SciHorizon-DataEVA: An Agentic System to Scalable AI-Readiness Evaluation of Heterogeneous Scientific Data

Dianyu Liu*
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
liudianyu00@gmail.com

Chuan Qin*
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
chuanqin0426@gmail.com

Xi Chen
University of Science and Technology of China
Hefei, China
Computer Network Information Center, Chinese Academy of Sciences
Beijing, China
chenxi0401@mail.ustc.edu.cn

Xiaohan Li
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
laughan123@gmail.com

Wenxi Xu
Computer Network Information Center, Chinese Academy of Sciences
School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
Beijing, China
wxxu@cnic.cn

Yuyang Wang
University of Science and Technology of China
Hefei, China
Computer Network Information Center, Chinese Academy of Sciences
Beijing, China
wyy1@mail.ustc.edu.cn

Xin Chen
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
chx@cnic.cn

Yuanchun Zhou
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
zyc@cnic.cn

Hengshu Zhu†
Computer Network Information Center, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
zhuhengshu@gmail.com

Abstract

AI-for-Science (AI4Science) is increasingly transforming scientific discovery by embedding machine learning models into prediction, simulation, and hypothesis generation workflows across domains. However, the effectiveness of these models is fundamentally constrained by the AI-readiness of scientific data, for which no scalable and systematic evaluation mechanism currently exists. In this work, we propose SciHorizon-DataEVA, a novel agentic system to scalable AI-readiness evaluation of heterogeneous scientific data. At the evaluation-criteria level, we introduce the Sci-TQA² principles, which organize AI-readiness into four complementary dimensions: Governance Trustworthiness, Data Quality, AI Compatibility, and Scientific Adaptability. Each dimension is decomposed into measurable atomic elements that enable fine-grained and executable assessment. To operationalize these principles at scale, we develop Sci-TQA²-Eval, a hierarchical multi-agent evaluation approach orchestrated through a directed, cyclic workflow. Our Sci-TQA²-Eval dynamically constructs dataset-aware evaluation specifications by combining lightweight dataset profiling, applicability-aware metric activation, and knowledge-augmented planning grounded in domain constraints and dataset–paper signals. These specifications are executed through an adaptive, tool-centric evaluation mechanism with built-in verification and self-correction, enabling scalable and reliable assessment across heterogeneous scientific data. Extensive experiments on scientific datasets spanning multiple domains demonstrate the effectiveness and generality of SciHorizon-DataEVA for principled AI-readiness evaluation.

CCS Concepts

• **Computing methodologies** → **Multi-agent planning**; • **General and reference** → **Evaluation**.

Keywords

Scientific data; agentic systems; AI-readiness; automated evaluation

ACM Reference Format:
Dianyu Liu, Chuan Qin, Xi Chen, Xiaohan Li, Wenxi Xu, Yuyang Wang, Xin Chen, Yuanchun Zhou, and Hengshu Zhu. 2026. SciHorizon-DataEVA: An Agentic System to Scalable AI-Readiness Evaluation of Heterogeneous Scientific Data. In *Proceedings of the 32nd ACM SIGKDD Conference on*

*Both are co-first authors and contribute equally to this work.
†Corresponding author.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818996

Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818996

1 Introduction

AI-for-Science (AI4Science) has emerged as a core research paradigm in which machine learning models are embedded within scientific analysis and discovery workflows [4, 23, 44, 53]. Across domains such as structural biology and materials science, AI methods are increasingly used for prediction, simulation, and data-driven hypothesis generation [20, 28]. The effectiveness of these methods, however, is fundamentally constrained by the AI-readiness of the scientific data they operate on. Scientific datasets vary widely in quality, governance, and structural characteristics, and such heterogeneity directly affects their suitability for AI training, inference, and downstream scientific reasoning [2]. Despite the rapid growth of large-scale scientific data repositories—such as ScienceDB, which hosts over 15 million datasets as of January 2026 [38]—there is no scalable mechanism to assess AI-readiness, leaving data selection largely ad hoc and limiting effective use at scale.

Data evaluation has long been studied in the data mining literature. Classical data quality frameworks primarily assess generic properties such as completeness, consistency, and timeliness [43], while FAIR-oriented tools focus on findability, accessibility, interoperability, and reusability of data [45]. Recently, SciHorizon framework [33] introduced a novel "Data + Large Language Model" dual-perspective approach to assess AI4Science readiness, including an AI-readiness assessment framework for scientific data. However, the current evaluation process still relies heavily on manual participation and lacks an automated, intelligent workflow for scalable assessment. Moreover, the indicator system remains insufficient for comprehensively evaluating AI-readiness in scientific contexts, and the framework primarily focuses on Earth and Life Sciences, limiting its extensibility across diverse scientific domains. Consequently, existing approaches exhibit fundamental limitations in **what they evaluate**. *One key limitation is the lack of assessment of AI compatibility.* Current methods do not assess whether data representations, feature structures, and implicit inductive assumptions are compatible with the requirements of learning and inference [5, 32, 47, 56]. As a result, datasets that satisfy traditional quality criteria may still be unsuitable for modern AI models [6, 18, 34, 39, 42]. For example, many molecular property predictors rely on 3D conformations and equivariant representations, rendering datasets that provide only SMILES strings or 2D graphs effectively incompatible despite being structurally complete [22, 25]. *Equally important, existing approaches rarely evaluate scientific adaptability.* They do not determine whether datasets contain the contextual and experimental variables required for scientifically meaningful reasoning about underlying mechanisms, rather than mere statistical fitting. Consequently, models trained on such data may produce predictions that appear accurate yet fail to support scientific interpretation or generalization. In clinical datasets, for instance, missing socioeconomic variables can lead models to attribute patient outcomes to hospital choice rather than underlying social or clinical factors, yielding predictions that are statistically plausible but scientifically invalid [31].

Beyond these limitations in existing evaluation criteria, current tools face substantial practical challenges in **how** the AI-readiness of scientific data **is evaluated**. *A central difficulty stems from the heterogeneous and multimodal nature of scientific data.* Scientific datasets span diverse modalities, including genomic sequences, medical images, molecular graphs, and high-dimensional simulation outputs, each governed by distinct structural and semantic constraints. As a result, most existing evaluation tools are designed for a narrow set of data types and fail to generalize across modalities without extensive manual adaptation. *These challenges are further exacerbated by limited scalability.* Existing evaluation mechanisms are predominantly static, relying on manually specified rules or rigid scripts that are typically designed for specific data formats or structured representations. Applying such tools to heterogeneous scientific datasets often requires domain experts to develop specialized evaluation logic for each new data type or discipline. As scientific data continues to grow in both volume and diversity, this expert-driven, format-specific process becomes increasingly costly and difficult to sustain at scale.

To address the above challenges, we propose a novel agentic system, called **SciHorizon-DataEVA**, for scalable AI-readiness evaluation of heterogeneous scientific data. At the evaluation-criteria level, we introduce **Sci-TQA² principles**, which organizes AI-readiness into four complementary dimensions: *Governance Trustworthiness* (T), *Data Quality* (Q), *AI Compatibility* (A), and *Scientific Adaptability* (A). Each dimension is further decomposed into measurable atomic elements that serve as executable units for systematic evaluation. Building on this foundation, we develop **Sci-TQA²-Eval**, a hierarchical multi-agent evaluation approach that operationalizes these principles at scale through a directed, cyclic workflow. Specifically, **Sci-TQA²-Eval** first generates dataset-aware evaluation specifications: a profile-oriented data inspector constructs a lightweight dataset profile without full data loading; an applicability-aware metric selector activates only fine-grained evaluation metrics that are feasible for the given dataset; and a knowledge-augmented evaluation specification planner instantiates each activated atomic element into executable data loading and scoring strategies, guided by retrieved domain constraints (Type-I knowledge) and dataset–paper relational signals (Type-II knowledge). These specifications are then executed by an adaptive, tool-centric evaluation module that integrates a structured tool library, persistent tool memory, and extensible tool orchestration to support scalable execution and cold-start tool generation, while a review-driven verification and self-correction mechanism ensures runtime correctness and semantic validity. Finally, element-level results are aggregated into multi-dimensional AI-readiness scores and synthesized into an AI-Ready Report, enabling robust and generalizable readiness assessment across heterogeneous scientific data. Extensive experiments on scientific datasets spanning multiple scientific domains demonstrate the effectiveness of our approach.

