JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

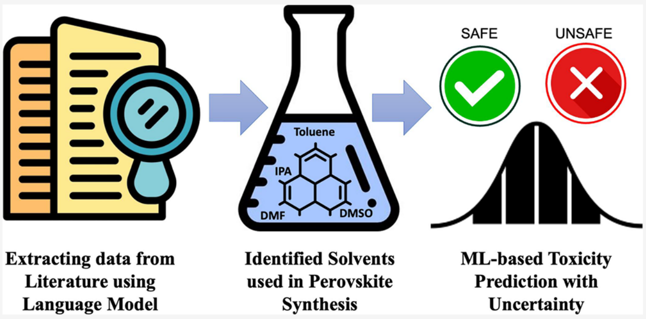
Uncertainty-Informed Screening for Safer Solvents Used in the Synthesis of Perovskites via Language Models

Arpan Mukherjee, Deepesh Giri, and Krishna Rajan*

Cite This: J. Chem. Inf. Model. 2025, 65, 7901−7918 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via UNIV BREMEN on January 10, 2026 at 15:53:00 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.


Toluene
IPA
DMF
DMSO
SAFE
UNSAFE
Extracting data from Literature using Language Model
Identified Solvents used in Perovskite Synthesis
ML-based Toxicity Prediction with Uncertainty

ABSTRACT: Automated data curation for niche scientific topics, where data quality and contextual accuracy are paramount, poses significant challenges. Bidirectional contextual models such as BERT and ELMo excel in contextual understanding and determinism. However, they are constrained by their narrower training corpora and inability to synthesize information across fragmented or sparse contexts. Conversely, autoregressive generative models like GPT can synthesize dispersed information by leveraging broader contextual knowledge and yet often generate plausible but incorrect ("hallucinated") information. To address these complementary limitations, we propose an ensemble approach that combines the deterministic precision of BERT/ELMo with the contextual depth of GPT. We have developed a hierarchical knowledge extraction framework to identify perovskites and their associated solvents in perovskite synthesis, progressing from broad topics to narrower details using two complementary methods. The first method leverages deterministic models like BERT/ELMo for precise entity extraction, while the second employs GPT for broader contextual synthesis and generalization. Outputs from both methods are validated through structure-matching and entity normalization, ensuring consistency and traceability. In the absence of benchmark data sets for this domain, we hold out a subset of papers for manual verification to serve as a reference set for tuning the rules for entity normalization. This enables quantitative evaluation of model precision, recall, and structural adherence while also providing a grounded estimate of model confidence. By intersecting the outputs from both methods, we generate a list of solvents with maximum confidence, combining precision with contextual depth to ensure accuracy and reliability. This approach increases precision at the expense of recall—a trade-off we accept given that, in high-trust scientific applications, minimizing hallucinations is often more critical than achieving full coverage, especially when downstream reliability is paramount. As a case study, the curated data set is used to predict the endocrine-disrupting (ED) potential of solvents with a pretrained deep learning model. Recognizing that machine learning models may not be trained on niche data sets such as perovskite-related solvents, we have quantified epistemic uncertainty using Shannon entropy. This measure evaluates the confidence of the ML model predictions, independent of uncertainties in the NLP-based data curation process, and identifies high-risk solvents requiring further validation. Additionally, the manual verification pipeline addresses ethical considerations around trust, structure, and transparency in AI-curated data sets.

1. INTRODUCTION

Automated data curation using advanced NLP techniques and language models offers a promising solution for managing and extracting insights from the vast data in materials science.[1−3] This field has been significantly influenced by the development and application of various language models, including both nongenerative prelarge language models (pre-LLMs), such as BERT and ELMo,[4−7] and contemporary generative LLMs like GPT-3.5 and GPT-4.0.[8−12] BERT and ELMo fill in missing information using context, while GPT generates new text by predicting the next token sequentially. BERT has been shown to achieve higher accuracy rates compared to models such as ELMo in numerous NLP tasks, including sentiment analysis, question answering, and named entity recognition, further solidifying its reputation as the preferred model among researchers.[7,13−15] BERT works by using bidirectional attention to capture context from both directions in a text sequence, making it highly effective for understanding nuanced language.[16] Variations of BERT, such as MatSciBERT,[7,15] OpticalBERT,[15] and BatteryBERT,[14] differ from other models by tailoring pretraining objectives, architecture, or domain-specific and have significantly advanced the field of materials science by enhancing the extraction and organization of information from scientific literature. However, even such

Received: March 20, 2025
Revised: May 16, 2025
Accepted: July 16, 2025
Published: July 22, 2025

ACS Publications

© 2025 The Authors. Published by American Chemical Society

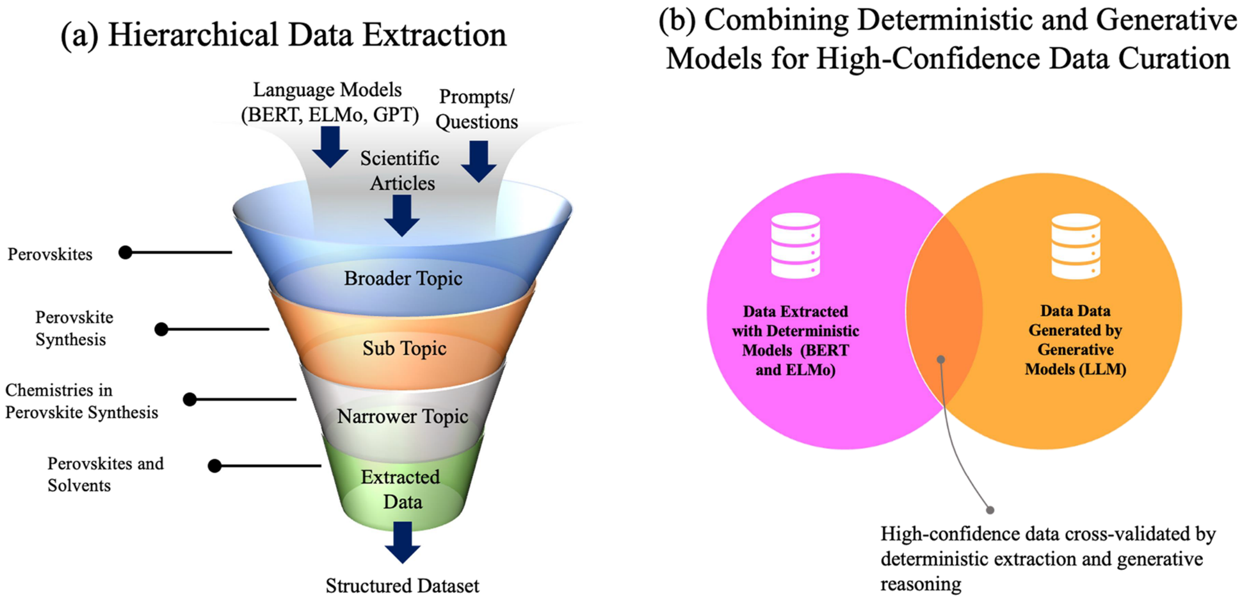

(a) Hierarchical Data Extraction
Language Models (BERT, ELMo, GPT)
Prompts/ Questions
Scientific Articles
Broader Topic
Sub Topic
Narrower Topic
Extracted Data
Structured Dataset
Perovskites
Perovskite Synthesis
Chemistries in Perovskite Synthesis
Perovskites and Solvents
(b) Combining Deterministic and Generative Models for High-Confidence Data Curation
Data Extracted with Deterministic Models (BERT and ELMo)
Data Data Generated by Generative Models (LLM)
High-confidence data cross-validated by deterministic extraction and generative reasoning

Figure 1. A framework for hierarchical data extraction method using an ensemble approach of deterministic pre-LLMs and generative LLMs in data extraction. (a) A hierarchical framework for extracting knowledge from scientific articles, narrowing broader topics into subtopics and refined extracted data through iterative questioning and processing. (b) The integration of deterministic outputs from models like BERT and ELMo with generative insights from LLMs, where the intersection represents high-confidence, verified data combining precision and contextual depth.

models are limited by their smaller training corpus and narrower knowledge base, which restrict their ability to capture the broader contextual nuances necessary for addressing sparse and highly specialized data, such as that found in perovskite synthesis. In contrast, LLMs such as GPT 3.5, 4.0, Llama or Gemini offer enhanced capacity for generalization across diverse contexts due to their larger training data sets and architectures.[9,10,17] Furthermore, LLMs can automate the data curation process by extracting and analyzing data from multiple sources, including product specifications and scientific articles.[18] However, apart from the resource demands of using an LLM, the application of these models in scientific data curation comes with its challenges, such as "hallucinations," where the model generates plausible but incorrect information.[19,20] This issue is particularly critical in scientific contexts, where accuracy is paramount. The underlying mathematical frameworks, including the optimization objectives, architectural design, and statistical properties of BERT/ELMo and GPT, are fundamentally distinct, and one cannot replace the other.

