

Article

Skill-Based Autonomous Agents for Material Creep Database Construction

Yue Wu ^{1,†}, Tianhao Su ^{1,†}, Shunbo Hu ^{1,2,3,4,*} and Deng Pan ^{1,*}¹ Materials Genome Institute, Shanghai University, Shanghai 200444, China; repeatsongyue@gmail.com (Y.W.); thsu0407@gmail.com (T.S.)² Institute for the Conservation of Cultural Heritage, Shanghai University, Shanghai 200444, China³ Institute for Quantum Science and Technology, Shanghai University, Shanghai 200444, China⁴ Key Laboratory of Silicate Cultural Relics Conservation (Ministry of Education), Shanghai University, Shanghai 200444, China

* Correspondence: shunbohu@shu.edu.cn (S.H.); dpan_mgi@shu.edu.cn (D.P.)

† These authors contributed equally to this work.

Abstract

The advancement of data-driven materials science is currently constrained by a fundamental bottleneck: the vast majority of historical experimental data remains locked within the unstructured text and rasterized figures of the legacy scientific literature. Manual curation of this knowledge is prohibitively labor-intensive and prone to human error. To address this challenge, we introduce an autonomous, agent-based framework powered by Large Language Models (LLMs) designed to excavate high-fidelity datasets from scientific PDFs without human intervention. By deploying a modular “skill-based” architecture, the agent orchestrates complex cognitive tasks, including semantic filtering, multi-modal information extraction, and physics-informed validation. We demonstrate the efficacy of this framework by constructing a physically self-consistent database for material creep mechanics, a domain characterized by complex graphical trajectories and heterogeneous constitutive models. Applying the pipeline to 243 publications, the agent achieved a verified extraction success rate exceeding 90% for graphical data digitization. Crucially, we introduce a cross-modal verification protocol, demonstrating that the agent can autonomously align visually extracted data points with textually extracted constitutive parameters ($R^2 > 0.99$), ensuring the physical self-consistency of the database. This work not only provides a critical resource for investigating time-dependent deformation across diverse material systems but also establishes a scalable paradigm for autonomous knowledge acquisition, paving the way for the next generation of self-driving laboratories.

Keywords: creep mechanics; large language models; autonomous agents; literature mining; multi-modal data extraction; constitutive models



Academic Editor: Giancarlo Fortino

Received: 24 June 2026

Revised: 8 July 2026

Accepted: 19 July 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The landscape of scientific research is undergoing a profound transformation, driven by the rapidly advancing capabilities of Artificial Intelligence (AI). The concept of “AI for Science” (AI4S) has shifted from a theoretical possibility to a practical paradigm, fundamentally altering how hypotheses are generated and validated [1,2]. Central to this revolution are Large Language Models (LLMs), which have demonstrated remarkable proficiency not only in natural language processing but also in logical reasoning and code generation [3,4].

While early models focused primarily on text generation, the recent integration of Chain-of-Thought (CoT) reasoning [5] and the ReAct paradigm [6] has evolved LLMs from passive generators to autonomous Agents. These Agents, now equipped with modular “skills” and tool-use capabilities, can interact with external environments [7,8]. This evolution promises to automate complex cognitive tasks that were previously the exclusive domain of human experts, potentially accelerating scientific discovery at an unprecedented pace [9–11].

Despite these technological strides, the application of AI in many specialized scientific domains faces a persistent impediment: the scarcity of high-quality data [12]. While general-purpose models benefit from internet-scale datasets, domain-specific tasks, such as those in materials science, mechanics, and chemistry, rely on data that is often fragmented, unstructured, and hidden within the text and figures of the academic literature [13]. Constructing high-fidelity datasets in these fields is notoriously labor-intensive and prohibitively expensive. Previous efforts like ChemDataExtractor [14] and MatScholar [15] made significant progress using rule-based or LSTM-based approaches, but they often struggle with complex semantic reasoning and cross-modal alignment. Manual curation remains the norm, yet it is prohibitively expensive and prone to human error [16]. This “data bottleneck” serves as the primary obstacle preventing the widespread adoption of data-driven machine learning (ML) techniques in these critical fields [17].

However, the rapid maturation of Agent frameworks and Skill-based architectures offers a novel solution to this dilemma [18]. Notably, the acquisition of the relevant literature in specialized fields is far from a trivial retrieval task; it entails significant complexity, requiring deep domain knowledge to discern valid experimental contexts from tangential mentions and evaluate the methodological relevance of each source. Unlike traditional Natural Language Processing (NLP) or Optical Character Recognition (OCR) pipelines, which lack the capacity for cross-modal reasoning to resolve the semantic coupling between graphical data points and textual physical parameters (e.g., temperature T , stress σ) [19–24], modern Agents can emulate expert-level cognition. We posit that the current state of these frameworks has reached a level of sophistication sufficient to autonomously navigate the vast ocean of scientific literature. By defining specific skills, we empower agents to achieve end-to-end autonomy, operating entirely as human-out-of-the-loop systems, to automatically collect relevant papers, filter for quality with professional discernment, extract distinct data points, and structure this information into queryable databases. To validate this hypothesis, we focus on the domain of material creep mechanics. We selected this field not merely as a testbed for our framework but because it represents a quintessential class of high-dimensional, non-linear, and scarce scientific data [25]. This unique characteristic renders it an ideal touchstone for validating the physics-informed reasoning capabilities of autonomous agents, rather than merely serving as a stochastic test case. While authoritative repositories such as the NIMS Creep Data Sheets [26] provide high-fidelity benchmarks, they fundamentally rely on meticulous manual curation and standardized testing protocols. This dependence renders them inherently static and limited in scope, unable to keep pace with the vast, heterogeneous experimental results continuously emerging in the academic literature. Consequently, the absence of a dynamic, large-scale database containing both creep curves and their explicit constitutive expressions severely limits research progress. On the one hand, from an AI perspective, this scarcity hinders the advancement of symbolic regression tasks [27,28] and physics-informed neural networks (PINNs) [29,30]; without a rich repository of known functional forms and experimental data, algorithms struggle to discover or refine physical laws governing material deformation. On the other hand, for experimentalists, the absence of a centralized library makes benchmarking laborious, prohibiting the efficient comparison of new experimental results against historical data.

Establishing a database in this domain would thus facilitate data-driven discovery of constitutive models and accelerate the assessment of high-temperature material performance.

Here, we demonstrate a novel workflow for the automated construction of domain-specific databases using an LLM-based Agent and Skills framework. Employing creep mechanics as a representative case study, we design a set of specialized skills that enable the agent to identify the relevant literature and extract creep curves alongside their explicit mathematical expressions. Our work illustrates how the synergy between LLMs and skill-based architectures can break the data bottleneck, providing a scalable pathway for building the foundational datasets necessary for the next generation of AI4Science.

