Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

From Raw Scientific Data to Trusted Discovery: Agentic AI for FAIR and Reproducible Data Science: A Systematic Review of Data Engineering, Semantic Data Models, Multimodal Analytics, Provenance, and Human Oversight

Jamal Shah[1]*, Sunila Kiran[2], Muhammad Hasnain[3], Amir Sajjad[4], Inam Ullah Khan[5], Ali Hamza[6], Arsalan Ahmad Khan[7]

[1]Computer Science, Center for Excellence in IT, Institute of Management Sciences, Peshawar 25130, Pakistan
[2]Riphah International University, Islamabad, Pakistan
[3]University of Hull, Cottingham Road, Hull HU6 7RX, United Kingdom
[4]Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, 71 Al-Farabi Avenue, Almaty 050040, Kazakhstan
[5]Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, 71 Al-Farabi Avenue, Almaty 050040, Kazakhstan
[6]Department of Computer Science, National College of Business Administration & Economics (NCBA&E), East Canal Campus, Lahore, Punjab 53400, Pakistan
[7]Department of Computer Science, Faculty of Computing, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan

DOI: https://doi.org/10.36347/sjet.2026.v14i09.006 | **Received:** 28.07.2026 | **Accepted:** 16.09.2026 | **Published:** 17.09.2026

***Corresponding author:** Jamal Shah
Computer Science, Center for Excellence in IT, Institute of Management Sciences, Peshawar 25130, Pakistan

Abstract | **Review Article**

Agentic artificial intelligence (AI) is beginning to change scientific data work from a sequence of manually connected tasks into a set of tool-mediated, partially autonomous workflows. That shift creates a practical question: can agents accelerate analysis without weakening the findability, accessibility, interoperability, reusability (FAIR), and reproducibility on which scientific claims depend? We systematically reviewed 218 peer-reviewed papers, preprints, and community standards published from 2016 to September 2026. The evidence was organized into five linked domains: data engineering and AutoML; semantic data models and knowledge graphs; multimodal analytics and retrieval-augmented generation; provenance and reproducibility; and human oversight, trust, and governance. Across these domains, the literature shows clear progress in execution-grounded data preparation, workflow orchestration, semantic retrieval, and multimodal evidence synthesis. It also shows a consistent boundary: systems can generate plausible analyses faster than they can establish that those analyses are complete, reproducible, and scientifically warranted. The most persistent weaknesses are incomplete lineage, unstable dependencies, weak cross-modal semantics, inconsistent evaluation, and poorly calibrated human reliance. We therefore frame trustworthy agentic science as a data-to-decision problem rather than a model-selection problem. Agents should produce claim-level provenance, machine-actionable semantic metadata, executable environments, uncertainty-aware outputs, and explicit escalation points for human review. The review concludes with a staged research agenda and an integrated framework that links FAIR practice, reproducibility engineering, and human governance across the scientific lifecycle.

Keywords: Agentic AI; FAIR data principles; reproducible data science; data engineering; semantic data models; knowledge graphs; multimodal analytics; data provenance; human oversight; systematic review.

1. INTRODUCTION

Scientific research now produces data at a scale and heterogeneity that challenge the conventions of file-based storage and manually curated analysis. Advances in instrumentation, simulation, and networked sensing have expanded the volume, velocity, and variety of observations, while the FAIR Guiding Principles established a machine-actionable vocabulary for making digital resources findable, accessible, interoperable, and reusable [1, 2]. FAIR has since become a shared reference point for funders, publishers, and research infrastructures, but implementation remains uneven:

metadata are frequently incomplete or non-standard, links between data and the activities that produced them are lost, and many datasets cannot be used reliably without substantial human reconstruction [3-5].

Agentic AI has emerged in response to this gap. Large language model (LLM)-based agents can call software tools, inspect intermediate results, revise a plan, and coordinate multiple specialised components over a long task horizon [6]. Recent surveys show rapid growth in application-oriented systems, but they also report limited coverage of the complete scientific lifecycle and immature mechanisms for accountability and adaptation [7-9]. This distinction matters: an agent that can produce a polished output is not necessarily an agent that can expose the assumptions, data transformations, and execution conditions required to reproduce that output.

The central problem is therefore not whether an agent can perform an isolated analysis, but whether it can preserve the chain of meaning that connects an observation to a decision. A trustworthy system must be able to prepare data without silent leakage, interpret heterogeneous measurements without collapsing their semantics, retrieve evidence without fabricating support, and explain which person, model, tool, and dataset contributed to a claim. These requirements connect data engineering to semantic representation, multimodal learning, provenance, and human oversight rather than treating them as independent software features.

The stakes are heightened by well-established weaknesses in computational reproducibility. Replication studies have shown that published findings can be difficult to verify even when no autonomous system is involved [10]. Recent open-reproduction efforts in machine learning report incomplete claim-level verification, disagreement between independent reproduction teams, and substantial variation in whether generated code executes under a clean environment [11-14]. These results should not be read as a single estimate of scientific reliability; they are evidence that reproducibility is a layered property, and that failures in dependencies, citations, notebooks, or data lineage can compound before a reader ever evaluates the headline result.

The same technologies also offer practical routes to improvement. Community specifications such as the Research Data Alliance Data Director blueprint, provenance-first workflows such as F(AI)²R, and FAIR data services such as FOXDEN treat metadata, attribution, and evidence as operational artefacts rather than as after-the-fact documentation [15-18]. Their common premise is that automation becomes scientifically useful when it leaves a verifiable record of what was selected, transformed, executed, and checked.

Against this background, this review synthesizes the literature on agentic AI for FAIR and reproducible data science around five connected questions. First, how are agents used for data acquisition, preparation, quality assessment, feature engineering, and model development? Second, how do ontologies, knowledge graphs, semantic layers, and FAIR digital objects provide the context needed for grounded reasoning? Third, how are text, images, tables, signals, spectra, and other modalities aligned for scientific inference? Fourth, which provenance standards and reproducibility practices make agentic outputs auditable? Fifth, how should human oversight, uncertainty communication, and governance be designed when systems operate over long horizons?

The review makes four contributions. It develops a unified taxonomy that connects agentic capabilities to FAIR sub-principles; synthesizes 218 studies across architecture, evaluation, provenance, and governance; identifies the coupled bottlenecks of incomplete reproducibility, autonomy-auditability tension, limited novelty beyond training distributions, and miscalibrated human reliance; and proposes a provenance-first research agenda supported by ten figures and six comparative tables.

Section 2 describes the review protocol, search strategy, eligibility criteria, and quality assessment. Sections 3-7 examine data engineering, semantic infrastructure, multimodal analytics, provenance, and human oversight in turn. Section 8 integrates these findings into a cross-cutting research agenda, and Section 9 concludes with design principles for trustworthy agentic data science.

2. METHODOLOGY

2.1. Review Protocol

We followed the PRISMA 2020 reporting guidance for systematic reviews and documented the protocol, search sources, eligibility decisions, and synthesis procedures before analysis [19]. Because the literature includes AI-assisted review methods and rapidly changing technical standards, we also recorded the role of automated tools and retained a human-verifiable audit trail for every included record [20, 21].

2.2. Search Strategy

We searched Scopus, Web of Science, IEEE Xplore, the ACM Digital Library, PubMed, and arXiv, and used targeted searches of Google Scholar, Semantic Scholar, and Zenodo to identify preprints, standards, and community specifications. The search covered January 2016 through September 2026. Query terms combined three concept groups: agentic and autonomous AI; FAIRness, provenance, and reproducibility; and data engineering, knowledge graphs, multimodal analytics, and scientific discovery. Boolean operators and field filters were adapted to each database, and the complete strings are provided in the supplementary material.

2.3. Inclusion and Exclusion Criteria

We included a study when it proposed, implemented, or evaluated an agentic system that supported at least one stage of the data-science lifecycle and made a substantive contribution to FAIR data practice, provenance, reproducibility, trust, or governance. Eligible records included peer-reviewed articles, conference papers, preprints, and community standards published in English between 2016 and 2026. We excluded work focused solely on robotics, gaming, or unrelated applications; records that mentioned agents without a technical contribution; duplicates, editorials, and commentaries; and studies that did not provide enough methodological detail for comparison.

2.4. Screening and Selection

After deduplication, the search returned 4,259 records. Title and abstract screening retained 2,847 records for full-text assessment, and 1,413 records were excluded at that stage. Of the 2,647 full texts assessed, 2,429 were excluded because they were out of scope, lacked an agentic component, did not address FAIRness or reproducibility, or provided insufficient methodological detail. The final corpus comprised 218 studies. The PRISMA flow diagram in Figure 1 reports these decisions and their stated reasons.

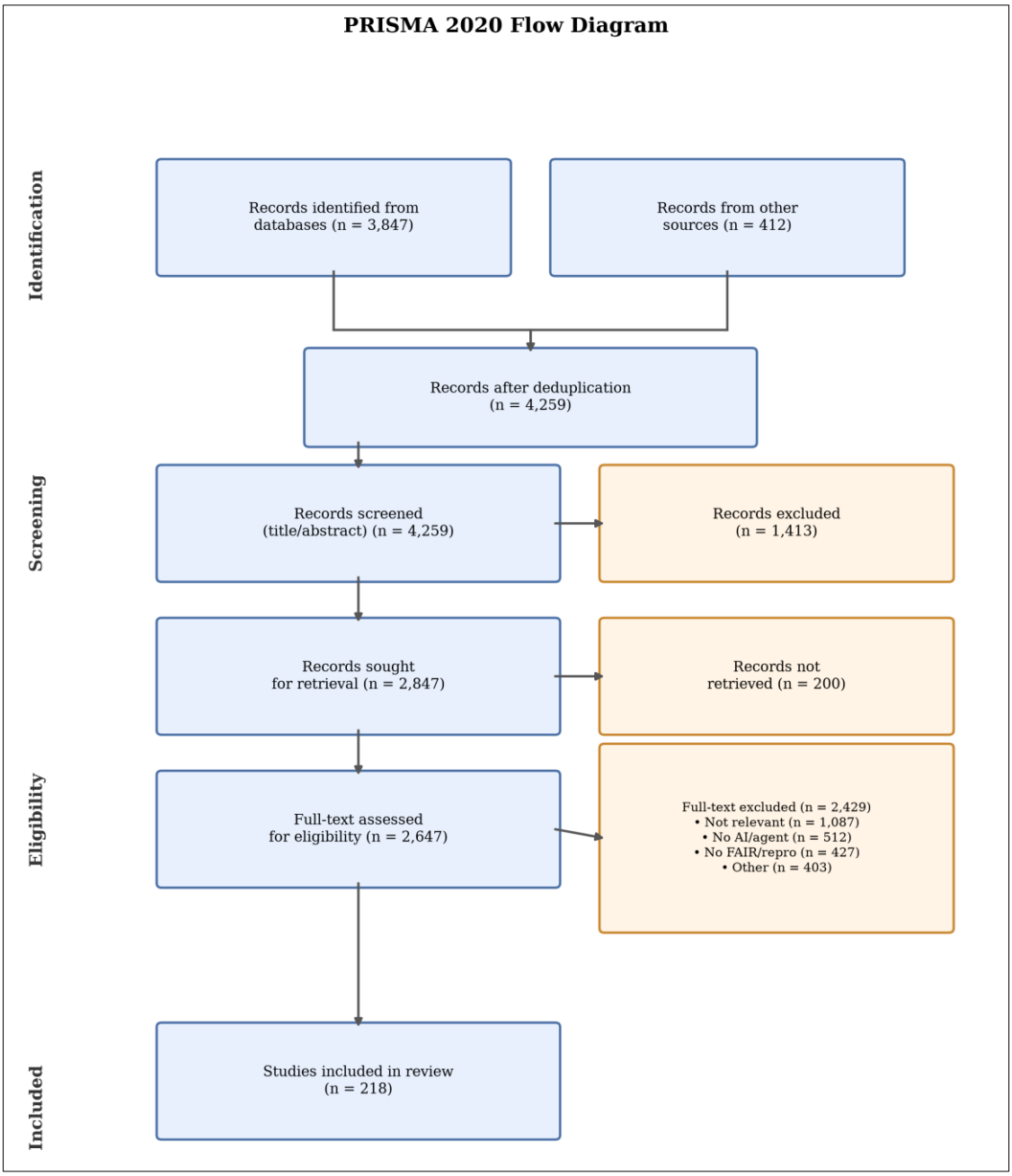

Figure 1: PRISMA 2020 flow diagram for study identification, screening, eligibility assessment, and inclusion. The final corpus comprised 218 studies

2.5. Data Extraction and Analysis

For each included record, we extracted bibliographic metadata, agent architecture, autonomy level, reasoning strategy, lifecycle coverage, modalities, FAIR principles addressed, provenance mechanisms, evaluation design, human-oversight pattern, and principal limitations. We combined descriptive statistics and temporal analysis with thematic coding and taxonomy construction. The resulting taxonomy separates lifecycle coverage from cross-cutting properties such as reasoning style, multimodality, verification, and governance, allowing a system to be recognised as broad in scope without being mistaken for fully autonomous.

2.6. Quality Assessment

Study quality was assessed with an adapted Mixed Methods Appraisal Tool (MMAT) [22]. The assessment considered the clarity of the system description, evaluation rigor, reproducibility of reported results, treatment of FAIR principles, and treatment of human oversight. Each dimension was rated high, medium, or low. Records with three or more low ratings were flagged in the extraction sheet but retained so that the synthesis would represent the field rather than only its best-documented examples.

2.7. Limitations of the Review

The review has five limitations. Rapid publication may have left very recent work unidentified; restricting the search to English introduces language bias; preprints and standards are not uniformly peer reviewed; evaluation protocols are too heterogeneous for a formal pooled effect estimate; and the results represent a time-bounded view of a field that is changing unusually quickly. Figure 2 provides the publication-year and venue distribution of the included corpus, which should be read as a description of this search rather than as a census of all agentic-AI research.