2 Related Work

2.1 Scientific Dataset Evaluation

Data evaluation is a cornerstone of building reliable data-driven systems. Early work largely framed it as a classical data quality problem, focusing on intrinsic dimensions such as accuracy, completeness, and consistency [13, 17]. In parallel, the FAIR principles (Findable, Accessible, Interoperable, Reusable) became the de facto

Table 1: Comparison of SciHorizon-DataEVA with existing data evaluation frameworks.

Methods	DQ Toolkit [40]	FAIR-Cookbook [36]	FAIR-Checker [10]	OpenDataVal [19]	AIDRIN [15]	SciHorizon [33]	Ours
Dimensions							
1. Trustworthiness	◐	●	●	○	●	●	●
2. General Quality	●	○	◐	●	●	●	●
3. AI Compatibility	◐	○	○	●	●	◐	●
4. Scientific Adaptability	○	○	○	○	○	◐	●
Capabilities							
Multi-Modal Support	○	○	◐	◐	○	●	●
Automated Tools	●	○	●	●	●	◐	●
Adaptability	○	◐	◐	○	○	○	●

Symbols: ● Fully Supported, ◐ Partially Supported, ○ Unsupported.

standard for scientific data stewardship [36, 46]. These paradigms have substantially shaped how datasets are curated and shared.

Recently, however, a growing body of work has argued that traditional quality notions and FAIRness, while necessary, are not sufficient for data used in AI model training and deployment [24, 35, 57]. Beyond being "high quality" or "FAIR," datasets must satisfy AI-specific requirements related to robustness, bias, generalization, and downstream utility [8, 9, 50, 54, 55]. This has led to the emergence of *AI-Readiness* as a distinct evaluation paradigm that explicitly targets the suitability of data for AI pipelines [14]. Several domain-specific AI-Readiness frameworks have been proposed in areas such as space science, medicine, and broader scientific data management. These frameworks introduce readiness levels, trustworthiness dimensions, and lifecycle-oriented maturity matrices [3, 30, 37]. Most notably, the SciHorizon project established a benchmark across Earth, Life, and Material sciences [33]. However, these approaches typically cover only a subset of relevant dimensions, rely heavily on predefined tools with limited coverage, and often require substantial expert intervention. In this paper, we propose holistic evaluation principles, Sci-TQA² principles, which explicitly address this gap.

2.2 Automated Evaluation Methods

Existing evaluation practices rely heavily on predefined tools, which can be grouped into three categories: traditional data quality tools, scientific metadata evaluators, and static AI-readiness toolkits. General tools such as DQLearn [40] and the Data Quality Toolkit [11] provide APIs for profiling structured data and detecting statistical anomalies (e.g., label noise), but lack the domain context required for scientific data. For scientific data management, tools such as FAIR-Checker [10] use semantic web technologies to audit metadata against community standards (e.g., DCAT, Bioschemas). They are effective for assessing FAIR compliance but do not inspect data content or its suitability for AI modeling. More recent tools such as AIDRIN [15, 16] and OpenDataVal [19] integrate AI-specific metrics (e.g., fairness, privacy, feature importance) into unified frameworks. However, they still rely on static, predefined rules, are mainly designed for structured data, and struggle to assess heterogeneous scientific modalities (e.g., molecular graphs, genomic sequences) or to generate domain-specific evaluation tools autonomously.

Recent advances in multi-agent systems (MAS) have demonstrated strong reasoning capabilities and sophisticated tool-use in scientific domains[7, 26, 27, 41, 49]. Building on these, we design Sci-TQA²-Eval, which is, to our knowledge, the first system to apply agentic tool generation to adaptively assess AI-readiness—including AI compatibility across heterogeneous scientific datasets.

3 Sci-TQA² Principles

We propose Sci-TQA² principles, which systematically characterizes the AI-readiness of scientific data into four dimensions: *Governance Trustworthiness*, *Data Quality*, *AI Compatibility*, and *Scientific Adaptability*, each further decomposed into sub-dimensions and atomic evaluation elements.

3.1 Top-level Dimensions

3.1.1 Governance Trustworthiness. This dimension characterizes whether a scientific dataset can be safely reused, shared, and redistributed under legal, ethical, and institutional constraints. It focuses on governance artifacts that render the dataset lifecycle transparent and auditable, which is particularly critical in high-stakes settings (e.g., clinical data or proprietary industrial experiments), where non-compliant reuse may invalidate results or incur legal risk.

This dimension is specified through three sub-dimensions: *FAIRness Principles*, which address the findability, accessibility, interoperability, and reusability of data and metadata; *Provenance & Licensing*, which ensure traceability of data origin and clarity of ownership and usage rights; and *Scientific Ethics*, which cover consent compliance and dual-use risk considerations. At a practical level, Governance Trustworthiness evaluates whether downstream users can unambiguously determine *what the data is*, *where it originates*, *under what terms it may be used*, and *whether its use is ethically permissible*. Strong performance on this dimension lowers non-technical barriers to adoption and supports cross-lab reproducibility through explicit and verifiable governance.

3.1.2 Data Quality. This dimension assesses whether a dataset is technically reliable and sufficiently standardized for ingestion by computational pipelines with minimal manual intervention. It reflects how faithfully the released datasets represent the intended observations and whether they conform to domain-appropriate representations, such as unit systems, coordinate conventions, and

identifier standards. Data quality has a direct impact on downstream engineering effort, as deficiencies often necessitate substantial reconciliation across files, instruments, or modalities.

This dimension comprises four sub-dimensions: *Completeness*, which addresses missing values and record integrity; *Accuracy*, which concerns the correctness of measurements and annotations; *Uniqueness*, which captures the presence of duplicate or near-duplicate samples; and *Consistency*, which ensures uniform formats, schemas, and encodings across artifacts. By systematically evaluating these sub-dimensions, practitioners can identify concrete quality deficiencies and prioritize targeted remediation actions, such as data collection or imputation, calibration and validation, deduplication, and unit or schema normalization, enabling curation-oriented improvements beyond coarse, one-shot quality indicators.

3.1.3 AI Compatibility. This dimension assesses the technical utility of a dataset for supporting learning and inference in AI models. Beyond basic data cleanliness, a dataset must provide sufficient learning signal, appropriate representational structure, and adequate scale to support common modeling paradigms. This dimension therefore examines whether a dataset can be effectively used by AI systems under realistic modeling assumptions.

We further decompose it into four sub-dimensions: *AI Model Adaptability*, which reflects support for diverse model families and learning setups; *Data Scale*, which captures effective sample size and dimensionality; *Class Balance*, which measures skew in label or outcome distributions; and *Feature Importance*, which assesses the strength of association between inputs and targets. By evaluating these sub-dimensions, practitioners can derive modeling-oriented recommendations, including adopting data-efficient learning when scale is limited, applying reweighting or resampling under severe class imbalance, and reconsidering modeling objectives when features are weakly informative. Such evaluation also helps identify bottlenecks that render standard learning and inference unreliable.

3.1.4 Scientific Adaptability. This dimension evaluates whether a dataset supports genuine scientific discovery rather than mere statistical fitting. Whereas AI Compatibility concerns the feasibility of learning and inference, Scientific Adaptability examines whether a dataset enables mechanism-aware reasoning, meaningful generalization across regimes, and scientifically interpretable conclusions.

This dimension is decomposed into four sub-dimensions. *Task Generalizability* captures the breadth of downstream scientific tasks supported by the dataset. *Data Scarcity* reflects the difficulty of data acquisition and the uniqueness of observations. *Causal Completeness* evaluates whether key variables and potential confounders are sufficiently represented to support causal reasoning. *Operating Condition Coverage* assesses the extent to which the dataset spans relevant physical or experimental parameter ranges. A dataset may be large and well annotated yet still score poorly on this dimension if it omits critical regimes, conflates confounding factors, or lacks the contextual variables required to interpret outcomes. By evaluating these sub-dimensions, Scientific Adaptability estimates a dataset's potential to support scientific discovery and signals when a benchmark risks encouraging narrow interpolation rather than robust scientific insight.

3.2 Atomic Evaluation Elements

To support fine-grained, actionable assessment, each sub-dimension in Sci-TQA² principles is decomposed into atomic evaluation elements. These elements represent the smallest measurable units of evaluation and, when feasible, are implemented as standalone quantitative checks. Some atomic elements are domain-agnostic (e.g., file integrity, schema validity, basic missingness rates, duplicate detection, and train–test leakage checks), whereas others require domain expertise for meaningful specification and quantification (e.g., acceptable value ranges, instrument-specific calibration constraints, and scientifically meaningful coverage of operating regimes).

