Bridging the Data Bottleneck: A Multi-Modal Generative Extraction Pipeline for Legacy Scientific Literature

Gaurav Kumar
New York University
gk2839@nyu.edu

Anasse Bari
New York University
abari@nyu.edu

David J. Nagel
George Washington University
nagel@gwu.edu

Abstract

The majority of historical experimental data across the empirical sciences remains locked in unstructured legacy formats, which slows data-driven discovery. We introduce an automated multi-modal pipeline that processes large literature corpora, filters non-empirical noise, and outputs structured datasets. Layout-aware document parsing is paired with a lightweight semantic triage gatekeeper to substantially reduce the computational overhead of applying frontier multi-modal models at scale. A generative model reconstructs experimental trajectories from text and figures, while a schema-constrained prompt and a strictly typed programmatic validation layer check physical bounds before serialization. We evaluate the pipeline on a 60-document corpus of extreme-condition thermodynamic literature drawn from Low Energy Nuclear Reactions (LENR) research, reporting concrete throughput, triage, and constraint-violation statistics. We distinguish safeguards enforced in code (physical-bound validators) from those requested through the prompt, such as source grounding, and outline the evaluation needed to verify the latter. The architecture is domain-agnostic by design, and we outline how it could transfer to catalysis, energy storage, and preclinical biomedical research.

1 Introduction

Data-driven science, from battery discovery to drug design, increasingly depends on machine learning. Yet decades of experimental findings sit inside unstructured legacy PDFs. Modern methods such as Graph Neural Networks and Bayesian optimizers need dense numeric tensors, while historical literature buries experimental conditions and kinetic rates inside prose, multi-column tables, and rasterized plots that are effectively invisible to large-scale computational analysis.

Naive OCR and text-extraction libraries destroy spatial structure, collapsing tables and charts into incoherent strings. LLMs parse this text more fluently, but applying them naively introduces new failure modes: hallucinated decimal points (Huang et al., 2025), inconsistent units, and a positivity bias that can silently drop failed experimental runs, precisely the negative examples a reinforcement-learning agent needs to learn the boundaries of a physical system.

We present a multi-modal, automated extraction pipeline designed for large, disordered collections of scientific papers. Its front end has two stages. First, layout-aware Intelligent Document Parsing (IDP) converts legacy publications into spatial Markdown that preserves the position of tables and figures through bounding-box data. Next, a two-stage semantic triage and visual screening step removes papers without scientific substance, along with decorative images, before the core generative model sees them.

To constrain hallucination in the generative step, we combine a prompt-level framework, requesting verbatim source-quote grounding and converting visual inflection points into temporal arrays, with validators that check simple physical bounds on whatever does get extracted. As the Limitations section discusses, we distinguish the two layers explicitly.

Our contributions:

- A multi-modal, dual-LLM architecture that reconstructs legacy scientific PDFs into spatial Markdown and visual tensors, cutting extraction cost through automated semantic triage.
- An anti-hallucination framework whose schema requests exact source-quote grounding for numeric claims and captures negative (null) results so that failed runs stay in the dataset.
- A bifurcated extraction design in which a

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 91–98
November 10, 2026 ©2026 Association for Computational Linguistics

schema-constrained generative model normalizes categorical entities and SI units, while a strictly typed validation layer independently checks physical bounds on the extracted numeric fields, rejecting grossly impossible values.

2 Related Work

Intelligent Document Processing (IDP) OCR-based extraction degrades sharply on documents with complex multi-column layouts or embedded figures. Early layout-aware models such as LayoutLM (Xu et al., 2020), together with datasets such as PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020), showed that combining text and spatial information substantially improves document understanding. More recent multi-modal models, including Nougat (Blecher et al., 2024), LayoutLMv3 (Huang et al., 2022), and the OCR-free Donut transformer (Kim et al., 2022), preserve structure such as mathematical formulas and tables, and benchmarks such as DocGenome (Xia et al., 2024) extend this to large-scale scientific documents. At scale, however, they trade speed against accuracy. We combine layout-aware parsing (Docling, Auer et al., 2024) with a lightweight triage stage that keeps a large share of documents away from the expensive model.

