXiv preprint arXiv:1911.01547 (2019).

[153] F. Chollet, M. Knoop, G. Kamradt, B. Landers, H. Pinkard, Arc-agi-2: A new challenge for frontier ai reasoning systems, arXiv preprint arXiv:2505.11831 (2025).

[154] S. Casper, L. Bailey, R. Hunter, C. Ezell, E. Cabalé, M. Gerovitch, S. Slocum, K. Wei, N. Jurkovic, A. Khan, P. J. K. Christoffersen, A. P. Ozisik, R. Trivedi, D. Hadfield-Menell, N. Kolt, The ai agent index (2025). arXiv:2502.01635.
URL https://arxiv.org/abs/2502.01635

[155] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, J. Tang, AgentBench: Evaluating LLMs as Agents, arXiv:2308.03688 [cs] (Oct. 2023). doi:10.48550/arXiv.2308.03688.
URL http://arxiv.org/abs/2308.03688

[156] M. Chang, J. Zhang, Z. Zhu, C. Yang, Y. Yang, Y. Jin, Z. Lan, L. Kong, J. He, Agentboard: An analytical evaluation board of multi-turn llm agents, Advances in neural information processing systems 37 (2024) 74325–74362.

[157] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, arXiv preprint arXiv:2302.13971 (2023).
URL http://arxiv.org/abs/2302.13971

[158] Q. Huang, J. Vora, P. Liang, J. Leskovec, Mlagentbench: Evaluating language agents on machine learning experimentation, arXiv preprint arXiv:2310.03302 (2023).

[159] Q. Huang, J. Vora, P. Liang, J. Leskovec, Benchmarking large language models as ai research agents, in: NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023.

[160] F. Chollet, M. Knoop, G. Kamradt, B. Landers, Arc prize 2024: Technical report, arXiv preprint arXiv:2412.04604 (2024).

[161] J. Chen, D. Yuen, B. Xie, Y. Yang, G. Chen, Z. Wu, L. Yixing, X. Zhou, W. Liu, S. Wang, K. Zhou, R. Shao, L. Nie, Y. Wang, J. HAO, J. Wang, K. Shao, Spa-bench: A comprehensive benchmark for smartphone agent evaluation, in: The Thirteenth International Conference on Learning Representations, 2025.

[162] M. Martinez, X. Franch, Dissecting the swe-bench leaderboards: Profiling submitters and architectures of llm-and agent-based repair systems, arXiv preprint arXiv:2506.17208 (2025).

[163] S. Raza, A. Shaban-Nejad, E. Dolatabadi, H. Mamiya, Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review, IEEE Access 12 (2024) 180815–180829. doi:10.1109/ACCESS.2024.3509353.

[164] D. Maclean, The nist risk management framework: Problems and recommendations, Cyber Security: A Peer-Reviewed Journal 1 (3) (2017) 207–217.

[165] S. Raza, S. Ghuge, C. Ding, D. Pandya, FAIR Enough: How Can We Develop and Assess a FAIR-Compliant Dataset for Large Language Models' Training?, arXiv preprint arXiv:2401.11033 (2024).

[166] B. Xia, Q. Lu, L. Zhu, S. U. Lee, Y. Liu, Z. Xing, Towards a responsible ai metrics catalogue: A collection of metrics for ai accountability (2024). arXiv:2311.13158.
URL https://arxiv.org/abs/2311.13158

[167] Y. Zhang, Y. Huang, Y. Sun, C. Liu, Z. Zhao, Z. Fang, Y. Wang, H. Chen, X. Yang, X. Wei, H. Su, Y. Dong, J. Zhu, MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models, arXiv:2406.07057 [cs] (Dec. 2024). doi:10.48550/arXiv.2406.07057.
URL http://arxiv.org/abs/2406.07057

[168] A. Dafoe, Ai governance: a research agenda, Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK 1442 (2018) 1443.