This atomic-element design ensures that evaluation remains both *mathematically well-defined* and *scientifically grounded*. Each element produces a normalized score accompanied by interpretable metrics and evidence traces (e.g., the fraction of records that violate unit conventions, the fields that lack provenance, and the subsets that exhibit distribution shift). Element-level scores can be aggregated to obtain sub-dimension and dimension-level summaries while preserving drill-down diagnostics for remediation. More details of Sci-TQA² principles are provided in [21].

4 Sci-TQA²-Eval

To operationalize the above principles for scientific data evaluation in an automated and scalable manner across heterogeneous datasets, we design a novel evaluation approach, Sci-TQA²-Eval. As shown in Figure 1, it is implemented as an MAS orchestrated by a directed cyclic graph (DCG) workflow. Unlike static, single-pass pipelines, Sci-TQA²-Eval tightly integrates *knowledge-grounded planning*, *tool-centric execution*, and *review-driven feedback*, enabling robust and adaptive evaluation over diverse scientific data representations and domain contexts. We first give an overview of the system architecture, followed by detailed descriptions of each module.

4.1 Overview

Our MAS consists of two coupled modules, transforming a raw dataset into executable quality assessments and an AI-ready report:

(I) Dataset-Aware Evaluation Specification Generation with Knowledge Augmentation. We first build a data profile $\mathcal{P}_{data}$ (file-tree, formats, headers, modality cues) without full loading by a Profile-Oriented Data **Inspector**. An Applicability-Aware Fine-grained Metric **Selector** then activates feasible metrics and outputs an Evaluation Manifest $\mathcal{M}_{eval}$. Finally, a Knowledge-Augmented Evaluation Specification **Planner** instantiates each active element $e \in \mathcal{M}_{eval}$ into an executable specification $\pi(e) = \{\pi_{load}, \pi_{eval}\}$, where π_{load} defines evidence-aware loading (targeting/sampling) and π_{eval} defines deterministic scoring grounded by retrieved Type-I domain constraints and informed by Type-II dataset–paper graph signals.

(II) Adaptive and Scalable Tool-Centric Evaluation. Given $\pi(e)$, the executor operationalizes evaluation via a structured **Tool Library** $\mathcal{L}$ and persistent **Tool Memory** $\mathcal{M}_{tool}$. For each element, it applies Adaptive **Tool Orchestration** (*retrieve–invoke–synthesize*): retrieve compatible tools, generate invocation code to run them on the loaded evidence, or synthesize and register new tools under cold-start. A **Review-Driven Verification** layer (runtime + semantic checks) triggers self-correction and, when needed, feeds

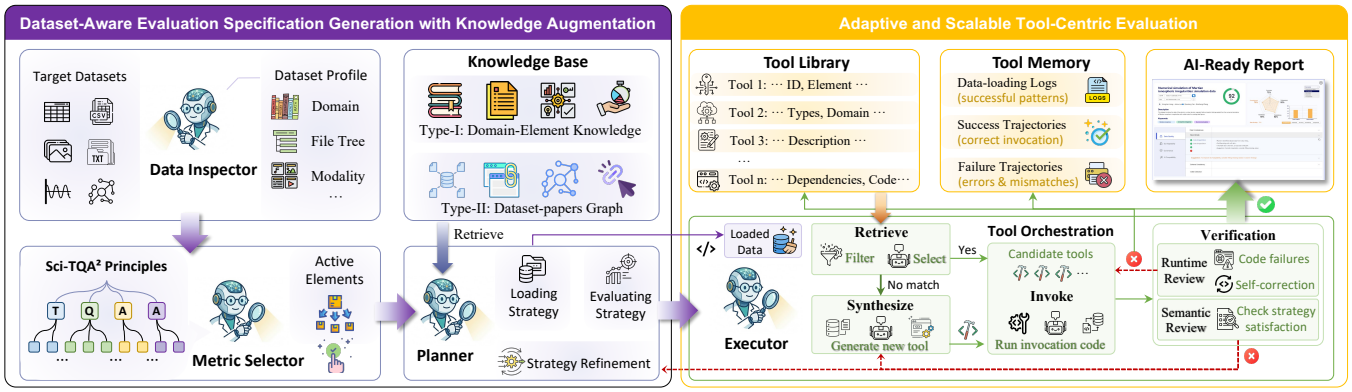

Figure 1: The architecture overview of our Sci-TQA²-Eval.

back to the planner for strategy-level refinement. Finally, we normalize tool outputs into a unified score format and aggregate them hierarchically (element → sub-dimension → dimension) to derive multi-dimensional quality scores.

4.2 Dataset-Aware Evaluation Specification Generation with Knowledge Augmentation

To accommodate heterogeneous datasets, we develop a domain-knowledge retrieval–augmented method for generating evaluation specifications. Given a target dataset, the method (i) constructs a dataset profile, (ii) selects applicable fine-grained evaluation metrics using an Applicability-Aware Metric Selector, and (iii) precisely defines *what evidence to load* and *how to score* each metric in a deterministic and scientifically grounded manner via a Knowledge-Augmented Evaluation Specification Planner.

4.2.1 Profile-Oriented Data Inspector. The Inspector functions as a front-end module that constructs a lightweight **Data Profile** $\mathcal{P}_{data}$ without fully loading the dataset. It scans the target path to extract the file-tree structure, file formats (e.g., `.csv`, `.pdb`, `.nc`), candidate primary structured files, and available metadata headers. Based on these observations, it infers the dataset modality (e.g., tabular, image, sequence, or multimodal) and summarizes salient characteristics, including storage organization, schema cues, and potential variable or attribute signatures. This metadata-first design improves scalability: by prioritizing header-level and structural probing, the Inspector remains practical for large scientific corpora where full loading is prohibitively expensive.

4.2.2 Applicability-Aware Fine-grained Metric Selector. Conditioned on $\mathcal{P}_{data}$, the Selector maps the observed dataset profile to a hierarchical taxonomy of evaluation metrics and activates only those fine-grained metrics that are both meaningful and feasible for the dataset. Leveraging retrieved Type-I knowledge, it determines the *applicability* of each metric and prunes irrelevant branches before triggering any expensive data loading or tool execution. For example, if the Inspector detects paired image–text artifacts, the Selector activates `ModalAlignment`; otherwise, that branch is disabled. The module outputs an Evaluation Manifest $\mathcal{M}_{eval}$, i.e., a dataset-conditioned set of active metrics to be evaluated. This early-stage filtering reduces unnecessary execution and ensures that subsequent planning is grounded in relevant evaluation objectives.

4.2.3 Knowledge-Augmented Evaluation Specification Planner. The Planner is the core component that converts the activated metric set into executable evaluation specifications. Conditioned on the data profile $\mathcal{P}_{data}$, the Evaluation Manifest $\mathcal{M}_{eval}$, and retrieved knowledge, it generates an element-level plan $\pi(e) = \{\pi_{load}, \pi_{eval}\}$ for each active element e, where π_{load} specifies *what evidence to load* and π_{eval} specifies *how to score* the metric deterministically. To support domain adaptation, we first construct a knowledge base and then describe the Planner's multi-stage specification procedure.

Knowledge Base Construction. Our Knowledge Base (KB) supports both domain-grounded metric instantiation and task-aware dataset reasoning via two components. **Type-I** is a semi-structured repository indexed by (`Domain`, `Element`) that records assessment-critical fields (e.g., standards, unit conventions, failure modes, and thresholds); the Planner retrieves it to specialize generic metrics into domain-specific criteria and to constrain π_{load} and π_{eval} for scientific validity and implementability. **Type-II** is a dataset–paper graph mined from scientific corpora and linked to open-science repositories (e.g., Hugging Face), encoding dataset–task–model–paper relations; given a dataset D, Sci-TQA²-Eval uses it to retrieve similar datasets and summarize downstream use cases, enabling *AI compatibility* and *scientific adaptability* analysis that informs task recommendations and diagnosis of low-scoring criteria.