Recent studies on responsible AI deployment emphasize the importance of transparency, privacy, verification and robustness in automated scientific systems, particularly when outputs influence safety-critical decisions.[21,22] However, the susceptibility of generative models to hallucination and output variability across runs limits their immediate applicability in high-stakes downstream tasks such as toxic chemical screening. Even when explicitly prompted to produce structured responses, generative models frequently return incomplete, ambiguous, or unstandardized entities, implying the need for deterministic postprocessing techniques such as entity normalization, synonym resolution, and role classification.[23,24] Thus, language models should be viewed as modular components embedded within broader, verifiable curation frameworks and not as an end-to-end solution. Furthermore, prompting techniques such as RAG (retrieval-augmented generation), CoT (chain-of-thought prompting) and CoVe (Chain-of-Verification) have shown promise in certain reasoning and QA benchmarks, no single prompting paradigm has yet emerged as consistently superior in domains characterized by ambiguity, sparse annotations, and structural diversity.[25,26] A combination of techniques such as self-checking and multipass evaluation is essential for the robustness of the overall data curation method using prompting. A model that identifies ″dimethylformamide″ in one run and ″DMF″ in another must resolve these as equivalent to ensure usability, consistency, deduplication, and traceability. Furthermore, scientific problems such as perovskite synthesis encompass a wide and variable set of chemistries, including solvents, precursors, processing conditions, and intermediate steps, many of which are described in inconsistent, domain-specific language. These concerns fuel the need for verifiable and structured knowledge extraction pipelines that mitigate hallucination, ensure interpretability, and support factual and ethical downstream applications.

A combination of both deterministic and generative types of models leverages the strengths of each model to enhance data processing, analysis, and knowledge extraction across various scientific domains. For instance, BERT-based scoring methods have been used to assess the efficiency of GPT models for text summarization and measuring hallucinations, establishing BERT output as more deterministic and reliable.[27,28] Foppiano et al.[29] have used output from BERT to benchmark the performance of GPT in Question-Answering and property prediction. The sequential stacking of GPT layers and BERT layers has been attempted to leverage the full potential of these language models[30−33] where BERT layers have been used for more deterministic tasks such as classification or semantic understanding, and GPT has been used for generative tasks such as generating material descriptors or new hypotheses. While these studies demonstrate significant progress through techniques such as fine-tuning[34] and domain-specific pretraining,[7,9] they are often constrained by either their focus on specific use cases or their inability to effectively integrate the

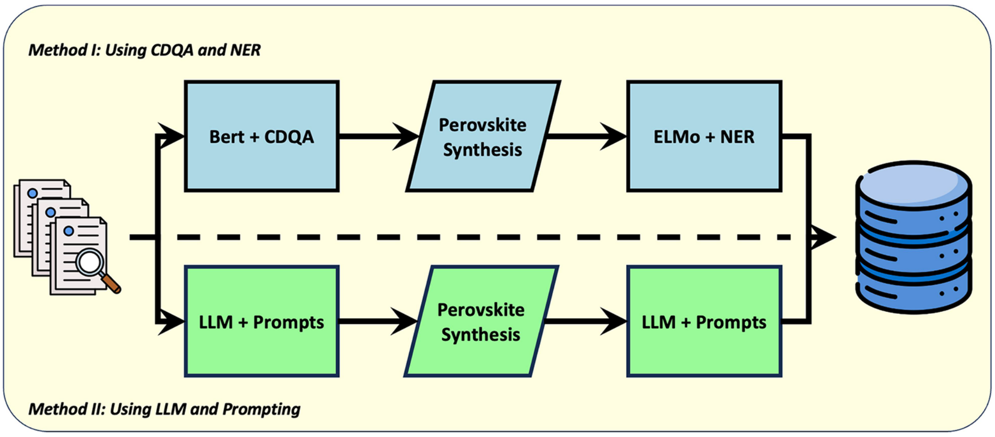


Figure 2. Automated data extraction and curation using language models. Two different methods for implementing the hierarchical extraction process. Method 1 uses a combination of CDQA and NER to extract and refine information. Method II employs LLMs with prompting to achieve the same objective, showcasing different approaches to achieve accurate data extraction from research articles.

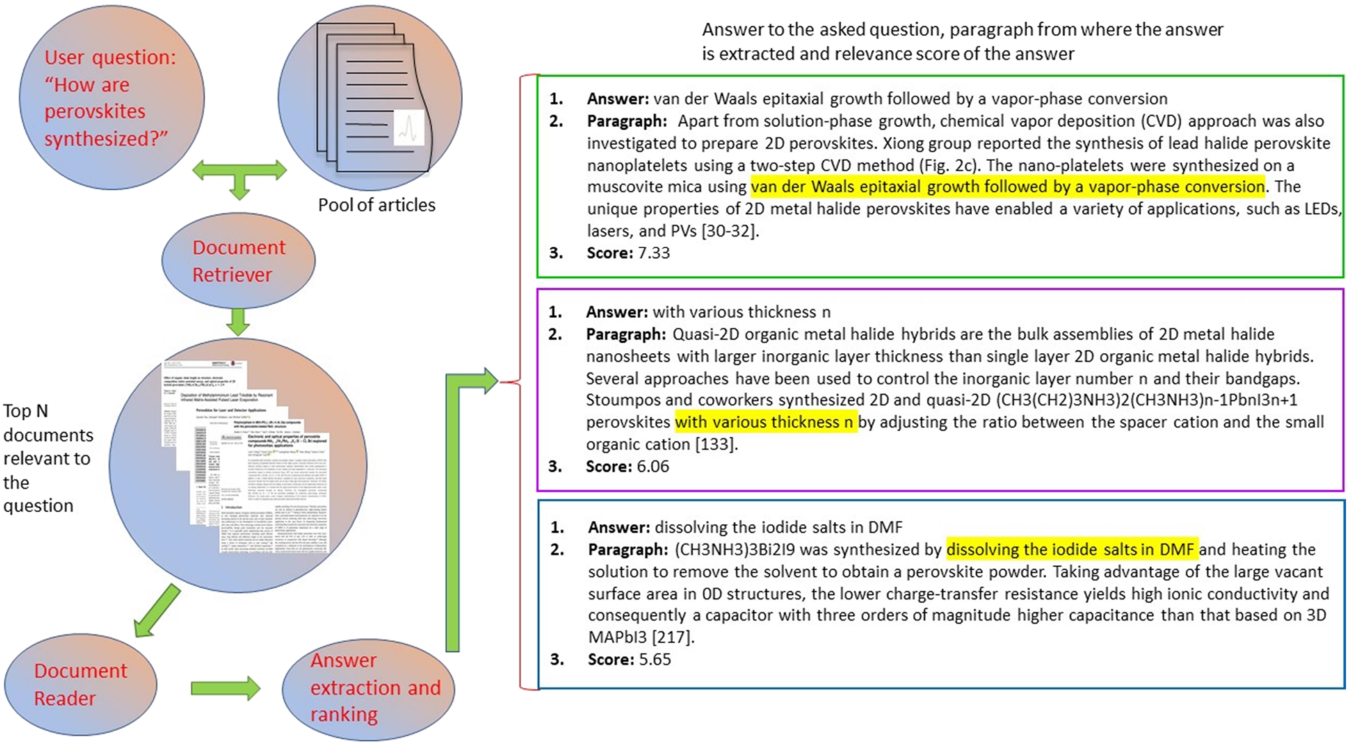
extracted data into a cohesive and hierarchical structure. Fine-tuning addresses task-specific needs, but the workflows largely remain one-directional, limiting the iterative refinement of outputs and their alignment with evolving domain-specific knowledge. Furthermore, the emphasis often remains on either deterministic tasks or generative capabilities without fully exploring their interplay in solving multilayered, complex problems.

In this study, we address the above-mentioned limitations by integrating hierarchical knowledge extraction with a novel ensemble framework, combining the precision and reliability of deterministic models with the contextual generation and broader knowledge base of LLMs to effectively capture both high-level context and specific granular details (see Figure 1). The hierarchical knowledge extraction follows a step-by-step refinement, starting from broader topics, narrowing down to subtopics, and finally extracting specific, granular details (see Figure 1(b)). At each stage, the framework ensures that the high-level context is retained while refining and verifying the extracted information. We have used an ensemble approach of combining both pre-LLM models, such as BERT and ELMo, and LLM models, such as GPT 3.5, 4.0, to facilitate this hierarchical knowledge extraction from scientific literature. To leverage effectively the precision, reliability, and domain-specific accuracy of BERT and ELMo and the contextual generation, broader knowledge-base, and sparse data handling of GPT, we have combined the data extracted by the two types of models (see Figure 1(b)). The intersection of the outputs from these models, where both agree or complement each other, represents high-confidence information. The mathematical justification for this approach lies in probabilistic intersection principles, which demonstrate that the probability of both models failing simultaneously is smaller than the probability of failure by either one individually. This error reduction enhances confidence in the extracted data, offering a mathematically grounded rationale for the ensemble approach. Furthermore, we have addressed the limitations of hallucination, omission, and lack of structured consistency by implementing a verification pipeline that combines paper-specific knowledge graph construction with multirun LLM sampling. Extracted entities such as perovskite names, solvents, and precursors are validated against manually curated knowledge graphs, representing the synthesis process's most chemically constrained and consistently reportable components. Given the inherent variability and complexity of perovskite synthesis, only a subset of entities can be modularized and verified via KG-based matching. The remaining synthesis-related descriptors are evaluated using cosine similarity between the extracted output and a curated ground truth to assess semantic fidelity. We apply this manual verification framework to a hold-out set of 50 articles, disjointed from the main extraction corpus, to ensure unbiased evaluation. Thus, we treat LLMs as probabilistic black-box extractors and apply structural verification to mitigate the inherent uncertainty of such systems.

