

Trustworthy Agentic AI: A Rapid, PRISMA-Informed Evidence Review of Safety, Explainability, Alignment, and Human Oversight in Autonomous AI Systems

Vivek Garike

vgarike27701@cumberlands.edu

University of the Cumberland

Research Article

Keywords: Agentic AI, Autonomous agents, LLM agents, Trustworthy AI, Explainable AI, AI alignment, Human oversight, AI safety, Retrieval-augmented generation, Multi-agent systems

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10582034/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Agentic artificial intelligence is rapidly emerging as a paradigm in which large language models, retrieval systems, memory modules, software tools, and multi-agent coordination mechanisms are integrated to support autonomous or semi-autonomous decision-making. Unlike conventional AI systems that primarily produce predictions or responses, agentic AI systems can interpret goals, decompose tasks, retrieve external evidence, invoke tools, interact with digital environments, communicate with other agents, and execute multi-step workflows. These capabilities introduce trustworthiness challenges because failures may propagate from generated text into planning, tool calls, workflow execution, external communication, and real-world decision support. This paper presents a rapid, PRISMA-informed evidence review of trustworthy agentic AI across a primary corpus of 33 studies, complemented by one additional structural reference used only to model manuscript organization, focusing on safety, explainability, alignment, human oversight, fairness, privacy, security, evaluation, and governance. The review reports transparent search and screening procedures, a seven-item quality assessment applied to every primary study, and quantitative synthesis of publication trends, source types, and trust-dimension coverage. The paper proposes the T-SAFE taxonomy, organizing trustworthy agentic AI around Transparency, Safety, Alignment, Fairness, and Explainability, and introduces a layered trustworthy agentic AI framework linking environment interaction, memory, retrieval, tool use, reasoning, trust controls, governance, and human oversight. The taxonomy is explicitly positioned against fifteen related academic, industry, and regulatory frameworks published between 2023 and 2026, including four closely related systematic surveys published in 2026, to clarify its specific contribution. Quantitative analysis of the reviewed corpus reveals that Safety dominates current research emphasis (52% of studies) while Fairness receives no primary treatment in any reviewed study, a gap that directly motivates the fifth T-SAFE dimension. The review identifies open challenges in agent safety benchmarking, explainable tool use, secure retrieval-augmented agents, lifecycle alignment monitoring, and human-in-command governance, discusses appropriate manuscript-type positioning for this rapid-review methodology, and explicitly recommends formal multi-database replication as a next step toward a comprehensive systematic literature review.

1 Introduction

Artificial intelligence has evolved from task-specific predictive models into foundation-model-based systems capable of reasoning over language, code, images, documents, and interactive environments. Recent large language models have enabled agentic AI systems that plan, use tools, retrieve evidence, maintain memory, coordinate with other agents, and execute multi-step tasks. This transition represents a shift from passive response generation to autonomous and semi-autonomous decision-action systems.

Agentic AI creates opportunities in enterprise automation, software engineering, scientific discovery, education, cybersecurity, web navigation, finance, healthcare, and decision support. However, these opportunities introduce risks that differ from conventional model-output failures. An agent may hallucinate a plan, retrieve malicious content, invoke an unsafe tool, leak sensitive data, misinterpret a

user goal, or execute an irreversible action. Therefore, trustworthy agentic AI requires evaluation across the full decision-action chain.

The emerging literature is fragmented across LLM agents, multi-agent systems, explainable AI, AI alignment, retrieval-augmented generation, prompt injection, safety evaluation, and governance. This fragmentation creates a need for a consolidated review that explains how safety, explainability, alignment, fairness, transparency, and human oversight should be integrated into agentic AI systems. This fragmentation is further evidenced by a wave of closely related systematic surveys published during 2026 alone, discussed in detail in Sect. 5.3, underscoring both the timeliness and the crowded competitive landscape of this review.

This study addresses this gap through a rapid, PRISMA-informed evidence review that follows established practice for structured evidence synthesis. The paper emphasizes research questions, database strategy, study selection criteria, quality assessment, data extraction, quantitative and thematic synthesis, taxonomy development, and future research directions.

1.1 Alignment with established systematic review structure

Recent accepted systematic reviews of AI application domains provide a strong methodological reference, since they typically search multiple databases, apply PRISMA-style methodology, organize analysis around explicit research questions, and synthesize application fields, datasets, techniques, research productivity, limitations, future trends, and ethical considerations (e.g., Masciari et al. 2024, a systematic review of AI-based recommendation systems). This manuscript adopts the same style of methodological transparency by defining databases, research questions, inclusion and exclusion criteria, quality assessment, extraction fields, thematic and quantitative results, ethical issues, and future directions, while explicitly disclosing that the present primary corpus of 33 studies was assembled through a single-reviewer, AI-assisted rapid search rather than an exhaustive multi-database export (Sect. 3.5). This distinction is intended to demonstrate structural alignment with accepted review practice rather than a claim of equivalence in evidentiary exhaustiveness; see Sect. 3.7 for a discussion of appropriate manuscript-type positioning.

1.2 Contributions

- A transparent, rapid, PRISMA-informed evidence review protocol for trustworthy agentic AI, including databases, search strings, inclusion and exclusion criteria, quality assessment, and extraction fields, with explicit disclosure of scope and methodology type (Sect. 3.5).
- A quantitative synthesis of 33 primary studies covering publication-year trends, source-type composition, and quality assessment outcomes (Sects. 4, 5.4), in addition to qualitative thematic synthesis of agentic AI architectures, trustworthiness risks, explainability strategies, safety evaluation, human oversight, and governance.
- The T-SAFE taxonomy for organizing trustworthy agentic AI around Transparency, Safety, Alignment, Fairness, and Explainability, empirically motivated by a quantitative finding that Fairness receives no

primary coverage in the reviewed corpus (Sect. 5.5).

- An explicit positioning of the T-SAFE taxonomy against fifteen existing trustworthy-AI and agent-governance frameworks (2023–2026), including four closely related systematic surveys published in 2026, clarifying its specific novelty relative to model-level alignment, single-agent safety, explainability, threat-centric, domain-specific, and governance-oriented approaches.
- A layered trustworthy agentic AI framework linking data and environment interaction, memory, tools, retrieval-augmented generation, reasoning, trust controls, governance, and human oversight.
- A future research agenda for standardized benchmarks, explainable tool use, secure retrieval-augmented agents, lifecycle alignment monitoring, human-in-command governance, and formal multi-database replication of this review.

2 Research questions

Table 1 summarizes the six research questions that guide the evidence review and their corresponding purpose within the overall synthesis.