Planning Procedure. For each active element $e \in \mathcal{M}_{eval}$, the Planner generates $\pi(e)$ in three stages as follows:

Stage 1: Evidence-Aware Loading Strategy. Scientific datasets are often too large for full loading and may require specialized parsers. The Planner therefore constructs π_{load} by balancing adequacy (sufficient evidence for assessment) and efficiency (minimal I/O). The strategy specifies (i) *targeting*—which artifacts to load (e.g., metadata only, selected variables/columns, or structured components such as coordinates) and (ii) *sampling*—how to select representative subsets (e.g., random sampling for distributional checks, time-slice sampling for temporal consistency, or stratified sampling

for rare-event coverage). The resulting specification is designed to be consumed by the execution for code generation and reuse.

Stage 2: Deterministic Evaluation Strategy. Given the evidence exposed by π_{load}, the Planner instantiates π_{eval} as a deterministic quantification recipe that produces a score in $[0, 1]$ together with interpretable quantities. Generic metric descriptions are grounded into dataset- and domain-specific criteria via Type-I retrieval. For example, in climate datasets, the generic check of "unit consistency" is instantiated as CMIP-style validation (e.g., verifying that temperature variables use Kelvin). This grounding yields evaluation strategies that are both computable and scientifically defensible.

Stage 3: Strategy-Level Refinement. Sci-TQA²-Eval maintains a semantic feedback loop between planning and execution. Beyond syntactic or runtime errors handled by tools, *semantic failures* (e.g., null outputs or scores outside $[0, 1]$) trigger *strategic reflection*. The Planner attributes the failure to either (a) π_{load} (e.g., incorrect file patterns, incompatible parsing, or insufficient sampling) or (b) π_{eval} (e.g., invalid assumptions or missing domain constraints), and revises $\pi(e)$ before requesting additional tool actions.

4.3 Adaptive and Scalable Tool-Centric Evaluation

Given $\pi(e)$, the execution subsystem operationalizes evaluation as a tool-centric pipeline that supports tool retrieval, adaptive invocation, and cold-start synthesis. It integrates a structured Tool Library $\mathcal{L}$, a persistent Tool Memory $\mathcal{M}_{tool}$, and a generation–verification loop. Rather than treating execution as a fixed runtime, Sci-TQA²-Eval casts it as a continual-learning process that expands executable coverage across diverse scientific datasets.

4.3.1 Structured Tool Library Construction. The Tool Library stores reusable evaluation tools as atomic functional units with rich metadata: (i) identity and scope (a unique tool ID and associated evaluation elements), (ii) operational constraints (accepted input types and domain applicability), (iii) semantic description (method summary and referenced standards), and (iv) encapsulated implementation (self-contained Python code with dependencies imported within the function). This design improves tool management and safe reuse while decoupling planning semantics from implementation details.

4.3.2 Adaptive and Extensible Tool Orchestration. To execute π_{eval}, the Executor uses a three-stage procedure: (i) *Retrieve*: apply hard filtering to keep tools compatible with the target element and the loaded artifacts, then perform LLM-based semantic selection over tool descriptions under the constraints of π_{eval}; (ii) *Invoke*: generate invocation code that binds the loaded data to the selected tool interface, with bounded retries to resolve parameterization errors; (iii) *Synthesize*: if no suitable tool is found (cold start), synthesize a new tool from π_{eval} and the loaded-data profile, verify it, and register it in $\mathcal{L}$. This orchestration favors reuse for efficiency while enabling synthesis to cover long-tail elements.

4.3.3 Tool Memory-Guided Execution. We maintain a persistent Tool Memory $\mathcal{M}_{tool}$ that records execution trajectories and is retrieved as contextual constraints during retrieval, invocation, and synthesis. It contains three streams: (i) *data-loading logs* (successful parsing and sampling patterns), (ii) *success trajectories* (correct invocation exemplars for specific tools and data types), and (iii) *failure trajectories* (runtime errors and semantic mismatches). Success trajectories provide few-shot guidance for parameterization, whereas failure trajectories act as negative constraints that reduce errors.

4.3.4 Review-Driven Verification and Self-Correction. The pipeline includes a two-layer review mechanism. *Runtime review* diagnoses code-level failures and guides self-correction (e.g., separating tool-internal faults from invocation errors). *Semantic review* checks whether outputs satisfy π_{eval}, including score-range constraints ($[0, 1]$), required output schema, and domain-consistent values. Semantic violations trigger tool regeneration or strategy-level reflection to the Planner, forming a closed-loop correction channel.

Finally, we synthesize element-level outputs into a cohesive scientific report. It normalizes tool outputs into a unified score representation and aggregates them hierarchically (element → sub-dimension → dimension) to form a multi-dimensional score vector. Beyond numeric scoring, Sci-TQA²-Eval leverages Type-II Dataset–Paper Pairs to infer latent potential: it traverses the relational graph to identify downstream AI tasks and experimental settings in which structurally similar datasets have been successfully used. Accordingly, the resulting **AI-Ready Report** not only diagnoses deficiencies and recommends remediation steps, but also suggests plausible AI-for-Science applications by explicitly linking low scores to constraints documented in prior successful studies (e.g., missing metadata fields required by a benchmark pipeline).

5 Experiments

We evaluated the performance of our SciHorizon-DataEVA on a diverse collection of heterogeneous scientific datasets. We first introduce the dataset preparation and implementation details. Subsequently, we present the main evaluation results, identifying key patterns in scientific data readiness, and provide a quantitative analysis of the system's effectiveness and tool generation capabilities.

5.1 Experimental Setup

5.1.1 Datasets Collection. To evaluate the generalizability of our framework across heterogeneous scientific contexts, we collected 80 datasets from authoritative open-science repositories, including ScienceDB, Zenodo, Figshare. These datasets span six distinct domains: *Astronomy*, *Biomedicine*, *Earth Science*, *Materials Chemistry*, *Physics & Engineering*, and *Socio-economics*. During collection, we prioritized diversity across both domain-specific governance standards and data modalities, ensuring comprehensive coverage of tabular and textual files, tensors and graphs, images, and several domain-specific modalities. This heterogeneity serves as a rigorous testbed for the system's adaptability in addressing diverse real-world challenges.

5.1.2 Implementation Details. Our multi-agent framework uses Gemini-3-pro as the backbone LLM. The execution environment for the generated evaluation tools is dynamic: standard evaluation scripts run in a local Python environment, while tools that require complex or conflicting dependencies are isolated in sandboxed Docker containers to ensure stability and security.

5.1.3 Metrics. In addition to the dataset-level evaluation based on Sci-TQA² principles in Section 3, we further design a set of metrics to systematically assess the reliability, scalability, and efficiency of

Sci-TQA²-Eval. Let $\mathcal{E}_d$ denote the set of atomic evaluation elements for dataset d, and $\mathcal{D}$ the set of all evaluated datasets.

- **System Correctness (SC).** SC measures the proportion of evaluation elements that yield a valid score within at most five debugging iterations:

$$\mathrm{SC} = \frac{\sum_{d\in\mathcal{D}} \sum_{e\in\mathcal{E}_d} \mathbb{I}(\mathrm{status}(e) = \mathrm{Success})}{\sum_{d\in\mathcal{D}} |\mathcal{E}_d|}, \tag{1}$$

  where $\mathbb{I}(\cdot)$ is the indicator function. A "Success" requires both successful execution of the underlying tool (retrieved or synthesized) and passage of semantic verification.
- **Tool Creation Success Rate (TCSR).** TCSR measures the fraction of synthesized tools that execute successfully and pass semantic verification:

$$\mathrm{TCSR} = \frac{|\mathcal{T}_{\mathrm{valid}}|}{|\mathcal{T}_{\mathrm{total}}|}, \tag{2}$$

  where $\mathcal{T}_{\mathrm{total}}$ is the set of all synthesis attempts triggered by retrieval failure, and $\mathcal{T}_{\mathrm{valid}}$ is the subset that completes the execution–verification loop within K iterations.
- **Tool Creation Efficiency (TCE).** For successfully synthesized tools, TCE measures the average number of debugging iterations required to pass verification:

$$\mathrm{TCE} = \frac{1}{|\mathcal{T}_{\mathrm{success}}|} \sum_{t\in\mathcal{T}_{\mathrm{success}}} \mathrm{Iter}(t). \tag{3}$$

  Lower TCE indicates higher synthesis precision and reduced computational overhead.