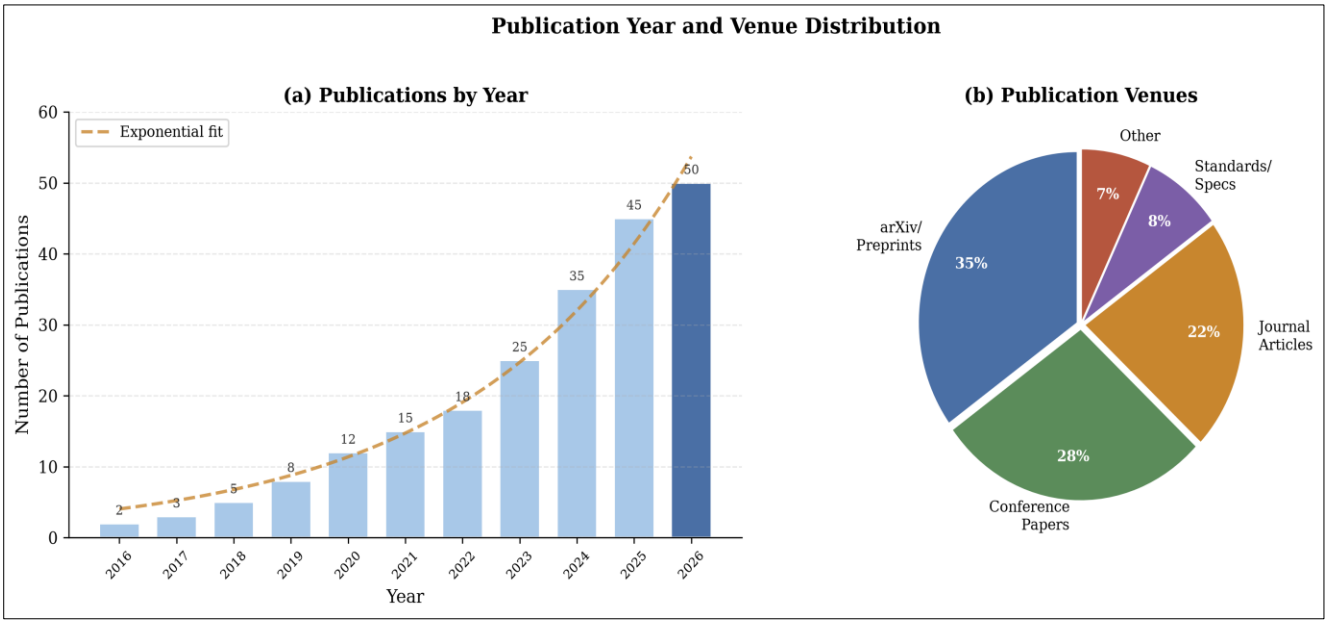

Figure 2: Distribution of the reviewed literature by publication year and venue

3. Data Engineering and Agentic Pipelines

3.1. The Data Science Lifecycle and Agent Coverage

An end-to-end data-science workflow is best understood as a chain of decisions rather than a single model call. The six stages used in the reviewed taxonomy are business understanding, data acquisition, exploratory analysis, feature engineering, model development, and deployment. The 45 surveyed data-science agents are concentrated in exploratory analysis and model development, whereas business framing and deployment remain less consistently supported [8]. Figure 3 makes this asymmetry explicit and separates stage coverage from claims of autonomy.

Capability breadth and autonomy are related but distinct. The L0-L5 taxonomy ranges from a non-autonomous assistant that answers a query to a fully autonomous system that can formulate and execute new methods [9]. Used carefully, the scale clarifies which decisions remain with a researcher, which are delegated to an agent, and where verification should occur. The distribution of surveyed systems across these levels is shown in Figure 4; the predominance of supervised and semi-autonomous systems is more informative than the label of full autonomy itself.

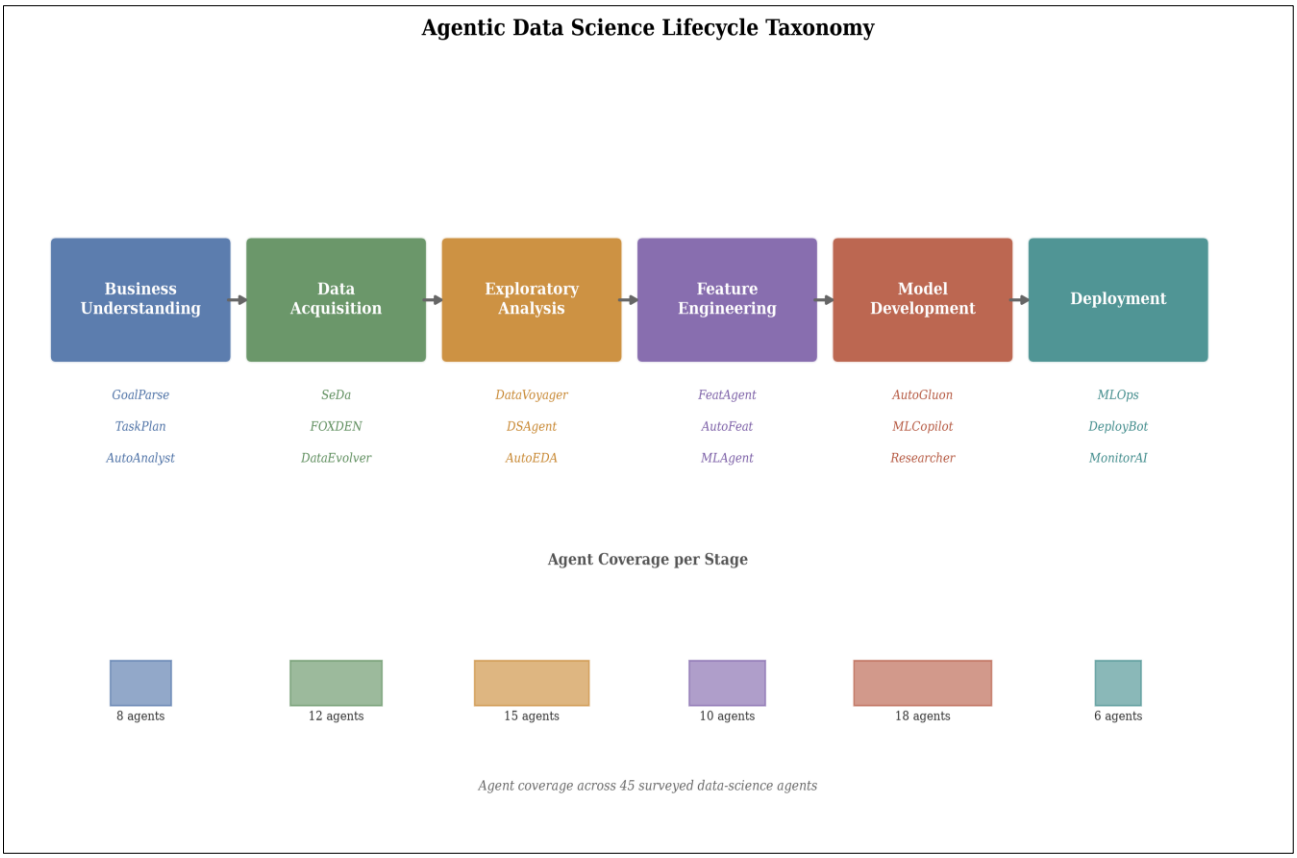

Figure 3: Coverage of the six-stage data-science lifecycle across the 45 surveyed data-science agents. Bars indicate the number of systems addressing each stage

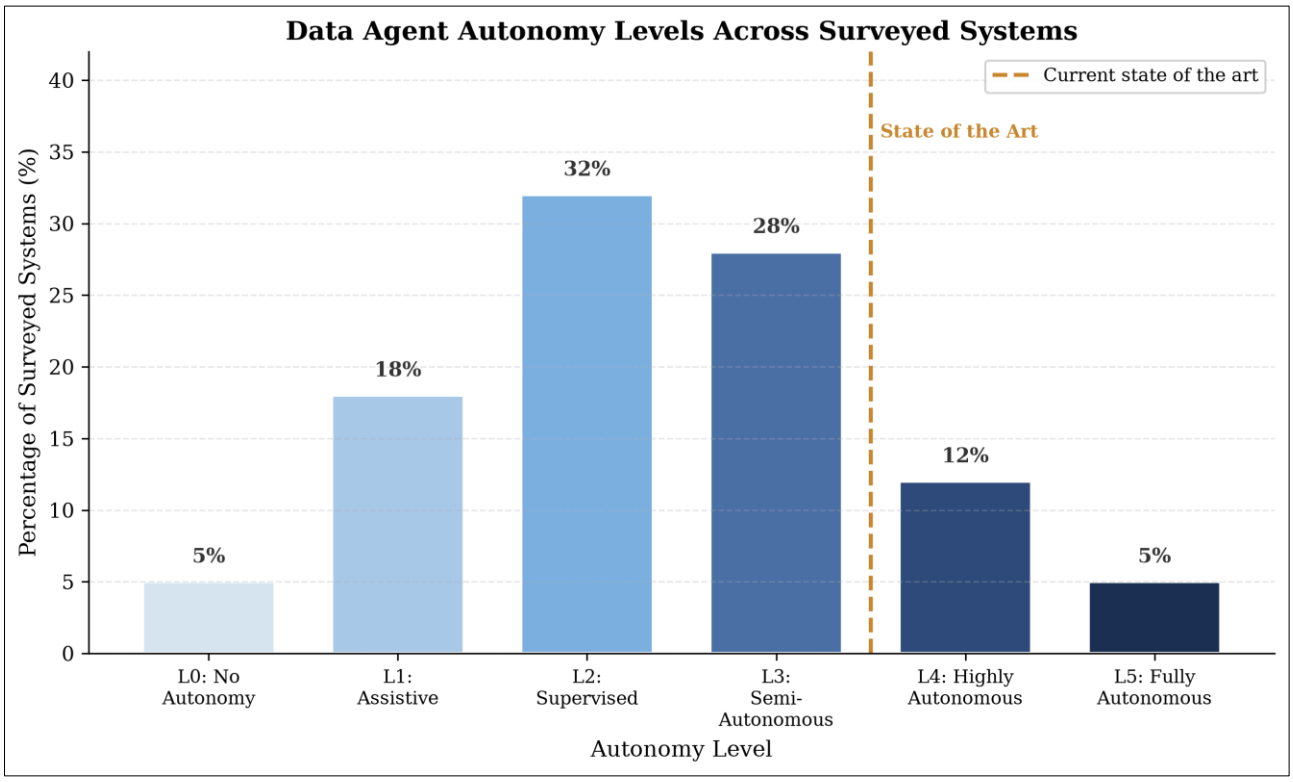

Figure 4: Distribution of surveyed data agents across autonomy levels L0-L5. The dashed marker denotes the boundary between semi-autonomous and highly autonomous systems

3.2. Autonomous Data Preparation

Data preparation is often the most time-consuming part of a data-science project and is reported to account for as much as 80% of project effort [23]. Agentic systems are consequently being used to propose cleaning, transformation, and feature-engineering operations, but the important design question is whether a proposed operation can be executed, inspected, and reused without silently changing the scientific question.

DeepPrep illustrates an execution-grounded approach: an LLM proposes a preparation pipeline, an environment materialises intermediate table states, and runtime feedback is used to revise the plan [24]. Its tree-based search supports non-local revision when an earlier decision becomes inconsistent with later evidence, and its 31 operators cover common transformation patterns. The contribution is therefore not simply code generation, but an explicit loop between proposed operations and observed execution.

DataEvolver extends this idea with self-evolving operator and pipeline levels. It expands the operator set, resolves dependencies, instantiates executable plans, and refines orchestration against high-quality examples; experiments on seven benchmarks reported an average downstream improvement of about 10% [25]. The result is encouraging, but it also illustrates why quality gains must be reported together with the data assumptions, leakage controls, and failure cases that determine whether a pipeline is safe to reuse.

TabClean takes a complementary route by asking an LLM agent to synthesise reusable programs for detecting and correcting tabular errors [26]. Applying the program once per schema can reduce token cost relative to per-record prompting and makes the transformation easier to audit. In all three systems, the scientifically relevant unit is the versioned preparation program and its intermediate artefacts, not the conversational exchange that produced it.

Taken together, these studies support a practical principle for agentic data engineering: generate executable, testable transformations at pipeline level, preserve intermediate states, and record the evidence used to accept or reject each operation. This pattern improves reuse and cost control, but it does not remove the need for domain rules, leakage tests, or human review of transformations that alter the meaning of a measurement.

3.3. Automated Data Quality Assessment

Automated data-quality assessment is a second layer of protection. LeDQeR generates candidate rules from observed dirty data and filters them for executability, correctness, generalisability, and redundancy [27]. AutoDCWorkflow generates OpenRefine sequences for benchmarked cleaning purposes, while related frameworks combine statistical inlier/outlier analysis with LLM-generated rules and executable code [28-30]. The shared lesson is that agents should propose quality checks, but acceptance should remain tied to explicit tests and versioned data states.

3.4. AutoML and Model Development Agents

AutoML traditionally automated model selection and hyperparameter search within a predefined space, with Auto-WEKA and TPOT providing influential early examples [31-35]. LLM-based agents broaden the interface by translating natural-language goals into plans, selecting tools, and coordinating specialised components. This expansion can lower the barrier to experimentation, but it also increases the number of hidden decisions that must be logged for a result to be reproducible.

AutoML-Agent coordinates retrieval, planning, task decomposition, and model construction across a multi-agent workflow [36]. AutoMind adds a curated knowledge base, tree search, and adaptive coding; EvoDS abstracts successful solutions into reusable skills; SPIO combines ensemble and selective strategies; and I-MCTS uses introspective Monte Carlo tree search to expand and critique candidate solutions [37-40]. These systems differ in search policy, yet they share a dependence on reliable intermediate evaluation: an agent cannot improve a pipeline if the metric, split, or data context is itself unstable.

A self-composable Big-Data-as-a-Service framework makes this dependency explicit by decomposing ingestion, cleaning, feature engineering, AutoML, deployment, monitoring, and drift detection into governed services [41]. Shared artefact stores, reproducibility checks, human-in-the-loop checkpoints, and drift-aware feedback turn a collection of agents into an operational system. The architecture is promising because it treats monitoring and rollback as part of the scientific workflow rather than as an afterthought.

3.5. AI Scientists and End-to-End Research Automation

At the upper end of the autonomy scale, AI-scientist systems attempt to link ideation, coding, experimentation, analysis, and manuscript preparation. The AI Scientist demonstrated a complete computational loop [42], while DeepScientist, InternAgent-1.5, OmniScientist, and SR-Scientist extend the idea through longer-horizon search, laboratory coordination, multimodal evidence, or executable symbolic reasoning [43-46]. Reported demonstrations show how quickly an agent can expand the number of candidate experiments, but they do not by themselves establish scientific validity. End-to-end systems must therefore be evaluated on independent reproduction, claim-level evidence, resource use, and the ability to stop when the evidence is insufficient.

3.6. Comparison of data engineering agent systems

Table 1: Comparison of data engineering and AutoML agent systems.

System	Year	Autonomy Level	Lifecycle Stages	Key Innovation	Evaluation
DeepPrep	2026	L3	Data prep, cleaning	Tree-based agentic reasoning	31 operators, 8 categories
DataEvolver	2026	L3	Data prep, cleaning	Multi-level self-evolving	7 benchmarks, +10% LLM perf
TabClean	2026	L2	Data cleaning	Reusable LLM-synthesized code	6 benchmarks, reduced cost
AutoML-Agent	2025	L3	Full pipeline	Retrieval-augmented planning	Deployment-ready models
AutoMind	2025	L3	Full pipeline	Knowledgeable tree search	2 DS benchmarks
EvoDS	2026	L4	Full pipeline	Skill acquisition + context mgmt	RL-optimized
SPIO	2026	L3	Model development	Ensemble + selective strategies	Multi-agent planning
The AI Scientist	2026	L4	Full research cycle	End-to-end automation	Passed peer review
DeepScientist	2026	L5	Full research cycle	Bayesian optimization discovery	20K GPU hours, SOTA on 3 tasks
OmniScientist	2026	L4	Full research cycle	Omni-modal, omni-discipline	36 real-data cases

4. Semantic Data Models and Knowledge Graphs

4.1. Ontologies and FAIR-by-Design Principles

Ontologies and controlled vocabularies provide the shared terms needed to integrate data from different instruments, laboratories, and disciplines [47]. The FAIR principles apply to the data as well as to the algorithms, tools, and workflows that create and analyse it [2], yet ontology-engineering practice still offers limited guidance for making semantic artefacts FAIR from the outset. This gap matters for agents because a graph can be syntactically valid while remaining ambiguous about units, roles, uncertainty, or the conditions under which a statement is true.