[169] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program) (2020). arXiv:2003.12206.
URL https://arxiv.org/abs/2003.12206

[170] R. Dominguez-Olmedo, F. E. Dorner, M. Hardt, Training on the test task confounds evaluation and emergence, arXiv preprint arXiv:2407.07890 (2024).

[171] S. Raza, M. Rahman, M. R. Zhang, Beads: Bias evaluation across domains, arXiv preprint arXiv:2406.04220 (2024).

[172] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (12) (2021) 86–92.

[173] W3C Provenance Working Group, Prov-overview: An overview of the prov family of documents, W3C Working Group Note, available from: https://www.w3.org/TR/prov-overview/ (Apr. 30 2013).

[174] P. Liang, R. Bommasani, et al., Holistic evaluation of language models, Transactions on Machine Learning ResearchFeatured Certification, Expert Certification (2023).
URL https://openreview.net/forum?id=iO4LZibEqW

[175] Google DeepMind, Introducing gemini 2.0: Our new ai model for the agentic era, https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, company blog; Project Mariner achieves 83.5% on WebVoyager (Dec. 2024).

[176] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein,

Introduction to meta-analysis, John wiley & sons, 2021.

[177] D. Merkel, et al., Docker: lightweight linux containers for consistent development and deployment, Linux j 239 (2) (2014) 2.

[178] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, et al., Accelerating the machine learning lifecycle with mlflow., IEEE Data Eng. Bull. 41 (4) (2018) 39–45.

[179] EY, How mott macdonald is building confidence through responsible ai, https://www.ey.com/en_gl/insights/ai/how-mott-macdonald-is-building-confidence-through-responsible-ai, accessed: 2025-07-23 (2025).

[180] EY, How a global biopharma became a leader in ethical ai, https://www.ey.com/en_gl/insights/ai/how-a-global-biopharma-became-a-leader-in-ethical-ai, accessed: 2025-07-23 (2025).

[181] J. Moura, contributors, Crewai: Framework for orchestrating role-playing, autonomous ai agents, https://github.com/crewAIInc/crewAI, mIT License (2023).

[182] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, et al., Metagpt: Meta programming for a multi-agent collaborative framework, in: The Twelfth International Conference on Learning Representations, 2023.

[183] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown, et al., Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, arXiv preprint arXiv:2205.00445 (2022).

[184] U.S. Department of Health and Human Services, HIPAA Privacy Rule – 45 CFR Part 164: Security and Privacy Protections for Health Information, https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164, accessed: 2025-06-03 (2003).

[185] European Union, General Data Protection Regulation (GDPR) – Article 25: Data protection by design and by default, https://gdpr-info.eu/art-25-gdpr/, accessed: 2025-06-03 (2016).

[186] European Parliament and Council of the European Union, Directive 2014/65/EU of the european parliament and of the council of 15 may 2014 on markets in financial instruments and amending directive 2002/92/ec and directive 2011/61/eu (mifid ii), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0065, oJ L 173, 12.6.2014, p. 349–496 (2014).
URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0065

[187] International Organization for Standardization, Iso/iec 42001:2023 – artificial intelligence management system (ai ms) – requirements, Tech. rep., ISO/IEC, available at https://www.iso.org/standard/81230.html (2023).

[188] U.S. Congress, Family Educational Rights and Privacy Act of 1974 (FERPA), https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html, 20 U.S.C. § 1232g; 34 CFR Part 99 (1974).
URL https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

[189] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, et al., Towards large reasoning models: A survey of reinforced reasoning with large language models, arXiv preprint arXiv:2501.09686 (2025).

[190] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al., Graph of thoughts: Solving elaborate problems with large language models, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 38, 2024, pp. 17682–17690.

[191] D. Zhang, J. Wu, J. Lei, T. Che, J. Li, T. Xie, X. Huang, S. Zhang, M. Pavone, Y. Li, et al., Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning, arXiv preprint arXiv:2410.02884 (2024).