5.2 Main Results

Table 2 presents the quantitative evaluation results for the top-performing datasets across six scientific domains. The complete per-dataset results are reported in the extended arXiv version [21]. We observe that scores for Governance Trustworthiness (T) and Data Quality (Q) generally surpass those for AI Compatibility (A^c) and Scientific Adaptability (A^s). This trend suggests that while open-science datasets often meet traditional stewardship standards regarding formatting and compliance, they frequently lack the semantic richness or structural alignment required for advanced AI modeling and mechanism discovery. Notably, the symbol "-" indicates metrics automatically pruned by the *Applicability-Aware Selector* as irrelevant to specific modalities or tasks (e.g., class balance for unsupervised learning or spatial coverage for non-geometric data), validating the framework's capacity to conduct context-dependent evaluation rather than applying rigid, static rules.

To validate the semantic correctness of these quantitative assessments, we cross-referenced the generated diagnostic reports with established findings in domain literature. For example, the framework's high A^s-score for *Ionospheric_Obs* [48] aligns with its status as a benchmark for analyzing rare "superstorm" events in recent space physics studies [52]. Crucially, the agent's specific critique of "tensor architecture mismatches" accurately reflects the practical engineering hurdles documented in data fusion methodologies, where converting heterogeneous meshgrids to AI-ready tensors remains a manual prerequisite. Similarly, for *BigSolDB v2.0* [1], our framework accurately decoupled its exceptional Scientific Adaptability ($A^c_2 = 100.0$) from its latent data noise; the agent's specific diagnosis of "inter-laboratory consistency conflicts" precisely captures the intrinsic variance introduced by aggregating experimental results from nearly 1,600 diverse literature sources. Comprehensive evidence mapping specific agent diagnoses to literature corroborations is provided in Table S2 in the Appendix.

We further conducted a blinded human expert validation to assess evaluation correctness. Specifically, we sampled 15 datasets from three domains (Biomedicine, Socio-economics, and Earth Science; 5 datasets per domain) and invited 9 AI-for-Science experts, with 3 experts per domain, to evaluate sub-dimension-level outputs. For each evaluated fragment, experts rated (i) *Score Accuracy*, i.e., whether the assigned score faithfully reflects the dataset on the corresponding sub-dimension, and (ii) *Analysis Quality*, i.e., whether the accompanying diagnostic analysis is reasonable and well justified, using a 5-point Likert scale. We compared SciHorizon-DataEVA with two ablation variants, w/o Knowledge and w/o Memory, under a blinded setting. As shown in Table 3, SciHorizon-DataEVA achieved higher Score Accuracy and Analysis Quality than both w/o Knowledge and w/o Memory variants, with good inter-rater agreement (ICC=0.742), supporting the reliability of the expert judgments.

5.3 Performance Analysis of Sci-TQA²-Eval

To further demonstrate the effectiveness of the proposed Sci-TQA²-Eval, we conducted quantitative evaluations based on the evaluation metrics introduced in Section 5.1.3. Additionally, we performed an ablation study to validate the necessity of its architectural components. As illustrated in Figure 2, the full Sci-TQA²-Eval demonstrates robust overall performance, achieving a framework evaluation success rate of 89.0% and a tool creation success rate of 97.4%, while maintaining a high generation efficiency with an average of only 1.19 iterations per tool.

Additional robustness analyses, including backbone LLM comparison, baseline comparison with existing tools, and analyses of scalability and cost, are reported in Appendix A.1.

To verify the necessity of each architectural component, we conducted an ablation study by systematically degrading key modules. We compared SciHorizon-DataEVA against three variants:

- **w/o Knowledge-Augmented Planning**: We removed the knowledge retrieval module from the *Planner*, forcing it to generate evaluation strategies based solely on the LLM's parametric knowledge. Crucially, the *Reviewer* retained access to domain knowledge (Type-I) to verify the results.
- **w/o Tool Memory**: We disabled the retrieval of existing tools and historical "success trajectories," forcing the system to synthesize every tool *de novo* without few-shot guidance.
- **w/o Self-Correction**: We disabled the *Review-Driven Verification and Self-Correction*, restricting the tool generation to a single-pass attempt ($K = 1$) without debug iterations.

The ablation results highlight the distinct contributions of each module. First, the removal of the *Iterative Review* mechanism causes the most significant performance degradation, precipitating a sharp drop in framework evaluation success to 33.0%. This indicates that the self-correction loop is critical for resolving runtime errors and ensuring the executability of synthesized tools. Second, the *w/o*

Table 2: Quantitative Evaluation Results — All Domains. Background intensity correlates with score magnitude. T: Governance Trustworthiness (T_1: FAIRness, T_2: Provenance, T_3: Ethics); Q: Data Quality (Q_1: Complete., Q_2: Acc., Q_3: Unique., Q_4: Consist.); A^c: AI Compatibility (A^c_1: Adapt., A^c_2: Scale, A^c_3: Bal., A^c_4: Feat.Imp.); A^s: Scientific Adaptability (A^s_1: Task Gen., A^s_2: Scarcity, A^s_3: Causal, A^s_4: Cond. Cov.). *Modality*: Tensor, Graph, Table, Image, Text.

Dataset	Mod.	Total	Gov. Trust. (T)			Data Qual. (Q)				AI Comp. (A^c)				Sci. Adapt. (A^s)			
			T_1	T_2	T_3	Q_1	Q_2	Q_3	Q_4	A^c_1	A^c_2	A^c_3	A^c_4	A^s_1	A^s_2	A^s_3	A^s_4
Astronomy																	
Ring_Moat_Dome_Str.	Image	88.6	76.0	100.0	100.0	100.0	100.0	100.0	99.8	67.5	100.0	73.4	45.7	93.8	93.3	89.0	87.0
Ionospheric_Obs.	Tensor	88.1	76.0	100.0	100.0	86.8	82.4	74.3	100.0	82.8	78.8	-	-	100.0	100.0	87.6	87.4
Martian_Aeol._Land.	Image	86.2	76.0	100.0	100.0	75.0	100.0	100.0	100.0	93.7	100.0	61.3	-	87.5	86.0	45.3	78.0
SOLAR_WIND_MODEL	Table	83.7	80.0	100.0	100.0	98.0	90.9	99.9	100.0	87.5	99.3	82.0	15.4	100.0	97.0	93.7	24.2
Solar_Act._AI_Fore.	Image,Table	81.8	76.0	100.0	100.0	93.5	88.5	0.0	100.0	73.3	100.0	83.4	-	100.0	94.5	61.3	46.7
Biomedicine																	
Flysta3D	Image	91.4	85.7	100.0	79.2	100.0	100.0	100.0	100.0	99.6	71.2	-	100.0	98.1	100.0	53.3	-
The-Gonzo-Dataset	Tensor,Table	86.8	100.0	55.6	83.3	100.0	48.8	-	100.0	100.0	78.3	-	-	100.0	100.0	82.7	100.0
Synth.-Skeletal-Mot.	Image,Table	81.4	100.0	44.4	83.3	100.0	50.0	96.8	100.0	87.7	68.0	84.5	91.0	92.2	60.0	86.3	87.1
MedMNISTv2	Tensor	80.2	100.0	55.6	83.3	100.0	100.0	92.1	100.0	70.0	75.6	80.3	-	85.0	74.7	6.0	85.2
Genomic-benchmarks	Text	79.9	60.0	66.7	66.7	90.6	78.1	70.0	100.0	94.0	79.9	98.6	-	62.5	80.0	0.0	96.1
Earth Science																	
Fengyun-3	Image	90.2	76.0	100.0	100.0	93.5	88.0	100.0	100.0	73.5	91.5	-	-	99.8	100.0	63.3	100.0
Space_Hurr._Recog.	Image	86.0	76.0	100.0	100.0	100.0	88.3	45.8	62.9	90.0	93.1	100.0	-	81.5	93.3	90.0	90.4
Turb._Heat_Flux_2023	Tensor	85.7	76.0	100.0	100.0	78.1	85.2	100.0	100.0	94.0	80.0	-	100.0	50.4	95.0	60.0	10.0
LoveDA	Image	85.5	100.0	55.6	83.3	100.0	89.3	100.0	100.0	87.5	89.8	94.0	-	77.5	71.7	55.0	93.9
EM_Field_Structure	Tensor	85.2	76.0	100.0	100.0	100.0	-	-	100.0	99.9	51.0	-	-	70.8	80.0	84.0	58.5
Materials & Chemistry																	
FreeSolv	Table,Graph	88.2	85.7	100.0	79.2	100.0	100.0	100.0	99.8	100.0	56.2	76.2	96.8	87.5	97.5	100.0	90.3
BigSolDBv2.0	Table	88.0	100.0	55.6	83.3	99.3	91.3	100.0	100.0	89.2	100.0	80.1	77.6	68.8	100.0	95.0	87.5
xxMD	Graph	86.3	76.0	100.0	100.0	100.0	99.3	100.0	100.0	90.6	84.0	71.8	47.6	100.0	100.0	48.2	85.4
Raw_Hyperspec._Refl.	Table	85.0	96.0	100.0	95.8	99.8	99.4	100.0	100.0	81.2	16.0	87.2	73.7	71.2	100.0	66.7	81.9
1018_Compounds	Table	85.0	96.0	100.0	95.8	100.0	71.0	100.0	100.0	99.9	68.1	99.4	-	62.5	90.0	40.0	72.0
Physics & Engineering																	
Multiharm._Feedback	Table	93.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	83.9	-	82.4	66.7	85.0	85.3	100.0
CdS_Nanowire_SHG	Table	88.9	92.0	100.0	100.0	100.0	55.6	100.0	100.0	96.0	49.4	-	86.6	87.5	90.0	90.0	100.0
Dusty_Plasma_Ratchet	Table	86.8	96.0	100.0	100.0	100.0	-	100.0	92.9	86.2	70.3	55.5	94.5	93.8	25.0	82.9	76.7
Tokamak_Dust_Dynamics	Table	85.1	76.0	100.0	100.0	100.0	100.0	100.0	100.0	65.1	55.0	-	100.0	70.0	100.0	83.3	47.0
Coherent_Light_Int.	Table	85.0	76.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	65.0	-	-	81.2	70.0	76.5	55.0
Socio-economics																	
Dig._Econ._Green_Dev.	Table	95.3	100.0	100.0	100.0	100.0	-	100.0	100.0	82.5	93.1	-	85.3	100.0	100.0	96.7	81.0
Env._Reg._Green_Inn.	Table	91.0	100.0	100.0	100.0	100.0	-	100.0	100.0	100.0	87.5	-	66.0	100.0	56.8	95.3	57.6
China_Cross_Bord._Ecom.	Table	89.2	100.0	100.0	100.0	74.1	100.0	100.0	100.0	99.8	96.4	49.4	72.6	100.0	99.7	100.0	85.0
China_282_Green_Econ.	Table	88.6	76.0	100.0	100.0	100.0	-	100.0	100.0	75.0	88.6	71.3	88.1	100.0	77.6	75.0	87.3
Mgmt._Tone_Stock_Pr.	Table	87.5	80.0	88.9	100.0	100.0	50.0	100.0	100.0	89.1	73.4	-	87.4	100.0	95.0	75.8	86.7
W._China_Farmer_Surv.	Table	86.7	100.0	100.0	100.0	98.0	-	100.0	99.9	90.2	45.6	100.0	98.6	76.7	90.0	37.0	-