The Semantic Units Framework addresses this problem with semantically coherent, independently addressable units that record what is asserted, how it is intended, and which epistemic assumptions apply [48, 49]. Its instance-quantified categories extend the usual RDF/OWL distinction between named individuals and classes and support assertional, contingent, prototypical, and universal statements in one representational scheme. For agentic systems, the value lies less in adding another vocabulary than in preserving the scope and status of a claim when it is retrieved, transformed, or combined with other evidence.

FAIR materials provides ontology tools and FAIRification utilities for scientific data workflows, while a provenance-aware semantic-integration framework links each experimental observation to the precursors and activities that produced it [50, 51]. These examples show why semantics and provenance should be designed together. The ontology states what an entity or measurement means; the provenance record states where that meaning came from and which transformations may have changed it. Figure 5 maps these relationships to the FAIR sub-principles considered in the review.

4.2. Knowledge Graph Construction Pipelines

Knowledge graphs (KGs) operationalise semantic integration by representing entities, relations, attributes, and evidence in a form that supports structured queries and validation. SPARQL enables expressive retrieval, while SHACL can encode constraints that expose missing, inconsistent, or ill-typed information [52]. A KG is therefore not only a storage layer for an agent; it can also act as a testable boundary between extracted statements and the claims that an agent is permitted to use.

KGpipe illustrates a modular construction pipeline for RDF, JSON, and text sources, with intermediate formats, static validation, and replaceable processing components [53]. Structured RDF inputs were the most stable in its evaluation, whereas JSON and text required more careful mapping, extraction, and entity linking. This result is consistent with a broader engineering rule: the less explicit the source schema, the more important it becomes to retain raw inputs, candidate mappings, and rejected extractions alongside the published graph.

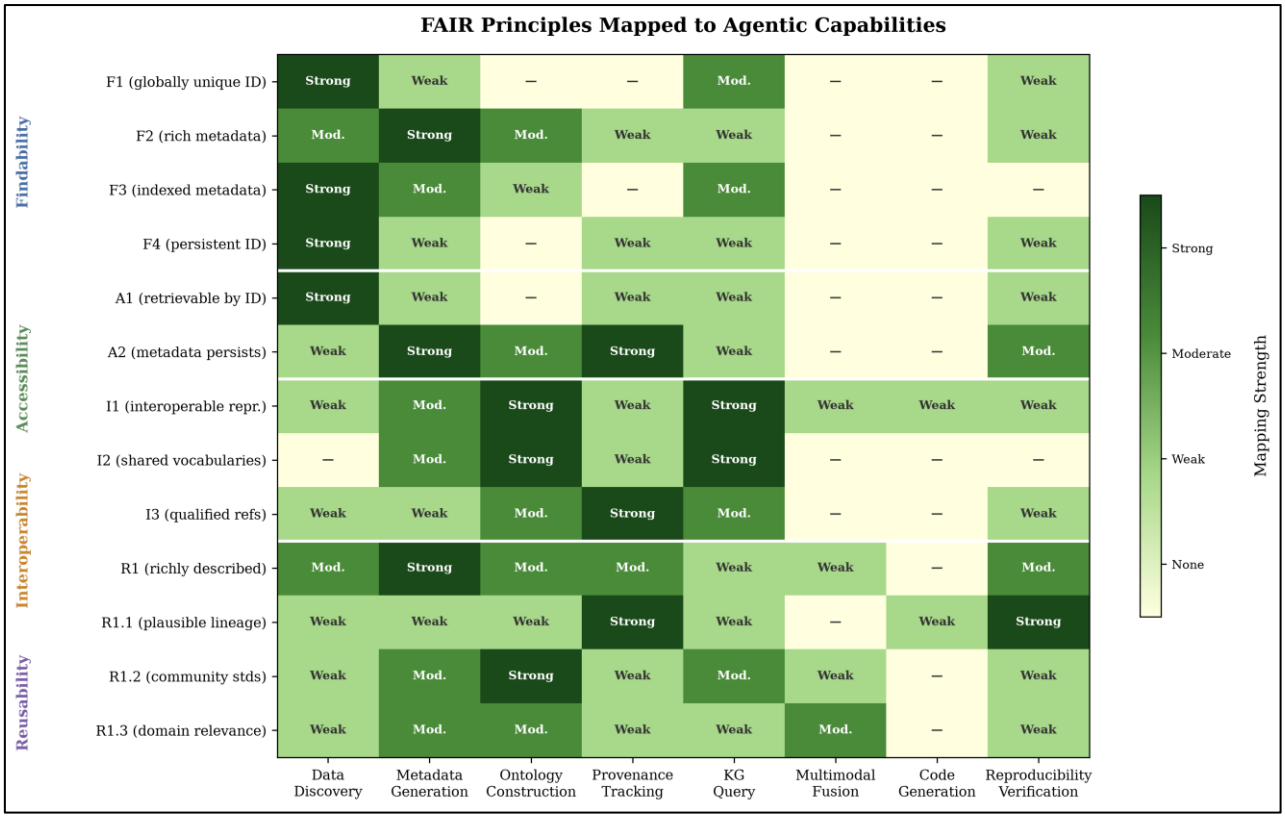

Figure 5: Alignment between agentic capabilities and FAIR sub-principles; darker cells indicate stronger evidence of alignment.

Wikontic combines LLM-based triplet extraction with entity normalisation, alias-aware deduplication, and ontology alignment [54]. Embedding retrieval narrows the candidate set before an LLM resolves synonymy, reducing the burden on the model while keeping the decision traceable. The workflow is representative of a useful hybrid pattern in which statistical retrieval proposes candidates, symbolic constraints eliminate impossible links, and a language model resolves residual ambiguity.

OntoKG introduces intrinsic-relational routing by separating node attributes from traversable graph relations [55]. Applied to a large Wikidata snapshot, it builds a core entity set and assigns properties to domain modules, improving category coverage. Such routing can make agent queries more efficient, but its scientific benefit depends on whether the module definitions, version of the source dump, and unresolved properties are recorded with the graph.

The Materials Science and Engineering Knowledge Graph uses Apache Airflow DAGs to automate acquisition, ontology population, reasoning, validation, and publication [56]. Because every stage is represented as a workflow component, the same pipeline can be inspected, rerun, or compared across versions. Figure 6 summarises this evidence-to-graph pattern and highlights validation as a publication gate rather than as a final cosmetic check.

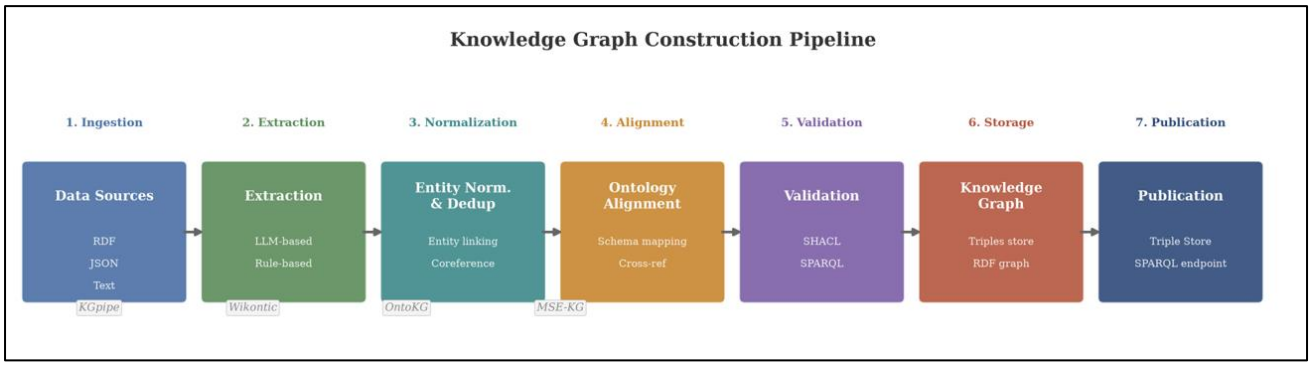

Figure 6: Evidence-to-knowledge-graph pipeline from heterogeneous sources to validated publication, with representative tools shown at each stage.

4.3. Semantic Layers for Agentic Reasoning

Semantic layers translate a graph into the definitions and relationships that downstream agents actually need. The Databricks Genie Ontology models trusted business definitions before adding inferred context, while Sema extracts a semantic ontology from a warehouse and serves query-relevant Semantic Context Objects to agents, catalogues, lineage tools, and natural-language interfaces [57, 58]. These abstractions are useful because an agent can receive the relevant slice of a graph without having to understand every storage detail.

The AWS Ontology Agent generates OWL ontologies from live schemas and persists them in a graph store, whereas Neocarta builds semantic-layer graphs from cloud catalogue information for access through agent-facing services [59, 60]. Their common challenge is governance: automatically inferred definitions must be versioned, reviewed, and linked to the source schema so that a change in an upstream table cannot silently change the meaning of a downstream answer.

4.4. FAIR Digital Objects and Persistent Identifiers

FAIR Digital Objects (FDOs) package data, software, metadata, and relationships behind persistent identifiers and explicit profiles [61]. The separation between abstract semantics and concrete implementation supports interoperability across repositories, provided that the profiles and registered attributes remain resolvable. Persistent identifiers for people, organisations, projects, and research objects can then be cross-linked at the metadata level, creating an attribution and discovery layer without requiring every provider to migrate into one database [62, 63].

4.5. Scientific Knowledge Graphs at Scale

Large scientific graphs demonstrate the potential scale of semantic infrastructure. OptimusKG integrates biomedical entities and relations from multiple ontologies and retains type-specific properties, cross-references, and provenance; its evaluation illustrates that evidence support and false-edge rejection can be measured rather than assumed [64]. Such measurements are especially important for agents, because graph density alone says little about whether the relations are current, justified, or fit for a particular question.

SciAtlas connects tens of millions of papers, entities, and relations across disciplines and uses neuro-symbolic retrieval to support cross-domain exploration [65]. Its value is not simply the number of nodes, but the possibility of traversing concepts that are separated by disciplinary vocabularies. The same breadth creates a governance burden: versioned ingestion, source-level provenance, and uncertainty labels are needed if an agent is to distinguish a well-supported bridge from an attractive but weak association.

SciMKG provides a multimodal educational graph in which text, images, video, and audio are aligned with scientific concepts; its reported cross-modal alignment improves concept-extraction F1 over the comparison systems [66]. This result reinforces the importance of representing modality-specific evidence explicitly. A graph that stores only the final concept loses the image region, transcript span, or audio segment that a reviewer would need to inspect.

4.6. Comparison of knowledge graph systems

Table 2: Comparison of knowledge graph construction and reasoning systems

System	Year	Scale	Modalities	Construction Method	Validation
KGpipe	2026	Modular	RDF, JSON, Text	Pipeline-based, LLM components	KGI-Bench
Wikontic	2026	Wikidata-aligned	Text	LLM triplet extraction + dedup	Cosine similarity + LLM
OntoKG	2026	34M entities	Wikidata	Intrinsic-relational routing	93.3% coverage
MSE-KG	2026	750K triples	Materials data	Airflow DAGs, ROBOT, SHACL	SPARQL validation
OptimusKG	2026	190K nodes, 22M edges	Multimodal biomedical	LPG from 18 ontologies	PaperQA3 (70% verified)
SciAtlas	2026	157M entities, 3B triplets	Academic papers	Neuro-symbolic retrieval	Tri-path recall
SciMKG	2026	34K concepts, 403K triples	Text, Image, Video, Audio	Cross-modal alignment	+9% F1 over baselines
EvoGraph-R1	2026	Dynamic hypergraph	Multimodal	MDP-based agentic retrieval	Self-evolving

5. Multimodal Analytics

5.1. The Multimodal Challenge in Scientific Data

Scientific evidence is intrinsically multimodal: a single conclusion may depend on prose, tables, images, time series, spectra, three-dimensional structures, and symbolic relations [45]. The omni-modal evaluation suite reviewed here spans five discipline groups and 36 cases, with most cases involving perceptual data [45]. The challenge is not to concatenate modalities, but to preserve their units, resolution, uncertainty, and experimental context while an agent moves between them.

5.2. Scientific Multimodal Foundation Models

Monkey King Bang (MKB) aligns scientific tokens with a language backbone through a modality-then-language curriculum [67]. A modality-specific stage learns representations while the core model is frozen, followed by consolidation with mixed scientific and general corpora. The design illustrates a broader strategy for scientific foundation models: specialised encoders can preserve measurement structure, while a shared backbone supports cross-domain reasoning.

FuXi-Uni uses domain-specific decoders within a unified multimodal model, Intern-S2-Preview combines long-sequence modelling with a memory path for scientific specialisation, and S1-Omni-Image couples image understanding, generation, and editing with a reasoning-before-generation procedure [68-70]. These systems make multimodal assistance more accessible, but their outputs still require modality-level validation. A fluent explanation cannot substitute for checking whether the model attended to the relevant image region, time interval, or numerical column.

5.3. Visual-Tabular and Multimodal Learning

VT-Bench evaluates visual-tabular prediction and reasoning over 14 datasets from nine domains and shows that strong performance on a single modality does not guarantee reliable fusion [71]. MMPFN addresses two recurrent failure modes, over-compressed tabular embeddings and attention imbalance caused by unequal token counts, with gated and cross-attention components [72]. For scientific use, benchmark design should also test unit consistency, missingness, measurement error, and whether the model can cite the exact visual and tabular evidence supporting a conclusion.

5.4. Retrieval-Augmented Generation for Science

Retrieval-augmented generation (RAG) reduces unsupported generation by coupling a language model to an external evidence store, but retrieval quality becomes part of the scientific method [73]. The literature increasingly treats RAG as a pipeline with indexing, retrieval, fusion, and generation stages. Vector, graph, multimodal, hybrid, and agentic variants differ in where they represent context and how they follow multi-hop evidence; all require records of corpus version, query, ranking, and cited passages if a result is to be reproduced.

SciRAG combines adaptive retrieval, citation-aware symbolic reasoning, and outline-guided synthesis [74]. LeanRAG aggregates knowledge hierarchically, TagRAG uses structured tags to guide graph construction and retrieval, and ParallaxRAG separates query and graph views to encourage diverse relational paths [75-77]. Their reported gains are best interpreted as evidence that retrieval structure matters, not as a guarantee of factuality. Each system can still inherit omissions or errors from its corpus, which is why evidence coverage and abstention should accompany answer-level metrics.

EvoGraph-R1 models the knowledge graph as a dynamic environment in which an agent can query, expand, refine, or terminate a retrieval episode [78]. MegaRAG extends graph-based retrieval to visually rich documents by aligning textual, visual, and spatial cues [79]. Together, these systems point toward a closed loop between representation and reasoning, but they also make graph evolution a provenance problem: every added, removed, or reweighted relation needs a traceable reason. Figure 7 places these retrieval mechanisms within a multimodal evidence pipeline.