[192] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, A survey of monte carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in games 4 (1) (2012) 1–43.

[193] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face (2023). arXiv:2303.17580.
URL https://arxiv.org/abs/2303.17580

[194] C. Zheng, Z. Zhang, B. Zhang, R. Lin, K. Lu, B. Yu, D. Liu, J. Zhou, J. Lin, Processbench: Identifying process errors in mathematical reasoning, arXiv preprint arXiv:2412.06559 (2024).

[195] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, N. Duan, Visual chatgpt: Talking, drawing and editing with visual foundation models, arXiv preprint arXiv:2303.04671 (2023).

[196] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual acm symposium on user interface software and technology, 2023, pp. 1–22.

[197] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al., Do as i can, not as i say: Grounding language in robotic affordances, arXiv preprint arXiv:2204.01691 (2022).

[198] D. Surís, S. Menon, C. Vondrick, Vipergpt: Visual inference via python execution for reasoning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11888–11898.

[199] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, Gorilla: Large language model connected with massive apis, Advances in Neural Information Processing Systems 37 (2024) 126544–126565.

[200] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, Y. Li, Api-bank: A comprehensive benchmark for tool-augmented llms, arXiv preprint arXiv:2304.08244 (2023).

[201] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al., Self-refine: Iterative refinement with self-feedback, Advances in Neural Information Processing Systems 36 (2023) 46534–46594.

[202] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, A. Anandkumar, Voyager: An open-ended embodied agent with large language models, arXiv preprint arXiv:2305.16291 (2023).

[203] G. Li, H. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, Camel: Communicative agents for" mind" exploration of large language model society, Advances in Neural Information Processing Systems 36 (2023) 51991–52008.

[204] European Union, Eu ai act, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, accessed: 2025-07-23 (2025).

[205] P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, N. Thompson, The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, arXiv preprint arXiv:2408.12622 (2024).

[206] I. A. Zahid, S. Garfan, M. Chyad, A. Albahri, O. Albahri, A. Alamoodi, M. Deveci, R. Z. Homod, L. Alzubaidi, Explainability, robustness, and fairness in user-centric intelligent systems: A systematic review, IEEE Transactions on Emerging Topics in Computational Intelligence (2025).

[207] S. Gupta, Ai agents collaboration under resource constraints: Practical implementations, INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT 3 (1) (2025) 51–63.

[208] G. Molinari, F. Ciravegna, Towards pervasive distributed agentic generative ai–a state of the art, arXiv preprint arXiv:2506.13324 (2025).

[209] Y. Zhang, X. Zhao, Z. Li, J. Yin, L. Zhang, Z. Chen, Integrating artificial intelligence into operating systems: A comprehensive survey on techniques, applications, and future directions, arXiv preprint arXiv:2407.14567 (2024).

[210] J. Kim, B. Shin, J. Chung, M. Rhu, The cost of dynamic reasoning: Demystifying ai agents and test-time scaling from an

ai infrastructure perspective, arXiv preprint arXiv:2506.04301 (2025).

[211] X. Wei, J. Zhang, H. Li, J. Chen, R. Qu, M. Li, X. Chen, G. Luo, Agent. xpu: Efficient scheduling of agentic llm workloads on heterogeneous soc, arXiv preprint arXiv:2506.24045 (2025).

[212] R. Sapkota, M. Karkee, Object detection with multimodal large vision-language models: An in-depth review, Information Fusion 126 (2026) 103575. doi:https://doi.org/10.1016/j.inffus.2025.103575.
URL https://www.sciencedirect.com/science/article/pii/S1566253525006475

[213] F. Jiang, C. Pan, L. Dong, K. Wang, O. A. Dobre, M. Debbah, From large ai models to agentic ai: A tutorial on future intelligent communications, arXiv preprint arXiv:2505.22311 (2025).