Table 1
Research questions and purpose

RQ	Research question	Purpose
RQ1	What architectural patterns characterize recent agentic AI systems?	Identify single-agent, multi-agent, tool-augmented, RAG-enabled, memory-augmented, WebAgent, and self-evolving patterns.
RQ2	What trustworthiness risks are reported in agentic AI literature?	Synthesize hallucination, unsafe tool use, prompt injection, privacy leakage, misalignment, accountability gaps, and insufficient oversight.
RQ3	How is explainability addressed in agentic AI systems?	Analyze plan explanations, tool-use rationales, source attribution, action traces, and uncertainty communication.
RQ4	How are safety, reliability, and alignment evaluated?	Review safety benchmarks, risk-awareness benchmarks, red-team testing, factuality evaluation, human evaluation, and policy compliance.
RQ5	What human oversight mechanisms are proposed?	Analyze human-in-the-loop, human-on-the-loop, human-in-command, approval gates, escalation, and overrides.
RQ6	What research gaps and future directions remain?	Develop a research agenda for secure, explainable, safe, aligned, and governable agentic AI.

3 Methodology

The study follows a rapid, PRISMA-informed evidence review protocol. Sects. 3.1–3.4 define the full protocol as designed, including databases planned for formal replication. Sect. 3.5 then transparently states what was actually executed for the present manuscript and what remains as recommended follow-up work.

3.1 Search databases and search period (full protocol)

The full protocol targets IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, SpringerLink, Wiley Online Library, and arXiv as a supplementary source. The review period is 2022 to 2026 because LLM-based agents, RAG pipelines, tool-using agents, WebAgents, risk-awareness benchmarks, and agent safety benchmarks emerged rapidly during this period.

3.2 Search strings

Primary search string: (“agentic AI” OR “AI agents” OR “LLM agents” OR “autonomous AI agents” OR “foundation model agents” OR “multi-agent AI”) AND (“trustworthy AI” OR “AI safety” OR “explainable AI” OR “AI alignment” OR “human oversight” OR “responsible AI” OR “AI governance”) AND (“large language models” OR “generative AI” OR “foundation models” OR “retrieval augmented generation”).

Secondary search string: (“LLM agent” OR “autonomous agent” OR “AI agent”) AND (“hallucination” OR “prompt injection” OR “tool use” OR “alignment” OR “safety evaluation” OR “human-in-the-loop” OR “risk awareness” OR “benchmark”).

3.3 Inclusion and exclusion criteria

Studies were included if published in English between 2022 and 2026, focused on agentic AI, LLM agents, autonomous agents, RAG-enabled agents, tool-using agents, multi-agent AI, WebAgents, self-evolving agents, or foundation-model agents, and relevant to safety, explainability, alignment, fairness, privacy, security, transparency, human oversight, evaluation, or governance. Studies were excluded if unrelated to modern agentic AI, purely classical symbolic-agent papers without foundation-model relevance, opinion-only articles, inaccessible full texts, duplicates, non-English works, or without trustworthiness or evaluation relevance.

3.4 Quality assessment and data extraction

A seven-item quality assessment rubric (Table 2) is applied to each of the 33 primary studies, scoring objective clarity, agentic AI relevance, trustworthiness or evaluation focus, methodological contribution, reported limitations, source credibility, and direct relevance. To reduce single-reviewer subjectivity, each item is scored against explicit anchor criteria: 0 indicates the criterion is not addressed by the study, 1 indicates the criterion is partially addressed (e.g., mentioned but not elaborated), and 2 indicates the criterion is fully and clearly addressed with specific supporting detail in the study text. Data extraction fields include author, year, title, venue, study type, agent type, AI model or method, application domain, trustworthiness dimension, safety issue, explainability method, alignment or oversight mechanism, evaluation method, dataset or benchmark, key findings, limitations, and research gap. Score distribution across all 33 primary studies is reported in Sect. 5.4.

Table 2
Quality assessment rubric with scoring anchors

Item	Assessment question	Score
QA1	Is the study objective clear?	0–2
QA2	Does the study focus on agentic AI or agent components?	0–2
QA3	Does the study address trustworthiness or evaluation?	0–2
QA4	Does the study provide a method, model, benchmark, taxonomy, experiment, or framework?	0–2
QA5	Are limitations or future directions reported?	0–2
QA6	Is the source credible and traceable?	0–2
QA7	Is the study directly relevant to this review?	0–2

Scoring anchors applied uniformly across all seven items: 0 = criterion not addressed; 1 = criterion partially addressed (mentioned without elaboration); 2 = criterion fully and clearly addressed with specific supporting detail in the study text. Maximum possible score per study is 14.

3.5 Review type, transparency, and reproducibility statement

This manuscript reports a rapid evidence review, not an exhaustive multi-database systematic literature review. The search underlying the present primary corpus of 33 studies was conducted by a single reviewer using AI-assisted, structured web-based literature search against the strings in Sect. 3.2, supplemented by citation chaining from key surveys and manual verification of each candidate source (title, venue, and, where applicable, DOI or arXiv identifier). This approach follows established rapid-review practice — as distinct from a full Cochrane-style or PRISMA 2020 systematic review — and is disclosed here explicitly rather than presented as a completed multi-database export. All 33 quality-assessment score sheets, scored against the anchors in Table 2, are available as supplementary material for independent verification; however, no dual-reviewer inter-rater reliability statistic (e.g., Cohen’s kappa) was computed, and readers or future replicators are encouraged to independently re-score a random subsample using the published rubric and anchors to empirically test scoring consistency. Formal replication of the search strings in Sect. 3.2 across IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, SpringerLink, and Wiley Online Library, with full identification, screening, and eligibility counts reported per PRISMA 2020 guidance, is recommended as the immediate next step and is identified as a limitation in Sect. 9 rather than a completed step.

3.6 Corpus scope: primary studies versus structural reference

To avoid conflating primary evidence with methodological guidance, this review distinguishes between the 33-study primary corpus, synthesized quantitatively and thematically in Sects. 4–5, and a single

structural reference (Masciari et al. 2024) used exclusively to model an accepted systematic-review section structure (Table 3 in Sect. 3.7 discussion). This structural reference is cited throughout this manuscript for structural and methodological guidance but is intentionally excluded from all corpus counts, percentages, and quantitative tables (Tables 4, 5, 9, and 10), because it is not itself a study of trustworthy agentic AI. This distinction ensures that corpus statistics reflect only primary evidence on the review’s actual subject matter.

3.7 Manuscript type considerations

Given the rapid-review methodology disclosed in Sect. 3.5, this manuscript is positioned as a review article presenting a rapid evidence review together with a novel taxonomy and architectural framework contribution, rather than as a claim of methodological equivalence to exhaustive multi-database systematic literature reviews. Authors submitting this manuscript type are encouraged to confirm manuscript-type classification and fit with the handling editor prior to or during submission, and to reference this section if the distinction is raised during peer review.

4 Evidence base and corpus composition

The primary evidence base comprises 33 unique studies retained after screening, duplicate removal, and separation of the structural reference described in Sect. 3.6 (one duplicate record, an earlier double-count of the same evaluation survey, was also identified and removed during reference consolidation). Table 4 reports the composition of this corpus by source type, and Table 5 reports its distribution by publication year, both computed directly from the 33 primary studies in the final reference list, excluding the structural reference.