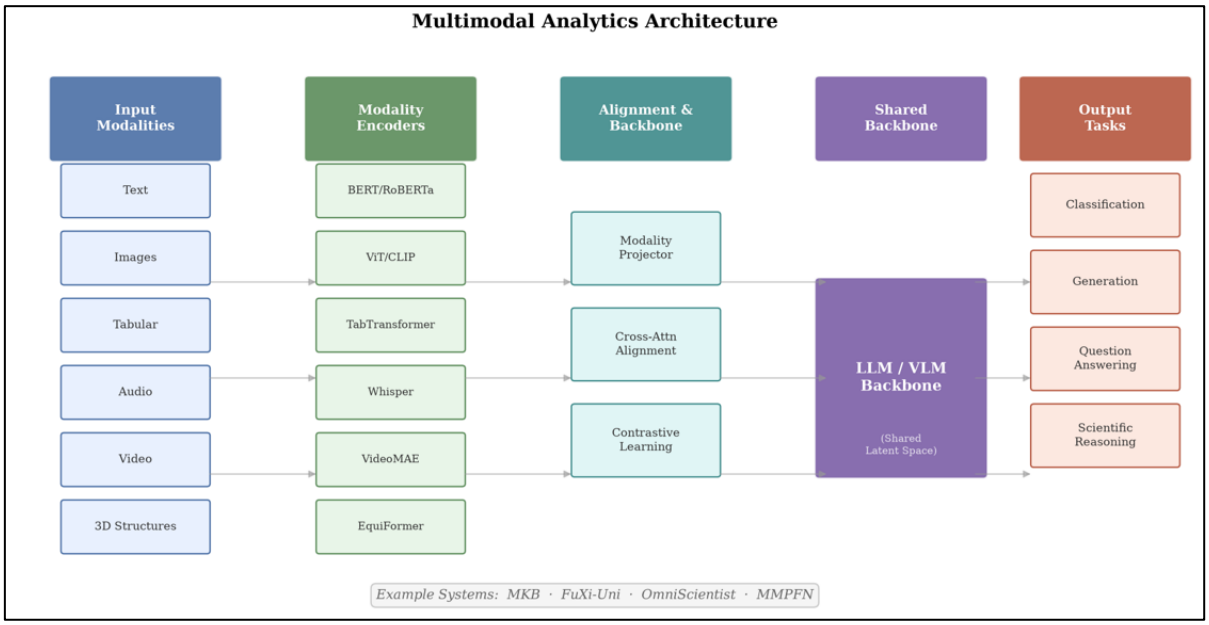

Figure 7: Multimodal analytics architecture linking heterogeneous scientific inputs, modality encoders, alignment, shared language/vision backbones, retrieval, and claim validation.

5.5. Comparison of retrieval-augmented generation systems

Table 3: Comparison of retrieval-augmented generation systems

System	Year	KG-Based	Multimodal	Key Innovation	Hallucination Reduction
SciRAG	2026	Yes (citation graph)	No	Citation-aware reasoning	Outline-guided synthesis
LeanRAG	2026	Yes (hierarchical)	No	Semantic aggregation	Cross-community reasoning
TagRAG	2026	Yes (tag hierarchy)	No	Tag-guided retrieval	78.36% win rate
ParallaxRAG	2026	Yes (multi-view)	No	Multi-hop, multi-view	Substantial reduction
EvoGraph-R1	2026	Yes (dynamic hypergraph)	Yes	MDP-based agentic retrieval	Self-evolving correction
MegaRAG	2026	Yes	Yes	Visual document reasoning	Structured concepts
GraLC-RAG	2026	Yes (UMLS)	No	Structure-aware chunking	KG-infused retrieval
EXYGEN	2026	Yes (SPARQL)	No	Conversational KG access	Schema-based grounding

6. Provenance and Reproducibility

6.1. The Reproducibility Crisis in AI-Generated Science

Computational reproducibility is a persistent concern, and agentic systems add new failure paths because they can generate code, data transformations, and interpretations at scale. An open replication effort for ICML 2026 oral papers reported that only a minority of fully replicated papers achieved high claim-level agreement, while a separate initiative examining 2,226 papers found both independently verified claims and contested or falsified claims [11, 12]. The important conclusion is not a single percentage; it is that reproducibility must be assessed at claim, data, code, and environment levels.

The cost and scale of contemporary machine-learning publication make independent checking difficult. Reports on the ICML corpus describe a sharp increase in submissions and estimate that reproducing all oral-paper experiments would require substantial computing resources [80]. Agentic systems may reduce the cost of running a particular experiment, but they also increase the number of intermediate states that a reproducer must recover. Reproduction budgets should therefore be treated as part of study design, not as an optional post-publication service.

6.2. Hallucinated Citations and the Integrity of AI-Generated References

Citation integrity is another reproducibility layer. Hallucinated references are dangerous because they can look plausible while providing no recoverable evidence. RefChecker found low rates at the individual-reference level but visible paper-level failures in large conference bibliographies, including papers with multiple likely hallucinations [13]. This pattern argues for claim-to-source verification rather than a superficial check that every bibliography entry has a DOI or a familiar venue name.

Recent tools operationalise that requirement. CiteCheck retrieves candidate publications and uses a structured verifier to distinguish exact, minor, and major mismatches; HalluCiteChecker supports lightweight offline checking; HalluPeer targets hallucinations in peer reviews; and PROBE decomposes detection into claim decomposition, evidence finding, evidence evaluation, and localisation [81-85]. These systems converge on a process principle: verification improves when the claim is made explicit before the evidence is scored.

6.3. LLM-Generated Code Reproducibility

Generated code is reproducible only when its execution context is reproducible. In a study of 300 projects produced by three coding agents, 68.3% executed out of the box, while the gap between declared and runtime dependencies expanded substantially [14]. The result exposes a common blind spot: a short script may depend on a long, undocumented chain of packages, operating-system behaviours, credentials, and data-access assumptions.

Notebook-based analyses show the same problem. An audit of repositories associated with Nature publications found that only two of 19 attempted notebooks were reproducible, largely because data and dependencies were missing [86]. Reusable workflow systems such as ReproAgent, DeepRepro, and Snakemaker respond by converting prose and ad hoc code into explicit implementation contracts, state-aware repair plans, or executable workflow descriptions [87-89]. Their success should be judged by independent reruns, not by whether a generated notebook appears complete.

The practical implication is to treat every agent-generated program as a research artefact: pin its environment, record input and output hashes, preserve execution logs, test failure paths, and publish the workflow with the data and configuration needed to rerun it. These controls also create a natural interface for human review because a reviewer can inspect a bounded artefact instead of reconstructing an entire conversation.

6.4. Provenance Standards and Frameworks

W3C PROV-O expresses the PROV data model in OWL2 and supplies a lightweight, extensible

vocabulary for entities, activities, agents, and their relations [90-92]. ProvONE extends the model for scientific workflows and captures the structure of workflow components and their execution relationships [93]. These standards do not make a result true, but they provide a shared language for stating what was used, what acted, and what was generated.

RO-Crate packages research artefacts with contextual metadata and links between files, people, software, and workflows [94-96]. Workflow Run RO-Crate extends this packaging model to execution provenance, bundling inputs, outputs, code, and run-level information in a portable research object [97, 98]. The standards become valuable for agentic work when a run can be opened by an independent tool without relying on the original agent's memory or service provider.

6.5. Provenance in Agentic Workflows

yProv provides multi-level provenance services and libraries for inspecting, navigating, and analysing provenance in AI-enabled workflows [99]. ReProv extends REANA with workflow-as-a-service capabilities and automated PROV-O capture, while Crystalia and OmicsLake address lineage for large or versioned datasets [100-102]. These systems reveal that lineage must be available at several resolutions: dataset-level history for reuse, record- or operation-level history for debugging, and claim-level history for scientific review.

6.6. Provenance as the Gateway to Trust

Provenance is the practical gateway to trust because it gives a reviewer something that can be reopened, checked, and corrected. A perspective on autonomous science argues that a trustworthy record should expose what was reasoned, executed, and measured rather than relying on a model's internal explanation [18]. Interpretability remains useful for diagnosis, but it cannot replace a recoverable chain from evidence to claim.

F(AI)²R turns this principle into an executable workflow: it records activities, claims, and sources, enforces a no-parentless-claim invariant, and treats a novelty statement as a bounded search claim with an explicit horizon [16, 103]. ZTEK adds an external verification boundary that prevents unverified claims from entering downstream workflows [104], while Traceable Trust asks what evidence supports an output, what agency was delegated, which threshold authorises action, and who can override it [105]. Together, these approaches distinguish documentation produced by the agent from assurance supplied by an independent check.

6.7. Artifact-Centered Observability

Artifact-centred, claim-aware observability makes the evidence relation itself a first-class object [106]. The proposed profile links individuals, operators, fitness records, lineage, archives, runs, streams, and steering commands. This is important because failures are distributed: a manuscript can cite the wrong source, a search can select a degenerate candidate, a laboratory claim can depend on an unstated rule, or a multi-agent plan can change without a visible trigger. A claim-aware record lets a reviewer locate the earliest decision that needs correction.

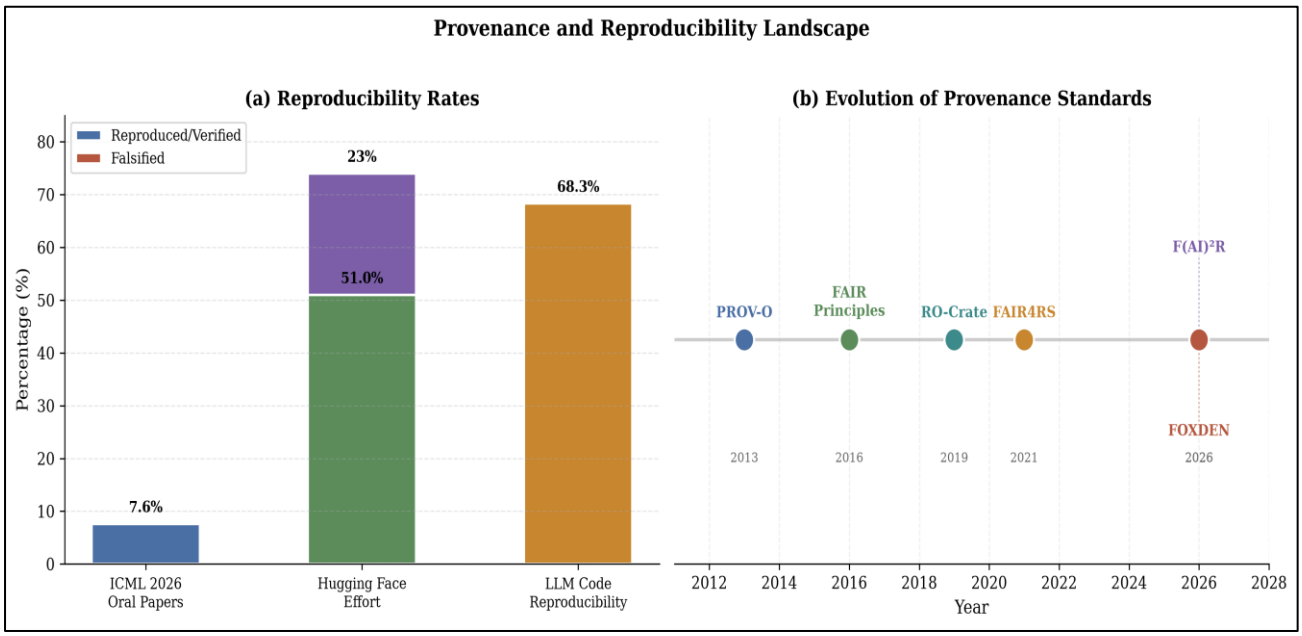

Figure 8: Provenance and reproducibility landscape: (a) reported reproducibility outcomes; (b) development of provenance standards from PROV-O to contemporary research-object profiles.

6.8. Comparison of provenance standards and tools

Table 4: Comparison of provenance standards and tools.

Standard/Tool	Year	Scope	Standard	Key Feature	Adoption
PROV-O	2013	General provenance	W3C Recommendation	OWL2 ontology, domain-extensible	Widely adopted
RO-Crate 1.3	2026	Research objects	JSON-LD, Schema.org	Packaging + metadata + workflows	Multi-community
Workflow Run RO-Crate	2023	Workflow provenance	RO-Crate profile	Execution provenance, interoperable	6+ workflow systems
yProv	2024	AI workflows	PROV-O extension	Multi-level provenance	yProv4ML, yProv4WFs
ReProv	2026	Workflow execution	PROV-O, CWL	WaaS + auto-capture	AIoD Platform
F(AI)²R	2026	AI-in-the-loop	PROV-O extension (aiprov)	Executable skill, no parentless claims	Self-case-study
ZTEK	2026	AI research workflows	Zero Trust protocol	5-pillar verification	Research program
Crystalia	2024	Large datasets	Custom	Parallelizable lineage	Life sciences, HEP
OmicsLake	2026	R/Bioconductor	Custom	Version-aware, agent-aware	Bioinformatics

7. Human Oversight and Trust

7.1. The Paradox of Human Oversight in AI-Driven Science

Human oversight becomes more difficult, not less important, as agents take on longer and more consequential tasks. The question is therefore not whether a person appears somewhere in the workflow, but whether that person can understand the current state, change the plan, and reject an unsupported conclusion. Discussions of the 'White Elephant in the Lab' emphasise the risk that convenient automation can erode scientific judgement when users cannot see where a recommendation came from [107].

7.2. The Novelty Ceiling and Minimum Oversight Rate

A PAC-theoretic analysis formalises one limit on autonomous discovery through a novelty ceiling: once a hypothesis lies beyond the structural support of the training corpus, a learned evaluator may provide no reliable signal [108]. The analysis derives a minimum oversight rate r* for routing hypotheses to a human expert and predicts that structurally diverse seeds can raise the ceiling. Whether the precise bound transfers across scientific domains remains an open empirical question, but the design implication is robust: novelty and uncertainty should influence escalation, not just model confidence.

7.3. Designing Human Oversight in Autonomous Discovery

A structured human-AI workflow embeds oversight at three points. Oversight by design resolves the goal, scope, constraints, and success criteria before execution; oversight by review checks plans and intermediate artefacts while work is under way; and oversight by escalation verifies outputs and triggers a refuse or stop signal when evidence is inadequate [109]. Figure 9 represents these patterns as control points rather than as a single approval step at the end.

MOOSE-Copilot makes the same principle operational by allowing scientists to provide an initial blueprint, decide when an exploration stage should give way to exploitation, and steer within a stage through structured feedback [110]. Such interaction reduces search-space explosion while preserving human authority over the transitions that change the scientific question.

A review of human-AI collaboration for scientific discovery identifies complementary roles across observation, hypothesis, and experiment [111]. The recurring pattern is not a fixed division of labour: humans contribute domain judgement, problem framing, and responsibility for consequences, whereas agents contribute search, synthesis, coding, and consistency checks. The interface between these roles should expose uncertainty and provenance at the moment a decision is made.