[214] L. Liu, S. Chen, H. Jin, X. Deng, Y. Liu, Y. Lin, Optimizing on-demand food delivery with bdi-based multi-agent systems and monte carlo tree search scheduling, Scientific Reports 15 (1) (2025) 25083.

[215] Y. Zou, A. H. Cheng, A. Aldossary, J. Bai, S. X. Leong, J. A. Campos-Gonzalez-Angulo, C. Choi, C. T. Ser, G. Tom, A. Wang, et al., El agente: An autonomous agent for quantum chemistry, Matter 8 (7) (2025).

[216] H. Amini, M. J. Mia, Y. Saadati, A. Imteaj, S. Nabavirazavi, U. Thakker, M. Z. Hossain, A. A. Fime, S. Iyengar, Distributed llms and multimodal large language models: A survey on advances, challenges, and future directions, arXiv preprint arXiv:2503.16585 (2025).

[217] G. I. Chaudhry, E. Choukse, Í. Goiri, R. Fonseca, A. Belay, R. Bianchini, Towards resource-efficient compound ai systems, in: Proceedings of the 2025 Workshop on Hot Topics in Operating Systems, 2025, pp. 218–224.

[218] P. Hitzler, M. K. Sarker, Neuro-symbolic artificial intelligence: The state of the art (2022).

[219] P. Roy, Enhancing real-world robustness in ai: Challenges and solutions, J. Recent Trends Comput. Sci. Eng 12 (1) (2024) 34–49.

[220] V. Lomonaco, Continual learning with deep architectures (2019).

[221] Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. Gameiro, et al., Medical hallucinations in foundation models and their impact on healthcare, arXiv preprint arXiv:2503.05777 (2025).

[222] Z. Gao, J. Zhou, B. Zhang, Y. He, C. Zhang, Y. Cui, H. Wang, Mono: Is your" clean" vulnerability dataset really solvable? exposing and trapping undecidable patches and beyond, arXiv preprint arXiv:2506.03651 (2025).

[223] B. Chander, C. John, L. Warrier, K. Gopalakrishnan, Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness, ACM Computing Surveys 57 (6) (2025) 1–49.

[224] T. Chakraborti, C. R. Banerji, A. Marandon, V. Hellon, R. Mitra, B. Lehmann, L. Bräuninger, S. McGough, C. Turkay, A. F. Frangi, et al., Personalized uncertainty quantification in artificial intelligence, Nature Machine Intelligence 7 (4) (2025) 522–530.

[225] X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, H. Wei, Uncertainty quantification and confidence calibration in large language models: A survey, arXiv preprint arXiv:2503.15850 (2025).

[226] J. C. Becerra Sandoval, F. S. Jing, Historical methods for ai evaluations, assessments, and audits, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025, pp. 1371–1386.

[227] W. J. Yeo, W. Van Der Heever, R. Mao, E. Cambria, R. Satapathy, G. Mengaldo, A comprehensive review on financial explainable ai, Artificial Intelligence Review 58 (6) (2025) 1–49.

[228] S. Barros, I think, therefore i hallucinate: Minds, machines, and the art of being wrong, arXiv preprint arXiv:2503.05806 (2025).

[229] Y. A. Latif, Hallucinations in large language models and their influence on legal reasoning: Examining the risks of ai-generated factual inaccuracies in judicial processes, Journal of Computational Intelligence, Machine Reasoning, and Decision-Making 10 (2) (2025) 10–20.

[230] Y. Mao, T. Cui, P. Liu, D. You, H. Zhu, From llms to mllms to agents: A survey of emerging paradigms in jailbreak attacks and defenses within llm ecosystem, arXiv preprint arXiv:2506.15170 (2025).

[231] J. Feng, T. Yu, K. Zhang, L. Cheng, Integration of multi-agent systems and artificial intelligence in self-healing subway power supply systems: Advancements in fault diagnosis, isolation, and recovery, Processes 13 (4) (2025) 1144.