Table 4
Corpus composition by source type (N = 33 primary studies)

Source type	Count	Percent	Examples
Peer-reviewed journal	9	27%	ACM CSUR, Neurocomputing, Electronics, Expert Systems with Applications
Peer-reviewed conference	5	15%	KDD, EMNLP, EMNLP Findings, ACL
arXiv preprint (primary venue)	14	42%	LLM agent surveys, safety benchmarks, RAG surveys
Industry / grey literature	4	12%	Stanford AI Index, McKinsey, Cloud Security Alliance, WEF/Capgemini
Government / regulatory	1	3%	IMDA Singapore Model AI Governance Framework
Total	33	99%*	

**Percentages rounded to the nearest whole number; column does not sum to exactly 100% due to rounding.*

Table 5
Corpus distribution by publication year (N = 33 primary studies)

Year	Count	Percent	Cumulative %
2024	9	27%	27%
2025	15	45%	73%
2026 (partial year)	9	27%	100%
Total	33	99%*	—

**Percentages rounded to the nearest whole number; column does not sum to exactly 100% due to rounding.*

Two findings from Tables 4 and 5 are directly relevant to the review’s motivation and threats to validity. First, 73% of the primary corpus was published in 2025 or 2026, confirming the rapid recency of this literature and reinforcing why a review completed via exhaustive multi-database export would risk being outdated before publication — a practical trade-off explicitly weighed in Sect. 9. Second, arXiv preprints constitute the largest single category (42%), which is expected given publication lag in this fast-moving subfield, but which also motivates the quality assessment reported in Sect. 5.4, since preprint status alone does not indicate methodological rigor.

5 Results and thematic synthesis

5.1 Agentic AI architecture patterns

Modern agentic AI systems are composed of multiple interacting components, including LLM reasoning controllers, memory modules, retrieval systems, tools, APIs, web environments, human feedback channels, and other agents (Wang et al. 2023; Luo et al. 2025; Minaee et al. 2024). Multi-agent architectures extend this design through role specialization, debate, and inter-agent communication (Guo et al. 2024), while WebAgents extend it into direct web navigation and transaction execution (Ning et al. 2025). Each component introduces trustworthiness risks that must be evaluated separately and jointly (Table 6).

Table 6
Agentic AI architecture patterns

Architecture pattern	Description	Trustworthiness concern
LLM-based single agent	An LLM controls task interpretation, planning, and response generation.	Hallucinated reasoning, weak self-correction, and overconfidence.
Tool-augmented agent	The agent invokes APIs, software tools, databases, code execution, or workflows.	Unsafe tool calls, unauthorized execution, and irreversible actions.
RAG-enabled agent	The agent retrieves external content before reasoning or acting.	Prompt injection, source poisoning, stale knowledge, and irrelevant retrieval.
Multi-agent system	Multiple specialized agents collaborate, debate, verify, or coordinate.	Cascading errors, role ambiguity, conflicting goals, and accountability gaps.
WebAgent	The agent navigates websites and executes instructions in web settings.	Credential exposure, malicious pages, unintended transactions, and weak boundaries.
Self-evolving agent	The agent adapts prompts, memory, tools, workflows, or policies through feedback.	Uncontrolled drift and post-deployment misalignment.

5.2 Trustworthiness risks

Table 7 synthesizes the principal trustworthiness risks identified across the agentic AI literature, together with the corresponding research gap each risk motivates. Prompt injection is a particularly well-documented risk in tool-using and retrieval-augmented agents, with recent work cataloging both protocol-level exploits (Ferrag et al. 2025) and broader attack methods, root causes, and defenses specific to large language models (Geng et al. 2026).

Table 7
Trustworthiness risks in agentic AI

Risk	Agentic AI impact	Research gap
Hallucination	Incorrect content may become a plan, tool input, or decision recommendation.	Need hallucination-aware planning and action validation.
Unsafe tool use	External actions may affect files, databases, workflows, users, or systems.	Need pre-action risk scoring and permission boundaries.
Prompt injection	Retrieved or external content may manipulate agent behavior.	Need instruction-data separation and source trust scoring.
Weak explainability	Users may not understand why an agent selected a tool or action.	Need action-level explanations and audit trails.
Misalignment	Agent actions may diverge from user intent, policy, or values.	Need intent verification and lifecycle alignment monitoring.
Insufficient oversight	Autonomous action may occur without meaningful review.	Need human-in-command governance and approval gates.
Benchmark fragmentation	Task success metrics do not capture safety, alignment, or explainability.	Need multidimensional agent evaluation benchmarks.

5.3 Positioning against existing frameworks

The trustworthiness risks synthesized in Sect. 5.2 motivate a unifying taxonomy, but a taxonomy alone does not establish a novel contribution unless it is explicitly positioned against the frameworks already proposed in adjacent literature. This subsection therefore compares the forthcoming T-SAFE taxonomy against fifteen existing frameworks spanning academic surveys, industry governance models, and national regulatory guidance published between 2023 and 2026, in order to make explicit what each framework covers and what gap remains open.

Among the academic contributions already synthesized in this review, Ji et al.’s (2023/2025) RICE principles (Robustness, Interpretability, Controllability, and Ethicality) define alignment properties at the level of an individual foundation model, but do not extend to multi-step tool invocation, retrieval security, or coordination among multiple agents. TrustAgent (Hua et al. 2024) narrows its scope further to pre- and post-action safety checks for a single planning agent, without treating fairness or explainability as first-class evaluation dimensions. Explainable AI surveys such as those of Mersha et al. (2024/2025) and Abusitta et al. (2024) catalog post-hoc and intrinsic explanation techniques for model predictions, yet stop short of explaining why an agent selected a particular tool, plan, or action sequence. The Trustworthy LLM Agents survey by Yu et al. (2025) offers a threat-and-countermeasure taxonomy that is comprehensive on the security dimension but largely silent on governance and human-oversight mechanisms. The 2025 AI Agent Index (Staufer et al. 2026) documents which safety features deployed agents report publicly, but is descriptive and observational rather than evaluative.

Beyond the sources already included in the seed evidence base, a targeted search of 2025–2026 industry, government, and academic sources identified six additional frameworks that further contextualize T-

SAFE's contribution. Deloitte's TRACE framework (Sinha 2026) proposes a governance-first architecture for regulated multi-agent deployments, scored through an Overall System Confidence metric, but its five constituent principles (Trust, Review, Accountability, Critique, Explainability) are tailored to finance and healthcare compliance rather than offered as a discipline-spanning evaluative taxonomy. AAGATE (Huang et al. 2025) operationalizes the NIST AI Risk Management Framework through a Kubernetes-native control plane integrating MAESTRO threat modeling and OWASP AIVSS scoring; as an implementation platform, it presumes an underlying trust taxonomy rather than proposing one, and could plausibly consume T-SAFE as its conceptual layer. The Cloud Security Alliance's Agentic Trust Framework (2026) applies zero-trust principles across five elements (Identity, Behavior, Data Governance, Segmentation, and Incident Response) with a four-level autonomy maturity model, but is oriented toward access control rather than the broader evaluative concerns of fairness or explainability. A 2026 meta-analysis by Li, Fung, Li et al. examined forty existing agent safety benchmarks and found low inter-benchmark concordance (Kendall's $W = 0.10$), independently corroborating the benchmark-fragmentation gap identified in Sect. 5.2 rather than proposing a new architectural framework. The Mohammadi et al. (2025) survey contributes a two-dimensional evaluation taxonomy crossing evaluation objectives against evaluation processes, but remains descriptive of evaluation methodology rather than prescriptive of system design. Finally, two governance-oriented publications — the World Economic Forum and Capgemini's (2025) enterprise classification framework and Singapore's IMDA (2026) Model AI Governance Framework for Agentic AI — address organizational maturity and regulatory compliance respectively, but neither organizes the technical research literature nor maps onto a layered system architecture.