LLMs in Scientific Extraction and Hallucination Large language models are increasingly used to extract structured data from unstructured text in materials science, chemistry, and pharmaceutical research, but they can produce factual errors (Huang et al., 2025). HalluMat (Vangala et al., 2025), for example, showed that LLMs often fabricate plausible but incorrect material formulations, and that multiple verification stages reduce this tendency without eliminating it. We call the analogous error "physical hallucination": extracting a value that contradicts established thermodynamics because the model has no firm grounding in physical law. Fact-checking systems such as FActScore (Min et al., 2023) and claim-verification benchmarks such as SciFact (Wadden et al., 2020) address a related problem by retrieving supporting evidence, and retrieval-augmented fact-checking (Lewis et al., 2020) is the prevailing approach, though it can struggle with very sparse datasets. Our approach instead relies on a prompt-level instruction to ground values in source quotes; the current state of independent verification of this step is described in the Limitations section.

Structured Schemas and Constraint Validation Schema-driven generation, such as asking an LLM for JSON or constraining the decoding process itself (Willard and Louf, 2023), yields a consistent format but does not guarantee physically correct values. We therefore add Python validators built on Pydantic (Colvin et al., 2024) that reject any output violating a small set of predefined physical limits, regardless of what the generative model produced.

Domain-Specific Scientific Extraction Closest to our setting: ChemDataExtractor mines chemical properties with classical NLP (Swain and Cole, 2016), MatSciBERT adapts a domain language model for materials-science text mining (Gupta et al., 2022), and Dagdelen et al. (2024) extract structured records from materials papers with instruction-tuned LLMs and a lightweight validation step. We target a harder regime on two axes at once: historical, multi-modal source documents (prose, tables, rasterized plots), and a target domain, Low Energy Nuclear Reactions, with unusually severe reporting variance and positivity bias.

3 The Generative Extraction Architecture

The system follows the six-step process shown in Figure 1, which separates spatial document handling from the LLM's reasoning. As a result, the expensive generative model sees only filtered input that has passed triage.

3.1 Document Parsing and Spatial Preservation

Simple extraction tools such as PyPDF2 and PDFMiner read documents as flat character streams, which merges table cells, destroys multi-column layouts, and causes errors in later inference stages. We instead use a layout-aware Document Layout Analysis system, IBM's Docling (Auer et al., 2024), to identify titles, paragraphs, and tables, and rebuild them as Markdown, keeping tables as Markdown tables rather than loose text strings.

The same pass performs visual extraction: rather than asking the LLM to infer data from a text description of a graph, the DLA engine locates every figure's bounding box and crops it to a standalone image. This bifurcated output, structured