The rest of the paper is organized as follows: Section 2 describes the implementation of the two methods used for automated data extraction, detailing how hierarchical knowledge extraction is performed using deterministic and generative models. Section 2 includes a description of the manual verification process, where a hold-out set of papers is used to evaluate extraction quality, normalize entities, and construct paper-specific knowledge graphs for benchmarking. Section 3 presents the results of the manual verification process on the hold-out set of papers, including precision, recall, and variability analyses across multiple LLM generations. Additionally, we visualize the curated data using the remaining corpus, focusing on keyword distributions and solvent-perovskite associations identified in the literature. This provides a structured data set for further analysis. In Section 4 applies the curated data set in a case study to explore the endocrine-disrupting potential of solvents using a deep learning-based uncertainty quantification (UQ) framework. Importantly, this UQ framework does not measure the uncertainty in the data extracted from the LLM but instead analyzes the epistemic uncertainty arising from training data limitations in the pretrained binary classification model by leveraging Shannon entropy, providing insights into prediction confidence and areas requiring further investigation. Finally, Section 5 concludes the paper by summarizing key contributions.

2. METHODOLOGY

In our work, a hierarchical knowledge extraction methodology using language models is implemented that progresses from broad to narrow topics. This approach ensures a comprehensive extraction of relevant information while maintaining


User question: "How are perovskites synthesized?"
Pool of articles
Document Retriever
Top N documents relevant to the question
Document Reader
Answer extraction and ranking
Answer to the asked question, paragraph from where the answer is extracted and relevance score of the answer

1. **Answer:** van der Waals epitaxial growth followed by a vapor-phase conversion
2. **Paragraph:** Apart from solution-phase growth, chemical vapor deposition (CVD) approach was also investigated to prepare 2D perovskites. Xiong group reported the synthesis of lead halide perovskite nanoplatelets using a two-step CVD method (Fig. 2c). The nano-platelets were synthesized on a muscovite mica using van der Waals epitaxial growth followed by a vapor-phase conversion. The unique properties of 2D metal halide perovskites have enabled a variety of applications, such as LEDs, lasers, and PVs [30-32].
3. **Score:** 7.33

1. **Answer:** with various thickness n
2. **Paragraph:** Quasi-2D organic metal halide hybrids are the bulk assemblies of 2D metal halide nanosheets with larger inorganic layer thickness than single layer 2D organic metal halide hybrids. Several approaches have been used to control the inorganic layer number n and their bandgaps. Stoumpos and coworkers synthesized 2D and quasi-2D (CH3(CH2)3NH3)2(CH3NH3)n-1PbnI3n+1 perovskites with various thickness n by adjusting the ratio between the spacer cation and the small organic cation [133].
3. **Score:** 6.06

1. **Answer:** dissolving the iodide salts in DMF
2. **Paragraph:** (CH3NH3)3Bi2I9 was synthesized by dissolving the iodide salts in DMF and heating the solution to remove the solvent to obtain a perovskite powder. Taking advantage of the large vacant surface area in 0D structures, the lower charge-transfer resistance yields high ionic conductivity and consequently a capacitor with three orders of magnitude higher capacitance than that based on 3D MAPbI3 [217].
3. **Score:** 5.65

Figure 3. In general, the question-answering (QA) system in NLP can be divided into two categories − Open Domain Question Answering (ODQA) and Closed Domain Question Answering (CDQA). The ODQA is capable of answering questions from any field, while the CDQA answers questions only from a specific domain of knowledge. Google Assistant, Amazon Alexa, etc., are examples of ODQA, while chatbots are examples of closed-domain systems. In this work, we use CDQA to identify the relevant paragraphs on perovskite synthesis that serve as metadata for further analysis. The 'Document Retriever' scans the given pool of articles to filter out the 'N' most relevant documents to the given question. The 'Document Reader' processes these documents to get the closest possible answers. In this work, we extracted three answers from each article. We also acquired the corresponding paragraphs where the answers are based and used them to get the perovskites and the solvents. Answers with higher scores appear more relevant than the others.

contextual accuracy and precision and is, hence, well-suited for sparse data.[35]

2.1. Data Curation.

We have downloaded 2000 peer-reviewed articles providing 30,000 paragraphs that serve as metadata for information retrieval. The DOIs for the articles were queried by searching for the phrases "halide perovskites," "hybrid organic, inorganic perovskites," "toxic perovskites," "perovskite solar cells," and "chemical synthesis of perovskites" on CrossRef.36 Following this, the articles were acquired from open-access journals such as Nature, American Chemical Society, Elsevier, and Royal Society of Chemistry. These articles form the metadata on which we implement contextual NLP to get data for further analysis.

2.2. Hierarchical Knowledge Extraction Process.

2.2.1. Method I: CDQA + NER Pipeline with Smaller Language Models. Method I has a straightforward sequence involving the use of a contextual model and a combination of Closed Document Question Answering (CDQA) and Named Entity Recognition (NER). Early contextual language models such as ELMo,[37] BERT,[16] and GPT-2[38] have significantly improved understanding of the sequence-level semantics and have shown state-of-the-art performance in several NLP tasks such as sequence classification,[39] question answering,[40,41] language modeling,[42] and translation[43,44] requiring fewer parameters and training time. Other NLP techniques, such as Closed Document Question Answering (CDQA) and Named Entity Recognition (NER), benefit from these advances, as data extraction has seen higher efficiency and accuracy (see Figure 2(b).). However, the reliance on specific contextual models integrated with CDQA and NER to identify chemical entities such as solvents presents challenges, primarily due to the scarcity of high-quality, chemically focused training data. This scarcity often results in a higher likelihood of type I errors (false positives) compared to type II errors (false negatives).

As per Figure 2, we have integrated BERT as a language model with Closed Document Question Answering (CDQA) followed by ELMo with Named Entity Recognition (NER) to automate the data extraction process,[45] enabling the hierarchical knowledge extraction from a broader topic to a structured data set. Figure 3 explains how the CDQA works. CDQA is an NLP subtask that involves asking context-specific questions within a closed domain, such as perovskite synthesis, extracting relevant paragraphs or sentences from a scientific article without having to manually annotate them. There are two main components of the CDQA system − Document Retriever and Document Reader. The Document Retriever identifies a list of 'Top N' candidate documents that are likeliest to the context of perovskite synthesis using similarity metrics. We have used cosine-similarity between the TF-IDF features of the documents and the phrase "perovskite synthesis." Next, these documents are divided into paragraphs and fed to the Document Reader, BERT, which gives the most probable paragraphs to the question "How is perovskite synthesized?" The answers were compared and ranked in the order of the model score, which is given by the softmax probability derived from the last layer of the BERT model. At the end of this step, three paragraphs most relevant to perovskite synthesis are extracted from each 'Top N' candidate document.

NER is the second subtask of our NLP pipeline that classifies keywords extracted from a given paragraph. Commonly available NER tools are ChemicalTagger,[46] OSCAR4,[47] Chemical Named Entities Recognition,[48] and ChemDataExtractor,[1] each trained for identifying specific terminologies and contexts within the materials science domain. In this work, to extract all the chemicals (perovskites, solvents, etc.), we used an ELMo-based NER tool developed by Kim et al.[6] The NER model developed by Kim et al.[6] uses a classification model that is trained on an internal database of over 2.5 million materials science articles. The details of the architecture of the NER model are provided in Table 1of the Supporting Information, while Table 2 presents a comprehensive list of all the training labels, which represent the specific category of chemistries that the NER is trained to identify and classify in the text. At the end of this step, a structured data set is formed by listing perovskites and their corresponding solvents that can be used for downstream tasks such as toxicity prediction.