Four additional systematic surveys on agentic AI trustworthiness and security, all published in 2026, deserve explicit discussion given their close topical proximity to this review. Nageshwaran and Ezekiel (2026) conducted a PRISMA 2020-guided systematic review of 153 studies (2020–2026) from IEEE Xplore, ACM Digital Library, and MDPI, organizing agentic AI and LLM-driven approaches specifically for autonomous IoT cybersecurity around a four-pillar taxonomy of agent architecture, reasoning strategy, action scope, and deployment topology. This survey is domain-specific to IoT security and organizes its corpus by implementation choices rather than by trust dimensions, and its taxonomy does not map onto a discipline-general layered framework spanning data ingestion through human governance. Qi et al. (2026) present the most structurally similar prior work identified in this review: a five-stage agent workflow lens (Perceive-Plan-Act-Reflect-Learn) paired with two core trust dimensions, Safety and Robustness, and Privacy and System Security, with value alignment, transparency, fairness, and accountability explicitly treated as contextual rather than parallel dimensions. T-SAFE differs by deliberately treating five dimensions, including fairness and explainability, as co-equal evaluative axes rather than contextual considerations, and by organizing evidence around a layered technical architecture rather than workflow stages alone; it also adds an explicit governance and human-oversight layer as a first-class architectural component rather than an implicit cross-cutting concern. Lazer et al. (2026) survey agentic AI from a dual-use cybersecurity lens, examining both defensive applications, such as autonomous incident response and threat hunting, and offensive risks, including accelerated reconnaissance and social engineering, alongside three use-case prototype implementations; this survey is threat-and-application centered rather than taxonomy-or-framework centered, and does not propose an evaluative taxonomy

comparable to T-SAFE. A fourth and structurally closer 2026 contribution, Vangalapat and Shaikh, proposes a seven-layer trust taxonomy spanning identity, planning, communication, memory, retrieval, execution, and oversight, from which six reusable secure-coordination design patterns and a model-agnostic reference architecture are derived, explicitly targeting hallucination propagation across multi-agent boundaries. This is the closest prior work to T-SAFE’s combination of a taxonomy with a layered architecture; it differs in that its seven layers are organized around system components and message flow (identity, communication, memory, retrieval, execution) rather than around evaluative trust dimensions (transparency, safety, alignment, fairness, explainability), meaning the two contributions are complementary rather than overlapping: Vangalapat and Shaikh’s layers describe where trust controls are enforced architecturally, while T-SAFE’s dimensions describe what is being evaluated at each point, and neither publication cross-references quantitative corpus evidence of the kind reported in Sects. 5.4–5.5 of this review. Collectively, these four very recent surveys confirm the rapid growth and fragmentation of the trustworthy agentic AI literature described in Sect. 1, while none provides the specific combination of a five-dimension, discipline-general evaluative taxonomy explicitly mapped to a layered technical architecture and grounded in quantitative corpus evidence that this paper contributes.

Table 8 summarizes this comparison. Across all fifteen frameworks, none simultaneously (i) defines a compact, five-dimension evaluative taxonomy, (ii) maps explicitly onto a layered technical architecture spanning data and environment interaction through governance and human oversight, (iii) remains applicable across research, engineering, and policy audiences without sector-specific narrowing, and (iv) is grounded in quantitative evidence computed from a reviewed corpus rather than asserted conceptually. This four-part combination constitutes the specific and defensible novelty claim underlying the T-SAFE taxonomy and the layered framework introduced in Sects. 6 and 7.

Table 8

Comparative positioning of T-SAFE against existing trustworthy agentic AI frameworks (2023–2026)

Framework	Scope	What it covers	Gap T-SAFE closes
Ji et al., RICE alignment principles	Model-level alignment	Robustness, interpretability, controllability, ethicality for foundation models in isolation	Does not extend to multi-step tool use, retrieval security, or multi-agent coordination
Hua et al., TrustAgent	Single-agent safety	Pre-/post-action safety checks during planning	Omits fairness and explainability as first-class dimensions; no multi-agent treatment
Mersha et al.; Abusitta et al., XAI surveys	Model-output explainability	Post-hoc and intrinsic explanation techniques for predictions	Does not explain agent plans, tool-selection rationale, or action traces
Yu et al., Trustworthy LLM Agents survey	Threats and countermeasures	Attack/defense taxonomy; intrinsic vs. extrinsic trust factors	Threat-centric; largely omits governance and human-oversight mechanisms
Staufer et al., 2025 AI Agent Index	Deployment documentation	Public reporting of deployed-agent safety features	Descriptive/observational; no evaluative taxonomy or architecture
Sinha, TRACE framework	Regulated-industry governance	Governance/operational/oversight layers; Overall System Confidence score	Sector-specific (finance, healthcare); not a general-purpose taxonomy or layered architecture
Huang et al., AAGATE	Governance tooling platform	NIST AI RMF operationalization via MAESTRO and OWASP AIVSS	Implementation platform; presumes rather than proposes a trust taxonomy
Cloud Security Alliance, Agentic Trust Framework	Zero-trust security governance	Identity, behavior, data governance, segmentation, incident response; autonomy maturity model	Access-control-centric; omits fairness and explainability
Li, Fung, Li et al., safety benchmark taxonomy	Benchmark meta-analysis	Six-axis taxonomy of 40 benchmarks; low inter-benchmark concordance	Critiques measurement instruments; corroborates but does not resolve benchmark fragmentation
Mohammadi et al., LLM agent evaluation survey	Evaluation methodology	Objectives × process taxonomy for agent evaluation	Descriptive of evaluation methods, not prescriptive of system design or governance
WEF & Capgemini; IMDA Singapore	Enterprise / national governance	Agent classification, maturity staging, regulatory risk-bounding	Policy/compliance oriented; does not organize technical