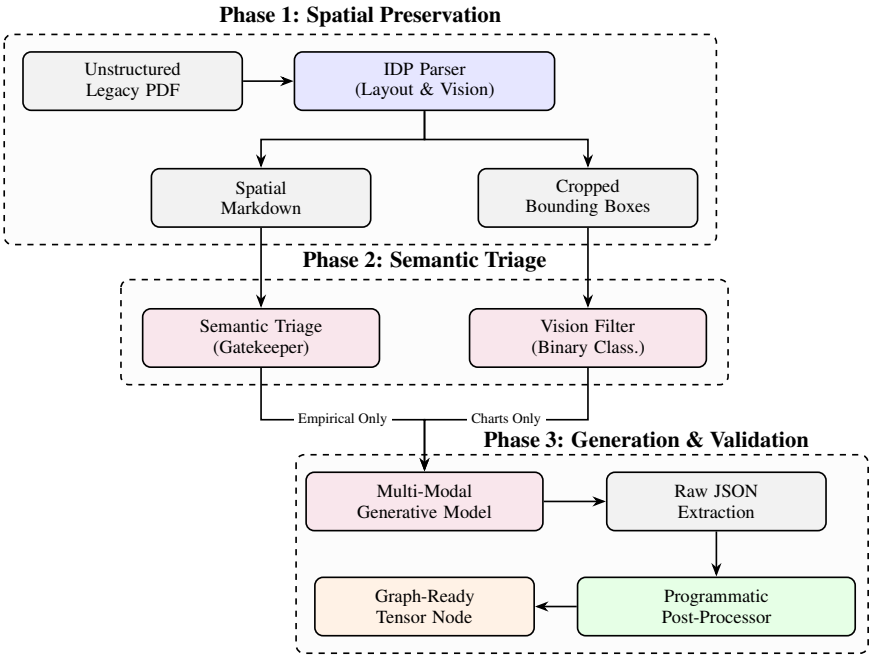


Figure 1: The Generative Extraction Architecture. Legacy PDFs are spatially decomposed into structural Markdown and visual elements. A dual-model triage system filters non-empirical noise before schema-guided multimodal extraction, and a typed post-processing layer rejects values that violate simple physical bounds before serialization.

Markdown plus isolated images, is what feeds the generative stage.

3.2 Semantic Triage (Cost and Token Optimization)

Running a frontier multi-modal model over an entire scraped corpus is expensive, and most scraped papers are theoretical or lack quantitative data, so they are useless to a downstream ML pipeline. We use a two-layer, low-cost triage gatekeeper to avoid paying for that noise.

The first layer performs document-level triage: a lightweight, high-speed LLM (`gemini-3.5-flash`) reads the first 30,000 characters (roughly 10 pages) and answers one boolean question: does this text contain extractable experimental data? Irrelevant or purely theoretical documents are routed out before the expensive model ever sees them.

The second layer filters images. Docling extracts every embedded rectangular region, including logos, headers, and portraits. A lightweight vision model (also `gemini-3.5-flash`), run at temperature 0.0 with output limited to 5 tokens, answers one yes/no question per image: is it a graph or a table? On our 60-document pilot (Section 6), the filter removed 79 of 332 extracted images (23.8%) as unsuitable for scientific analysis, which reduced the visual input volume (Table 1).

3.3 Multi-Modal Generative Extraction

The core generative model, Gemini (Gemini Team, Google, 2023), receives the structured Markdown together with the filtered high-resolution images and links the two modalities, for example when a value appears only on a plot's axis label or in a table's caption, which text-only extraction cannot handle. Its output is a structured but not yet validated JSON payload, which moves the information from natural language into a form that code can check. In the full pipeline, triage uses a lightweight Flash-class model and extraction uses a larger model (`gemini-3-pro-preview`); the 60-paper pilot reported here used `gemini-3.5-flash` for all stages.

4 The Anti-Hallucination Framework

LLMs can struggle with zero-shot numerical extraction: they may misplace decimal points, reverse logical relationships, and lose track of dependencies between variables in complex multi-dimensional settings. Figure 2 shows our approach, which combines three prompt-level instructions (source grounding, negative-result capture, and entity anchoring) with the programmatic validation described in Section 5. The Limitations

section separates what the model is only asked to do from what is checked independently in code.

4.1 Source Grounding and Confidence Scoring

The schema requires a `source_quote` for each key parameter. When the LLM proposes a value, such as a peak temperature of 450 K, it must also supply the exact wording that supports it. This requirement is implemented through the prompt and schema. The programmatic validation layer (Section 5) checks physical limits and data types, but it does not check that each `source_quote` is an exact excerpt of the source document. We describe the status of our audit in the Limitations section; a dedicated substring validator is the natural next step.

The model also emits a single `extraction_confidence_score` (0.0–1.0) per processed paper, reflecting its overall confidence in that document's legibility and completeness. This document-level score lets downstream models down-weight poorly digitized papers, but it cannot distinguish a confidently extracted value from an uncertain one within the same document; we discuss the case for per-parameter confidence in the Limitations section.

4.2 Negative Result Capture and Entity Anchoring

Scientific publications favor successful outcomes, and generative models inherit that bias by underreporting failed configurations, a serious problem when a reinforcement-learning agent needs negative boundaries to understand a system's limits. We counter this with an explicit Negative Bias Trap: the model is prompted to hunt for null runs and baseline controls, setting an `is_failed_replication` flag so downstream models learn from failures as well as successes.

Historical literature also re-publishes the same experiment across venues over many years; treating each paper as an isolated event inflates dataset scale and causes entity-resolution bleed. Entity Anchoring addresses this by capturing citations linking a run to prior publications into a `related_prior_experiments` array, letting the downstream graph merge duplicate runs.