A critical limitation of Method I is that the segmentation is typically conducted at the paragraph level rather than considering token-level constraints. This approach can overlook specific details that may span multiple sentences or paragraphs within a single article. Crucial information about the interaction of solvents with perovskite materials might be dispersed across several sentences or paragraphs within a single research paper, but the paragraph-level segmentation used in CDQA overlooks these interconnected details. This fragmented approach can lead to information loss, similar to the challenges encountered in Retrieval-Augmented Generation (RAG) models, which also struggle with integrating information across fragmented document sections. Furthermore, hallucination and omission manifest differently in such deterministic models. The solvents identified by the NER (ELMo) model can result in hallucinations when a solvent mentioned in the paper is extracted but not actually used in the perovskite synthesis context. Conversely, if a relevant solvent is present in the paper but underrepresented in the model's training data, it may be omitted entirely.

2.2.2. Method II: Prompting and Verification with Large Language Models. Method II uses Large Language Models, GPT 3.5, along with designed prompts for the hierarchical automated data extraction. LLMs have brought new capabilities that differ from earlier contextual models by utilizing a high number of self-attention layers and a more extensive training corpus. These features enable them to generate more accurate and diverse responses and better generalize across various tasks without the explicit need for task-specific downstream architectures like CDQA and NER. As shown in Figure 2(b), prompt engineering becomes essential when utilizing the in-built response generation capabilities of LLMs, as it replaces the role of traditional NLP tools by allowing the model to adapt its responses based on finely tuned prompts.[9] This method leverages the built-in response generation capabilities of the LLMs, enabling the identification and classification of chemical entities such as solvents directly through well-designed prompts rather than integrating them with separate tools. Furthermore, the use of domain knowledge is essential for designing and refining the prompts to evaluate the relevance and accuracy of the LLM's responses. During inference, LLMs process text at the token level, predicting the next token in a sequence given the preceding tokens. This capability allows them to assign probabilities to different tokens, including those corresponding to named entities like solvents, based on the context provided. Thus, LLMs are capable of performing both CDQA and NER tasks through their all-purpose design, eliminating the need for additional specialized tools.

While prompting strategies such as Retrieval-Augmented Generation (RAG),[49−51] Chain-of-Thought (CoT),[52,53] and Chain-of-Verification (CoVe)[54,55] have each demonstrated strengths in isolated settings, they also exhibit notable limitations when applied independently. RAG relies heavily on retrieving relevant text passages from a corpus, but it performs poorly when critical information is embedded in formats that are difficult to retrieve as plain text, such as tables and graphs.[51] CoT improves reasoning transparency but can propagate logically sound yet factually incorrect chains, especially in scientific domains lacking annotated reasoning paths.[52,53] CoVe addresses factuality through verification, but at high computational cost and with limited ability to disambiguate context without external support.[54,55] Across these works, a common theme emerges that no single prompting paradigm provides robust performance across accuracy, factual grounding, and computational efficiency. This has motivated recent research toward modular and hybrid strategies that integrate retrieval, stepwise reasoning, and verification in a context-sensitive pipeline.[50] Our prompting strategy synthesizes ideas from several leading prompting paradigms such as RAG, CoT, and CoVe, as shown in Figure 4.

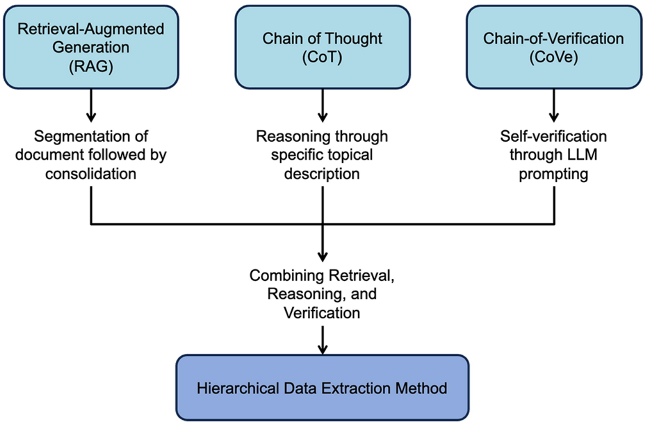

Retrieval-Augmented Generation (RAG)
Chain of Thought (CoT)
Chain-of-Verification (CoVe)
Segmentation of document followed by consolidation
Reasoning through specific topical description
Self-verification through LLM prompting
Combining Retrieval, Reasoning, and Verification
Hierarchical Data Extraction Method

Figure 4. Integration of prompting paradigms in Method II. The hierarchical data extraction framework draws from the segmentation and consolidation structure of RAG, the reasoning through specific topical description from CoT, and the iterative self-verification approach of CoVe into a domain-aware, structure-constrained pipeline for scientific information extraction.

Our method incorporates topic-based paper segmentation (inspired by RAG), hierarchical decomposition of queries (from CoT), and self-verification loops (from CoVe), but reconfigures them for the chemical domain.

Generative models, like GPT 3.5 or GPT 4.0, trained on vast corpora, have a broader knowledge base that enables them to synthesize answers by integrating information across entire texts and thereby establish connections between prompts and specific scientific concepts like perovskites, which is beyond the capability of Method I. As explained earlier, the hierarchical

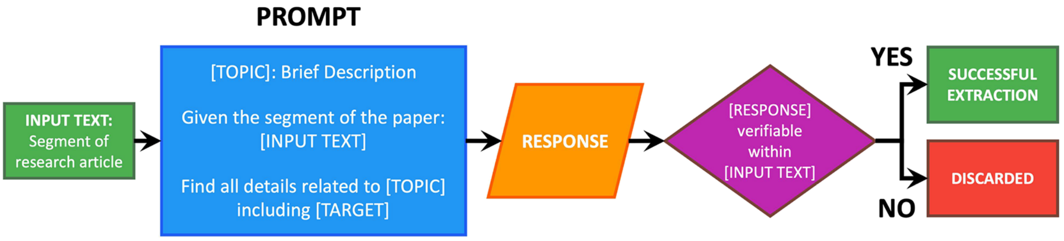

PROMPT
INPUT TEXT: Segment of research article
[TOPIC]: Brief Description
Given the segment of the paper: [INPUT TEXT]
Find all details related to [TOPIC] including [TARGET]
RESPONSE
[RESPONSE] verifiable within [INPUT TEXT]
YES
NO
SUCCESSFUL EXTRACTION
DISCARDED

Figure 5. Flowchart for information extraction from a research article using prompting and verification technique, starting with the "Input Text" box where the paper segment is specified, followed by a "Prompt" box detailing the search query. The process then moves to a "Response" diamond, indicating LLM response, which leads to either "Successful Extraction" or "Discarded" based on the verifiability within the input text.