The remainder of this paper is structured as follows: Section 2 details the methodology, including the skill-based agent architecture, custom multi-modal extraction pipeline, and evaluation metrics. Section 3 presents the quantitative results, database composition, and cross-modal physical validation. Finally, Section 4 concludes with a discussion of the framework's limitations and future perspectives.

2. Methodology

The overall architecture of the automated literature mining pipeline is illustrated in Figure 1. The system is designed to transform unstructured scientific PDFs into a structured, queryable database of creep behaviors and constitutive models. The workflow leverages a Foundation Large Language Model (LLM) as the central reasoning agent, orchestrating a five-stage process: (1) Literature Collection, (2) Automated Screening, (3) Multi-modal Information Extraction, (4) Formula Validation, and (5) Structured Storage. This modular design ensures that both textual metadata and graphical experimental data are processed with high fidelity and physical consistency. Specifically, the core cognitive engine of this framework is instantiated using the Qwen3-235B-A22B foundation model (Alibaba Cloud, Hangzhou, China) [31], which provides the advanced logical reasoning capabilities necessary to drive these complex autonomous workflows. It is worth noting that while the framework requires a model with reliable visual multimodal capabilities for accurate figure classification, its heavily constrained, skill-based architecture reduces the absolute dependency on ultra-large-scale parameters. The textual screening and orchestration tasks can be effectively handled by the lightweight or economical variants of mainstream multimodal models without significantly affecting the overall extraction performance. This flexibility allows researchers to easily adapt the pipeline based on their computational budgets.

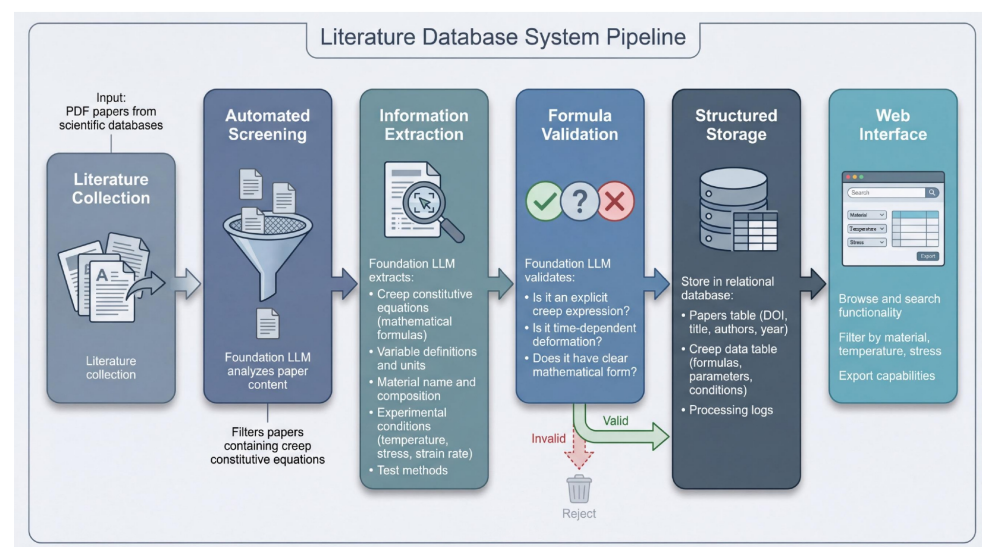


Figure 1. Architecture of the automated literature mining pipeline. The workflow proceeds in five distinct stages: (1) **Literature Collection**: Aggregation of source PDFs from scientific repositories.

(2) **Automated Screening:** An LLM-based filter analyzes document content to identify relevant publications containing creep constitutive models. (3) **Information Extraction:** A multi-modal extraction engine parses both unstructured text (for material metadata and formulas) and graphical plots (for experimental creep trajectories). (4) **Formula Validation:** A logic-based validation module checks the mathematical integrity and physical plausibility of the extracted expressions. (5) **Structured Storage & Interface:** Validated data is serialized into a relational database, accessible via a user-friendly web interface for searching, filtering, and exporting.

Google Gemini 2.5 Flash Image (also known as Nano Banana) was used only to generate illustrative graphic elements for the workflow diagram in Figure 1; it was not used to generate or modify any scientific data, results, or analytical conclusions.

2.1. Literature Collection and Automated Screening

The pipeline begins with the aggregation of a broad corpus of scientific literature. Potential candidate documents are sourced from major scientific repositories using a comprehensive set of keywords related to creep deformation and constitutive modeling.

Given the noise inherent in keyword-based searches, a rigorous **Automated Screening** module is employed to filter the corpus. Unlike traditional rule-based filters, this module utilizes the semantic understanding capabilities of the Foundation LLM. The agent analyzes the full text of each document to determine relevance based on specific criteria: the presence of experimental creep data and the discussion of constitutive equations. Papers that are purely theoretical (without experimental validation) or unrelated to time-dependent deformation are automatically rejected, ensuring that downstream processing is focused only on high-value content.

2.2. Multi-Modal Information Extraction

The core of the system is the extraction engine, which handles the heterogeneous nature of scientific documents by processing two distinct modalities:

- **Semantic extraction (Text & Formulas):** The agent scans the document to identify and extract critical metadata, including material designations (e.g., chemical composition, standard names), experimental conditions (temperature T , stress σ), and explicit mathematical expressions of creep constitutive laws. Special attention is paid to variable definitions and unit standardization to ensure consistency across the database.
- **Visual extraction (Graphical Data):** Unlike indiscriminate digitization, the agent is designed to prioritize representative experimental trajectories that can be strictly mapped to the constitutive models explicitly discussed in the text. Complementing the textual analysis, the agent identifies and processes experimental plots (e.g., Strain vs. Time curves) while filtering out supplementary or ambiguous data series. It detects coordinate systems, identifies axis scales, and digitizes the data points from relevant trajectories. This capability allows the system to capture raw experimental data that is often absent from the text but crucial for model validation.

The explicit semantic criteria for plot selection and the deterministic geometric algorithms used to discard artificial trajectories (e.g., grid lines and legend boxes) are comprehensively detailed in Section MultiModal Parser (Figure Classification) Prompt and Section S2 of the Supplementary Information.