5 Programmatic Validation and Strict Type Enforcement

A syntactically valid JSON payload is not necessarily a physically valid one: the LLM remains stochastic and can produce impossible numeric values even when formatting is perfect. We separate the two concerns: the generative model performs initial normalization under prompt constraints, and a strictly typed validation library, Pydantic (Colvin et al., 2024), independently rechecks the result before it is allowed into the database.

5.1 Entity Normalization and Categorical Enums

Different labs describe the same material differently (e.g., "heavy water bath" vs. "D_2O medium"); naive extraction preserves this fragmentation and floods a knowledge graph with duplicate nodes. We instead bind the model to predefined schema Enums, and Pydantic rejects any unmapped raw string in a categorical field. Since science continuously introduces materials no enum anticipates, unknown values fall back to a generic `OTHER` category paired with a free-text `_notes` field (e.g., `medium: OTHER`, `medium_notes: "5% D2SO4 in D2O"`), so novel materials are preserved without polluting the categorical structure.

5.2 Trajectory Serialization

Experimental parameters like temperature or pressure are rarely static, but naive extraction collapses them to a single scalar, discarding the temporal context time-series models need. We model these instead as `Trajectory` objects: the generative model reads inflection points off a plot (or computes them from a stated rate, e.g., 1 K/min) and serializes them as `[time_offset_seconds, value]` pairs. The validation layer then rejects any trajectory containing a negative `time_offset_seconds`, ruling out causally impossible timelines.

5.3 Enforcing Physical Constraints

Unconstrained LLMs can produce thermodynamically infeasible values or mishandle units, for example by writing a Celsius value into a Kelvin field. We use two layers. First, the prompt instructs the model to convert every measurement to SI units before emitting JSON. Second, Pydantic `@field_validator` functions independently

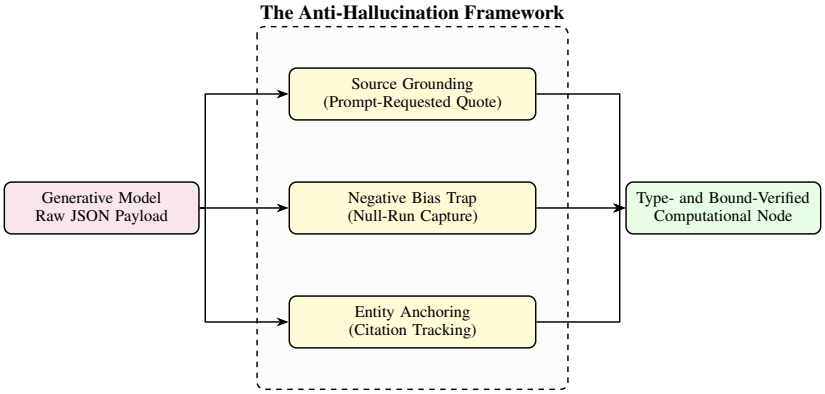


Figure 2: The Anti-Hallucination Framework. Source grounding, the negative-result trap, and entity anchoring are all implemented through the prompt and schema; only physical bounds and types are checked in code (Section 5).

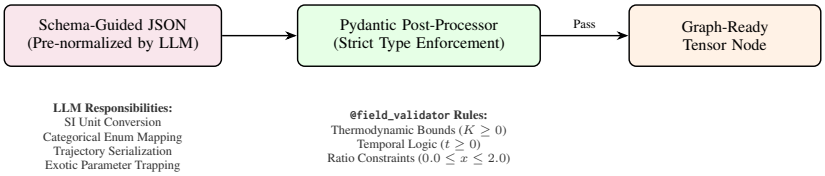


Figure 3: The Programmatic Validation Layer. The generative model handles unit conversion, time-series serialization, and entity normalization via prompt constraints; the Pydantic post-processor independently rejects values that violate physical bounds before serialization.

reject negative temperatures, pressures, and time offsets (along with negative molarity and coefficient of performance) and keep the deuterium-to-palladium loading ratio (`loading_ratio_D_Pd`) within 0.0–2.0. A violation raises a `ValueError`, the payload is rejected, and it is not inserted into the database. These are plausibility checks that catch grossly impossible values, such as a negative temperature in kelvin.