Table 3: Blinded human expert validation on 15 datasets.

System	Score Accuracy	Analysis Quality
SciHorizon-DataEVA	**4.15**	**4.11**
w/o Knowledge	3.52	3.26
w/o Memory	3.99	3.90

Knowledge variant exhibits a substantial decline in evaluation success (51.7%) and records the lowest tool creation efficiency (1.48 iterations). This suggests that without retrieved domain-knowledge constraints, the system struggles to generate valid evaluation logic efficiently, leading to higher trial-and-error costs. Finally, the *w/o Memory* setting results in a moderate performance decrease (Evaluation Success: 82.6%, Efficiency: 1.30). While the system remains functional, this confirms that the tool memory facilitates stable generation and reduces computational overhead by retrieving proven execution patterns.

5.4 Case Study

To illustrate the automated evaluation workflow of Sci-TQA²-Eval, we present a representative case study on DiTing [51], a large-scale dataset for seismic analysis. As depicted in Figure 3, the analysis focuses on the *Target Distribution Health* dimension, which determines the statistical suitability of ground truth labels for model convergence. The process begins with the **Planner** retrieving domain-specific knowledge, correctly identifying that seismic magnitudes inherently follow the exponential Gutenberg-Richter law [12]. Subsequently, the **Executor** invokes the statistical assessment tool to

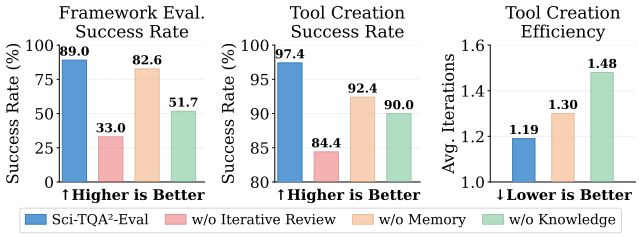

Figure 2: The performance of Sci-TQA²-Eval and its variants.

quantify distributional properties, calculating skewness penalties for regression targets and entropy for classification labels. The resulting score of 55.2/100 accurately captures the characteristics of the evaluated data subset: while the distribution is physically valid, the sampled partition exhibits a heavy-tailed skew and class imbalance that may destabilize standard loss functions. Finally, based on these diagnostics, the system synthesizes a report that maps the dataset to appropriate downstream tasks (e.g., Phase Picking) and recommends robust model architectures[29].

6 Limitations and Ethical Considerations

Our method was evaluated using open-source heterogeneous scientific datasets from six scientific domains. While these datasets span various fields, their scope remains somewhat limited. Future work could include datasets from additional domains to further validate and expand the method's applicability. Regarding ethics, all datasets used in this study are publicly available, and we have strictly followed the usage guidelines for each. These datasets were used solely for validating our proposed method, ensuring that no ethical concerns arise from their use. Another limitation is that the current AI Compatibility assessment is grounded in known model architectures; future work will extend the taxonomy as new AI paradigms emerge to avoid over-favoring any single modeling family.

7 Conclusion

In this work, we addressed the growing need for principled and scalable evaluation of AI-readiness in scientific data. We proposed SciHorizon-DataEVA, an agentic system that systematically assessed the AI-readiness of heterogeneous scientific datasets by jointly considering governance, quality, compatibility, and adaptability factors that are essential for effective learning and scientific reasoning. At the criteria level, we introduced the Sci-TQA² principles, which decomposed AI-readiness into four complementary dimensions—Governance Trustworthiness, Data Quality, AI Compatibility, and Scientific Adaptability—and further operationalized them through measurable atomic elements. To enable scalable evaluation in practice, we developed Sci-TQA²-Eval, a hierarchical multi-agent approach that dynamically constructed and executed dataset-aware evaluation specifications through a directed, cyclic workflow. By integrating lightweight profiling, applicability-aware metric selection, knowledge-augmented planning, and review-driven verification, the proposed system supported reliable and scalable assessment across diverse scientific data modalities. Extensive experiments on heterogeneous datasets spanning multiple scientific domains demonstrated the effectiveness and generality of our approach.

Case Study: Evaluation Trace for DiTing (Seismology)

Phase 1: Knowledge-Augmented Planning
[INPUT] Dataset: *DiTing (Subset: part_27.csv & part_27.hdf5, 3-4% of total)*
[Sub-Dimension] Class Balance
[RETRIEVAL] Domain: Seismology → Aspect: Target Distribution Health
[CONSTRAINT] Gutenberg-Richter Law: Earthquake magnitudes follow an exponential decay ($N \propto 10^{-bM}$). **Expectation**: High skewness in regression targets, dominated by small earthquakes and scarce large events, may hinder the accurate prediction of larger earthquakes if untreated.

Phase 2: Adaptive Tool Generation & Execution (Snapshot)
[TOOL] assess_target_distribution_health
[LOGIC] *Calculating statistical suitability for ML training:*

```
def assess_distribution(data, target_config):
    if col_type == 'continuous': # e.g., 'evmag'
        # Penalty for exponential distribution (G-R Law detection)
        ......
    elif col_type == 'categorical': # e.g., 'p_motion'
        # Normalized Shannon Entropy for Class Balance
        ......
    return raw_score * (1.0 - nan_ratio)
```

[RESULT] evmag skewness $\gamma > 2.5$ → Score penalized.
[AGGREGATED SCORE] **55.2 / 100 (Moderate)**

Phase 3: AI-Ready Report Generation
[DIAGNOSIS] **Target Distribution Health (55.2/100)**: The dataset exhibits two target-distribution issues: a heavy-tailed magnitude distribution dominated by small earthquakes, and imbalanced polarity labels.
[WARNING] Direct training with MSE loss will be dominated by small-magnitude events; Polarity classifiers may suffer from mode collapse.
[RECOMMENDED TASKS]
✓ **Seismic Phase Picking**: High suitability due to dense P/S wave labels.
✓ **Polarity Classification**: Feasible but requires re-balancing class weights.
[MODEL SUGGESTION] **EQTransformer** (Hierarchical Attention): Recommended to handle the multi-task nature (detection + picking) and mitigate noise in the subset.