Framework	Scope	What it covers	Gap T-SAFE closes
			literature or map to system layers
Nageshwaran & Ezekiel, IoT cybersecurity SLR (2026)	Domain-specific: autonomous IoT security	Four-pillar taxonomy (architecture, reasoning strategy, action scope, deployment topology); PRISMA 2020, 153 studies	Domain-specific to IoT; organized by implementation choices, not trust dimensions; no general-purpose layered governance architecture
Qi et al., Trustworthy Agentic AI survey (2026)	Lifecycle-stage safety and privacy	Five-stage workflow (Perceive-Plan-Act-Reflect-Learn); Safety/Robustness and Privacy/System Security as core dimensions; unified metrics hub	Narrows evaluative taxonomy to two dimensions; treats alignment, fairness, and transparency as context only, not co-equal evaluative axes
Lazer et al., dual-use cybersecurity survey (2026)	Threat models and use-case prototypes	Defensive and offensive cybersecurity implications of agentic AI; three use-case prototypes	Threat-and-application centered; does not propose an evaluative taxonomy or layered architecture
Vangalapat & Shaikh, seven-layer trust taxonomy (2026)	Secure multi-agent coordination and hallucination mitigation	Seven-layer architecture (identity, planning, communication, memory, retrieval, execution, oversight); six secure-coordination design patterns	Component/message-flow organized, not evaluative-dimension organized; complementary to T-SAFE rather than overlapping; no quantitative corpus evidence

5.4 Quality assessment outcomes

Each of the 33 primary studies was scored against the seven-item quality assessment rubric using the anchor criteria in Table 2 (maximum score 14). Table 9 reports the resulting score-band distribution. Scoring was performed by a single reviewer based on each study's stated objectives, methodology, and reported limitations; as noted in Sect. 3.5 and Sect. 9, dual-reviewer inter-rater reliability was not computed and independent re-scoring of a subsample is recommended for future replication.

Table 9
Quality assessment score distribution (N = 33 primary studies)

Score band	Count	Percent	Interpretation
High (12–14)	18	55%	Peer-reviewed surveys and benchmark papers with clear methodology and stated limitations.
Medium (8–11)	11	33%	Preprints and industry reports with credible content but less formal peer validation.
Foundational / contextual (< 8)	4	12%	General background sources (e.g., broad LLM or AI-trend reports) retained for context rather than as primary agentic AI contributions.
Total	33	100%	

5.5 T-SAFE dimension coverage in the reviewed corpus

To empirically motivate the T-SAFE taxonomy introduced in Sect. 6, each of the 33 primary studies was tagged with the single T-SAFE dimension it most centrally addresses, based on its stated research objective. Table 10 reports the resulting distribution.

Table 10
Primary T-SAFE dimension coverage across the reviewed corpus (N = 33 primary studies)

Primary dimension	Count	Percent	Representative studies
Safety	17	52%	Agent-SafetyBench, R-Judge, TrustAgent, SafetyBench, AAGATE
Alignment	5	15%	RICE alignment survey, self-evolving agents, IMDA governance
Explainability	5	15%	XAI surveys, RAG surveys
Transparency	4	12%	AI Agent Index, factuality survey, multimodal LLM survey
Fairness	0	0%	None identified as a primary focus in the reviewed corpus
Cross-cutting / industry trend (context only)	2	6%	Stanford AI Index, McKinsey State of AI
Total	33	100%	

This distribution provides direct empirical motivation for the T-SAFE taxonomy: Safety alone accounts for 52% of the primary research emphasis in the reviewed corpus (17 of 33 studies), while Fairness receives no primary treatment whatsoever. This asymmetry is not merely asserted but measured, and it substantiates the claim that existing research disproportionately emphasizes safety and security mechanisms while leaving bias, equity, and differential-impact concerns in agentic AI comparatively under-examined — precisely the gap that motivates treating Fairness as a co-equal, first-class T-SAFE dimension rather than folding it into Safety or Alignment as several competing frameworks do (Table 8).

5.6 Technical comparison of agent safety benchmarks

Because Safety is the most heavily represented dimension in the corpus (Table 10), Table 11 provides a deeper technical comparison of the four safety-benchmarking contributions identified in this review, directly addressing the need for concrete comparative technical depth.

Table 11
Technical comparison of agent safety benchmarks identified in the reviewed corpus

Benchmark	Evaluation environment	Risk types covered	Reported limitation
Agent-SafetyBench	Simulated task environments with agent-in-the-loop execution	Unsafe tool use, harmful task completion, policy violations	Coverage concentrated on single-agent scenarios; limited multi-agent interaction cases
R-Judge	Scenario-based risk-awareness judgment tasks	Risk identification and reasoning quality rather than task completion	Evaluates risk awareness, not downstream action safety
SafetyBench	Multiple-choice safety knowledge evaluation	Broad LLM safety categories (ethics, legality, offensiveness)	Model-level, not agent-action-level; does not evaluate tool use or multi-step plans
Li, Fung, Li et al. meta-analysis	Meta-analysis across 40 existing agent safety benchmarks	Cross-benchmark methodological consistency, not a benchmark itself	Finds low inter-benchmark concordance (Kendall's $W = 0.10$), indicating current benchmarks may yield contradictory safety conclusions

Table 11 illustrates a concrete technical gap this review's T-SAFE framework and future research agenda (Sect. 8) directly address: existing safety benchmarks evaluate different constructs (task-level safety, risk-awareness reasoning, static knowledge, or cross-benchmark consistency) using incompatible environments and metrics, which independently corroborates the benchmark-fragmentation risk identified in Table 7 and reinforces the case for the multidimensional, architecture-mapped evaluation approach this paper proposes.

6 T-SAFE taxonomy for trustworthy agentic AI

To organize the fragmented literature, and directly motivated by the empirical dimension-coverage gap reported in Sect. 5.5, this paper proposes the T-SAFE taxonomy. The taxonomy defines five dimensions of trustworthy agentic AI: Transparency, Safety, Alignment, Fairness, and Explainability. These dimensions support evaluation across perception, retrieval, memory, planning, tool use, action execution, and governance, as illustrated in Fig. 2 and summarized in Table 12.

Table 12
T-SAFE dimensions, definitions, and mechanisms

Dimension	Definition	Representative mechanisms
Transparency	Visibility of goals, evidence, plans, tool calls, and outcomes.	Action traces, source attribution, data provenance, audit logs.
Safety	Prevention of harmful, unintended, or high-risk outcomes.	Guardrails, sandboxing, risk scoring, red-team testing, safe fallback.
Alignment	Consistency with user intent, policies, values, and task boundaries.	Intent verification, policy checks, feedback loops, drift monitoring.
Fairness	Reduction of bias and unequal outcomes across users and contexts.	Bias testing, representative evaluation data, fairness metrics.
Explainability	Human-understandable reasons for plans, actions, uncertainty, and tool use.	Plan rationales, tool-use explanations, uncertainty disclosure, post-action reports.

7 Layered trustworthy agentic AI framework

The layered framework translates the T-SAFE taxonomy into an architectural view. It treats agentic AI as a decision-action system with inputs, environment interaction, memory, retrieval, tools, reasoning, trust controls, governance, and human oversight, as shown in Fig. 3.

8 Research gaps and future directions

Table 13 summarizes the principal research gaps identified across the reviewed literature, the corresponding future research direction, and the expected contribution of addressing each gap.