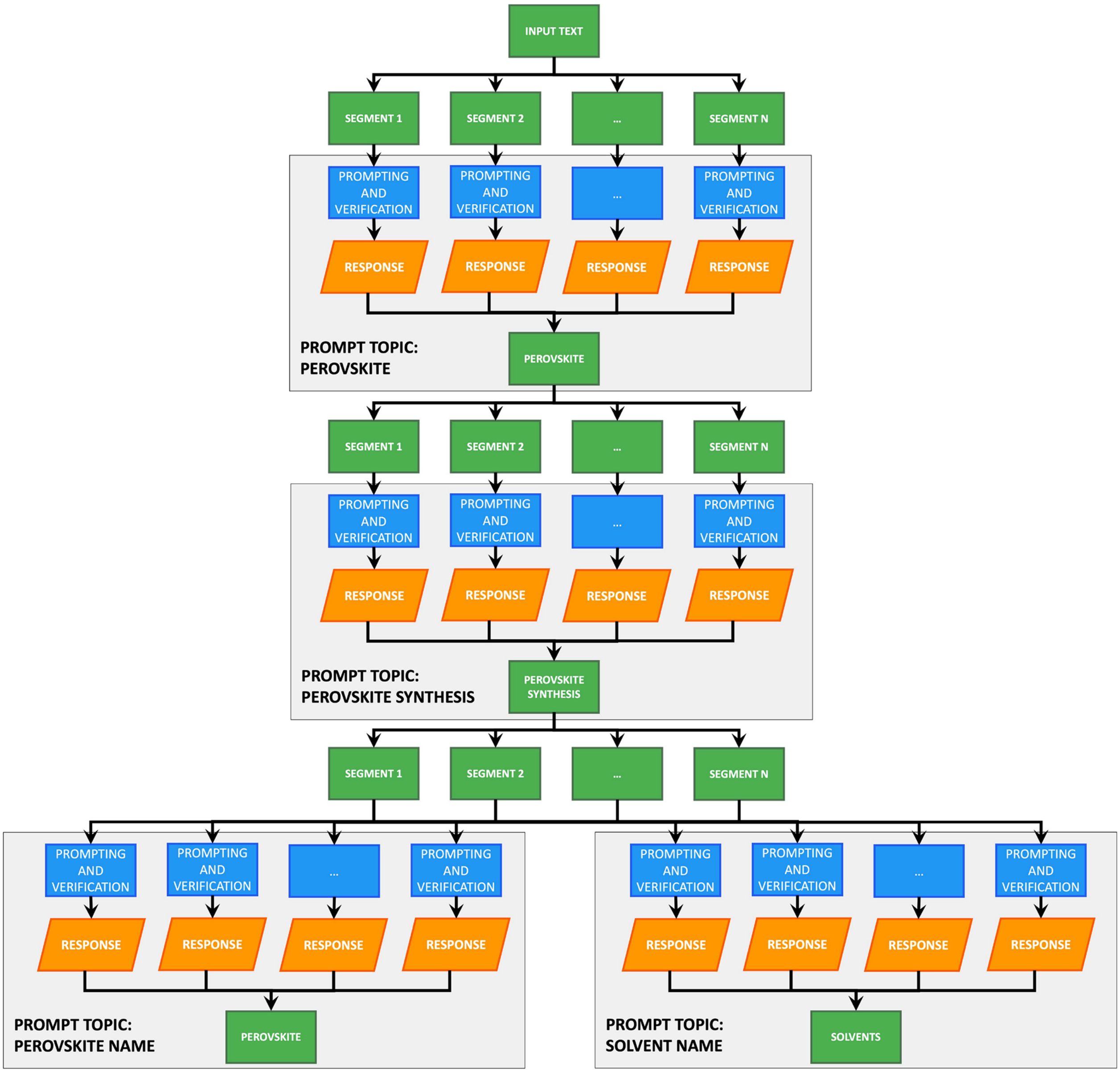


Figure 6. Iterative hierarchical knowledge extraction process using LLMs. The input text is segmented into smaller chunks, each undergoing prompting and verification to extract responses relevant to the broad topic (Perovskite). These responses are then combined and resegmented for the next level of specificity (Perovskite Synthesis), where the process is repeated. Finally, the combined responses are further segmented and processed at the narrowest levels, which include both the Perovskite Name and Associated Solvent, ensuring accurate and detailed extraction of specific information.

information extraction using LLM requires the careful design of prompts. We first explain the method of using prompts and LLMs for a particular level by detailing the steps involved in extracting and verifying information from research articles (see Figure 5).

Figure 5 shows that we employ a structured *prompting and verification* process to extract and verify specific information from a predefined segment of a research article. Responses from Method I, which provides the most relevant paragraphs, are used to design prompts through a trial-and-error process. OpenAI Playground[†] offers an interactive dashboard to experiment with various models and parameters, allowing users to fine-tune and test prompts in real time. Although models such as GPT 3.5, 4.0, and 4.5 incorporate progressive enhancements in reasoning, contextual understanding, and computational efficiency, they are all built upon the same underlying transformer-based architecture and mathematical principles. The details of the transformer-based architecture are given in the Supporting Information. Given an input text segment, a prompt is generated to find all details related to the topic. While extracting information from a text segment on a specific [TOPIC], the LLM is prompted with a brief [DESCRIPTION] of the [TOPIC], along with the text segment [INPUT TEXT]. The [TARGET] denotes the type of information to be extracted from a given segment. This differentiates our approach from traditional prompting by explicitly contextualizing the query within the prompt, ensuring that the LLM search is focused and relevant to the specific topic.[8] Since scientific texts often contain complex syntactic structures, nested entities, and domain-specific terminologies, it is important to include details related to questions in the prompt to extract the correct information.[56] This step is followed by a verification through subsequent prompting,[10] where the LLM checks if the response details from the previous prompt are explicitly found within the provided input text segment. This strategy helps mitigate hallucinations by increasing specificity until the LLM produces the correct answer that is guided by accurate responses known from previous steps. This verification and refinement in Method II are performed probabilistically using the LLM. While a more deterministic and less resource-intensive approach, such as leveraging BERT-score,[28] could have been used, we intentionally avoided this to preserve the independence of the two pipelines, each based on fundamentally different methodologies—deterministic models (BERT/ELMo) and generative LLMs.

The prompting and verification technique is applied iteratively at each level, progressively narrowing down from broad topics to specific details by refining prompts and verifying responses (see Figure 6). Too many promptings can be cost-intensive; thus, care is given so that the target data set can be obtained without excessive prompting. At each layer, the text from the previous layer is segmented based on the token limit of the LLM model. This segmentation approach utilizes the analytical capabilities of the LLM to interpret complex scientific data by concentrating on a smaller window for contextual understanding. The responses from multiple segments of a single paper are then consolidated using the LLM to form a coherent and comprehensive summary, which streamlines the relevant sparse and disparate information into an easily accessible form. The [TOPIC]s and their brief [DESCRIPTION]s for each layer are given in Table 1. Domain expertise, along with trial-and-error and the responses from Method I, have been used to come up with the descriptions. The first TOPIC is ‘Perovskite,’ where the description is targeted to establish a foundational understanding of the material.

Table 1. TOPICS and Their Brief Descriptions Used for Prompting and Extraction of Data Using the Layer-wise Prompting and Verification Process Shown in Figure 5

TOPIC	description	targeted information
level 1: perovskite	Perovskite has a unique crystal structure with the formula ABX3, where ‘A’ and ‘B’ are cations and ‘X’ is an anion, forming a three-dimensional network that contributes to the unique properties of perovskites, such as their excellent electronic and ionic conductivity.	perovskites, including their chemical compositions, synthesis processes, and various applications
level 2: perovskite synthesis	Perovskite synthesis involves steps such as precursor preparation, dissolution in solvents, deposition, and subsequent annealing and crystallization to form the ABX3 crystal structure.	chemistries related to perovskite synthesis, such as precursors, perovskite, and solvents
level 3: perovskite name	specific form of the ABX3 crystal, where ‘A’ and ‘B’ are cations and ‘X’ is an anion	name of the perovskite crystal in ABX3 form
level 3: solvent name	Solvents in perovskite synthesis are organic chemicals used to dissolve the precursors.	name of the organic solvent

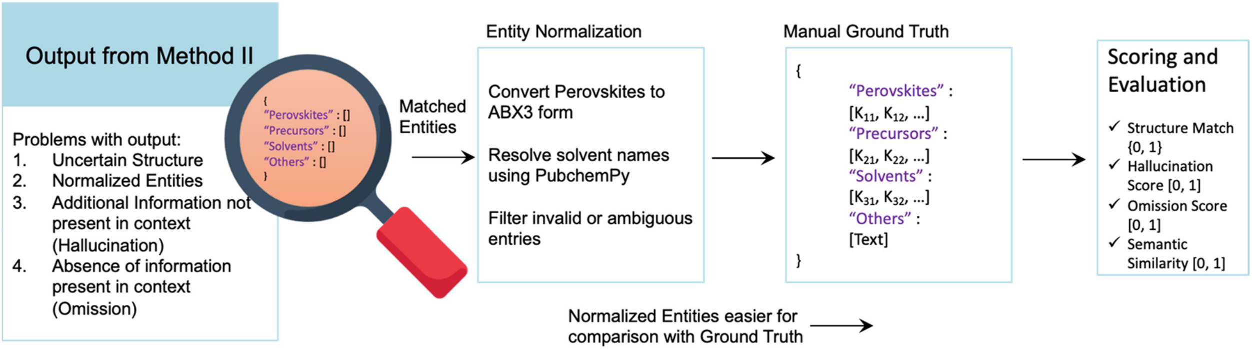

Output from Method II
Problems with output:
1. Uncertain Structure
2. Normalized Entities
3. Additional Information not present in context (Hallucination)
4. Absence of information present in context (Omission)
{ "Perovskites" : [] "Precursors" : [] "Solvents" : [] "Others" : [] }
Matched Entities
Entity Normalization
Convert Perovskites to ABX3 form
Resolve solvent names using PubchemPy
Filter invalid or ambiguous entries
Manual Ground Truth
{ "Perovskites" : [K11, K12, ...] "Precursors" : [K21, K22, ...] "Solvents" : [K31, K32, ...] "Others" : [Text] }
Scoring and Evaluation
✓ Structure Match {0, 1}
✓ Hallucination Score [0, 1]
✓ Omission Score [0, 1]
✓ Semantic Similarity [0, 1]
Normalized Entities easier for comparison with Ground Truth

Figure 7. Manual verification framework for evaluating LLM outputs (Method II). Extracted entities are first checked for structural conformance (first panel). If an output does not conform to the desired structure, then it is discarded. This is followed by an entity normalization step that standardizes perovskite formulas to the ABX_3 structure and resolves solvent names using PubChem (second panel). The normalized outputs are then compared against a paper-specific knowledge graph (third panel) to assess hallucination, omission, and entity consistency. The scoring module includes a structure match indicator (binary value in {0, 1}), hallucination and omission scores (real-valued between 0 and 1), and semantic similarity (real-valued between 0 and 1) for contextual synthesis.