2.3. Physics-Informed Validation Protocol

To mitigate the risk of “hallucinations” common in generative AI models, a Formula Validation module is integrated into the workflow as a physics-informed logic gate. This module rigorously evaluates each extracted entry against a triad of criteria to ensure scientific rigor. Specifically, the agent verifies whether the extracted content constitutes a

complete, explicit mathematical equation rather than a mere descriptive fragment; checks for physical relevance to ensure the formula describes time-dependent deformation consistent with established creep mechanics; and confirms mathematical integrity by ensuring all variables are clearly defined and unit-consistent. Only entries that successfully pass this comprehensive validation check are marked as valid and proceed to the database, while non-compliant entries are automatically flagged for rejection or manual review.

2.4. Database Construction and Web Interface

Verified data is serialized into a relational database structure comprising two primary entities: a Papers Table for bibliographic metadata (DOI, authors, year) and a Creep Data Table for the extracted physics parameters and experimental conditions. Crucially, the preservation of the Digital Object Identifier (DOI) is implemented to guarantee data provenance, ensuring that every extracted data point is strictly traceable to its original source publication. To facilitate accessibility, a user-friendly Web Interface was developed on top of the database. As depicted in Figure 2, the interface provides functionality for browsing and searching. Users can filter the database by material type, temperature range, or stress levels, and export the structured datasets for further analysis. This end-to-end implementation ensures that the extracted knowledge is not only stored but readily available for data-driven AI4S discovery.

2.5. Skill-Based Cognitive Architecture and Constraint Mechanisms

To execute the aforementioned workflow with high autonomy and reliability, we engineered a modular “Skill-Based” cognitive architecture. This framework departs from monolithic, conversational agent designs by decomposing complex reasoning tasks into specialized, executable units defined as Skills. To ensure full reproducibility and system robustness, we formally model an Agent Skill (S) as a tightly constrained cognitive tuple $S = \{I, T, C\}$:

- I (Instruction Injection): Represents domain-specific, schema-guided prompts that “prime” the LLM with expert knowledge. Rather than relying on unconstrained generation or extensively hand-tuned few-shot examples, our system employs zero-shot instructions strictly bounded by physical domain logic. To avoid overfitting and ensure broad generalizability, all prompt templates were iteratively refined on a distinct validation subset of documents before being rigidly frozen for the final pipeline execution.
- T (Scoped Tools): A restricted set of deterministic executable functions (e.g., PyMuPDF parsers, custom OpenCV vision algorithms) available exclusively to that specific skill. This architectural boundary ensures that the Foundation LLM/VLM solely proposes structured semantic labels, whereas all actual numerical extraction, curve filtering, and physical validations are executed by deterministic code.
- C (Hard Constraints): Mandatory output schemas that enforce machine-actionable results. To eliminate stochastic hallucinations, the temperature parameter for all OpenAI-compatible LLM and VLM API calls is strictly set to 0. The agent is forced to respond in predefined JSON structures; any output containing markdown fences or explanatory text that fails to parse against the JSON schema is programmatically stripped or rejected.

As illustrated in Table 1, this architecture allows the agent to dynamically instantiate specialized personas—ranging from a “Bibliographer” to a “Physicist”—at different stages of the pipeline, enforcing strict logic gates to improve data fidelity. All exact prompt templates and machine-readable schemas are comprehensively documented in the Supplementary Information (SI) and our open-source repository.

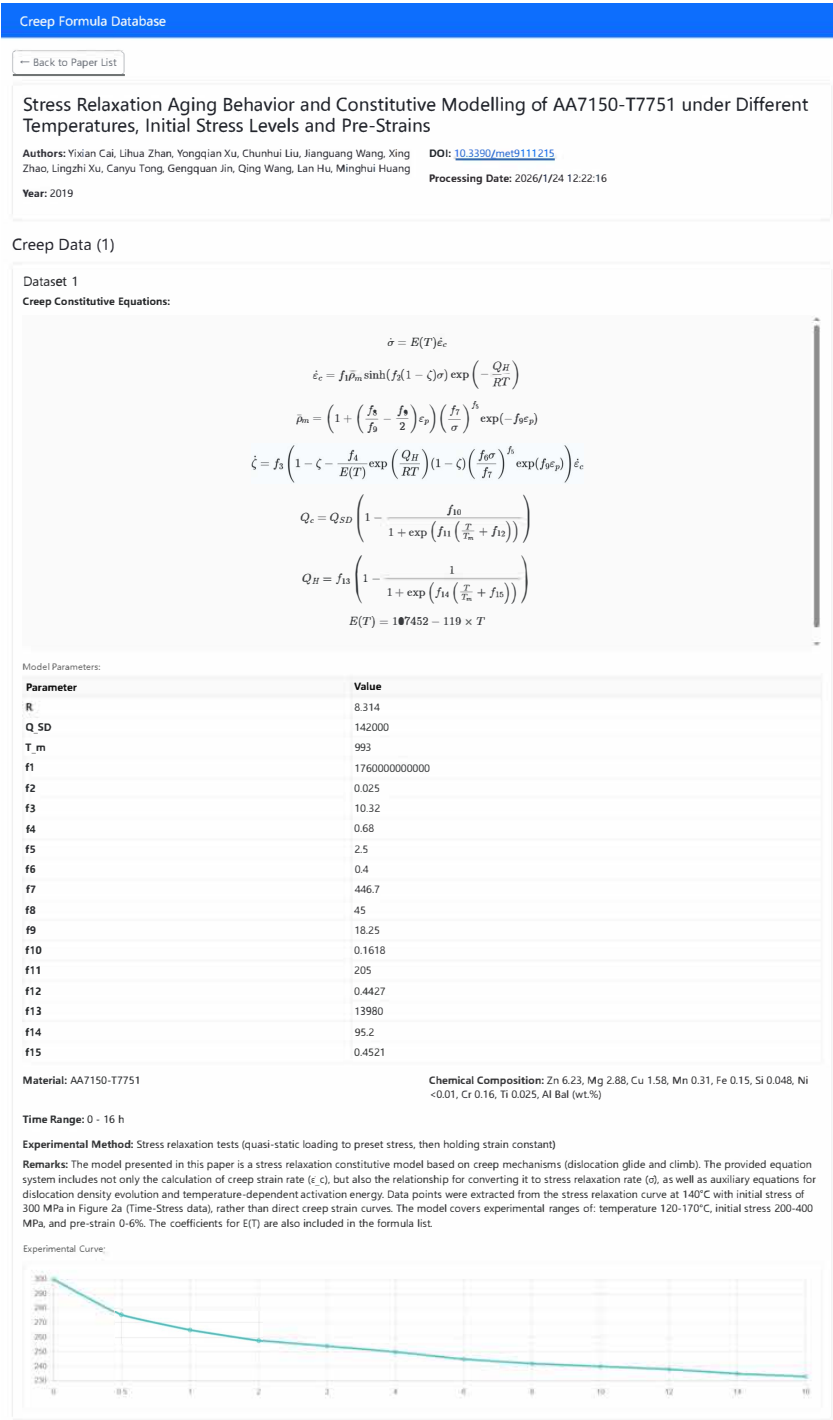


Table 1. Mapping of pipeline stages to agent skills and constraint logic.