Figure 3: An example of the evaluation trace and results of Sci-TQA²-Eval on DiTing dataset.

Acknowledgments

This work is partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDB1350102) and the National Natural Science Foundation of China (No.62506352, No.92470204). The development of this project was supported by the SciHorizon platform.

References

[1] Emad Al Ibrahim, Nathan Morgan, Simon Muller, Saikiran Motati, and William H Green. 2025. Accurately predicting solubility curves via a thermodynamic cycle, machine learning, and solvent ensembles. *Journal of the American Chemical Society* 147, 49 (2025), 45057–45069.

[2] Wesley Brewer, Patrick Widener, Valentine Anantharaj, Feiyi Wang, Tom Beck, Arjun Shankar, and Sarp Oral. 2025. Data Readiness for Scientific AI at Scale. In *Workshop Proceedings of the 54th International Conference on Parallel Processing*. 18–24.

[3] Wesley Brewer, Patrick Widener, Valentine Anantharaj, Feiyi Wang, Tom Beck, Arjun Shankar, and Sarp Oral. 2025. Data Readiness for Scientific AI at Scale. In *Workshop Proceedings of the 54th International Conference on Parallel Processing*. 18–24.

[4] Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. 2025. AI4Research: A Survey of Artificial Intelligence for Scientific Research. *arXiv preprint arXiv:2507.01903* (2025).

[5] Xi Chen, Chuan Qin, Jinpeng Li, Shasha Hu, Chao Wang, Hengshu Zhu, and Hui Xiong. 2026. GenDis: Generative-Discriminative Dual-View Co-Training for Generalized Category Discovery. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*.

[6] Xi Chen, Chuan Qin, Ziqi Wang, Shasha Hu, Chao Wang, Hengshu Zhu, and Hui Xiong. 2026. Beyond the Known: An Unknown-Aware Large Language Model for Open-Set Text Classification. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=BqLGlQF46f

[7] Jihye Choi, Nils Palumbo, Prasad Chalasani, Matthew M Engelhard, Somesh Jha, Anivarya Kumar, and David Page. 2024. MALADE: orchestration of LLM-powered agents with retrieval augmented generation for pharmacovigilance. *arXiv preprint arXiv:2408.01869* (2024).

[8] Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu, Fuzhen Zhuang, and Hui Xiong. 2023. RecruitPro: A Pretrained Language Model with Skill-Aware Prompt Learning for Intelligent Recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 3991–4002. doi:10.1145/3580305.3599894

[9] Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. 2021. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 364–373.

[10] Alban Gaignard, Thomas Rosnet, Frédéric De Lamotte, Vincent Lefort, and Marie-Dominique Devignes. 2023. FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *Journal of Biomedical Semantics* 14, 1 (2023), 7.

[11] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, et al. 2021. Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935* (2021).

[12] Beno Gutenberg and Carl F Richter. 1956. Earthquake magnitude, intensity, energy, and acceleration: (Second paper). *Bulletin of the seismological society of America* 46, 2 (1956), 105–145.

[13] Anders Haug. 2025. Intrinsic data quality dimensions: expanding on Wand and Wang's data quality model. *Industrial Management & Data Systems* 125, 1 (2025), 238–261.

[14] Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2025. Data readiness for AI: A 360-degree survey. *Comput. Surveys* 57, 9 (2025), 1–39.

[15] Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2024. Ai data readiness inspector (aidrin) for quantitative assessment of data readiness for ai. In *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*. 1–12.

[16] Kaveen Hiniduma, Dylan Ryan, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2025. AIDRIN 2.0: A Framework to Assess Data Readiness for AI. *arXiv preprint arXiv:2505.18213* (2025).

[17] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3561–3562.

[18] Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing Question Answering for Enterprise Knowledge Bases using Large Language Models. In *Database Systems for Advanced Applications: 29th International Conference, DASFAA 2024, Gifu, Japan, July 2–5, 2024, Proceedings, Part IV*. Springer-Verlag, Berlin, Heidelberg, 273–290. doi:10.1007/978-981-97-5562-2_18

[19] Kevin Jiang, Weixin Liang, James Y Zou, and Yongchan Kwon. 2023. Opendataval: a unified benchmark for data valuation. *Advances in Neural Information Processing Systems* 36 (2023), 28624–28647.

[20] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

[21] Dianyu Liu, Chuan Qin, Xi Chen, Xiaohan Li, Wenxi Xu, Yuyang Wang, Xin Chen, Yuanchun Zhou, and Hengshu Zhu. 2026. SciHorizon-DataEVA: An Agentic System for AI-Readiness Evaluation of Heterogeneous Scientific Data. arXiv:2604.26645

[22] Yunchao Lance Liu, Yu Wang, Oanh Vu, Rocco Moretti, Bobby Bodenheimer, Jens Meiler, and Tyler Derr. 2023. Interpretable chirality-aware graph neural network for quantitative structure activity relationship modeling in drug discovery. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 14356–14364.

[23] Qingqing Long, Haotian Chen, Chenyang Zhao, Xiaolei Du, Xuezhi Wang, Pengyao Wang, Chengzan Li, Yuanchun Zhou, and Hengshu Zhu. 2026. ScienceDB AI: An LLM-Driven Agentic Recommender System for Large-Scale Scientific Data Sharing Services. *arXiv preprint arXiv:2601.01118* (2026).

[24] Qingqing Long, Shuai Liu, Ning Cao, Zhicheng Ren, Xiao Luo, Wei Ju, Chen Fang, Zhihong Zhu, Hengshu Zhu, and Yuanchun Zhou. 2026. A survey of large language models for traffic forecasting: Methods and applications. *IEEE Transactions on Big Data* (2026).

[25] Qingqing Long, Yuchen Yan, Wentao Cui, Wei Ju, Zhihong Zhu, Yuanchun Zhou, Xuezhi Wang, and Meng Xiao. 2024. MOAT: Graph prompting for 3D molecular graphs. In *Proceedings of the 33rd ACM International conference on information and knowledge management*. 1586–1596.

[26] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460* (2025).

[27] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451* (2024).

[28] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature* 624, 7990 (2023), 80–85.

[29] S Mostafa Mousavi, William L Ellsworth, Weiqiang Zhu, Lindsay Y Chuang, and Gregory C Beroza. 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications* 11, 1 (2020), 3952.

[30] Bala Poduval, RL McPherron, R Walker, MD Himes, KM Pitman, AR Azari, C Shneider, AK Tiwari, S Kapali, G Bruno, et al. 2023. AI-ready data in space science and solar physics: problems, mitigation and action plan. *Frontiers in Astronomy and Space Sciences* 10 (2023), 1203598.

[31] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.

[32] Chuan Qin, Xi Chen, Jinpeng Li, and Hengshu Zhu. 2026. BOLT: Benchmarking Open-World Learning for Text Classification. In *Findings of the Association for Computational Linguistics: ACL 2026*.

[33] Chuan Qin, Xin Chen, Chengrui Wang, Pengmin Wu, Xi Chen, Yihang Cheng, Jingyi Zhao, Meng Xiao, Xiangchao Dong, Qingqing Long, et al. 2025. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5754–5765.

[34] Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Fuzhen Zhuang, and Hengshu Zhu. 2025. COTR: Efficient Job Task Recognition for Occupational Information Systems with Class-Incremental Learning. *ACM Transactions on Management Information Systems* 16, 2 (2025), 1–30. doi:10.1145/3712306

[35] Chuan Qin, Le Zhang, Yihang Cheng, Rui Zha, Dazhong Shen, Qi Zhang, Xi Chen, Ying Sun, Chen Zhu, Hengshu Zhu, and Hui Xiong. 2025. A Comprehensive Survey of Artificial Intelligence Techniques for Talent Analytics. *Proc. IEEE* 113, 2 (2025), 125–171. doi:10.1109/JPROC.2025.3572744

[36] Philippe Rocca-Serra, Wei Gu, Vassilios Ioannidis, Tooba Abbassi-Daloii, Salvador Capella-Gutierrez, Ishwar Chandramouliswaran, Andrea Splendiani, Tony Burdett, Robert T Giessmann, David Henderson, et al. 2023. The FAIR Cookbook-the essential resource for and by FAIR doers. *Scientific data* 10, 1 (2023), 292.