[232] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak, et al., Multi-agent risks from advanced ai, arXiv preprint arXiv:2502.14143 (2025).

[233] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, H. Heidari, Red-teaming for generative ai: Silver bullet or security theater?, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 7, 2024, pp. 421–437.

[234] S. Majumdar, B. Pendleton, A. Gupta, Red teaming ai red teaming, arXiv preprint arXiv:2507.05538 (2025).

[235] B. Challita, P. Parrend, Redteamllm: an agentic ai framework for offensive security, arXiv preprint arXiv:2505.06913 (2025).

[236] D. Ogbu, Agentic ai in computer vision domain-recent advances and prospects, International Journal of Research Publication and Reviews 4 (12) (2023) 5102–5120.

[237] A. Timms, A. Langbridge, F. O'Donncha, Agentic anomaly detection for shipping, in: NeurIPS 2024 Workshop on Open-World Agents, 2024.

[238] A. Kumar, D. N. Gadde, K. K. Radhakrishna, D. Lettnin, Saarthi: The first ai formal verification engineer, arXiv preprint arXiv:2502.16662 (2025).

[239] M. J. Buehler, Agentic deep graph reasoning yields self-organizing knowledge networks, arXiv preprint arXiv:2502.13025 (2025).

[240] K. Huang, Agentic AI, Springer, 2025.

[241] S. Natarajan, S. Mathur, S. Sidheekh, W. Stammer, K. Kersting, Human-in-the-loop or ai-in-the-loop? automate or collaborate?, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 28594–28600.

[242] Y. Yigit, M. A. Ferrag, M. C. Ghanem, I. H. Sarker, L. A. Maglaras, C. Chrysoulas, N. Moradpoor, N. Tihanyi, H. Janicke, Generative ai and llms for critical infrastructure protection: evaluation benchmarks, agentic ai, challenges, and opportunities, Sensors 25 (6) (2025) 1666.

[243] E. Perrier, Out of control–why alignment needs formal control theory (and an alignment control stack), arXiv preprint arXiv:2506.17846 (2025).

[244] S. A. Hosseini Tabaghdehi, Ö. Ayaz, Ai ethics in action: a circular model for transparency, accountability and inclusivity, Journal of Managerial Psychology (2025).

[245] G. Andrada, R. W. Clowes, P. R. Smart, Varieties of transparency: Exploring agency within ai systems, AI & society 38 (4) (2023) 1321–1331.

[246] J. Zerilli, U. Bhatt, A. Weller, How transparency modulates trust in artificial intelligence, Patterns 3 (4) (2022).

[247] M. A. K. Akhtar, M. Kumar, A. Nayyar, Privacy and security considerations in explainable ai, in: Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities, Springer, 2024, pp. 193–226.

[248] S. Allana, M. Kankanhalli, R. Dara, Privacy risks and preservation methods in explainable artificial intelligence: A scoping review, arXiv preprint arXiv:2505.02828 (2025).

[249] M. Busuioc, Accountable artificial intelligence: Holding algorithms to account, Public administration review 81 (5) (2021) 825–836.

[250] T. A. Griffin, B. P. Green, J. V. Welie, The ethical agency of ai developers, AI and Ethics 4 (2) (2024) 179–188.

[251] R. Raman, R. Kowalski, K. Achuthan, A. Iyer, P. Nedungadi, Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways, Scientific