Stage	Agent Skill	Scoped Tools (T)	Constraint Mechanism (C)
Collection	Bibliographic Navigator	RAG-based Retrieval	Semantic Disambiguation: Transforms ambiguous NL descriptors into precise, syntax-compliant Boolean logic; expands implied keywords (e.g., “superalloy” → “Ni-based OR Co-based”) to maximize recall.
Screening	Domain Filter	Zero-shot Classification	Logical Gating: Implements a strict conjunctive filter ($Pass \iff Has_Data \wedge Has_Equation$); automatically rejects partial matches (e.g., pure simulation studies) regardless of keyword relevance.
Extraction	MultiModal Parser	Vision-Language Alignment	Ontological & Geometric Alignment: Enforces a unified variable schema to resolve nomenclatural heterogeneity, while imposing strict topological constraints (e.g., monotonicity) to guarantee the structural integrity of digitized trajectories.
Validation	Physics Guardrail	Symbolic Reasoning	Physics-Informed Integrity: Mandates adherence to dimensional homogeneity laws and enforces a rigorous statistical threshold ($R^2 > 0.9$) for cross-modal alignment between textual parameters and visual data.
Storage	Data Serializer	Structured Serialization	Schema Compliance & Provenance: Enforces strict data typing to ensure machine-readability, while embedding immutable metadata (e.g., source DOI, visual evidence links) to establish a verifiable audit trail for every entry.

2.6. Scoped Tools and Custom Digitization Pipeline

To address the limitations of generic Optical Character Recognition (OCR) or manual digitization utilities, our framework equips the agent with a highly specialized, custom-developed toolchain. Within our skill-based architecture, these traditional tools serve as the deterministic “executors” that realize the high-level semantic intentions generated by the Foundation LLM. The foundational document parsing is handled via mature libraries (e.g., PyMuPDF version 1.26.1; Artifex Software Inc., San Francisco, CA, USA), while complex graphical vectorization relies on a bespoke computer vision (CV) pipeline built upon OpenCV version 4.12.0 (Open Source Vision Foundation, Palo Alto, CA, USA).

Rather than relying on the LLM to directly process raw visual pixels—which is computationally inefficient and prone to hallucination—the digitization process offloads pixel-level extraction to advanced deterministic CV algorithms. Coordinate axes are isolated using Canny edge detection and Probabilistic Hough Line Transforms, effectively filtering out non-data panels. For multi-series entanglement, a custom color-series digitizer separates distinct trajectories utilizing HSV color space masking and K-Means clustering. To mitigate interference from legend boxes or grid lines, we implemented a continuous-segment filtering algorithm. Furthermore, for scatter plots, connected-component analysis combined with RANSAC-like filtering isolates valid experimental markers from background noise.

For textual extraction, the pipeline is reinforced by a deterministic Regular Expression (Regex) fallback mechanism. This ensures that while the LLM conducts the primary semantic structuring of formulas, the system can rely on rigorous conjunctive matches between explicit mathematical markers and domain-specific terminology (e.g., Arrhenius, Norton) to guarantee the capture of complex constitutive parameters.

The mathematical basis for this physical data reconstruction relies on a multi-point affine transformation matrix. The custom OpenCV digitizer calculates the rigorous pixel-to-physical coordinate mapping by deriving transformation parameters directly from the coordinate axis ticks identified by the vision module, robustly accommodating both linear and logarithmic scales.

2.7. Agent-Tool Integration Mechanism

The core innovation of this framework lies in how the Foundation LLM orchestrates these mature programmatic tools to bridge the gap between unstructured scientific literature and queryable machine-actionable databases. We resolve the inherent tension between the stochastic nature of language generation and the rigorous precision required for scientific data by establishing a strict architectural boundary: the Foundation LLM acts exclusively as the cognitive orchestrator—proposing structured semantic labels and conceptual linkages—whereas all numerical extraction, spatial filtering, and validation metrics are executed by deterministic Python (version 3.10; Python Software Foundation, Beaverton, OR, USA) modules.

This Agent-Tool integration operates through a tightly constrained, five-stage workflow:

1. **Document Parsing:** The backend parses the academic PDF, systematically extracting unstructured text blocks and rasterizing image candidates.
2. **Semantic Gating & Classification:** The LLM exercises its cognitive capacity to execute an initial textual prescreening, discerning relevant experimental contexts from tangential mentions. Subsequently, visual candidates are processed by the Vision-Language Model (VLM) via strict zero-shot prompting for precise figure classification, aggressively rejecting microstructural or purely theoretical plots.
3. **Cross-Modal Orchestration:** Exploiting its advanced semantic reasoning capacity, the LLM outputs heavily constrained JSON parameters that map a specific visual curve to its corresponding textual constitutive formula. This is achieved by autonomously correlating visual features—such as extracted legend labels and color spaces—with the physical variables (e.g., temperature T , stress σ) explicitly defined in the text.
4. **Deterministic Execution:** The LLM's semantically linked JSON output acts as a precise functional trigger for the custom CV digitization tools, translating the targeted pixel trajectories into physical numerical datasets without human intervention.
5. **Validation and Serialization:** The digitized curves and extracted formula parameters undergo physics-informed cross-modal validation. Verified equation-data pairs are then permanently serialized into a normalized relational SQLite database, establishing a fully traceable provenance audit trail from the raw source document to the extracted constitutive parameters.

2.8. Quality Assessment of the Automated Workflow

To evaluate the efficacy of the automated screening module, a quantitative assessment was performed against a human-verified ground truth. In response to the need for robust statistical validation, we established an expanded validation subset consisting of 100 randomly selected documents from the initial corpus. We manually reviewed these documents to categorize them as either “Relevant” or “Irrelevant”. The detailed confusion matrix yielded 68 True Positives, 12 False Positives, 18 True Negatives, and 2 False Negatives. Based on this 100-document set, the system achieved a Precision of 85.00% (95% CI: 75.6–91.2%), Recall of 97.14% (95% CI: 90.2–99.2%), overall Accuracy of 86.00% (95% CI: 77.9–91.5%) and F1-Score of 90.67% (95% CI: 84.9–94.4%). This intentionally high Recall reflects our conservative design philosophy: prioritizing the prevention of data loss while effectively filtering out purely theoretical noise.