The second [TOPIC] is 'Perovskite Synthesis,' aimed at understanding the processes involved in creating perovskites. The prompt at this level extracts detailed information about the synthesis steps, including precursor preparation, dissolution in solvents, deposition, and subsequent annealing and crystallization. The responses from Level 2 are manually compared against the responses from the CDQA in Method I to check for the correctness of the prompting method. The third level focuses on more specific details, divided into two subtopics: 'Perovskite Name' and 'Solvent Name.' This step is similar to the NER step of the previous method, where instead of using a classification model, we rely on the LLM's inherent understanding of context and scientific terms. The 'Perovskite Name' prompt seeks to identify specific forms of the ABX3 crystal by listing the various cations and anions that define different perovskite compounds. It is to be noted that at any level, there can be multiple subdivisions based on the specific information needed, where subdivisions refer to narrower topics or categories derived from the broader topic to extract detailed and relevant data. The 'Solvent Name' prompt extracts information on the organic chemicals used in the synthesis process to dissolve precursors. The division into 'Perovskite Name' and 'Solvent Name' has been deliberately done to ensure that the LLM can accurately identify the named entities by using separate prompts and descriptions for each. Additionally, as explained earlier, the larger training corpus for GPT 3.5 eliminates the need for a separate NER component for identifying perovskites and solvents. The description of the terms added to the prompts aids in better identifying the context of these terms, while the [TARGET] targets the LLM toward specific data to be extracted. Furthermore, the hierarchical extraction allows data to be extracted at each level, and the data from each level can be repurposed for other research objectives, such as identifying precursor materials from the 'Level 2: Perovskite Synthesis' responses or evaluating device performance from the 'Level 1: Perovskite' responses.

At the end of this step, we prompt the LLM to return its output in a structured JSON format. This serves two complementary purposes: when benchmark data is available, it enables direct alignment for evaluation; in its absence, it ensures syntactic consistency that facilitates downstream normalization, filtering, and validation.

2.3. Manual Extraction and Evaluation Framework.

To evaluate the reliability of the extracted entities and quantify hallucinations and omissions, we designed a structured manual verification protocol using a held-out set of 50 scientific articles, distinct from those used in the main data set construction (Figure 7). These papers span a range of research contexts: while many contain detailed information on perovskite synthesis, others discuss perovskite materials more generally without providing full synthesis protocols or specific chemical entities such as solvents and precursors. This variability reflects the real-world heterogeneity of the literature and ensures that our evaluation framework captures the challenges of entity extraction under both high-information and sparse-information conditions. For each paper, we manually extracted the ground truth entities and recorded them in a structured JSON format with the keys: 'perovskite', 'solvent', and 'others.' The 'others' category is also a text variable containing all synthesis-relevant descriptors such as antisolvents, deposition methods (e.g., spin-coating), annealing conditions, temperature ranges, and procedural steps that may not conform to a fixed schema but are critical for capturing the synthesis context. We used the ChatGPT Plus interface (GPT-4o) to generate structured outputs from a held-out set of 50 scientific articles, and each response was manually verified against expert-curated ground truth to evaluate accuracy, consistency, and contextual fidelity. For Method I, the comparison is straightforward because the output follows a fixed schema guaranteed by deterministic parsing rules from the BERT + NER pipeline, ensuring consistent entity boundaries and types.

For Method II, the comparison is not straightforward due to the unstructured and variable nature of the generative outputs. A self-check by GPT is integrated into our prompting + verification method as explained in the previous section. This prompting + verification is executed iteratively and repeated multiple times per paper to capture variability and enhance robustness. To assess correctness, we integrate this output with a knowledge graph constructed for each paper based on its manually extracted ground truth entities, which serves as a structured reference for verifying perovskite names and solvents.

Knowledge Graph Construction. For each paper, we curated a small-scale knowledge graph (KG) that captured chemically significant entities relevant to perovskite synthesis, such as

1. **Perovskites:** (e.g., $MAPbI_3$, $FAPbBr_3$)
2. **Solvents:** (e.g., DMF, DMSO, GBL)

These graphs were constructed manually using a combination of full-text reading and entity normalization via domain-specific naming conventions to ensure consistency and grounding in structure and context-based factual accuracy. The benchmark data set is constructed as a paper-specific knowledge graph for each article, enabling fine-grained and localized comparison. This design allows us to assess whether the extraction method accurately captures contextual usage, for example, distinguishing solvents merely mentioned in the text from those used in perovskite synthesis. An example of a paper-specific KG is given in the Supporting Information.

For each of the 50 papers, we applied our GPT-based extraction pipeline (Method II) and generated 10 independent outputs per paper to capture variability in generative behavior. Let each paper P have the ground truth entity set

$$E = E_{KG} \cup E_{Other}$$

Where E_{KG} are structured essential information, such as perovskites and solvents (e.g., perovskites and solvents) and E_{Other} are structured contextual information (e.g., procedures, temperatures, deposition steps). The ground truth for a paper P is structured as

$$E_{KG,P}\colon \{K_{KG,P,1}, K_{KG,P,2}, \cdots\}$$

$$E_{other,P}\colon K_{other,P}$$

Where each $K_{KG,\,P,1}$, $K_{KG,\,P,2}$, ⋯ corresponds to the perovskite, or solvent present in the paper P. $K_{other,\,P}$ includes free-form synthesis descriptors such as processing steps, antisolvents, or temperatures. We do not expect a structured format for this part in the LLM output. Let the prompt output be $\hat{E}$ which can be unstructured and unpartitioned. Within the 10 sampled outputs generated per paper, we apply a structure-matching filter to identify those outputs that adhere to the expected schema, specifically by checking for the presence of the keys defined in E_{KG}(e.g., perovskite and solvent). Only outputs that include this structured representation are retained for KG-based scoring, while the rest are discarded. This approach ensures that the evaluation is performed on syntactically consistent outputs and mirrors the strategy used in the actual corpus, where no ground truth is available, by enforcing the structure as a proxy for format correctness. At the same time, this filtering step provides insight into the model's compliance with the JSON formatting instructions, specifically, how often the LLM adheres to the expected structure, despite the prompt explicitly requesting it. This serves as an indirect measure of the model's reliability in producing syntactically usable outputs under controlled prompting.

Entity Normalization. Once the JSON structure is detected in an LLM output, we apply entity normalization to perovskites and solvents to ensure consistency and comparability across model outputs and ground truth. For perovskites, we restrict our analysis to compounds with the general formula ABX3, where A and B are cations and X is an anion (typically O, Cl, Br, or I). We first extract perovskite candidates from the GPT output using a prompt that identifies and returns chemical formulas matching the ABX_3 pattern. We then apply a rule-based filter that hard-codes a structural check, ensuring that only compounds with exactly three elements and the ABX3 stoichiometry are retained. We have used the Materials Project database to verify the structure and composition of extracted perovskites, ensuring that each candidate matches a known compound entry.

For solvents, normalization is performed using PubChemPy, which queries the PubChem database via the PUG REST API to retrieve standardized chemical identifiers. PubChemPy returns a list of synonyms for each compound, and we verify whether the extracted name appears within this list to confirm the match and ensure robust normalization. This process resolves synonyms, abbreviations (e.g., "DMF" vs "dimethylformamide"), and minor variations in naming, ensuring that each solvent is matched to its canonical representation in the PubChem repository. This normalization step is essential for reliable entity comparison, deduplication, and scoring. The normalization methods for the generative outputs of Method II were implemented as hard-coded rules, but were iteratively adapted based on patterns observed in the model's output during evaluation. These rules account for inconsistencies in entity formatting, such as variations in casing, spacing, and chemical notation, and were specifically crafted to align with how the model tended to represent perovskites and solvents. While not learned, the normalization logic was guided by empirical analysis of the generative behavior.

The response is partitioned into two segments for scoring: the entries corresponding to the expected keys in E_{KG} which are evaluated using knowledge graph matching, and the remainder of the output, which is treated as E_{Other} and assessed using semantic similarity. The keys of the structured essential information, such as perovskites and solvents, are then verified using knowledge graph matching. To evaluate model performance against the structured ground truth, we compute KG-based precision and recall scores. Precision reflects the proportion of predicted entities that are correct, serving as an inverse measure of hallucination. Recall measures the proportion of ground-truth entities that are successfully recovered, serving as an inverse measure of omission. These metrics allow for intuitive, interpretable evaluation of structured entity extraction aligned with standard information retrieval principles.