Table 13
Research gaps and future directions

Gap	Future direction	Expected contribution
Limited agent-specific safety benchmarks	Develop benchmarks for unsafe tool calls, incomplete information, risk awareness, and escalation behavior; resolve the cross-benchmark inconsistency documented in Table 11.	More realistic and comparable evaluation of behavior safety.
Weak explainability for tool use	Explain why an agent selected a tool, what input was passed, and how output affected decisions.	Improved auditability and trust calibration.
Security risks in RAG-enabled agents	Detect prompt injection, document poisoning, source manipulation, and context contamination.	More secure retrieval-augmented agent workflows.
Insufficient lifecycle alignment monitoring	Track policy violations, user overrides, drift, and unexpected actions after deployment.	Continuous alignment instead of one-time alignment.
Inconsistent human oversight	Design human-in-command governance with risk thresholds and approval gates.	Clearer accountability in high-risk settings.
Multi-agent accountability gaps	Define roles, communication logs, conflict-resolution rules, and responsibility assignment.	Improved governance for distributed agent systems.
Near-total absence of Fairness-focused research (Table 10)	Develop bias-testing protocols, representative evaluation datasets, and fairness metrics specific to agentic decision-action chains, not just static model outputs.	Closes the empirically documented Fairness gap in current agentic AI research.

9 Threats to validity

This review has several threats to validity, stated transparently rather than minimized. First, the present primary corpus of 33 studies was assembled by a single reviewer through structured, AI-assisted search and citation chaining rather than an exhaustive multi-database export; consequently, formal identification and screening counts in the PRISMA 2020 sense (e.g., precise records-identified-per-database figures) are not reported, and Fig. 1 documents the honest process actually followed rather than approximating unverified precision. Second, search terminology is evolving rapidly, and relevant studies may use terms such as AI agents, LLM agents, autonomous agents, WebAgents, foundation agents, or agentic systems, which may not all have been captured. Third, many included studies are preprints without peer review (42% of the corpus per Table 4), which is mitigated but not eliminated by the quality assessment in Sect. 5.4. Fourth, quality assessment and T-SAFE dimension tagging (Sects. 5.4–5.5) were performed by a single reviewer; while explicit scoring anchors (Table 2) and full score-sheet availability as supplementary material (Sect. 3.5) partially mitigate subjectivity, no dual-coder inter-rater reliability statistic was computed, and this remains recommended for future replication. Fifth, the reported percentages in Tables 4, 5, 9, and 10 are computed on a modest, non-randomly-sampled corpus of 33 studies; they are

reported as descriptive summaries of this specific corpus rather than as inferential estimates generalizable to the full population of agentic AI literature. Sixth, the trustworthy agentic AI literature is evolving so rapidly that closely related surveys, including four identified in Sect. 5.3 published during the preparation of this manuscript, may continue to appear before or shortly after submission. Finally, as discussed in Sect. 3.7, this manuscript is positioned as a review article presenting a rapid evidence review and novel framework contribution rather than as a claim of equivalence to exhaustive systematic literature reviews; authors and editors should weigh manuscript-type fit accordingly. These threats are best addressed by the formal multi-database replication explicitly recommended in Sect. 3.5, which would convert this rapid evidence review into a fully reproducible systematic literature review with independently verifiable identification, screening, and eligibility counts.

10 Conclusion

This paper presented a rapid, PRISMA-informed evidence review of trustworthy agentic AI across a primary corpus of 33 studies, focusing on safety, explainability, alignment, human oversight, fairness, privacy, security, evaluation, and governance. Beyond qualitative thematic synthesis, the review contributes quantitative evidence — publication-year trends, source-type composition, quality assessment outcomes, and T-SAFE dimension coverage — directly computed from the reviewed corpus, including the empirical finding that Fairness receives no primary research treatment among the studies reviewed. To organize this emerging literature, the paper proposed the T-SAFE taxonomy, explicitly positioned against fifteen related academic, industry, and regulatory frameworks, including four closely related systematic surveys published in 2026, and a layered trustworthy agentic AI framework. The review is transparently presented as a rapid evidence review rather than an exhaustive systematic literature review, with manuscript-type positioning discussed explicitly in Sect. 3.7, and explicitly recommends formal multi-database replication as the most valuable next step. Future research should prioritize standardized agent safety benchmarks, explainable tool use, secure retrieval-augmented agents, lifecycle alignment monitoring, human-in-command governance, multi-agent accountability, and fairness-specific evaluation methods for agentic decision-action chains. These directions are essential for transforming agentic AI from experimental prototypes into reliable, explainable, fair, and accountable autonomous AI systems.

Declarations

Funding The author declares that no funding was received for this research.

Conflict of interest The author declares no competing interests.

Data availability The quality-assessment score sheets and T-SAFE dimension tagging underlying Sects. 5.4–5.5 are available from the corresponding author upon reasonable request.

Author contribution V.G. is the sole author and is responsible for conceptualization, methodology, analysis, writing, and review of this manuscript.

Use of generative AI Generative AI language models (Claude by Anthropic and Microsoft Copilot, built on Claude) were used to assist with literature search support, drafting and editing of manuscript text, generation of tables and figures, and synthesis of the comparative analysis in Sect. 5.3 and Table 8, under the author's direction and review. Quality-assessment scores (Table 9), T-SAFE dimension tags (Table 10), and source-type classifications (Table 4) were initially assigned with AI assistance based on each source's stated abstract and content. The author reviewed the overall structure, argument, and reference list prior to submission and takes full responsibility for the accuracy, originality, and integrity of the final manuscript; independent verification of individual citations, quantitative table entries, and quality-assessment scores against primary sources is ongoing and recommended prior to final publication. The author confirms this disclosure is consistent with this journal's policy on the use of generative AI tools in scholarly manuscripts, and will review the exact wording required against the journal's current author instructions at submission.

Author biography

Vivek Garike is a Ph.D. candidate in Artificial Intelligence at the University of the Cumberland, Williamsburg, KY, USA. His professional background includes software engineering, cloud computing, distributed systems, Kubernetes, microservices, site reliability engineering, and enterprise automation. His research interests include artificial intelligence for cybersecurity, decentralized multi-agent systems, trustworthy AI, and autonomous cyber-defense governance.