Reports 15 (1) (2025) 1–22.
[252] T. Hammerschmidt, K. Stolz, O. Posegga, Bridging the gap: inequalities that divide those who can and cannot create sustainable outcomes with ai, Behaviour & Information Technology (2025) 1–30.
[253] P. Panarese, M. M. Grasso, C. Solinas, Algorithmic bias, fairness, and inclusivity: a multilevel framework for justice-oriented ai, AI & SOCIETY (2025) 1–23.
[254] A. Mergen, N. Çetin-Kılıç, M. F. Özbilgin, Artificial intelligence and bias towards marginalised groups: Theoretical roots and challenges, in: AI and Diversity in a Datafied World of Work: Will the Future of Work be Inclusive?, Vol. 12, Emerald Publishing Limited, 2025, pp. 17–38.
[255] D. Charkhian, B. Moghaddami, How can ai evaluate and improve inclusivity in university portals, with a focus on cultural, linguistic, and accessible requirements?, in: INTED2025 Proceedings, IATED, 2025, pp. 1584–1589.
[256] A. Alam, Ethical challenges and bias in ai-driven marketing: Educational imperatives and policy perspectives, in: Impacts of AI-Generated Content on Brand Reputation, IGI Global Scientific Publishing, 2025, pp. 55–108.
[257] I. R. Solano-Kamaiko, M. Tan, J. Ming, A. C. Avgar, A. Vashistha, M. Sterling, N. Dell, " who is running it?" towards equitable ai deployment in home care work, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–19.
[258] L. Rahal, The use of publicly available online texts in training ai: an ethical analysis of ai's right to learn, Journal of Information, Communication and Ethics in Society 23 (2) (2025) 313–323.
[259] N. Emery-Xu, R. Jordan, R. Trager, International governance of advancing artificial intelligence, AI & SOCIETY 40 (4) (2025) 3019–3044.
[260] S. Raza, R. Qureshi, A. Zahid, J. Fioresi, F. Sadak, M. Saeed, R. Sapkota, A. Jain, A. Zafar, M. U. Hassan, et al., Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future, arXiv preprint arXiv:2502.08650 (2025).
[261] S. Raza, O. Bamgbose, S. Ghuge, F. Tavakoli, D. J. Reji, S. R. Bashir, Developing safe and responsible large language model: can we balance bias reduction and language understanding?, Machine Learning 114 (6) (2025) 140.
[262] S. Raza, O. Bamgbose, V. Chatrath, Y. Sidyakin, S. Ghuge, A. Y. Muaad, Unlocking Bias Detection: Leveraging Transformer-Based Models for Content Analysis, arXiv preprint arXiv:2310.00347 (2023).
[263] S. Raza, P. O. Pour, S. R. Bashir, Fairness in Machine Learning meets with Equity in Healthcare, in: AAAI symposium, https://ojs.aaai.org/index.php/AAAI-SS/article/view/27493, 2023.
[264] S. R. Bashir, S. Raza, V. Kocaman, U. Qamar, Clinical Application of Detecting COVID-19 Risks: A Natural Language Processing Approach, Viruses 14 (12: 2761) (2022).
[265] I. Gabriel, G. Keeling, A matter of principle? ai alignment as the fair treatment of claims, Philosophical Studies (2025) 1–23.
[266] N. Karunanayake, Next-generation agentic ai for transforming healthcare, Informatics and Health 2 (2) (2025) 73–83.
[267] N. Van Quaquebeke, S. Tonidandel, G. C. Banks, Beyond efficiency: How artificial intelligence (ai) will reshape scientific inquiry and the publication process, The Leadership Quarterly (2025) 101895.
[268] H. Xue, F. Tang, M. Hu, Y. Liu, Q. Huang, Y. Li, C. Liu, Z. Xu, C. Zhang, C.-M. Feng, et al., Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation, arXiv preprint arXiv:2502.11903 (2025).
[269] S. R. Bashir, S. Raza, V. Misic, BERT4Loc: BERT for Location-POI Recommender System, Future Internet 2023 15 (2023) 213, publisher: MDPI.