Given that this process functions as an information retrieval task, we adopted standard evaluation metrics used in Retrieval-Augmented Generation (RAG) systems to quantify performance: Precision, Recall, F1-Score, and Accuracy. To ensure clarity for readers across disciplines, we define these metrics based on the following classification outcomes:

- **True Positive (TP):** The Agent correctly identified a relevant paper containing creep data.
- **False Positive (FP):** The Agent incorrectly classified an irrelevant paper as relevant (noise).
- **True Negative (TN):** The Agent correctly rejected an irrelevant paper.
- **False Negative (FN):** The Agent failed to identify a relevant paper (missed data).

Based on these counts, the metrics are calculated as follows: **Precision** measures the reliability of the agent's selection, indicating the percentage of retrieved documents that are actually relevant:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Recall measures the completeness of the search, indicating the percentage of all existing relevant documents that the agent successfully found:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

The **F1-Score** provides a harmonic mean of Precision and Recall, offering a balanced view of the agent's robustness:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Finally, **Accuracy** represents the overall correctness of the classifier across the entire validation set:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

These metrics provide a comprehensive framework to benchmark the agent's ability to maximize data acquisition while minimizing the inclusion of irrelevant literature.

3. Results

To ensure the reliability of the constructed database, we implemented a rigorous "human-in-the-loop" verification protocol. Following the automated extraction phase, a random subset of the processed literature was subjected to manual inspection. Out of the 243 processed documents, the agent achieved a verified success rate exceeding 90% in correctly identifying and digitizing target creep curves. This high metric is particularly significant given the heterogeneous nature of the source documents, which range from high-resolution modern digital PDFs to scanned legacy publications spanning several decades, often characterized by varying clarity and rasterization artifacts. Figure 3 presents a qualitative comparison that highlights the agent's robustness in handling complex extraction scenarios. A major challenge in automated literature mining is the high density of visual information; a single publication typically contains various experimental visualizations—ranging from tensile stress-strain curves to microstructural micrographs—that must be distinguished from the target creep data. Furthermore, even within an identified creep plot, multiple trajectories often coexist, representing different stress or temperature conditions (as shown in Figure 3a, where curves for 52.7 MPa and 31.6 MPa overlap). As illustrated in the reconstructed result (Figure 3b), the agent successfully performed semantic disentanglement: it not only filtered out graphical artifacts like grid lines and

background noise but also accurately isolated the specific trajectory corresponding to the target experimental condition ($\sigma = 31.6$ MPa), effectively ignoring the extraneous data series. This demonstrates a context-aware extraction capability that transcends simple visual segmentation. Beyond the graphical data, the semantic accuracy of the extracted metadata—specifically the experimental conditions (temperature, stress) and material identifiers—was also validated. The consistency of these textual extractions is evidenced by the logical clustering observed in the previously discussed parameter space, where no distinct physical outliers (e.g., ice appearing in high-temperature clusters) were detected. This dual validation of both visual (curve) and semantic (text) information confirms that the automated workflow is sufficiently precise to serve as a foundation for data-driven scientific inquiry.

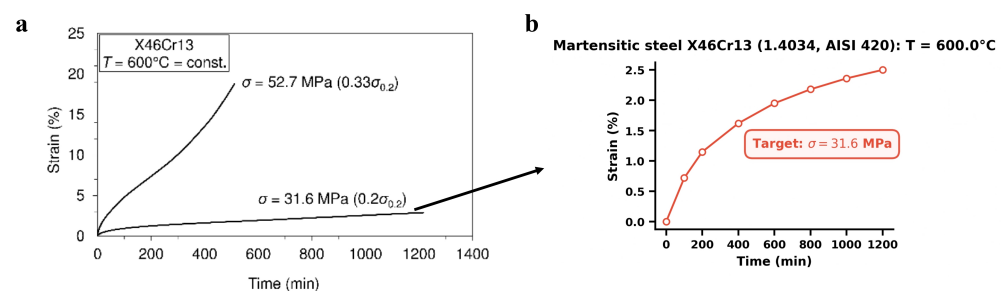


Figure 3. Demonstration of selective curve extraction from complex multi-variable plots. (a) Original source image (Material: X46Cr13, $T = 600$ °C) containing multiple creep trajectories under different stress levels [32]. (b) The digitized result reconstructed from the agent’s output. The agent successfully disentangled the visual information, accurately isolating the specific curve corresponding to the target condition ($\sigma = 31.6$ MPa) while ignoring the extraneous data series (52.7 MPa). Note that the high fidelity of the extracted data points preserves the underlying creep trend.

3.1. Overview and Composition of the Constructed Database

To evaluate the efficacy of the proposed Agent-Skills framework, we deployed the system to perform an automated extraction task on a massive collection of unstructured academic literature. Our initial keyword-retrieved raw pool consisted of 1344 documents. After filtering for accessible and machine-parsable PDFs, a corpus of 361 publications was passed to the Agent’s automated screening module. The LLM successfully filtered out purely theoretical and irrelevant noise (a semantic rejection rate of approximately 32.7%), retaining 243 highly relevant experimental publications. From these, the agent successfully constructed 353 high-fidelity ‘Equation-Data’ pairs. While a single publication may contain multiple graphical datasets, our framework focuses on isolating the principal trajectories used by the original authors to derive or validate their constitutive models. This rigorous filtering resulted in a focused dataset where each curve is accompanied by its explicit mathematical parameters, prioritizing physical interpretability over raw data volume. A subsequent manual verification of the extracted data revealed a high fidelity in the automated workflow, with a validated success rate exceeding 90%. This high precision underscores the robustness of the defined skills in handling the heterogeneous formatting and nomenclature found across different journals and decades of research. Crucially, this >90% verified extraction success rate was evaluated strictly at the Equation-Data-Pair level (per target curve), rather than per document. If a single graphical figure contained multiple distinct experimental data series, each individual curve was evaluated as an independent extraction attempt. A single curve extraction was recorded as a “Success” only if it simultaneously satisfied four stringent criteria: (1) The specific target experimental creep curve was correctly identified and isolated. (2) The coordinate axes and their physical units were correctly parsed. (3) The corresponding theoretical constitutive equation was successfully

extracted and associated. (4) The quantitative reconstruction error (Normalized RMSE) of the digitized trajectory was strictly less than 15%. To further validate the necessity of our LLM-driven semantic gating, we benchmarked this pipeline against a deterministic rule-based OCR/CV baseline (detailed in Supplementary Information Section S7). The baseline failed to produce any valid cross-modal equation-data pairs, quantitatively demonstrating that semantic reasoning, rather than mere digitization, is the primary bottleneck in legacy literature mining.