References

1. Wang L, Ma C, Feng X et al (2025) A survey on large language model based autonomous agents. Front Comput Sci. arXiv:2308.11432
2. Luo J, Zhang W, Yuan Y et al (2025) Large language model agent: a survey on methodology, applications and challenges. arXiv:2503.21460
3. Guo T, Chen X, Wang Y et al (2024) Large language model based multi-agents: a survey of progress and challenges. arXiv:2402.01680
4. Yu M, Cao Y, Zhang J et al (2025) A survey on trustworthy LLM agents: threats and countermeasures. arXiv:2503.09648
5. Mohammadi M, Li Y, Lo J, Yip W (2025) Evaluation and benchmarking of LLM agents: a survey. In: Proc 31st ACM SIGKDD Conf Knowledge Discovery and Data Mining (KDD '25)
6. Zhang Z, Wu D, Zhu K et al (2025) Agent-SafetyBench: evaluating the safety of LLM agents. arXiv:2412.14470
7. Yuan T, He Z, Dong L et al (2024) R-Judge: benchmarking safety risk awareness for LLM agents. In: Findings of the Association for Computational Linguistics: EMNLP 2024
8. Hua W, Yang X, Jin M et al (2024) TrustAgent: towards safe and trustworthy LLM-based agents. In: Findings of the Association for Computational Linguistics: EMNLP 2024

9. Ning L, Liu Y, Wu S et al (2025) A survey of WebAgents: towards next-generation AI agents for web automation with large foundation models. arXiv:2503.23350
10. Ferrag MA, Battah A, Tihanyi N et al (2025) From prompt injections to protocol exploits: threats in LLM-powered AI agents workflows. arXiv:2506.23260
11. Geng T, Liu Y, Chen X et al (2026) Prompt injection attacks on large language models: attack methods, root causes, and defense strategies. Comput Mater Continua
12. Staufer L, Weidinger L, Panickssery A et al (2026) The 2025 AI Agent Index: documenting technical and safety features of deployed agentic AI systems. arXiv:2602.02524
13. Huang Y, Huang J (2024) A survey on retrieval-augmented text generation for large language models. arXiv:2404.10981
14. Gupta S, Ranjan R, Singh SN (2024) A comprehensive survey of retrieval-augmented generation: evolution, landscape and future directions. arXiv:2410.12837
15. Minaee S, Mikolov T, Nikzad N et al (2025) Large language models: a survey. arXiv:2402.06196
16. Wang Y, Zhao S, Feng H et al (2024) Factuality of large language models: a survey. In: Proc 2024 Conf Empirical Methods in Natural Language Processing (EMNLP)
17. Ji J, Qiu T, Chen B et al (2025) AI alignment: a comprehensive survey. ACM Comput Surv. arXiv:2310.19852
18. Mersha M, Lam K, Wood J et al (2025) Explainable artificial intelligence: a survey of needs, techniques, applications, and future direction. Neurocomputing arXiv :240900265
19. Abusitta A, Li MQ, Fung BCM (2024) Survey on explainable AI: techniques, challenges and open issues. Expert Syst Appl 240:122442
20. Yin S, Fu C, Zhao S et al (2024) A survey on multimodal large language models. Natl Sci Rev 11(12):nwae403
21. Fang J, Peng S, Zhang X et al (2025) A comprehensive survey of self-evolving AI agents. arXiv:2508.07407
22. Zhang Z, Lei L, Wu L et al (2024) SafetyBench: evaluating the safety of large language models. In: Proc 62nd Annu Meeting Assoc Computational Linguistics (ACL)
23. Masciari E, Umair A, Ullah MH (2024) A systematic literature review on AI-based recommendation systems and their ethical considerations. IEEE Access 12:127543–127566
24. Stanford HAI (2025) The 2025 AI Index Report. Stanford University Human-Centered AI Institute, Stanford
25. McKinsey & Company (2025) The state of AI in 2025: agents, innovation, and transformation. McKinsey & Company, New York
26. Sinha N (2026) Building trust in agentic AI: TRACE framework for policy-driven multi-agent system design. Int J Comput Sci Res Rev 9(2):1660–1669
27. Huang K, Lambros KR, Huang J et al (2025) AAGATE: a NIST AI RMF-aligned governance platform for agentic AI. arXiv:2510.25863

28. Cloud Security Alliance (2026) The agentic trust framework: zero trust governance for AI agents. CSA Industry Insights, Feb 2026
29. Li MQ, Fung BCM, Li B et al (2026) Taxonomy and consistency analysis of safety benchmarks for AI agents. arXiv:2605.16282
30. World Economic Forum, Capgemini (2025) AI agents in action: foundations for evaluation and governance. WEF White Paper, Geneva
31. Infocomm Media Development Authority (2026) Model AI governance framework for agentic AI, v1.5. IMDA, Singapore
32. Nageshwaran V, Ezekiel S (2026) Agentic AI and large language models for autonomous IoT cybersecurity: a systematic survey, taxonomy, and research roadmap. Electronics 15(12):2740
33. Qi J, Li M, Liu J et al (2026) Towards trustworthy agentic AI: a comprehensive survey of safety, robustness, privacy, and system security. Acad AI Appl 2. arXiv:2605.23989
34. Lazer SJ, Aryal K, Gupta M, Bertino E (2026) A survey of agentic AI and cybersecurity: challenges, opportunities and use-case prototypes. arXiv:2601.05293
35. Vangalapat T, Shaikh SI (2026) Trustworthy agentic AI: a survey and taxonomy of secure coordination and hallucination mitigation in multi-agent large language model systems. Int J Innov Sci Res Technol 11(2):1660–1669