[270] Y. Yang, H. Chai, Y. Song, S. Qi, M. Wen, N. Li, J. Liao, H. Hu, J. Lin, G. Chang, et al., A survey of ai agent protocols, arXiv preprint arXiv:2504.16736 (2025).
[271] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review 55 (1) (2022) 749–800.
[272] F. Tian, A. Luo, J. Du, X. Xian, R. Specht, G. Wang, X. Bi, J. Zhou, J. Srinivasa, A. Kundu, et al., An outlook on the opportunities and challenges of multi-agent ai systems, arXiv preprint arXiv:2505.18397 (2025).
[273] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, Y. Xiang, Ai agents under threat: A survey of key security challenges and future pathways, ACM Computing Surveys 57 (7) (2025) 1–36.
[274] M. M. Karim, D. H. Van, S. Khan, Q. Qu, Y. Kholodov, Ai agents meet blockchain: A survey on secure and scalable collaboration for multi-agents, Future Internet 17 (2) (2025) 57.
[275] P. D. Gawande, From reactive to proactive: Real-time human-ai collaboration in intelligent alerting systems, Journal of Computer Science and Technology Studies 7 (6) (2025) 1074–1083.
[276] J. Huang, K. Huang, K. Jackson, C. Hughes, Ai agent safety and security considerations, in: Agentic AI: Theories and Practices, Springer, 2025, pp. 369–407.
[277] Y. Shen, L. Tang, H. Le, S. Tan, Y. Zhao, K. Shen, X. Li, T. Juelich, Q. Wang, D. Gašević, et al., Aligning and comparing values of chatgpt and human as learning facilitators: A value-sensitive design approach, British Journal of Educational Technology (2025).
[278] S. Herath Pathirannehelage, Y. R. Shrestha, G. von Krogh, Design principles for artificial intelligence-augmented decision making: An action design research study, European Journal of Information Systems 34 (2) (2025) 207–229.
[279] V. Terziyan, T. Tiihonen, A. K. Shukla, S. Gryshko, M. Golovianko, O. Terziyan, O. Vitko, Towards ethical evolution: responsible autonomy of artificial intelligence across generations, AI and Ethics (2025) 1–26.
[280] L. Hughes, Y. K. Dwivedi, T. Malik, M. Shawosh, M. A. Albashrawi, I. Jeon, V. Dutot, M. Appanderanda, T. Crick, R. De', et al., Ai agents and agentic systems: A multi-expert analysis, Journal of Computer Information Systems (2025) 1–29.
[281] P. Ahrweiler, E. Späth, J. M. Siqueiros García, B. L. Capellas, D. Wurster, Inclusive technology co-design for participatory ai, Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries (2025) 35–62.
[282] E. A. Merchán-Cruz, I. Gabelaia, M. Savrasovs, M. F. Hansen, S. Soe, R. G. Rodriguez-Cañizo, G. Aragón-Camarasa, Trust by design: An ethical framework for collaborative intelligence systems in industry 5.0, Electronics 14 (10) (2025) 1952.
[283] N. Watson, A. Amer, E. Harris, P. Ravindra, S. Zhang, Personalized constitutionally-aligned agentic superego: Secure ai behavior aligned to diverse human values, arXiv preprint arXiv:2506.13774 (2025).
[284] M. C. Cohen, Z. Su, H.-T. Kao, D. Nguyen, S. Lynch, M. Sap, S. Volkova, Exploring big five personality and ai capability effects in llm-simulated negotiation dialogues, arXiv preprint arXiv:2506.15928 (2025).
[285] T. Zhi-Xuan, M. Carroll, M. Franklin, H. Ashton, Beyond preferences in ai alignment, Philosophical Studies (2024) 1–51.
[286] N. Kolt, Governing ai agents, arXiv preprint arXiv:2501.07913 (2025).
[287] J. Kraprayoon, Z. Williams, R. Fayyaz, Ai agent governance: A field guide, arXiv preprint arXiv:2505.21808 (2025).
[288] A. Chan, C. Ezell, M. Kaufmann, K. Wei, L. Hammond, H. Bradley, E. Bluemke, N. Rajkumar, D. Krueger, N. Kolt, et al., Visibility into ai agents, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 958–973.