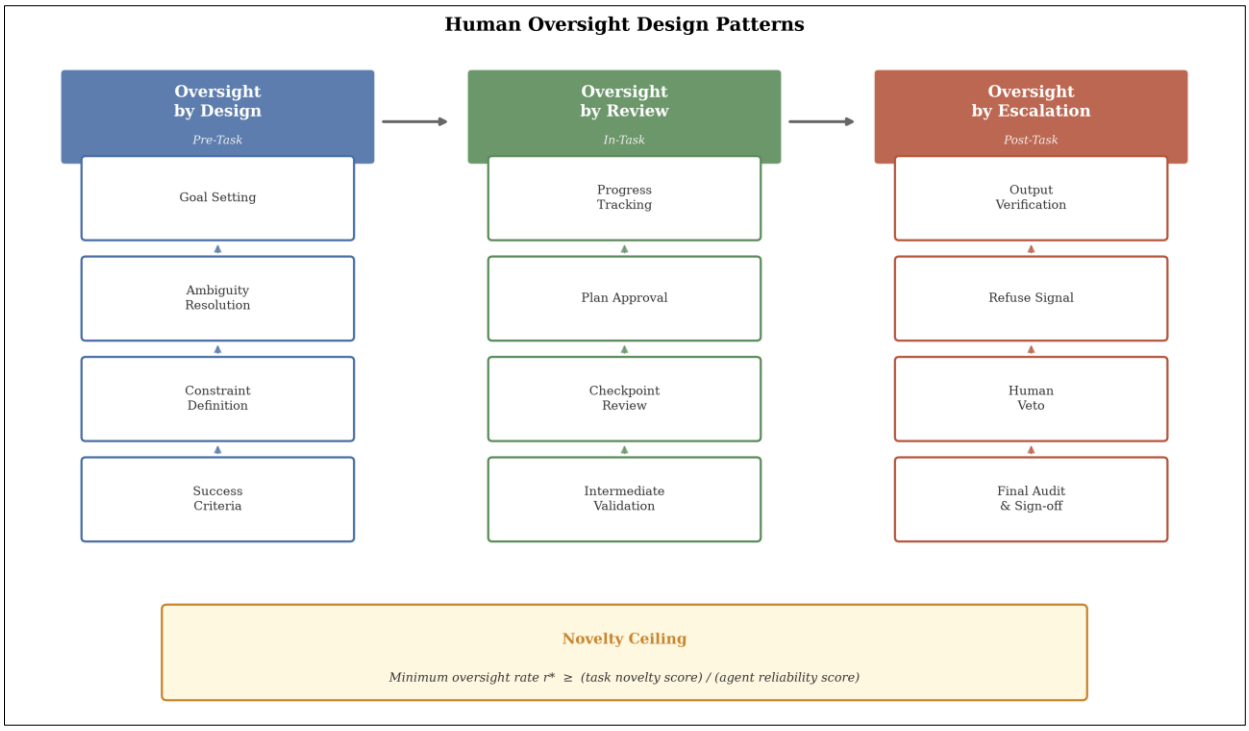

Figure 9: Human-oversight design patterns spanning pre-task design, in-task review, and post-task escalation, together with the novelty ceiling and minimum oversight rate r*

7.4. When Should an AI Scientist Stop?

Verifiable experiment steering and refusal places an auditable stop-and-escalate signal inside the loop that selects the next experiment [112]. Its revocation tests are instructive: an early confident identification is withdrawn when additional evidence reveals structural mismatch, while an in-library control remains identified. This ability to retract a claim is a stronger indicator of governed autonomy than an uninterrupted sequence of successful-looking actions.

7.5. Trust Calibration and Automation Bias

Trust calibration aligns reliance with demonstrated capability. Poor calibration produces automation misuse when people accept an output beyond its competence and automation disuse when they ignore a useful system [113]. A review of automation bias treats trust as a response to uncertainty and vulnerability, which means that interface design, evidence visibility, and escalation rules are as important as model accuracy.

Experiments on conflicting information show that the same cue can calibrate some users and destabilise others: moderate conflict can reduce over-reliance among younger adults, whereas intense conflict can impair older adults, and high automation can amplify over-trust [114]. These results caution against a single universal reliance policy. Agentic systems should support user- and context-sensitive checks without using demographic categories as a substitute for measuring actual performance.

A calibrated adaptive framework reports low failure rates under stress when uncertainty is quantified, probabilities are calibrated, and reliance adapts to observed performance [115]. Its ablations indicate that calibration, rather than raw model capability alone, is essential. In scientific workflows, the analogous control is to expose confidence with its reference distribution, error history, and evidence coverage rather than displaying a bare score.

A complex-adaptive account of over-reliance further shows that social proof can turn individual shortcuts into a collective verification failure [116]. Visible, unverified use suppresses checking, whereas making verification visible can stabilise appropriate reliance. This finding extends trust calibration beyond the human-agent pair to the laboratory, team, or institution in which the agent is embedded.

7.6. Governance Frameworks and Responsible AI Documentation

Model cards and data cards provide structured disclosures about a model or dataset, including intended use, construction, performance, limitations, and conditions of use [117-119]. MLOps systems can populate these records from training runs and evaluation reports, but automated completion should not be mistaken for independent assurance [120, 121]. Documentation is useful only when a reviewer can trace each field to a source and identify what has not been measured.

For high-risk AI systems, the EU AI Act requires technical documentation covering the system, intended purpose, data, validation, and testing, together with long-term retention obligations [121]. Scientific agents may not fall into one regulatory category, but the underlying practice is transferable: retain the configuration, data basis, evaluation evidence, and change history long enough for an independent party to reconstruct a consequential decision.

7.7. The RDA Data Director Blueprint

The RDA Data Director blueprint is a technology-agnostic specification for helping researchers prepare, document, and deposit data in alignment with FAIR principles [15, 122]. It responds to limited researcher capacity, inconsistent metadata, uneven data-support access, policy complexity, uncertain data quality, and governance of sensitive material. A key requirement is that agentic outputs identify their sources, model, processing date, persistent identifier, and human-curation status. This turns a generic assurance statement into a set of fields that can be checked.

7.8. Human oversight design patterns

Table 5: Human oversight design patterns in agentic data science.

Pattern	Stage	Mechanism	Systems	Key Property
Oversight by Design	Pre-task	Goal setting, ambiguity resolution	ASD workflow【 109】	Human decision authority
Oversight by Review	In-task	Progress tracking, plan approval	MOOSE-Copilot【 110】	Inter-stage routing
Oversight by Escalation	Post-task	Verification, refuse signal	Verifiable Steering【 112】	Self-correction, revocation
Trust Calibration	Continuous	Uncertainty quantification	Calibrated framework【 115】	Adaptive reliance
Novelty Ceiling	Continuous	Minimum oversight rate r*	PAC theory【 108】	Diversity seeds reduce r*
Provenance Audit	Post-hoc	Complete, re-openable records	F(AI)²R【 16】, ZTEK【 104】	No parentless claims
Claim Verification	Post-hoc	Citation checking, fact verification	CiteCheck【 81】, RefChecker【 13】	Falsifiable novelty claims

8. DISCUSSION: CROSS-CUTTING CHALLENGES AND FUTURE DIRECTIONS

8.1. The Provenance Paradox: Autonomy vs. Auditability

The review reveals a provenance paradox: as an agent becomes more autonomous, its chain of planning, tool use, experimentation, and revision becomes harder to capture, even though the need for auditability increases. Long-horizon systems such as DeepScientist illustrate the scale of the problem, because thousands of decisions can be made before a human sees the final claim [43]. A linear activity log is not enough when plans branch, evidence is discarded, and later actions depend on earlier model judgements.

Two complementary responses are emerging. Self-generated provenance records the agent's actions at the granularity needed for debugging, whereas an external verification layer checks whether claims and evidence satisfy an independent contract [16, 104]. Neither is sufficient alone: a self-report can omit an unobserved dependency, while an external checker may lack the context needed to understand why a decision was made. A robust architecture therefore joins detailed run provenance to claim-level verification.

Claim-aware observability adds a third layer by making the links among a claim, its supporting evidence, the search that selected that evidence, and the plan that governed the search explicit [106]. This representation gives reviewers a practical route from a disputed sentence to the artefact and decision that produced it.

8.2. The FAIR-Autonomy Tension

FAIRness and autonomy are often treated as aligned because both favour machine-actionable resources, but they impose different demands. FAIR metadata must remain understandable to the community that curates and reuses a resource, whereas an autonomous agent may produce a technically valid record that is opaque to a human reviewer. The Data Director blueprint records whether human curation occurred [122], but quality still depends on explicit definitions, validation rules, and evidence for each metadata field.

The FAIR+COPE proposal extends the conversation toward comparable, organised, predictive, and engaged data [123]. Its broader lesson is that a property such as privacy or FAIRness should be stated as an evidence-bearing claim, not as an implicit label [124]. Agents should therefore report which tests were passed, which were not run, and which judgements remain provisional.

8.3. The Reproducibility Deficit

Reproducibility deficits accumulate across the data-to-claim chain. At claim level, open reproduction studies report incomplete verification and contested findings [12]; at code level, only 68.3% of generated projects in one evaluation ran out of the box [14]; at notebook level, an audit found two reproducible repositories among 19 attempted executions [86]; and at citation level, conference bibliographies contained a visible tail of likely hallucinated references [13]. These are not interchangeable metrics, but together they show why a failure in dependencies, evidence, or data lineage can invalidate a conclusion even when the prose appears coherent.

Agentic AI can amplify the deficit by producing more analyses with less verification, but it can also supply countermeasures. MLReplicate benchmarks autonomous research systems on standardised reproduction tasks, while AgentActionBench records the actions taken during paper-to-experiment reproduction [125, 126]. The next generation of benchmarks should score not only whether a result matches, but also whether the system identifies an irreproducible step, abstains when evidence is missing, and leaves a portable record for a human reproducer.

8.4. The Semantic Grounding Problem

Agentic reasoning requires semantic grounding: data, metadata, and claims must be connected to shared definitions and relations. The Semantic Units Framework addresses representational limitations in conventional OWL/RDF models, while OptimusKG and SciAtlas demonstrate the scale at which structured evidence can be assembled [48, 64, 65]. Yet graph construction and agent reasoning remain separate activities in much of the literature. A graph can be large and well populated without making it clear which path an agent should trust for a particular claim.

Self-evolving approaches such as EvoGraph-R1 close part of this gap by treating retrieval as a Markov decision process: the agent observes graph state and chooses whether to query, expand, refine, or stop [78]. This creates a useful feedback loop, but it also makes graph evolution itself a scientific object. Versioned relation histories, source-level evidence, and reversible updates are required if a mutable graph is to support reproducible reasoning.

8.5. The Multimodal Integration Challenge

Scientific evidence spans text, images, tables, signals, audio, video, three-dimensional structures, spectra, and symbolic graphs [45]. MKB, FuXi-Uni, and related models demonstrate that unified scientific reasoning is feasible [67-70], but three challenges remain central. First, alignment must preserve scientific semantics rather than only statistical correlation. Second, benchmarks must test units, missingness, uncertainty, and evidence localisation rather than rewarding fluent summaries. Third, provenance must follow a claim across every modality and transformation, as early multimodal-graph systems such as MegaRAG begin to do [79].

8.6. The Trust Calibration Dilemma

Trust calibration presents a structural dilemma: the systems that most need oversight are often the systems whose long-horizon behaviour is hardest to inspect. The novelty-ceiling analysis formalises one part of this problem through a minimum oversight rate r* [108]. In practice, escalation should also depend on evidence coverage, distribution shift, irreversible actions, and the cost of being wrong.

The finding that human bias can dominate model capability shifts attention from accuracy alone to interaction design [115]. Social proof can suppress verification and produce collective over-reliance, whereas interfaces that expose checks and dissent can support more appropriate use [116]. Trustworthy agents should make uncertainty, provenance, and refusal behaviour visible enough that a team can challenge an output before it becomes institutional memory.

8.7. Governance and Ethics

Governance for agentic science is evolving alongside the technology. Documentation obligations under the EU AI Act, model and data cards, and research-integrity guidance all point toward the same operational requirement: preserve intended use, data provenance, validation evidence, limitations, and change history [118-121]. Synthetic-data research makes the point especially clearly: privacy is not guaranteed by realism or by a generator label, but must be evaluated against a threat model and a stated use case [124, 127-129].

Contextualised scientific decision workflows extend governance from artefacts to choices. By representing dataset selection, algorithm choice, parameterisation, and interpretation as linked, verifiable assets, they preserve attribution and make collective decisions inspectable [130]. This is important for agentic systems because a final model may be reproducible while the earlier decision to use a dataset, metric, or threshold remains undocumented.

8.8. Infrastructure and Ecosystem

The European Open Science Cloud (EOSC) illustrates the infrastructure needed to connect repositories, software, and services across organisational boundaries [131-134]. At the dataset level, SeDa harmonises metadata across millions of records, while the Leibniz Data Manager supports federated discovery and links research data to external knowledge bases [135, 136]. Such platforms can become an operating environment for agentic science only if access policies, identifiers, provenance, and service versions are exposed to the agent and to the human reviewer.

The resulting architecture is not a single model but a governed stack: data sources feed versioned engineering pipelines; semantic layers expose context; multimodal analytics generate candidate claims; provenance services record the run; and human oversight controls escalation and release. Figure 10 summarises this stack and shows why FAIR compliance and human oversight must cut across every layer rather than appear only at the end.

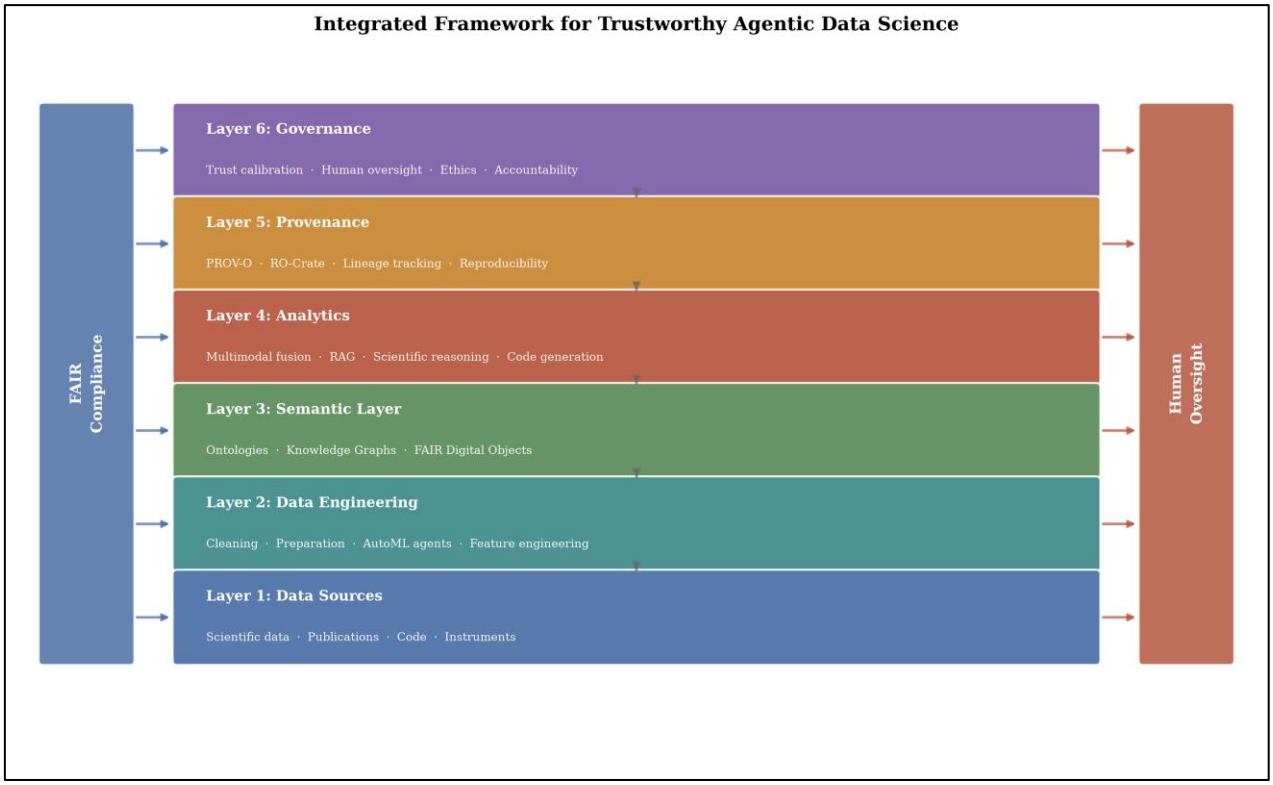

Figure 10: Integrated framework for trustworthy agentic data science. Data sources, engineering, semantics, analytics, provenance, and governance are coupled by FAIR compliance and human oversight.