5.4 The Exotic Parameter Escape Hatch

A rigid schema risks silently discarding genuinely novel variables: an unusual stimulation method or dopant with no predefined field. Our `exotic_parameters` escape hatch routes such values into key-value pairs (e.g., `{"key": "electron_beam_energy_MeV", "value": "1.0"}`) instead of dropping them. To keep this from becoming a dumping ground, the prompt explicitly forbids placing standard variables (temperature, voltage, current, pressure) here; a stated "temperature range 9 to 270 K" must be formatted as a `Trajectory`, not an exotic key-value pair.

6 Case Study: Extreme-Condition Thermodynamic Literature

We applied the pipeline to Low Energy Nuclear Reactions (LENR) literature (Fleischmann and Pons, 1989; Storms, 2010), which spans three decades. The field has high reporting variance, non-standard units, and strong positivity bias, which makes it a hard test of the "data bottleneck" problem. We benchmarked throughput and constraint enforcement on a randomly selected, unannotated subset of 60 historical papers.

Table 1: Empirical Evaluation Metrics (60-Document Corpus)

Category	Metric	Value
Semantic Triage	Documents Processed	60
	Theoretical Papers Filtered	35 (58.3%)
	Bounding Boxes Processed	332
	Non-Sci. Images Pruned	79 (23.8%)
	Tokens Saved (est.)	**314,416**
Validation	Payloads Rejected (Pydantic)	1
Throughput	Total Latency	48.5 min
	Latency / Document	48.5 sec

6.1 Throughput and Token Optimization

Triage flagged 35 of the 60 papers (58.3%) as theoretical and removed them before generative extraction. The vision filter removed 79 of 332 images (23.8%) as unsuitable for scientific analysis (Sec-

tion 3.2). We estimate that the two filters together avoided about 314,416 tokens. This is a heuristic estimate, not billed usage: it counts 258 tokens per rejected image plus, for each skipped paper, the smaller of one quarter of its character count and 32,000, and about 94% of the total comes from skipped papers. Processing took 48.5 seconds per document on average (2,910.7 seconds in total). Triage, image filtering, and pilot extraction all used `gemini-3.5-flash`.

6.2 Anti-Hallucination and Constraint Enforcement

In the pilot, the validation layer rejected one extraction payload. Source grounding is requested through the prompt. Our evaluator checks it by substring match for two `Measurement` fields (peak coefficient of performance and total excess energy), but neither was populated in this sample, so the check is inconclusive (0 of 0). Extending it to all `Measurement` fields and making grounding a validation rule is the next step (see the Limitations section).

The enum-constrained schema also mapped descriptions like "heavy water bath" and "deuterium oxide medium" onto a single `D2O` value.

6.3 Temporal Reconstruction and Exotic Variables

Trajectory serialization matters for LENR because the field relies on temporal thermodynamics: the model reconstructed temporal arrays by interpolating calorimetry charts and by deducing offsets from stated rates in text. The Negative Bias Trap is designed to flag `is_failed_replication` on null runs so the resulting graph captures negative boundaries as well as successes. The `exotic_parameters` field was used to store non-standard conditions such as external magnetic fields, laser stimulation, and unusual lattice dopants as key-value pairs, keeping them out of the strictly typed continuous variables.

7 Generalization to Broader Scientific Domains

The pipeline separates spatial parsing (IDP) and programmatic validation (Pydantic) from the generative model, so adapting it to another field would in principle require new enumerations and new dimensional limits rather than model retraining. This is a design argument: we have not yet run the pipeline outside LENR, and the examples below are possible applications rather than results.

- **Energy Storage and Battery Materials:** Trajectory serialization could extract capacity and Coulombic-efficiency timelines from cycling plots, with the validation layer enforcing voltage bounds.
- **Catalysis and Chemical Synthesis:** Kinetic parameters such as turnover frequency, reaction temperature, and yield often appear in tables, which spatial preservation is meant to keep intact (Swain and Cole, 2016); novel dopants could go into the exotic-parameter field.
- **Preclinical Biomedical Research:** Dosage schedules and pharmacokinetic profiles could in principle be extracted the same way, with validation appropriate to clinical use.