The resulting database exhibits a remarkable diversity in material composition, demonstrating the agent’s ability to generalize across various scientific sub-domains. As illustrated in Figure 4 (Material Distribution), the dataset is predominantly composed of high-performance engineering materials, with Nickel-based alloys (19.0%) and Steel/Iron (12.5%) constituting the largest shares. These are critical materials for high-temperature applications, reflecting the literature’s focus on industrial relevance. Notably, the agent also successfully retrieved data for less conventional categories, including Polymers/Plastics (8.5%), Rocks/Geological materials (5.9%), and even Ice/Glaciers (4.0%). This inclusion of diverse material classes—ranging from soft matter to geological composites—validates that the agent is not over-fitted to specific metallic keywords but possesses a semantic understanding of the “creep” phenomenon broadly defined.

Material Distribution in Creep Database

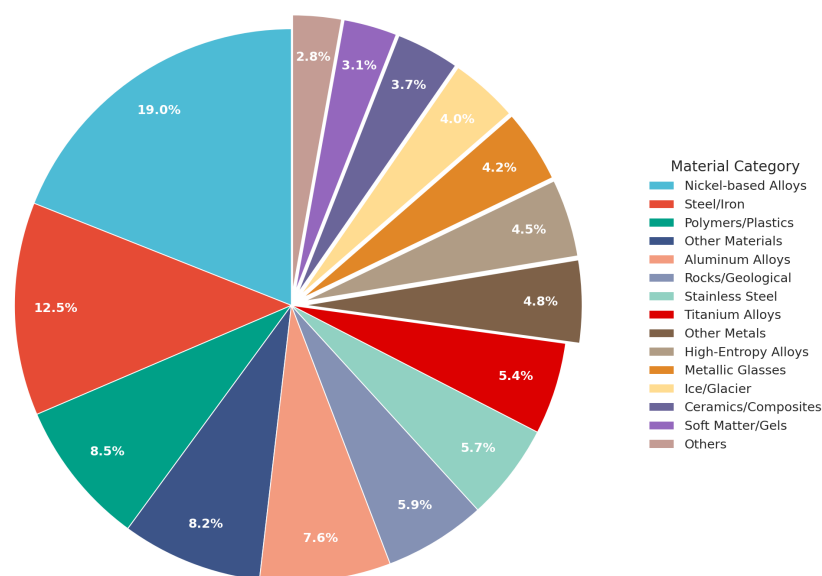


Figure 4. Diversity of material composition within the constructed creep database. The sector chart illustrates the distribution of material categories across the 353 extracted creep curves. The dataset demonstrates broad domain coverage, encompassing critical industrial materials (e.g., Nickel-based alloys, 19.0%; Steel/Iron, 12.5%) as well as non-metallic systems (e.g., Polymers, 8.5%; Ice/Glaciers, 4.0%). This diversity confirms the agent’s semantic ability to identify relevant “creep” behaviors across heterogeneous material science sub-domains.

Furthermore, the database covers a comprehensive range of experimental conditions, providing a rich parameter space for future modeling. Figure 5 visualizes the distribution of key physical parameters: temperature and stress. The temperature distribution (Figure 5a) spans from cryogenic and room-temperature environments (typical for polymers and ice) to extreme conditions exceeding 1000 °C (characteristic of superalloys and ceramics). Similarly, the stress levels (Figure 5b) range from low-magnitude geological

pressures to high-stress engineering loads (>1000 MPa). The interaction between these variables is mapped in the Temperature-Stress parameter space shown in Figure 6. This scatter plot reveals distinct clusters corresponding to different material capabilities—for instance, the high-temperature/high-stress region occupied by superalloys versus the low-temperature/low-stress region of soft matter. The successful population of this broad physical landscape confirms that the constructed database is not only sizable but also physically representative, effectively capturing the constitutive behaviors of materials across multiple scales.

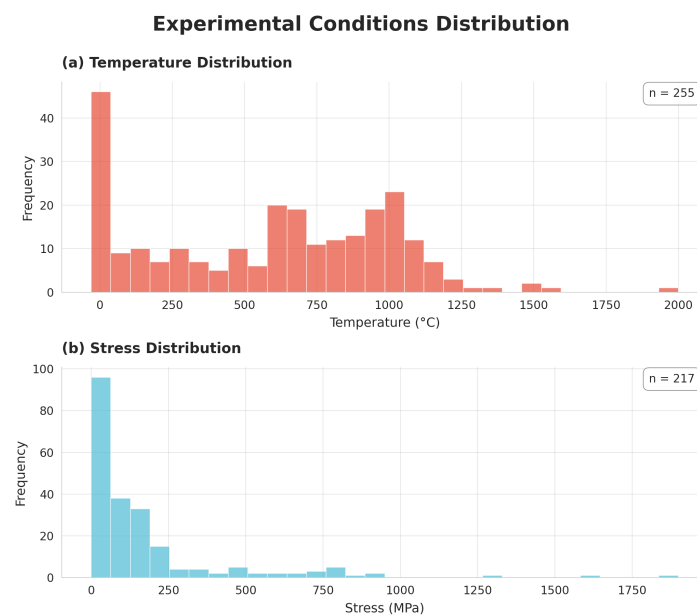


Figure 5. Statistical distribution of experimental conditions extracted from the literature. (a) Frequency distribution of experimental temperatures (°C), spanning from cryogenic environments to ultra-high temperature regimes (>1000 °C). (b) Frequency distribution of applied stress levels (MPa). The broad range of operating conditions indicates that the database captures a comprehensive snapshot of material performance boundaries documented in the historical literature.

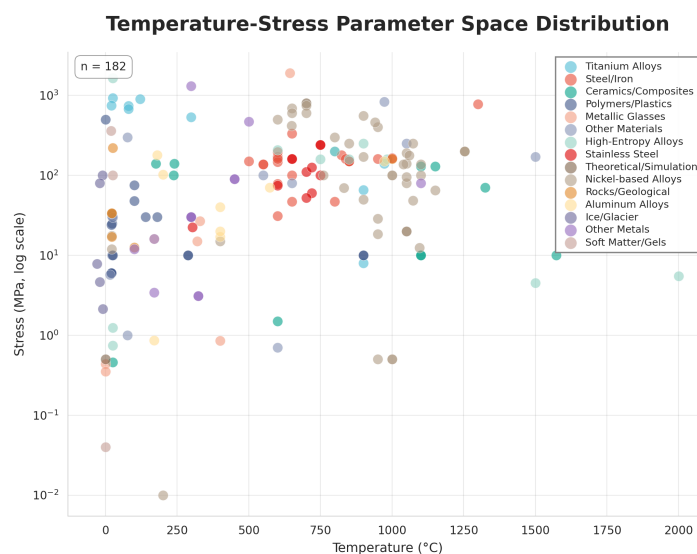


Figure 6. Visualization of the Temperature-Stress parameter space. A scatter plot mapping the experimental conditions for all extracted curves ($n = 353$), color-coded by material category. The visualization reveals distinct physical clusters corresponding to material capabilities—such as the high-temperature/high-stress dominance of superalloys versus the low-energy regimes of soft matter—demonstrating the physical consistency and logical distribution of the automatically constructed dataset.