8.9. Cross-cutting challenges and research directions

Table 6: Summary of challenges, current state, and future directions.

Challenge	Current State	Gap	Future Direction
Provenance paradox	Self-gen + external verification	Autonomy reduces auditability	Dual-layer provenance, claim-aware lineage
Reproducibility deficit	51% claims verified, 68% code runs	Compounding deficits at multiple levels	Automated reproduction, mutation testing
FAIR-autonomy tension	RDA Data Director Blueprint	Machine metadata not human-reviewable	Evidence-based FAIR claims, FAIR+COPE
Semantic grounding	OptimusKG, SciAtlas, EvoGraph-R1	KG construction ≠ agent reasoning	Self-evolving KGs, MDP-based retrieval
Multimodal integration	MKB, FuXi-Uni, OmniScientist	Alignment, evaluation, multi-modal provenance	Cross-modal semantic alignment, VT-Bench
Trust calibration	Novelty ceiling, r* formula	Human bias dominant failure mode	Verification visibility, adaptive reliance
Hallucinated citations	RefChecker, CiteCheck, HalluCiteChecker	~5% papers have hallucinated refs	Pre-publication auditing, $0.04/paper
Code reproducibility	ReproAgent, DeepRepro	68.3% LLM code runs	Contract-guided generation, dependency tracking
Governance	EU AI Act, model cards, ZTEK	Rapidly evolving, incomplete	Zero-trust protocols, automated compliance
Infrastructure	EOSC, SeDa, LDM	Federation, interoperability	Federated KGs, cross-platform provenance

9. CONCLUSION

This systematic review examined 218 studies on agentic AI for FAIR and reproducible data science across five linked domains: data engineering, semantic infrastructure, multimodal analytics, provenance and reproducibility, and human oversight. The synthesis shows a field moving from isolated assistants toward tool-using systems that can plan, execute, inspect, and revise long workflows. Progress is clearest in execution-grounded data preparation, automated pipeline construction, semantic retrieval, and multimodal evidence handling.

The same synthesis sets a boundary on what can currently be claimed. End-to-end demonstrations can generate ideas, code, analyses, and manuscripts, but scale does not guarantee valid evidence. Knowledge graphs provide a useful substrate only when their relations are versioned and supported; RAG improves grounding only when retrieval and corpus changes are recorded; and multimodal models remain vulnerable to alignment, unit, and localisation errors. The strongest systems therefore combine model capability with explicit tests, structured artefacts, and reversible decisions.

Reproducibility remains the decisive constraint. Independent claim verification is incomplete, generated code often depends on hidden runtime conditions, notebooks fail when data and environments are missing, and plausible citations can be unsupported [12-14, 86]. These weaknesses compound across the data-to-decision chain. The provenance paradox follows directly: autonomy increases the number of actions that need to be audited, while the same autonomy can reduce the visibility of those actions.

We therefore recommend a provenance-first design philosophy. An agentic system should emit a complete, reopenable record of inputs, transformations, tool calls, model versions, evidence, uncertainty, and human interventions. The record should be machine-actionable through standards such as PROV-O and RO-Crate, semantically grounded in shared vocabularies, and checked at the claim boundary by an independent verifier [90, 94, 104, 106]. Human oversight should be staged across goal setting, intermediate review, and escalation, with novelty, uncertainty, distribution shift, and consequence determining when a person must intervene.

FAIR principles remain a durable foundation for this agenda because they connect discovery and reuse to the conditions under which evidence can be interpreted. Agentic AI can make FAIRification and reproducibility more scalable, but it can also multiply unsupported claims if verification is treated as optional. The field will earn scientific trust when agents are evaluated not only on what they can produce, but on what they can document, reproduce, retract, and hand back to a human decision-maker.

Acknowledgments

This review was conducted using open-access databases and community standards. We acknowledge the Research Data Alliance, GO FAIR, the European Open Science Cloud, and the W3C for foundational work on FAIR principles, provenance standards, and research-data infrastructure.

REFERENCES

1. Wilkinson, M.D. *et al.,*(2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. DOI: 10.1038/sdata.2016.18.
2. Wilkinson, M.D. *et al.,*(2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. DOI: 10.1038/sdata.2016.18.
3. Mons, B. (2026). How FAIR data are helping to build trust in science. Nature, d41586-026-01982-y. DOI: 10.1038/d41586-026-01982-y.
4. INGV Data Registry (2026). The INGV data registry as a curated metadata infrastructure for Earth Science data stewardship. Scientific Data, 13, 6980. DOI: 10.1038/s41597-026-06980-3.
5. SetGo (2026). SetGo: Metadata Readiness for Scientific AI Datasets. arXiv:2607.22677.
6. Sun, Z. *et al.,*(2026). From Language Models to Agentic AI: A Survey of Autonomous, Action-Enabled, and Collaborative LLM Agents. Cognitive Computation. DOI: 10.1007/s12559-026-10619-1.
7. IEEE Access (2026). A Systematic Literature Review of Agentic AI: Definitions, Architectures, and Challenges. IEEE Access. DOI: 10.1109/access.2026.3668138.
8. Sahu, A.K. *et al.,*(2025). LLM-Based Data Science Agents: A Survey of Capabilities, Challenges, and Future Directions. arXiv:2510.04023. DOI: 10.48550/arxiv.2510.04023.
9. Data Agents (2026). Data Agents: Levels, State of the Art, and Open Problems. arXiv:2602.04261. DOI: 10.48550/arxiv.2602.04261.
10. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. DOI: 10.1126/science.aac4716.
11. SAI (2026). How much science is verifiable? Results from replicating ICML 2026 oral papers. SAI Blog. Available: https://sai.science/blog/how-much-science-is-verifiable.
12. Hugging Face (2026). ICML 2026 Open Reproductions. GitHub. Available: https://github.com/huggingface/blog/blob/main/icml-2026-open-reproductions.md.
13. RefChecker (2026). Phantom References: Hallucinated Citations That Survive Peer Review at Top-Tier Conferences. arXiv:2607.00738. DOI: 10.48550/arxiv.2607.00738.
14. Gehani, A. *et al.,*(2026). AI-Generated Code Is Not Reproducible (Yet). RAI 2026. DOI: 10.48550/arxiv.2512.22387.

15. RDA (2026). RDA Data Director Agentic AI Blueprint WG Recommendation. Research Data Alliance. Available: https://www.rd-alliance.org/wp-content/uploads/2026/06/SEPTEMBER2026_v1.0_ENDORSED_DataDirectorAgenticAIBlueprint.pdf.
16. F(AI)²R (2026). F(AI)2R: Who Did What, and Who Checked? Verifiable AI Provenance as an Executable Skill. arXiv:2607.25637.
17. FOXDEN (2026). FOXDEN: FAIR Services for AI-Ready Scientific Datasets. arXiv:2609.03105. DOI: 10.48550/arxiv.2609.03105.
18. Provenance grounds trust in autonomous science (2026). Nature Computational Science, 6, 1035. DOI: 10.1038/s43588-026-01035-4.
19. Page, M.J. *et al.,*(2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71. DOI: 10.1136/bmj.n71.
20. ARISMA (2026). ARISMA: Guidelines for AI- and LLM-Assisted Systematic Reviews, Scoping Reviews, and Mapping Studies. arXiv:2608.25050.
21. RevAIse (2026). RevAIse: A Data Model and Protocol Conformance Framework for AI-Assisted Systematic Literature Review Workflows. Zenodo. DOI: 10.5281/zenodo.21078213.
22. Hong, Q.N. *et al.,*(2018). The Mixed Methods Appraisal Tool (MMAT) version 2018. ICEP, 20, 285–291. DOI: 10.1186/s40900-018-0091-3.
23. DeepPrep (2026). DeepPrep: An LLM-Powered Agentic System for Autonomous Data Preparation. VLDB PVLDB, 19, 3371. Available: https://www.vldb.org/pvldb/vol19/p3371-fan.pdf.
24. DeepPrep (2026). DeepPrep: An LLM-Powered Agentic System for Autonomous Data Preparation. arXiv:2602.07371.
25. DataEvolver (2026). DataEvolver: Automatic Data Preparation for Large Language Models through Multi-Level Self-Evolving. arXiv:2606.07001. DOI: 10.48550/arxiv.2606.07001.
26. TabClean (2026). TabClean: Scalable Tabular Data Cleaning via Reusable LLM-Synthesized Programs. TaDA 2026. Available: https://tabular-data-analysis.github.io/tada2026/papers/TaDA26_2.pdf.
27. LeDQeR (2026). Data Quality Rule Generation with LLMs. arXiv:2609.06053. DOI: 10.48550/arxiv.2609.06053.
28. AutoDCWorkflow (2025). AutoDCWorkflow: LLM-based Data Cleaning Workflow Auto-Generation and Benchmark. EMNLP Findings 2025. Available: https://aclanthology.org/anthology-files/anthology-files/pdf/findings/2025.findings-emnlp.410.pdf.
29. Quality Assessment (2025). Quality Assessment of Tabular Data using Large Language Models and Code Generation. EMNLP Industry 2025. DOI: 10.18653/v1/2025.emnlp-industry.183.
30. Quality Assessment (2025). Quality Assessment of Tabular Data using Large Language Models and Code Generation. arXiv:2509.10572. DOI: 10.48550/arxiv.2509.10572.
31. He, X. *et al.,*(2021). Automated Machine Learning (AutoML): A survey. TKDD, 15(2), 1–37.
32. Thornton, C. *et al.,*(2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization. AutoML Workshop, ICML 2013.
33. Olson, R.S. *et al.,*(2016). Evaluation of a Tree-based Pipeline Optimization Tool. GECCO 2016. DOI: 10.1145/2908812.2908918.
34. LeDell, E. *et al.,*(2020). H2O AutoML: Scalable Automatic Machine Learning. ICML AutoML Workshop 2020.
35. Kanter, J.M. & Veeramachaneni, K. (2015). Deep Feature Synthesis. IDDA 2015. DOI: 10.1007/s10115-014-0772-3.
36. Trirat, P. *et al.,*(2025). AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML. PMLR v267. Available: https://proceedings.mlr.press/v267/trirat25a.html.
37. AutoMind (2025). AutoMind: Adaptive Knowledgeable Agent for Automated Data Science. arXiv:2506.10974.
38. EvoDS (2026). EvoDS: Self-Evolving Autonomous Data Science Agent with Skill Learning and Context Management. arXiv:2606.03841.
39. SPIO (2026). SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science. ACL 2026. Available: https://aclanthology.org/2026.acl-long.1039.pdf.
40. I-MCTS (2026). I-MCTS: Enhancing Agentic AutoML via Introspective Monte Carlo Tree Search. EACL Findings 2026. Available: https://aclanthology.org/2026.findings-eacl.11.pdf.
41. BDaaS (2026). Trustworthy Self-Composable Big-Data-as-a-Service: An LLM-Orchestrated Multi-Agent Framework. arXiv:2606.17915.
42. Lu, C. *et al.,*(2026). Towards end-to-end automation of AI research. Nature, 641, 10265. DOI: 10.1038/s41586-026-10265-5.
43. DeepScientist (2026). DEEPSCIENTIST: Goal-Oriented Fully Autonomous Scientific Discovery. ICLR 2026. Available: https://proceedings.iclr.cc/paper_files/paper/2026/file/4f64494ecc3442f1c9261baa036378bc-Paper-Conference.pdf.
44. InternAgent (2026). InternAgent-1.5: A Unified Agentic Framework for Long-Horizon Autonomous Scientific Discovery. arXiv:2602.08990. DOI: 10.48550/arxiv.2602.08990.
45. OmniScientist (2026). OmniScientist: An Omni-Modal Omni-Discipline AI Scientist. arXiv:2608.13558. DOI: 10.48550/arxiv.2608.13558.
46. SR-Scientist (2026). SR-SCIENTIST: AI Agent for Autonomous Symbolic Regression. ICLR 2026. Available: https://proceedings.iclr.cc/paper_files/paper/2026/file/7a9d48d13056b6b6a7f5bce1c3ba8f2d-Paper-Conference.pdf.