Limitations

We note the following limitations.

Source grounding is prompt-level, not validator-enforced. The schema requests a verbatim `source_quote` alongside every numeric claim (Section 4.1), but the validation layer does not yet run an automatic substring check confirming the quote is a literal excerpt; only physical bounds and types are independently enforced (Section 5). Our evaluator currently audits grounding for two `Measurement` fields (peak coefficient of performance, total excess energy); on our 60-paper pilot both were unpopulated, so the check returned 0 of 0 rather than a meaningful pass rate. Extending programmatic grounding verification to all `Measurement` fields, and re-running the audit where those fields are populated, is the most important follow-up implied by this paper.

Confidence scoring is document-level, not parameter-level. One `extraction_confidence_score` per paper is enough to down-weight an entire document, but cannot distinguish a confidently extracted value from an uncertain one within it. Per-parameter confidence would need a schema change and additional prompt engineering, left to future work.

Evaluation is a small, single-domain pilot. The quantitative results come from a pilot of 60 randomly selected LENR papers (Section 6), and

the pipeline has been run only on LENR (Section 7). We did not compare extracted values against a human-verified reference set, so we report no precision or recall; the metrics measure throughput and constraint behavior rather than extraction correctness. The validator bounds are coarse, and the token savings are a heuristic estimate.

Single generative backbone. All pilot stages used one model family (Gemini, `gemini-3.5-flash`); we have not tested whether the prompt-level safeguards hold under a different backbone LLM.

8 Conclusion

We presented an architecture for the "data bottleneck" in the empirical sciences: legacy PDFs digitized via decoupled spatial parsing, schema-guided extraction, and independent programmatic validation. The two-layer triage and vision-filtering system keeps the expensive generative model off irrelevant material, and the typed validation layer independently rejects values that violate simple physical bounds (negative temperature, pressure, or time, and an out-of-range loading ratio). Source-quote grounding is currently requested through the prompt, and verifying it programmatically is the natural next step.

Applied to extreme-condition LENR thermodynamics, this pipeline turns non-standardized legacy literature into structured data at about 49 seconds per document on average, including papers removed by triage. We believe that separating what is programmatically checked from what is merely requested of the model is useful more broadly for reliable LLM-based extraction from the scientific literature.

Acknowledgments

The research presented in this paper is part of a broader research effort led by Professor David J. Nagel of George Washington University and Professor Anasse Bari of New York University, under whose joint supervision this work was carried out. The overarching project aims to advance Artificial Intelligence (AI) and Predictive Analytics tools to support the understanding of Low Energy Nuclear Reactions (LENR) and, more broadly, scientific discovery. The program builds an end-to-end pipeline that turns unstructured scientific literature into structured, machine-readable data, organizes that data into knowledge graphs, and trains predictive models on it, with the longer-term goal of supporting autonomous scientific discovery. This paper presents the data-extraction stage of that work. Gaurav Kumar, lead researcher at the Predictive Analytics and AI Research Lab within the Courant Institute of Mathematical Sciences at New York University, played a key role in the design and development of the pipeline described here. This study was also supported in part by the Suzanne McIntosh Master's Research Fellowship awarded to him. We also acknowledge the entire research team at the Predictive Analytics and AI Research Lab at NYU for their continued collaboration and assistance, and we thank the anonymous reviewers for their constructive feedback.

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling technical report. *Preprint*, arXiv:2408.09869.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

Samuel Colvin et al. 2024. Pydantic: Data validation using python type hints. `https://docs.pydantic.dev`.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

Martin Fleischmann and Stanley Pons. 1989. Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 261(2):301–308.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tanishq Gupta, Mohd Zaki, N M Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8:102.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Edmund Storms. 2010. Status of cold fusion (2010). *Naturwissenschaften*, 97(10):861–881.

Matthew C Swain and Jacqueline M Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Bhanu Prakash Vangala, Sajid Mahmud, Pawan Neupane, Joel Selvaraj, and Jianlin Cheng. 2025. HalluMat: Detecting hallucinations in LLM-generated materials science content through multi-stage verification. *arXiv preprint arXiv:2512.22396*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. DocGenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.