To maximize the practical usability of these 353 verified pairs, the integrated web interface (Figure 2) allows researchers to dynamically query the dataset. Users can apply multi-dimensional filters based on specific alloy compositions or thermo-mechanical boundaries, visually compare the overlapping reconstructed trajectories, and instantly export the structured data for downstream machine learning and constitutive modeling.

3.2. Verification of Physical Consistency via Cross-Modal Alignment

Beyond graphical accuracy, a critical requirement for a scientific database is the preservation of physical consistency between the extracted raw data and the theoretical models described in the source text. Scientific publications often exhibit information redundancy, presenting experimental results in graphical forms while simultaneously reporting the fitted constitutive parameters in the accompanying text. This offers a unique opportunity for cross-modal validation. To demonstrate this, we analyzed a representative case study on AMAG-183 metallic glass (Figure 7). The agent independently executed two tasks: (1) digitizing the raw deformation-time trajectory from a complex plot characterized by colored backgrounds and multiple data series (Figure 7a) and (2) extracting the mathematical modeling parameters from the document's discussion section. In this specific case, the source literature described the deformation behavior using a non-linear Duffing oscillator model to capture the chaotic dynamics of shear band evolution. Figure 7b visualizes the alignment between these two modalities. The open circles represent the visually digitized experimental data, while the solid red line represents the theoretical trajectory reconstructed by substituting the textually extracted parameters into the governing differential equation. As quantified in the inset, the theoretically reconstructed model shows exceptional agreement with the digitized data points, achieving a coefficient of determination (R^2) of 0.9999. This near-perfect synchronization confirms that the agent effectively bridged the semantic gap between unstructured textual parameters and rasterized visual data, ensuring the database is physically self-consistent even for complex, non-linear material behaviors. This rigorous cross-modal verification protocol implicitly acts as a quality filter; while it may limit the total number of extracted curves per publication compared to indiscriminate extraction, it guarantees that every entry in the database possesses a verifiable high-fidelity link between experimental data and physical theory.

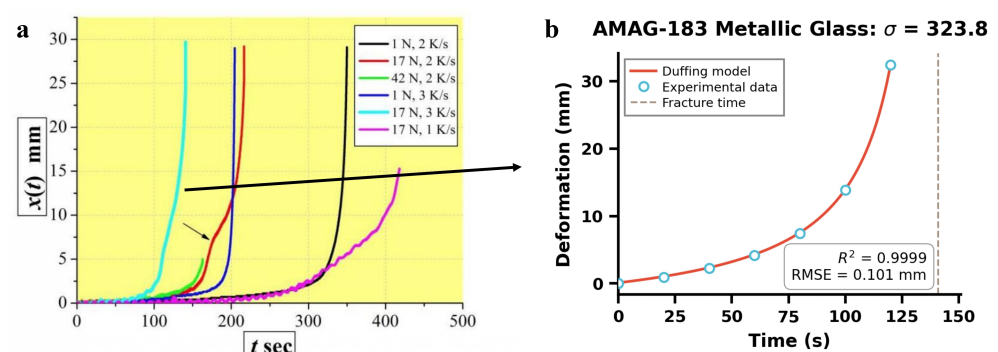


Figure 7. Cross-modal verification of physical consistency for complex non-linear behavior. (a) Original experimental plot for AMAG-183 metallic glass, characterized by a non-standard background and multiple overlapping series [33]. (b) Validation of the extracted data against the theoretical physics model. The symbols (blue circles) represent the data digitized from the image, while the solid line (red) represents the trajectory calculated using the Duffing model parameters extracted from the text. The exceptional correlation ($R^2 = 0.9999$) demonstrates the agent's ability to correctly map textual mathematical descriptions to their corresponding visual representations, validating the physical integrity of the constructed database.

4. Conclusions

In this study, we presented a novel, autonomous framework capable of transforming the fragmented landscape of the scientific literature into structured, machine-actionable knowledge. By synergizing the reasoning capabilities of Foundation LLMs with precise visual extraction skills, we successfully breached the “data bottleneck” that has long impeded data-driven research in mechanics and materials science.

The construction of the creep mechanics database served as a rigorous testbed for this approach. Our results demonstrate that the agent transcends traditional OCR-based limitations, exhibiting the semantic intelligence necessary to disentangle complex experimental plots and validate data against physical laws. The achievement of a near-perfect correlation ($R^2 = 0.9999$) between digitized curves and theoretical models in our case study highlights the system’s ability to perform cross-modal reasoning, effectively bridging the gap between visual observations and mathematical descriptions.

While this work focused on creep deformation, the foundational architecture—including the LLM orchestration logic, visual classification, and color-series disentanglement—is inherently domain-agnostic. Conversely, the physics-based validation gates are domain-specific. However, the integration of Foundation LLMs introduces a paradigm shift in cross-domain migration: rather than requiring experts to write complex, hard-coded extraction rules, the burden is drastically reduced to merely auditing and updating natural language prompts. This fundamentally lowers the barrier for migrating the framework to other specialized fields, such as battery degradation cycling or catalytic reaction kinetics.

Nevertheless, we acknowledge certain limitations in the current pipeline. The geometric reconstruction heavily relies on the clarity of the coordinate axis ticks; highly degraded legacy scans lacking explicit tick marks may require human intervention. Future research will focus on enhancing the agent’s robustness against extreme visual degradation and extending its cognitive skills to autonomously cross-reference multiple PDFs simultaneously. As LLMs continue to evolve, we envision such autonomous data agents becoming a standard component of the scientific infrastructure, accelerating the loop of hypothesis generation and discovery in the era of AI-driven science.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16157819/s1>, References [34,35] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, T.S. and D.P.; methodology, Y.W. and S.H.; software, Y.W. and T.S.; validation, Y.W., T.S. and S.H.; formal analysis, Y.W. and T.S.; writing—original draft preparation, Y.W. and T.S.; writing—review and editing, D.P. and S.H.; supervision, D.P. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Innovation Action Plan, grant number 25TS1402500.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The complete automated extraction pipeline, including the core logic, prompt templates, and custom OpenCV digitization scripts, is open-sourced and available on GitHub at <https://github.com/shuWuYue123/creep-literature-pipeline> (accessed on 8 July 2026). Due to copyright and licensing restrictions associated with the original legacy publications, the extracted numerical datasets and rasterized source images are not included in the public repository.