47. Ontology FAIR (2026). Ontology Engineering and the FAIR principles: A Gap Analysis toward a FAIR-by-design methodology. CEUR-WS, 3882. Available: https://ceur-ws.org/Vol-3882/foam-4.pdf.
48. Semantic Units (2026). The Semantic Units Framework, a technology-agnostic representational approach to FAIR and CLEAR knowledge infrastructures. Scientific Data, 13, 7588. DOI: 10.1038/s41597-026-07588-3.
49. Semantic Units (2026). The Semantic Units Framework. Scientific Data. DOI: 10.1038/s41597-026-07588-3.
50. Bradley, A.H. *et al.,*(2026). FAIRmaterials: Ontology Tools with Data FAIRification in Development. JOSS, 11(117), 7287. DOI: 10.21105/joss.07287.
51. Ontology Framework (2026). An Ontology-Based Framework for Semantic Integration and Interoperable Assessment of Green-Synthesized Nanomaterials. Applied Sciences, 16(3), 1539. DOI: 10.3390/app16031539.
52. RDM Ontology (2026). Integrating Semantics into Research Data Management: Modelling and Validating Materials Science Experiment Workflows. arXiv:2608.20879.
53. KGpipe (2026). KGpipe: Generation of Pipelines for Data Integration into Knowledge Graphs. University of Leipzig. Available: https://dbs.uni-leipzig.de/files/research/publications/2026-9/pdf/KGpipe_Framework.pdf.
54. Wikontic (2026). Wikontic: Constructing Wikidata-Aligned, Ontology-Aware Knowledge Graphs with Large Language Models. EACL 2026. Available: https://aclanthology.org/2026.eacl-long.388.pdf.
55. OntoKG (2026). OntoKG: Ontology-Oriented Knowledge Graph Construction with Intrinsic-Relational Routing. arXiv:2604.02618. DOI: 10.48550/arxiv.2604.02618.
56. MSE-KG (2026). The Materials Science and Engineering Knowledge Graph (MSE-KG): Apache Airflow–Orchestrated Construction Pipeline. Zenodo. DOI: 10.5281/zenodo.19484379.
57. Databricks (2026). Operationalizing Genie Ontology in Your Data Stack. Databricks Blog. Available: https://www.databricks.com/blog/operationalizing-genie-ontology-your-data-stack.
58. Sema (2026). Nine-Sigma/sema: Semantic ontology extraction from data warehouses. GitHub. Available: https://github.com/Nine-Sigma/sema.
59. AWS (2026). sample-semantic-layer-structured/agents/ontology_agent. GitHub. Available: https://github.com/aws-samples/sample-semantic-layer-structured/tree/main/agents/ontology_agent.
60. Neo4j (2026). Build a Semantic Layer from GCP with Neocarta. Neo4j Blog. Available: https://neo4j.com/blog/genai/build-a-semantic-layer-from-gcp-with-neocarta/.
61. FDO Architecture (2026). Fair Digital Objects Architecture DRAFT. FDO Forum. Available: https://fairdo-org.github.io/fdo-architecture-spec/.
62. CASRAI (2026). Persistent identifiers in 2026: ORCID + ROR + RAiD + DOI. CASRAI. Available: https://casrai.org/news/persistent-identifiers-2026-orcid-ror-raid-doi.
63. CASRAI (2026). PID Federation Across Research Infrastructure. CASRAI. Available: https://casrai.org/guides/persistent-identifier-federation-across-research-infrastructure.
64. OptimusKG (2026). OptimusKG: Unifying biomedical knowledge in a modern multimodal graph. arXiv:2604.27269. DOI: 10.48550/arxiv.2604.27269.
65. SciAtlas (2026). SciAtlas: A Large-Scale Knowledge Graph for Automated Scientific Research. arXiv:2605.22878.
66. SciMKG (2026). SciMKG: A Multimodal Knowledge Graph for Science Education with Text, Image, Video and Audio. AAAI 2026. DOI: 10.1609/aaai.v40i18.38574.
67. MKB (2026). Monkey King Bang: A Unified Scientific Multimodal Foundation Model. arXiv:2607.20557. DOI: 10.48550/arxiv.2607.20557.
68. FuXi-Uni (2026). FuXi-Uni: A Native Unified Multimodal Model for Scientific Understanding. arXiv:2601.01363. DOI: 10.48550/arxiv.2601.01363.
69. Intern-S2 (2026). Intern-S2-Preview: Scientific Agentic Foundation Model. arXiv:2608.13505. DOI: 10.48550/arxiv.2608.13505.
70. S1-Omni-Image (2026). S1-Omni-Image: Unified Multimodal Model for Scientific Image Understanding, Generation, and Editing. GitHub. Available: https://github.com/ScienceOne-AI/S1-Omni-Image.
71. VT-Bench (2026). VT-Bench: A Unified Benchmark for Visual-Tabular Multi-Modal Learning. arXiv:2605.08146. DOI: 10.48550/arxiv.2605.08146.
72. MMPFN (2026). MMPFN: Multimodal Prior Fitted Networks for Visual-Tabular Learning. arXiv:2602.20223. DOI: 10.48550/arxiv.2602.20223.
73. RAG Survey (2026). From vectors to knowledge graphs: A comprehensive survey of RAG. Computer Science Review. DOI: 10.1016/j.cosrev.2026.100341.
74. SciRAG (2026). SciRAG: Adaptive, Citation-Aware, and Outline-Guided Retrieval and Synthesis for Scientific Literature. EACL 2026. Available: https://aclanthology.org/2026.eacl-long.303/.
75. LeanRAG (2026). LeanRAG: Knowledge-Graph-Based Generation with Semantic Aggregation and Hierarchical Retrieval. AAAI 2026. Available: https://ojs.aaai.org/index.php/AAAI/article/view/40789.

76. TagRAG (2026). TagRAG: Tag-guided Hierarchical Knowledge Graph Retrieval-Augmented Generation. ACL Findings 2026. Available: https://aclanthology.org/2026.findings-acl.321.pdf.
77. ParallaxRAG (2026). Think Parallax: Solving Multi-Hop Problems via Multi-View Knowledge-Graph-Based Retrieval-Augmented Generation. ACL 2026. Available: https://aclanthology.org/2026.acl-long.1226.pdf.
78. EvoGraph-R1 (2026). EvoGraph-R1: Self-Evolving Multimodal Knowledge Hypergraphs for Agentic Retrieval. CVPR 2026. Available: https://openaccess.thecvf.com/content/CVPR2026/papers/Lin_EvoGraph-R1_Self-Evolving_Multimodal_Knowledge_Hypergraphs_for_Agentic_Retrieval_CVPR_2026_paper.pdf.
79. MegaRAG (2026). MegaRAG: Multimodal Knowledge Graph-Based Retrieval Augmented Generation. ACL 2026. Available: https://aclanthology.org/2026.acl-long.2218.pdf.
80. Science (2026). Researchers presented 6000 papers at a major meeting. Could AI reproduce their findings? Science. DOI: 10.1126/science.adq1234.
81. CiteCheck (2026). CiteCheck: Retrieval-Grounded Detection of LLM Citation Hallucinations in Scientific Text. arXiv:2605.27700. DOI: 10.48550/arxiv.2605.27700.
82. HalluCiteChecker (2026). HalluCiteChecker: A Lightweight Toolkit for Hallucinated Citation Detection and Verification. arXiv:2604.26835. DOI: 10.48550/arxiv.2604.26835.
83. HalluPeer (2026). HalluPeer: A Taxonomy-driven Benchmark for Detecting Hallucinations in Scientific Peer Reviews. arXiv:2609.03580. DOI: 10.48550/arxiv.2609.03580.
84. PROBE (2026). PROBE: PROcess-Based BEnchmark for Hallucination Detection. ACL Findings 2026. Available: https://aclanthology.org/2026.findings-acl.2099.pdf.
85. Enoki (2026). Enoki: Efficient Multi-Level Hallucination Detection. arXiv:2609.00581. DOI: 10.48550/arxiv.2609.00581.
86. Notebook Quality (2026). A Study of Scientific Computational Notebook Quality. arXiv:2603.22726.
87. ReproAgent (2026). ReproAgent: Contract-Guided Paper-to-Code Reproduction. arXiv:2608.24291. DOI: 10.48550/arxiv.2608.24291.
88. DeepRepro (2026). DeepRepro: State-Aware Subplanning for Paper-to-Code Reproduction in Evolving Repositories. arXiv:2608.26557. DOI: 10.48550/arxiv.2608.26557.
89. Snakemaker (2025). Snakemaker: Seamlessly transforming ad-hoc analyses into sustainable Snakemake workflows with generative AI. arXiv:2505.02841. DOI: 10.48550/arxiv.2505.02841.
90. W3C PROV-O (2013). PROV-O: The PROV Ontology. W3C Recommendation. Available: https://www.w3.org/TR/prov-o/.
91. W3C PROV (2013). The PROV Namespace. W3C. Available: https://www.w3.org/ns/prov.
92. PROV-BFO Mapping (2025). A semantic approach to mapping the Provenance Ontology to the Basic Formal Ontology. Scientific Data, 12, 4580. DOI: 10.1038/s41597-025-04580-1.
93. ProvONE (2026). The ProvONE Data Model for Scientific Workflow Provenance. DataONE. Available: http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html.
94. RO-Crate (2026). RO-Crate Metadata Specification 1.3.0. Zenodo. DOI: 10.5281/zenodo.3406497.
95. RO-Crate 1.3 (2026). Announcing RO-Crate 1.3. Research Object Crate. Available: https://www.researchobject.org/ro-crate/blog/2026-06-23/announcing-ro-crate-1-3.
96. RO-Crate Spec (2026). RO-Crate 1.3 Specification. Available: https://www.researchobject.org/ro/crate/specification/1.3/index.html.
97. Workflow Run RO-Crate (2023). Recording provenance of workflow runs with RO-Crate. PLOS ONE. DOI: 10.1371/journal.pone.0309210.
98. Workflow Run RO-Crate (2023). Making workflow provenance FAIR across workflow systems with Workflow Run RO-Crate. Zenodo. DOI: 10.5281/zenodo.8004793.
99. yProv (2024). A software ecosystem for multi-level provenance management in large-scale scientific workflows for AI applications. IEEE SCW 2024. DOI: 10.1109/scw63240.2024.00253.
100. ReProv (2026). ReProv: Extending REANA with WaaS capabilities and data provenance on the European AI-on-Demand Platform. University of Glasgow. Available: https://eprints.gla.ac.uk/391558/1/391558.pdf.
101. Crystalia (2024). Crystalia: Flexible and Efficient Method for Large Dataset Lineage Tracking. IEEE BigData 2024. DOI: 10.1109/bigdata62323.2024.10826067.
102. OmicsLake (2026). OmicsLake: Versioned, Agent-Aware Data Lineage for R/Bioconductor Workflows. bioRxiv. DOI: 10.64898/2026.07.17.739088.
103. F(AI)²R (2026). F(AI)2R: Verifiable AI Provenance as an Executable Skill. arXiv:2607.25637v1.
104. ZTEK (2026). ZTEK: A Zero Trust Epistemic Kernel for AI-Assisted Research Workflows. Zenodo. DOI: 10.5281/zenodo.19591564.
105. Traceable Trust (2026). Traceable Trust for action-ready artificial intelligence in bioscience. arXiv:2608.17997. DOI: 10.48550/arxiv.2608.17997.
106. Artifact Observability (2026). Artifact-centered Claim-aware Observability for Autonomous

Scientific Agents. arXiv:2608.18312. DOI: 10.48550/arxiv.2608.18312.
107. Digital Discovery (2026). Scientists might be reaffirming the relevance of human oversight as AI lands in labs. Digital Discovery. DOI: 10.1039/D6DD00030D.
108. Novelty Ceiling (2026). The Novelty Ceiling: PAC-Theoretic Bounds on Autonomous Scientific Discovery and the Minimum Oversight Rate. ICML 2026. Available: https://icml.cc/virtual/2026/73443.
109. ASD Oversight (2026). Designing Human Oversight in Autonomous Scientific Discovery. Bar-Ilan University. Available: https://cris.biu.ac.il/en/publications/designing-human-oversight-inautonomous-scientific-discovery/.
110. MOOSE-Copilot (2026). MOOSE-Copilot: A Web-Based Interactive Assistant for Unified Exploratory and Fine-Grained Scientific Hypothesis Discovery. ACL Demo 2026. DOI: 10.18653/v1/2026.acl-demo.85.
111. HAI Survey (2026). A Survey of Human-AI Collaboration for Scientific Discovery. TechRxiv. DOI: 10.36227/techrxiv.176823034.44078313/v2.
112. Verifiable Steering (2026). When Should an AI Scientist Stop? Verifiable Experiment Steering and Refusal for Autonomous Discovery. arXiv:2606.07576. DOI: 10.48550/arxiv.2606.07576.
113. Automation Bias (2025). Exploring automation bias in human–AI collaboration: a review and meta-analysis. AI & Society. DOI: 10.1007/s00146-025-02422-7.
114. Overreliance (2026). Conflicting Information and AI Capabilities in Calibrating Overreliance: Age-Differentiated Mechanisms. IJHCI. DOI: 10.1080/10447318.2026.2686773.
115. Calibrated Framework (2026). Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems. Scientific Reports, 16, 65730. DOI: 10.1038/s41598-026-65730-y.
116. Overreliance CAS (2026). Modeling AI Overreliance as a Complex Adaptive System. arXiv:2608.19616. DOI: 10.48550/arxiv.2608.19616.
117. Mitchell, M. *et al.*,(2019). Model Cards for Model Reporting. FAT 2019. DOI: 10.1145/3287560.3287596.
118. Model Cards 2026 (2026). Model Cards and System Cards: What Regulators Expect by Default. CallSphere. Available: https://callsphere.ai/blog/model-cards-system-cards-2026-regulators-expect-by-default.md.
119. Data Cards (2026). The Data Cards Playbook. Google Research. Available: https://sites.research.google/datacardsplaybook/.
120. [120] MLOps Documentation (2026). MLOps & Automated Documentation: Meeting the AI Act's Logging Requirements. IMT Solutions. Available: https://m.imt-soft.com/ja/2026/06/05/mlops-automated-documentation-meeting-the-ai-acts-logging-requirements/.
121. Model Cards Datasheets (2026). Model Cards and Datasheets: Documentation That Matters. AIPMO. Available: https://aipmo.co/model-cards-datasheets/.
122. RDA Blueprint (2026). Data Director Agentic AI Tool Blueprint. RDA. Available: https://www.rd-alliance.org/groups/data-director-agentic-ai-blueprint/outputs/data-director-agentic-ai-tool-blueprint/.
123. FAIR+COPE (2026). Advancing FAIR data towards comparable, organized, predictive AI-ready data for community validation. Communications Biology. DOI: 10.1038/s42003-026-10694-y.
124. Privacy Claim (2026). Position: Privacy Is a Claim, Not a Property of Synthetic Data. arXiv:2609.01273. DOI: 10.48550/arxiv.2609.01273.
125. MLReplicate (2026). MLReplicate: Benchmarking Autonomous Research Systems for Machine Learning Reproducibility. arXiv:2605.16616. DOI: 10.48550/arxiv.2605.16616.
126. AgentActionBench (2026). Overview of the NLPCC 2026 Shared Task 11: Agent-Based Experiment Reproduction from Scientific Papers. arXiv:2609.11117. DOI: 10.48550/arxiv.2609.11117.
127. Synthetic Data Ethics (2025). GenAI synthetic data create ethical challenges for scientists. PNAS. DOI: 10.1073/pnas.2409182122.
128. Anonymity Claims (2026). Rethinking Anonymity Claims in Synthetic Data Generation: A Model-Centric Privacy Attack Perspective. arXiv:2601.22434. DOI: 10.48550/arxiv.2601.22434.
129. Utility-Fidelity-Privacy (2026). Beyond Realism: A Utility-Fidelity-Privacy Framework for Trustworthy Synthetic Data. JDIs. DOI: 10.1515/jdis-2026-0038.
130. Decision Workflows (2026). Contextualized, Trustworthy, and Collective Scientific Decision Workflows. SN Computer Science. DOI: 10.1007/s42979-026-05224-w.
131. EOSC (2026). European Open Science Cloud - EOSC. European Commission. Available: https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud.
132. EOSC Data Commons (2026). EOSC - EOSC Data Commons. Available: https://www.eosc-data-commons.eu/eosc.
133. EOSC EU Node (2026). EOSC EU Node. European Commission. Available: https://open-science-cloud.ec.europa.eu/about/eosc-eu-node.
134. EOSC Data Commons (2026). EOSC Data Commons: Building Europe's Next-Generation Research Data Infrastructure. ERCIM News, 144. Available: https://ercim-news.ercim.eu/en144/special/eosc-data-commons-