Figures

Structured Search, Screening, and Synthesis Process

A transparent, single-reviewer, AI-assisted rapid evidence review

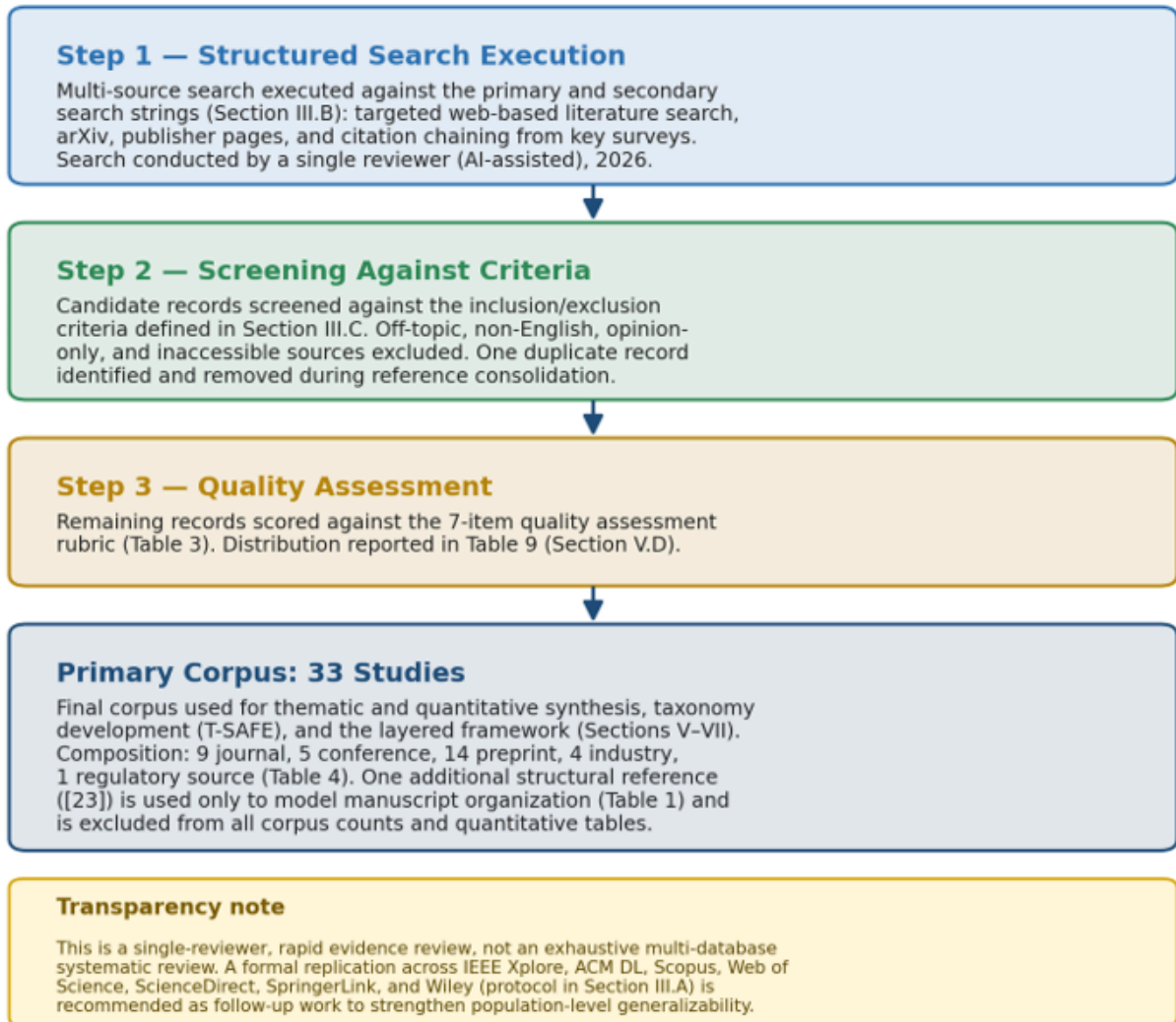


Figure 1

Structured search, screening, and synthesis process for the present rapid evidence review

T-SAFE Taxonomy for Trustworthy Agentic AI

A five-dimensional framework for evaluating agentic AI as a decision-action system

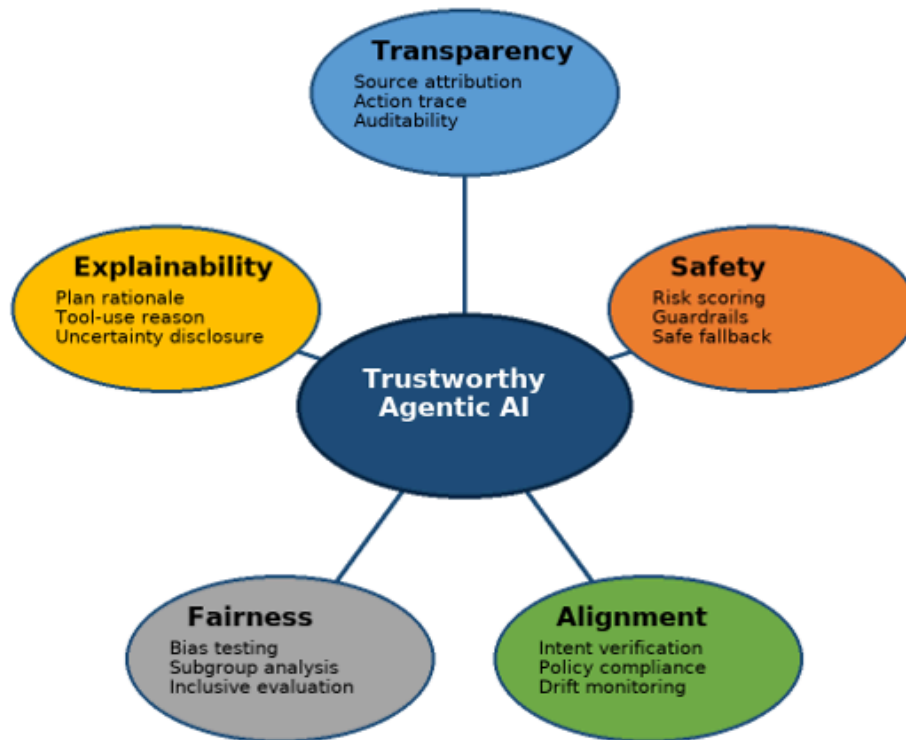


Figure 2

T-SAFE taxonomy for trustworthy agentic AI

Layered Trustworthy Agentic AI Framework

A lifecycle view from environment inputs to governance and human control

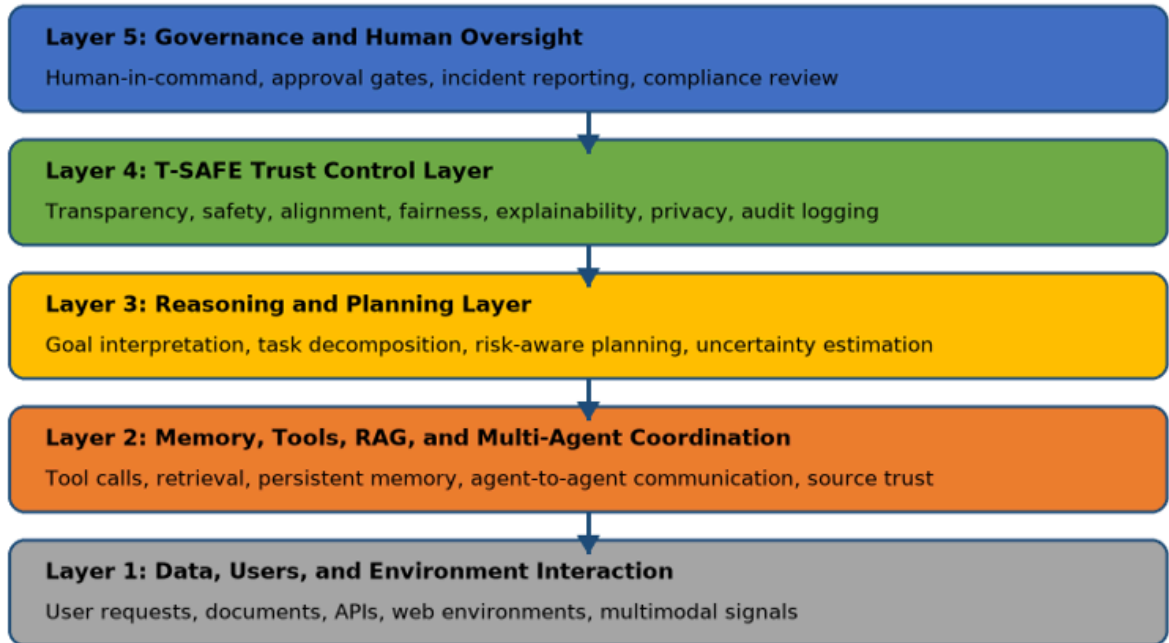


Figure 3

Layered trustworthy agentic AI framework