Precision (KG-Based Hallucination Indicator). We define the KG-based Hallucination score for the Structured Essential entities Score P_{KG}. This reflects how many of the predicted entities were correct

$$P_{KG} = \bigcup_{E_{KG}} \frac{|K_{KG,P} \cap \widehat{K_{KG,P}}|}{|\widehat{K_{KG,P}}|}$$

Recall (KG-Based Omission Indicator). We define the KG-based Omission score for the Structured Essential entities Score R_{KG}. This reflects how many of the ground-truth entities were recovered.

$$R_{KG} = \bigcup_{E_{KG}} \frac{|K_{KG,P} \cap \widehat{K_{KG,P}}|}{|K_{KG,P}|}$$

For the contextual synthesis section ("others"), which is not required for downstream tasks but relevant for assessing overall model fidelity, we compute a semantic similarity score between the model's output and the ground-truth description. Specifically, we embed both the extracted and reference texts using the GPT-3.5 encoder using the OpenAI library, and compute the cosine similarity between the two embedding vectors. This score captures high-level semantic alignment even when surface forms differ, and serves as a soft metric to evaluate how well the model reproduces relevant synthesis context in a free-form generation setting. We define the semantic similarity score as

$$S_{Other} = \cos(GPT(E_{Other}, K_{Other}), GPT(\widehat{E_{Other}}, \widehat{K_{Other}}))$$

Here, GPT(·,·) refers to the joint text embedding of the value—key pair using the GPT-3.5 embedding model accessed via the OpenAI library. A threshold of $S_{Other} \geq 0.5$ is used to indicate semantic alignment between the model output and the ground truth; values below this threshold are flagged as divergent or low-fidelity reproductions.

3. RESULTS

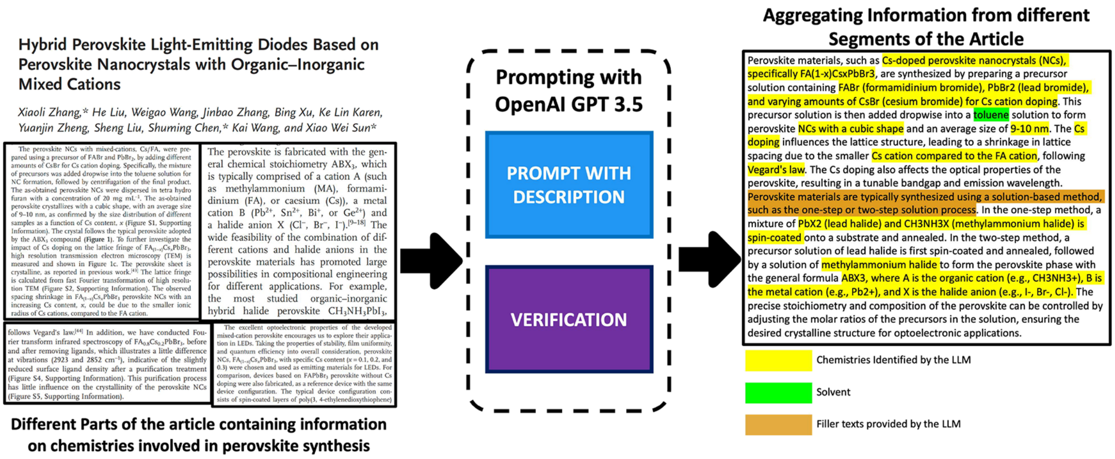
Our manual evaluation across 50 papers and 10 GPT outputs per paper (500 total generations) reveals several key trends in the performance of Method II (LLM-based extraction), Method I (BERT + NER), and their intersection. We computed the mean and standard deviation of precision and recall for each paper across 10 LLM generations, separately for perovskite and solvent entities. The resulting distributions are visualized as histograms and included in the Supporting Information (See SI Figures 1-4) to illustrate trends in model performance, stability, and entity-specific variation.

First, GPT-based outputs demonstrated complete structural consistency: 100% of responses adhered to the expected JSON format, with no mislabeling of keys. Across all generations, the LLM correctly distinguished entity types, e.g., never labeling a solvent as a perovskite or vice versa, indicating a strong internal understanding of categorical boundaries of the LLM. However, while the structure was stable, the content of extracted entities varied modestly, with different subsets of solvents or perovskites included in each run. This led to run-to-run variation in precision and recall. (See SI Figure 1 for perovskite precision and recall variation, showing tight structural agreement but measurable score dispersion) Second, Method I showed higher precision but lower recall than Method II (See SI 5). This is expected for deterministic models like BERT + NER, which tend to extract entities only when they match known patterns or training examples. As a result, Method I introduces lesser hallucinations, and subsequently higher precision. However, Method I often misses contextually relevant entities, especially when those entities are not a part of the training data. SI Figure 5 illustrates this gap, where perovskite extractions under Method I show higher precision (lesser hallucination) but a lower recall (higher omission) as compared to Method II.

Third, Method II exhibited higher recall but lower precision. Its hierarchical prompting allows broader context comprehension, enabling it to identify more valid entities. However, this comes with a slight increase in hallucinations. Notably, most hallucinated solvents were chemically plausible and commonly associated with perovskite synthesis (e.g., DMF, DMSO, GBL), suggesting the influence of the model's training priors rather than random error. In SI Figure 3, the solvent-related precision and recall histograms show greater dispersion than those of perovskites, with a lower mean and wider standard deviation, highlighting the inherent difficulty of solvent identification. Fourth, both methods performed better on perovskites than on solvents. Precision and recall scores were consistently higher for perovskite entities, likely due to their more formulaic and structured representation (e.g., $CsPbI_3$), which lends itself to both pattern-based and generative extraction. Solvents, by contrast, are linguistically diverse and context-sensitive, making them harder to extract reliably, particularly when mentioned in tables or nonsynthesis contexts. As shown in SI Figure 4, the solvent recall distribution is broader and more symmetric, indicating that solvent extraction is both less reliable and more sensitive to document structure. Fifth, recall scores were inversely correlated with the number of ground truth entities in a paper. When more entities are present, both methods, particularly Method I, struggle to maintain complete coverage. This effect is magnified in documents that embed key synthesis information in tables, which generative models process less reliably. These trends reinforce the importance of integrating layout-aware tools like Tabula into future iterations of the pipeline to extract information that escapes both LLM comprehension and sequence labeling models.

The two methods of hierarchical data extraction and curation have estimated different numbers of solvents used for perovskite synthesis. In the Supporting Information, we have shown how the outputs are generated using both methods for ref 57. A detailed working example comparing the outputs of BERT/ELMo (Method I) and GPT 3.5 (Method II) illustrates their contrasting strengths and weaknesses. In this example, Method I identified four perovskites (FA0.7Cs0.3PbBr3, FA0.8Cs0.2PbBr3, Cs4PbBr6, and CsPbBr3) and no solvents. At the same time, Method II generalized the perovskites as belonging to the class FA(1−x)CsxPbBr3 and identified Toluene as the primary solvent. However, the general formula was not specifically requested, and it would have been more valuable if the exact perovskites were identified by the LLM instead. BERT-based CDQA provides unrefined, noisy text directly from the PDF file, extracting entire paragraphs verbatim. While this ensures no information is lost, the extracted content is often dense, fragmented, and not readily machine-readable. In contrast, GPT 3.5 refines the extracted information, synthesizing it into structured and concise outputs. For example, instead of delivering lengthy paragraphs, Method II generates a list of relevant entities (e.g., Toluene as the solvent) and connects them to broader processes (e.g., its use in synthesizing FA(1−x)CsxPbBr3 perovskites). Method I's reliance on paragraph-level segmentation results in fragmented data extraction, often limiting its scope to predefined entity types like perovskites and failing to capture solvents, which may not be explicitly tagged in the training data set. For instance, the solvent Toluene, which appears sparsely, was overlooked by Method I but captured by Method II's capability to study a larger context window. Interestingly, neither method correctly identified the exact perovskite-solvent pair, suggesting that both approaches have limitations that need further investigation. This highlights the need for an ensemble method to leverage the strengths of each model while addressing their weaknesses.

We have attached a spreadsheet, 'solvents_list_output.xlsx', containing a table of all solvents identified by the two methods. We have identified 35 different solvents using Method I and 54 solvents using Method II that are used during perovskite synthesis. A key distinction between the two methods lies in the flexibility of the prompting technique employed in Method II, which can be further refined and segmented into multiple iterative steps to enhance data extraction. In contrast, Method


Hybrid Perovskite Light-Emitting Diodes Based on Perovskite Nanocrystals with Organic–Inorganic Mixed Cations
Xiaoli Zhang,* He Liu, Weigao Wang, Jinbao Zhang, Bing Xu, Ke Lin Karen, Yuanjin Zheng, Sheng Liu, Shuming Chen,* Kai Wang, and Xiao Wei Sun*
Different Parts of the article containing information on chemistries involved in perovskite synthesis
Prompting with OpenAI GPT 3.5
PROMPT WITH DESCRIPTION
VERIFICATION
Aggregating Information from different Segments of the Article
Chemistries Identified by the LLM
Solvent
Filler texts provided by the LLM

The perovskite NCs with mixed-cations, Cs/FA, were prepared using a precursor of FABr and $PbBr_2$, by adding different amounts of CsBr for Cs cation doping. Specifically, the mixture of precursors was added dropwise into the toluene solution for NC formation, followed by centrifugation of the final product. The as-obtained perovskite NCs were dispersed in tetra hydro furan with a concentration of 20 mg mL^{-1}. The as-obtained perovskite crystallizes with a cubic shape, with an average size of 9–10 nm, as confirmed by the size distribution of different samples as a function of Cs content, x (Figure S1, Supporting Information). The crystal follows the typical perovskite adopted by the ABX_3 compound (**Figure 1**). To further investigate the impact of Cs doping on the lattice fringe of $FA_{(1-x)}Cs_xPbBr_3$, high resolution transmission electron microscopy (TEM) is measured and shown in Figure 1c. The perovskite sheet is crystalline, as reported in previous work.[43] The lattice fringe is calculated from fast Fourier transformation of high resolution TEM (Figure S2, Supporting Information). The observed spacing shrinkage in $FA_{(1-x)}Cs_xPbBr_3$ perovskite NCs with an increasing Cs content, x, could be due to the smaller ionic radius of Cs cations, compared to the FA cation.