[37] Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. 2024. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ digital medicine* 7, 1 (2024), 203.

[38] Science Data Bank. 2026. Science Data Bank (ScienceDB): Data Browsing Page. https://www.scidb.cn/en/list. Accessed January 2026.

[39] Tingjia Shen, Hao Wang, Chuan Qin, Ruijun Sun, Yang Song, Defu Lian, Hengshu Zhu, and Enhong Chen. 2026. Prompting is not Enough: Exploring Knowledge Integration and Controllable Generation on Large Language Models. *Big Data Mining and Analytics* 9, 2 (2026), 563–579. doi:10.26599/BDMA.2025.9020052

[40] Shrey Shrivastava, Dhaval Patel, Nianjun Zhou, Arun Iyengar, and Anuradha Bhamidipaty. 2020. Dqlearn: A toolkit for structured data quality learning. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1644–1653.

[41] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* 646, 8085 (2025), 716–723.

[42] Zhenyu Tong, Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Jingshuai Zhang, Chen Zhu, and Hengshu Zhu. 2025. From Missteps to Mastery: Enhancing Low-Resource Dense Retrieval through Adaptive Query Generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*. ACM, 1373–1384. doi:10.1145/3690624.3709225

[43] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[44] Jiaqi Wei et al. 2025. From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. *arXiv preprint arXiv:2508.14111* (2025).

[45] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[46] Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific data* 5, 1 (2018), 1–4.

[47] Wenxi Xu, Chuan Qin, Xi Chen, Chuyu Fang, Yuanchun Zhou, and Hengshu Zhu. 2026. TLSA: LLM-Guided Text-Label Space Alignment with Contrastive Learning for Generalized Category Discovery. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*.

[48] Yuyan Yang, Wenbo Li, Wei Yuan, Wenjie Sun, and Xiukuan Zhao. 2025. Ionospheric Observations Over the Asia-Australia Sector During the May 2024 Superstorm Recovery Phase. doi:10.57760/sciencedb.space.02830

[49] Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. 2024. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135* (2024).
[50] Haodong Zhang, Xinyue Wang, Tao Ren, Yifan Wang, Siyu Yi, Fanchun Meng, Zeyu Ma, Qingqing Long, and Wei Ju. 2026. FairGC: Fostering Individual and Group Fairness for Deep Graph Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 28194–28202.
[51] Ming Zhao, Zhuowei Xiao, Shi Chen, and Lihua Fang. 2023. DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. *Earthquake Science* 36, 2 (2023), 84–94.
[52] Qiang Zhao and Le Yu. 2025. Advancing sustainable development goals through earth observation satellite data: Current insights and future directions. *Journal of Remote Sensing* 5 (2025), 0403.
[53] Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259* (2025).
[54] Qi Zhou, Xi Chen, Chuyu Fang, Jianji Wang, Chuan Qin, and Fuzhen Zhuang. 2025. Enhancing Dual-Target Cross-Domain Recommendation via Similar User Bridging. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, 4487–4497. doi:10.1145/3746252.3761356
[55] Chen Zhu, Xiao Hu, Han Wu, Chuan Qin, Hengshu Zhu, and Hui Xiong. 2025. Enhancing Job Recommendations with LLM-based Resume Completion: A Behavior-Denoised Alignment Approach. *Information Processing & Management* 62, 6 (2025), 104261. doi:10.1016/j.ipm.2025.104261
[56] Zhihong Zhu, Fan Zhang, Yunyan Zhang, Jinghan Sun, Zhiqi Huang, Qingqing Long, Bowen Xing, and Xian Wu. 2025. A Survey on Multi-modal Intent Recognition: Recent Advances and New Frontiers. *Findings of the Association for Computational Linguistics: EMNLP 2025* (2025), 15223–15236.
[57] Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*. 6748–6769.

A Appendix

A.1 Additional Robustness Analyses

Backbone LLM comparison. We evaluated Sci-TQA²-Eval with alternative backbone LLMs on 18 datasets across all six domains. Table S1 shows that most strong backbones achieve comparable SC, while the large drop without self-correction confirms that the system design is more critical than the specific backbone.

Table S1: Backbone LLM comparison on 18 datasets.

Backbone	SC (%)
Gemini-3-Pro	89.0
Claude Sonnet 4.6	88.9
GLM-5	82.5
GPT-5.1	79.7
DeepSeek-v3.2	73.2
Gemini-3-Pro w/o Self-Corr.	33.0

Baseline comparison. We compared SciHorizon-DataEVA with two open-source tools, FAIR-Checker and AIDRIN, on the same 15-dataset subset used in the expert validation. FAIR-Checker primarily evaluates metadata encoding compliance, while AIDRIN focuses on tabular data-quality profiling. Existing tools cover only partial dimensions of Sci-TQA²: at least 80% of our fine-grained sub-dimensions cannot be assessed by either tool, and AIDRIN failed on 8/15 datasets due to format limitations. FAIR-Checker mainly checks RDF-level metadata, while AIDRIN provides generic statistical snapshots; in contrast, SciHorizon-DataEVA performs context-aware assessment using LLM reasoning and retrieved domain knowledge.

Scalability and cost. Across approximately 80 datasets, the average wall-clock time is 22m31s and the average API cost is $7.22 per evaluation. Cost scales mainly with the number of active evaluation elements rather than raw dataset size, since the Inspector avoids full data loading. The Tool Library also amortizes cold-start synthesis: average evaluation time decreases from 29m49s for the first 20 datasets to 16m35s for the last 20.

A.2 Evidence Collection and Traceability

To ensure transparency and auditability, we provide an evidence table linking representative scores to justifications in source publications. Due to space constraints, representative evidence is shown in Table S2, while additional evidence is provided in [21]. Each entry includes: (**Dataset**, **Dimension**, **Score**, **Article**, **Evidence**). The **Evidence** field contains minimal quoted excerpts used by SciHorizon-DataEVA to justify the score, enabling manual verification and future re-annotation as scoring rubrics evolve. Note that this table is not an exhaustive execution log; rather, it provides human-readable traces for selected (dataset, metric) pairs that are representative or potentially contentious.

A.3 GenAI Disclosure

In this work, we developed an agentic system based on LLMs to assess the AI-readiness of scientific data. The operation of this system relies on LLMs as core components for reasoning, analysis, and interaction. Additionally, during the preparation of this manuscript, we used generative artificial intelligence–based tools to support language-related tasks, such as grammar checking and improving writing clarity. These tools were not used to generate scientific content, interpret results, or draw conclusions. All scientific ideas, analyses, and interpretations presented in this paper are solely the responsibility of the authors.

Table S2: Evidence table of datasets, dimensions, scores, and supporting quotations.

Dataset	Dimension	Score	Article	Evidence
BigSolDBv2	Class Balance	80.1	BigSolDB 2.0, dataset of solubility values for organic compounds in different solvents at various temperatures, *Scientific Data*, 2025	"In this study, we present a dataset containing 103944 experimental solubility values within a temperature range from 243 to 425 K for 1448 organic compounds measured in 213 individual solvents extracted from 1595 peer-reviewed articles."
	AI Model Adaptability	89.2		"Moreover, this dataset paves the way to train machine learning models for solubility prediction in a wide variety of organic solvents which can be extremely useful in the design of many chemical and technological processes."
FreeSolv	Completeness	100.0	FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des. 2014	"The starting point in constructing the FreeSolv database was to pull together all of the lead author's previous work calculating hydration free energies in explicit solvent. This included calculated values, experimental values, and structures and input files from several previous studies."
	Accuracy	100.0		"we had retained not only calculated and experimental hydration free energies and original coordinate files (.mol2 format) containing geometries and partial charges, but also input files in the form of GROMACS topology and coordinate files."
MedMNISTv2	Completeness	100.0	MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification, *Scientific Data*, 2023	"we provide an official train-validation-test split for each subset... to avoid data leakage"
	Accuracy	100.0		"Three neuroscience experts segment a pyramidal neuron within the whole volume and proofread all the synapses"
	AI Model Adaptability	70.0		"For MedMNIST2D, we first implement ResNets10…as baseline methods"
	Task Generalizability	85.0		"MedMNIST has been particularly useful for machine learning and computer vision research…"
	Consistency	80.3		"All images are pre-processed into a MNIST-like format…"