building-europes-next-generation-research-data-infrastructure.
135. SeDa (2026). SeDa: A Unified System for Dataset Discovery and Multi-Entity Augmented Semantic Exploration. arXiv:2603.07502. DOI: 10.48550/arxiv.2603.07502.
136. Leibniz Data Manager (2026). Enabling Federated Data Services: Semantic Research Data Management with the Leibniz Data Manager. TIB. Available: https://oa.tib.eu/renate/items/adefe42d-c83f-4d7b-95b5-bbb451fbd158.
137. FAIRsFAIR Metrics (2025). FAIRsFAIR Data Object Assessment Metrics. Zenodo. DOI: 10.5281/zenodo.15045911.
138. RDA Maturity Model (2020). FAIR Data Maturity Model: specification and guidelines. RDA. Available: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg/outputs/fair-data-maturity-model-specification-and-guidelines/.
139. F-UJI (2026). F-UJI: Automated FAIR Data Assessment. CASRAI. Available: https://casrai.org/guides/f-uji-automated-fair-assessment-tool.
140. FIP Check (2026). FIP Check: A Rubric-Based Tool for Assessing FAIR Implementation Profiles and Enabling Resources. Data Science Journal. DOI: 10.5334/dsj-2026-008.
141. FAIR Semantic Metrics (2026). A Semantic Web Model for Standardizing FAIR Metrics and FAIR Assessment Results. Data Science Journal. DOI: 10.5334/dsj-2026-030.
142. FAIR4RS (2022). FAIR Principles for Research Software (FAIR4RS Principles). RDA. Available: https://www.rd-alliance.org/wp-content/uploads/2022/03/FAIR4RS20principles20v1.0.pdf.
143. FAIR Workflows (2026). Applying the FAIR Principles to computational workflows. PMC. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11850811/.
144. Workflow Knowledge Base (2026). A Standards-Based Knowledge Graph that Bridges Scientific Workflows, Run-Time Provenance, and Tool Registries. HAL. Available: https://hal.science/hal-05517640.
145. FAIR Research Objects (2026). FAIR Research Objects and Computational Workflows. University of Amsterdam. Available: https://dare.uva.nl/id/d7fcc474-fcd1-4f2e-b5c0-1fd53a88f11d.
146. Omics Workflow (2026). A metadata-managed FAIR end-to-end workflow for microbial community omics data analysis. Scientific Data, 13, 8224. DOI: 10.1038/s41597-026-08224-w.
147. FDO Specifications (2026). Specifications – FAIR Digital Objects. FDO Forum. Available: https://fairdo.org/specifications/.
148. FDO Framework (2026). FAIR Digital Object Framework Documentation. Available: https://fairdigitalobjectframework.org/.
149. CASRAI FDO (2026). FAIR Digital Objects (FDO) Explained. CASRAI. Available: https://casrai.org/guides/fair-digital-object-fdo.
150. CASRAI PID (2026). Persistent identifiers. CASRAI. Available: https://casrai.org/for-authors/persistent-identifiers.
151. Mutation Testing (2026). Mutation Testing for Reproducibility Safeguards in Machine Learning Research Software: An Empirical Study. arXiv:2608.27100. DOI: 10.48550/arxiv.2608.27100.
152. ReproNLP (2026). The Shared Task on Reproducibility of Evaluations in NLP (ReproNLP) 2026. ACL Anthology. Available: https://aclanthology.org/2026.gem-main.83/.
153. NeurIPS Reproducibility (2026). NeurIPS 2026 Call For Reproducibility. NeurIPS. Available: https://neurips.cc/Conferences/2026/CallForReproducibility.
154. CyAgent (2026). From SQL to Knowledge Graphs: An LLM-Driven Multi-Agent Approach with Data Schema Improvement. arXiv:2608.26117. DOI: 10.48550/arxiv.2608.26117.
155. Multi-Layer KG (2026). From Document Silos to Process Intelligence: A Multi-Layer Knowledge Graph for CMC Process Development. arXiv:2609.11493. DOI: 10.48550/arxiv.2609.11493.
156. REMAP (2026). Multimodal Learning on Graphs for Disease Relation Extraction. Harvard Zitnik Lab. Available: https://zitniklab.hms.harvard.edu/projects/REMAP/.
157. Atomistic KG (2026). Ontology-based knowledge graph infrastructure for interoperable atomistic simulation data. arXiv:2604.06230. DOI: 10.48550/arxiv.2604.06230.
158. RDM KG (2026). Integrating Semantics into Research Data Management: Modelling and Validating Materials Science Experiment Workflows. arXiv:2608.20879. DOI: 10.48550/arxiv.2608.20879.
159. Data Governance (2026). Construction of a Theoretical Framework for Scientific Data Governance. Scientific Data, 13, 6525. DOI: 10.1038/s41597-025-06525-0.
160. Executable Metadata (2026). Executable Metadata for Governed, Interoperable Data Platforms. IEEE ICIPTM 2026. DOI: 10.1109/iciptm69057.2026.11466030.
161. FRED (2026). FRED enables standardized FAIR metadata generation and management for omics research. Scientific Reports, 16, 61886. DOI: 10.1038/s41598-026-61886-9.
162. DRAI (2026). Data Readiness for AI principles (DRAI). Referenced in SetGo, arXiv:2607.22677.

163. REDI (2026). Readiness Engine for Data Integration (REDI). Referenced in SetGo, arXiv:2607.22677.
164. LACE (2026). Evolving Executable Pipeline Programs for AutoML with Language Models. arXiv:2608.16416. DOI: 10.48550/arxiv.2608.16416.
165. Agentic AI Survey (2026). Agentic AI: a comprehensive survey of architectures, applications, and future directions. AI Review. DOI: 10.1007/s10462-025-11422-4.
166. Holistic Review (2026). A Holistic Review of Agentic AI Frameworks, Applications, and Research Trajectories. Archives of Computational Methods in Engineering. DOI: 10.1007/s11831-026-10675-8.
167. Agentic Systems Survey (2026). Agentic AI systems: A systematic survey of multi-agent architectures, cognitive foundations, interaction, explainability, security, and performance evaluation. Neurocomputing, 696, 134049. DOI: 10.1016/j.neucom.2026.134049.
168. Nature Robin (2026). A multi-agent system for automating scientific discovery. Nature, 641, 10652. DOI: 10.1038/s41586-026-10652-y.
169. ProvDeploy (2024). ProvDeploy: Provenance-oriented Containerization of High Performance Computing Scientific Workflows. arXiv:2403.15324. DOI: 10.48550/arxiv.2403.15324.
170. Kubernetes Reproducibility (2020). Reproducibility of Computational Experiments on Kubernetes-Managed Container Clouds with HyperFlow. DOI: 10.1007/978-3-030-50371-0_16.
171. [171] KubeAdaptor (2023). KubeAdaptor: A docking framework for workflow containerization on Kubernetes. Future Generation Computer Systems. DOI: 10.1016/j.future.2023.06.022.
172. AI Migration (2026). An AI-Assisted Migration Framework for Transforming Legacy Scientific Applications into Reusable Cloud-Based Workflows. arXiv:2608.23146. DOI: 10.48550/arxiv.2608.23146.
173. NextGen Containerized (2026). Portable Reproducibility for NextGen Water Modeling Using Containerized HPC and Cloud Workflows. CIROH. Available: https://ciroh.ua.edu/abstracts/portable-reproducibility-for-nextgen-water-modeling-using-containerized-hpc-and-cloud-workflows/.
174. Kubernetes Quantum (2026). Kubernetes-Orchestrated Hybrid Quantum–Classical Workflows. arXiv:2603.24206. DOI: 10.48550/arxiv.2603.24206.
175. Sarus Suite (2026). Sarus Suite: Multi-container workloads for HPC. arXiv:2604.17064. DOI: 10.48550/arxiv.2604.17064.
176. Evidence-Aligned Entity (2026). Evidence-Aligned Entity Verification for Hallucination Detection in Retrieval-Augmented Generation. arXiv:2609.08267. DOI: 10.48550/arxiv.2609.08267.
177. Hallucination Survey (2026). Hallucinations in generative artificial intelligence and large language models: A survey. Language Resources and Evaluation. DOI: 10.1007/s10579-026-09938-4.
178. GraLC-RAG (2026). Graph-Aware Late Chunking for Retrieval-Augmented Generation in Biomedical Literature. arXiv:2603.22633. DOI: 10.48550/arxiv.2603.22633.
179. EXYGEN (2026). Enabling Knowledge Graph Understanding at Scale with the EXplore Your Graphs ENgine (EXYGEN). arXiv:2609.11569. DOI: 10.48550/arxiv.2609.11569.
180. LLMDapCat (2026). LLMDapCat: An LLM-based Data Catalogue System for Data Sharing and Exploration. CEUR-WS, 4085. Available: https://ceur-ws.org/Vol-4085/paper80.pdf.
181. Vocabulary Hub (2026). The Vocabulary Hub as a Catalog for Semantic Artifacts for Discovery and Alignment of Datasets. ESWC 2026. Available: https://pietercolpaert.be/papers/eswc2026-deploy-emds/SDS2026.pdf.
182. Dataset Explorer (2026). Semantic search across 200K+ public dataset schemas. GitHub. Available: https://github.com/peter-products/dataset-explorer.
183. Datum (2026). Datum: A Scientific Metadata Catalog. INL. Available: https://inlsoftware.inl.gov/product/datum.
184. Trust Transparency (2026). Building trustworthy Artificial Intelligence through transparency explainability uncertainty and trust calibration. Discover AI. DOI: 10.1007/s44163-026-01219-x.
185. Trust Calibration Design (2026). Trust Calibration in Agentic AI: Designing for Appropriate Reliance, Not Blind Trust. Designative. Available: https://www.designative.info/2026/05/21/trust-calibration-in-agentic-ai-designing-for-appropriate-reliance-not-blind-trust/.
186. Human Reliance (2026). Examining human reliance on artificial intelligence in decision making. PMC. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12881489/.
187. Synthetic Data Policy (2026). Synthetic Data in Research: Generation, Documentation, and Sharing Policy. CASRAI. Available: https://casrai.org/guides/synthetic-data-in-research.
188. DP Synthetic Data (2026). How to DP-Fy Your Data: A Practical Guide to Generating Synthetic Data With Differential Privacy. JAIR, 86, Article 17. DOI: 10.1613/jair.1.21111.
189. Synthetic Clinical Data (2026). Ethic and legal management of high fidelity synthetic clinical data. OSF. DOI: 10.31219/osf.io/wbvrm_v1.
190. FAIR Checklist (2026). How to Make Your Dataset FAIR: A Checklist. CASRAI. Available: https://casrai.org/guides/how-to-make-your-dataset-fair-a-step-by-step-checklist.
191. FAIR Assessment Guide (2026). FAIR data assessment frameworks: a buyer's guide for institutions. CASRAI. Available:

https://www.casrai.org/news/fair-data-assessment-frameworks-2026.
192. FAIR Trust (2026). How FAIR data are helping to build trust in science. Nature. DOI: 10.1038/d41586-026-01982-y.
193. RDA EOSC (2026). Enabling an operational, open and FAIR EOSC ecosystem (INFRAEOSC). European Research Executive Agency. Available: https://rea.ec.europa.eu/funding-and-grants/horizon-europe-research-infrastructures/enabling-operational-open-and-fair-eosc-ecosystem-infraeosc_en.
194. EOSC Association (2026). EOSC EU Node - EOSC Association. Available: https://eosc.eu/building-the-eosc-federation/eosc-eu-node.
195. EOSC Managed Services (2026). Managed Services for the EOSC EU Node 2026. European Commission. Available: https://digital-strategy.ec.europa.eu/en/funding/managed-services-eosc-eu-node-2026.
196. PRISMA Toolkit (2026). PRISMA: A Python Toolkit for Systematic Literature Reviews. Zenodo. DOI: 10.5281/zenodo.19883809.
197. prismAId (2026). prismAId - Open Science AI Tools for Systematic, Protocol-Based Literature Reviews. Zenodo. DOI: 10.5281/zenodo.19607410.
198. PRISMA-LLM (2026). PRISMA-LLM Pipeline: Five-Layer Human-in-the-Loop Systematic Review Screening. Zenodo. DOI: 10.5281/zenodo.20395205.
199. Systematic Review Skill (2026). AngelChen-HC/systematic-review-skill. GitHub. Available: https://github.com/angelchen-hc/systematic-review-skill.
200. Jupyter Studio (2026). deepelementlab/jupyter-studio: AI-native JupyterLab. GitHub. Available: https://github.com/deepelementlab/jupyter-studio.
201. My-Jogyo (2026). Yeachan-Heo/My-Jogyo: End-to-end research automation. GitHub. Available: https://github.com/Yeachan-Heo/My-Jogyo/.
202. GO FAIR (2026). FAIR Principles. GO FAIR. Available: https://www.go-fair.org/fair-principles/.
203. FORCE11 (2026). The FAIR Data Principles. FORCE11. Available: https://force11.org/info/the-fair-data-principles.
204. CASRAI Identifiers (2026). Identifier Crosswalk: Connecting DOI, ORCID, ROR, and Institutional IDs in Practice. CASRAI. Available: https://casrai.org/guides/identifier-crosswalk-connecting-doi-orcid-ror-institutional-ids.
205. ProvCaRe (2026). An Extensible Ontology Modeling Approach Using Post Coordinated Expressions for Semantic Provenance in Biomedical Research. NSF PAR. Available: https://par.nsf.gov/servlets/purl/10067792.
206. RDA D4.5 (2021). D4.5 Report on FAIR Data Assessment Toolset and Badging Scheme. Zenodo. DOI: 10.5281/zenodo.5336159.
207. Sakana AI (2026). The AI Scientist: Towards Fully Automated AI Research. Sakana AI. Available: https://sakana.ai/ai-scientist-nature/.
208. Agentic AI Taxonomy (2026). Agentic AI: a comprehensive survey of architectures, applications, and future directions. AI Review. DOI: 10.1007/s10462-025-11422-4.
209. Model Documentation Standard (2026). AI Model Documentation Standard. TheProductionLine. Available: https://www.theproductionline.ai/knowledge-base/governance/ai-model-documentation.
210. LACE Pipeline (2026). LACE: Evolving Executable Pipeline Programs for AutoML with Language Models. arXiv:2608.16416.
211. Cosmological Simulation KG (2026). Ontology-based knowledge graph infrastructure for interoperable atomistic simulation data. arXiv:2604.06230.
212. Bioscience Trust (2026). Traceable Trust for action-ready artificial intelligence in bioscience. arXiv:2608.17997.
213. ProvONE Best Practices (2026). PROV Best Practices. W3C. Available: https://dvcs.w3.org/hg/prov/raw-file/tip/bestpractices/BestPractices.html.
214. Bioschemas (2026). Bioschemas Computational Workflow profile. Referenced in RO-Crate 1.3 specification.
215. Amazon SageMaker (2026). Amazon SageMaker Model Cards. AWS. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html.
216. Synthetic Data PNAS (2025). GenAI synthetic data create ethical challenges for scientists using synthetic data in research. PNAS. DOI: 10.1073/pnas.2409182122.
217. ProvCaRe Ontology (2026). ProvCaRe: Extending PROV-O for biomedical research provenance. Referenced in NSF PAR.
218. EOSC Federation (2026). EOSC EU Node - EOSC Association. Available: https://eosc.eu/building-the-eosc-federation/eosc-eu-node.