Acknowledgments: During the preparation of this work, the authors used Google Gemini 2.5 Flash Image (also known as Nano Banana; <https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-image>; accessed on 1 February 2026) to generate specific illustrative graphic elements and

visual assets used in the workflow diagram (Figure 1). After using this tool, the authors manually assembled, rigorously reviewed, and edited the generated visual elements to construct the final schematic. The authors take full responsibility for the final content of the publication and confirm that no generative AI technologies were used to generate, alter, or manipulate any scientific data, experimental results, or analytical conclusions presented in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
VLM	Vision-Language Model
RMSE	Root Mean Square Error
CV	Computer Vision
OCR	Optical Character Recognition

References

1. Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60. <https://doi.org/10.1038/s41586-023-06221-2>.
2. Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **2016**, *4*, 053208. <https://doi.org/10.1063/1.4946894>.
3. Chen, M.; Tworek, J.; Jun, H.; Johnson, Q.; Pinto, H.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. Evaluating large language models trained on code. *arXiv* **2021**, arXiv:2107.03374.
4. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* **2023**, arXiv:2303.12712.
5. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 24824–24837. <https://doi.org/10.52202/068431-1800>.
6. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, 1–5 May 2023.
7. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023. <https://doi.org/10.52202/075280-2997>.
8. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The rise and potential of large language model based agents: A survey. *Sci. China Inf. Sci.* **2025**, *68*, 121101. <https://doi.org/10.1007/s11432-024-4222-0>.
9. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578. <https://doi.org/10.1038/s41586-023-06792-0>.
10. Szymanski, N.J.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.; Cubuk, E.D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86–91. <https://doi.org/10.1038/s41586-023-06734-w>.
11. Merchant, A.; Batzner, S.; Schoenholz, S.S.; Aykol, M.; Cheon, G.; Cubuk, E.D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85. <https://doi.org/10.1038/s41586-023-06735-9>.
12. Wang, Z.; Sun, Z.; Yin, H.; Liu, X.; Wang, J.; Zhao, H.; Pang, C.H.; Wu, T.; Li, S.; Yin, Z.; et al. Data-driven materials innovation and applications. *Adv. Mater.* **2022**, *34*, 2104113. <https://doi.org/10.1002/adma.202104113>.
13. Fan, V.; Qian, Y.; Wang, A.; Wang, A.; Coley, C.W.; Barzilay, R. OpenChemIE: An information extraction toolkit for chemistry literature. *J. Chem. Inf. Model.* **2024**, *64*, 5521–5534. <https://doi.org/10.1021/acs.jcim.4c00572>.
14. Swain, M.C.; Cole, J.M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904. <https://doi.org/10.1021/acs.jcim.6b00207>.
15. Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K.A.; Ceder, G.; Jain, A. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **2019**, *59*, 3692–3702. <https://doi.org/10.1021/acs.jcim.9b00470>.

16. Wang, Z.; Kononova, O.; Cruse, K.; He, T.; Huo, H.; Fei, Y.; Zeng, Y.; Sun, Y.; Cai, Z.; Sun, W.; et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Sci. Data* **2022**, *9*, 231. <https://doi.org/10.1038/s41597-022-01317-2>.
17. Zhang, Y.; Ling, C. A strategy to apply machine learning to small datasets in materials science. *Npj Comput. Mater.* **2018**, *4*, 25. <https://doi.org/10.1038/s41524-018-0081-z>.
18. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525–535. <https://doi.org/10.1038/s42256-024-00832-8>.
19. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. <https://doi.org/10.1038/s41586-019-1335-8>.
20. Luo, J.; Li, Z.; Wang, J.; Lin, C.Y. ChartOCR: Data extraction from charts images via a deep hybrid framework. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2021; pp. 1916–1924. <https://doi.org/10.1109/WACV48630.2021.00196>.
21. Masry, A.; Do, X.L.; Tan, J.Q.; Joty, S.; Hoque, E. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 2263–2279. <https://doi.org/10.18653/v1/2022.findings-acl.177>.
22. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. <https://doi.org/10.1145/3571730>.
23. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474.
24. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 34892–34916. <https://doi.org/10.52202/075280-1516>.
25. Kassner, M.E. *Fundamentals of Creep in Metals and Alloys*, 3rd ed.; Butterworth-Heinemann: Oxford, UK, 2015. <https://doi.org/10.1016/B978-0-08-099427-7.00001-3>.
26. Sawada, K.; Kimura, K.; Abe, F.; Taniuchi, Y.; Sekido, K.; Nojima, T.; Ohba, T.; Kushima, H.; Miyazaki, H.; Hongo, H.; et al. Catalog of NIMS creep data sheets. *Sci. Technol. Adv. Mater.* **2019**, *20*, 1131–1149. <https://doi.org/10.1080/14686996.2019.1697616>.
27. Udrescu, S.M.; Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **2020**, *6*, eaay2631. <https://doi.org/10.1126/sciadv.aay2631>.
28. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. <https://doi.org/10.1038/s42254-021-00314-5>.
29. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>.
30. Zhang, X.; Hu, S. A Multitask Dynamically Interconnected Learning Approach for Physics-Informed Neural Networks. *Mater. Today Commun.* **2025**, *47*, 113054. <https://doi.org/10.1016/j.mtcomm.2025.113054>.
31. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
32. Brnic, J.; Krscanski, S.; Lanc, D.; Brcic, M.; Turkalj, G.; Canadija, M.; Niu, J. Analysis of the Mechanical Behavior, Creep Resistance and Uniaxial Fatigue Strength of Martensitic Steel X46Cr13. *Materials* **2017**, *10*, 388. <https://doi.org/10.3390/ma10040388>.
33. Berezner, A.; Fedorov, V.; Perov, N.; Grigoriev, G. Novel exact solutions of the Duffing equation: Stability analysis and application to real non-linear deformation tests. *arXiv* **2025**, arXiv:2512.24774.
34. Izuno, H.; Yokokawa, T.; Koizumi, Y.; Odaka, S.; Harada, H. A Creep Constructive Equation of Ni-Base Superalloys Using Structural Parameters. *J. Jpn. Inst. Met. Mater.* **2004**, *68*, 526–529. <https://doi.org/10.2320/jinstmet.68.526>.
35. Yang, X.; Xu, Q.; Lu, Z. The Development and Validation of the Creep Damage Constitutive Equations for P91 Alloy. In *Proceedings of the 2013 World Congress in Computer Science and Computer Engineering and Application*; CSREA Press: Las Vegas, NV, USA, 2013, pp. 121–127.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.