The perovskite is fabricated with the general chemical stoichiometry ABX_3, which is typically comprised of a cation A (such as methylammonium (MA), formamidinium (FA), or caesium (Cs)), a metal cation B (Pb^{2+}, Sn^{2+}, Bi^{+}, or Ge^{2+}) and a halide anion X (Cl^-, Br^-, I^-).[9–18] The wide feasibility of the combination of different cations and halide anions in the perovskite materials has promoted large possibilities in compositional engineering for different applications. For example, the most studied organic–inorganic hybrid halide perovskite $CH_3NH_3PbI_3$,

follows Vegard's law.[44] In addition, we have conducted Fourier transform infrared spectroscopy of $FA_{0.8}Cs_{0.2}PbBr_3$, before and after removing ligands, which illustrates a little difference at vibrations (2923 and 2852 cm^{-1}), indicative of the slightly reduced surface ligand density after a purification treatment (Figure S4, Supporting Information). This purification process has little influence on the crystallinity of the perovskite NCs (Figure S5, Supporting Information).

The excellent optoelectronic properties of the developed mixed-cation perovskite encourages us to explore their application in LEDs. Taking the properties of stability, film uniformity, and quantum efficiency into overall consideration, perovskite NCs, $FA_{(1-x)}Cs_xPbBr_3$, with specific Cs content (x = 0.1, 0.2, and 0.3) were chosen and used as emitting materials for LEDs. For comparison, devices based on $FAPbBr_3$ perovskite without Cs doping were also fabricated, as a reference device with the same device configuration. The typical device configuration consists of spin-coated layers of poly(3, 4-ethylenedioxythiophene)

Perovskite materials, such as Cs-doped perovskite nanocrystals (NCs), specifically FA(1-x)CsxPbBr3, are synthesized by preparing a precursor solution containing FABr (formamidinium bromide), PbBr2 (lead bromide), and varying amounts of CsBr (cesium bromide) for Cs cation doping. This precursor solution is then added dropwise into a toluene solution to form perovskite NCs with a cubic shape and an average size of 9-10 nm. The Cs doping influences the lattice structure, leading to a shrinkage in lattice spacing due to the smaller Cs cation compared to the FA cation, following Vegard's law. The Cs doping also affects the optical properties of the perovskite, resulting in a tunable bandgap and emission wavelength. Perovskite materials are typically synthesized using a solution-based method, such as the one-step or two-step solution process. In the one-step method, a mixture of PbX2 (lead halide) and CH3NH3X (methylammonium halide) is spin-coated onto a substrate and annealed. In the two-step method, a precursor solution of lead halide is first spin-coated and annealed, followed by a solution of methylammonium halide to form the perovskite phase with the general formula ABX3, where A is the organic cation (e.g., CH3NH3+), B is the metal cation (e.g., Pb2+), and X is the halide anion (e.g., I-, Br-, Cl-). The precise stoichiometry and composition of the perovskite can be controlled by adjusting the molar ratios of the precursors in the solution, ensuring the desired crystalline structure for optoelectronic applications.

Figure 8. Data Extraction using Method II demonstrates the ability of our method to fuse information from different sections of a research paper to extract detailed chemical information related to perovskite synthesis. The highlighted sections show various mentions of solvents, cations, and synthesis methods scattered throughout the document. The Method II method successfully integrates these disparate pieces of information. Results from Methods I and II are reported in the Supporting Information.

Table 2. Frequently Used Organic Solvents in Perovskite Synthesis Are Categorized into Two Subclasses (Agonist and Binding) of Active/inactive Endocrine Disruptors (EDs)[a]

			ED subclasses		
index	solvents	SMILES	agonist	binding	reference
1	dimethylformamide (DMF)	CN(C)C = O	active	active	ref 68,69
2	dimethysulfoxide (DMSO)	CS(=O)C	inactive	inactive	
3	toluene	CC1 = CC = =CC = C1	active	active	ref 70,71
4	oleic acid (OA)	CCCCCCCCC = CCCCCCCCC(═O)O	inactive	inactive	
5	oleylamine (OLA)	CCCCCCCCC = CCCCCCCCCN	inactive	inactive	
6	octadecene (ODE)	CCCCCCCCCCCCCCCCC = C	inactive	inactive	
7	acetone	CC(═O)C	inactive	inactive	ref 72,73
8	chloroform	C(Cl)(Cl)Cli	inactive	inactive	
9	chlorobenzene (CB)	C1 = CC = C(C = C1)Cl	active	inactive	ref 74
10	isopropanol (IPA)	CC(C)O	inactive	inactive	
11	ethanol	CCO	inactive	inactive	
12	benzyl alcohol	C1 = CC = C(C = C1)CO	inactive	inactive	
13	acetonitrile	CC#N	inactive	inactive	
14	*n*-hexane	CCCCCC	inactive	inactive	ref 75,76
15	cyclohexane	C1CCCCC1	inactive	inactive	
16	diethyl ether	CCOCC	active	active	
17	*γ*-butyrolactone (GBL)	C1CC(═O)OC1	inactive	inactive	
18	methyl acetate	CC(═O)OC	active	active	
19	ethyl acetate	CCOC(═O)C	inactive	inactive	
20	ethylene glycol	C(CO)O	inactive	inactive	
21	*n*-octane	CCCCCCCC	active	active	ref 77
22	pyridine	C1 = CC = NC = C1	inactive	inactive	
23	diethylene glycol (DEG)	C(COCCO)O	inactive	inactive	
24	tetrahydrofuran	C1CCOC1	inactive	inactive	
25	trioctylphosphine (TOP)	CCCCCCCCP(CCCCCCCC)CCCCCCCC	active	active	

[a] These two subclasses denote a molecule's ability to interact with the estrogen receptor (ER).[64] For a chemical, the state of being active or inactive in one of the subclasses is independent of its nature in the other subclass. However, if the chemical is "Active" in any of the subclasses, then it's potentially an EDC. This classification is done with the help of a deep-learning model that takes SMILES as the inputs and gives a multi-output binary classification. The studies that back up our data for this classification are mentioned in the last column.

I is constrained by its reliance on specific NLP tasks, such as Closed Domain Question Answering (CDQA) and Named Entity Recognition (NER). Consequently, its performance is inherently limited by these tools' predefined architectures and capabilities, restricting its ability to adapt to more nuanced or complex data extraction scenarios. A larger number of solvents identified by Method II is a probable outcome because the NER model used in Method I has limitations due to its dependency on the training data set. On the contrary, LLMs leverage contextual understanding and the brief descriptions provided with the prompts to better identify solvents. Additionally, the LLM can fuse information from different sections of a paper, while Method I relies on paragraph-level segmentation and extraction, which may miss solvents mentioned across different sections or in less explicit contexts. Figure 8 demonstrates an example of how our proposed method can fuse data from different parts of a paper, as given in ref 57. Information on chemistries related to perovskite synthesis, such as such as solvents, cations, and synthesis methods, is scattered throughout various sections of the paper. The paragraph on the right represents comprehensive information about perovskite synthesis, which can be used to identify relevant chemicals and processes. The solvent Toluene appears just once in the paper but has been identified by the prompting method, demonstrating its efficiency in fusing sparse information.

While Method II identified more solvents overall, there are notable solvents that were exclusively identified by Method I but missed by Method II, including 1-butanol, Dimethyl ether, Sodium hypochlorite, Benzene, Trioctylphosphine oxide, and Dichloromethane. The solvents in our list that were not identified by Method I are − Dichlorobenzene, 2-Methoxyethanol, Ethylenediamine, Ethanethiol, and 1-Methyl-2-pyrrolildinone (commonly known as the NMP solvent). A total of **25 solvents** that both methods have unanimously identified are listed in Table 2. These are the solvents with maximum confidence as they are extracted by both the deterministic and the generative approaches. We have used a Chord Diagram that represents the conditional probability distribution of solvents given perovskites in Figure 9. This conditional distribution quantifies the preferred solvent choices for synthesizing specific perovskites to devise solvent substitution strategies. The direction of the plot follows the arc from perovskites (sources) to solvents (targets), where the thickness of the chords is proportional to the frequency of their co-occurrence in synthesis literature. The top three strongest perovskite-solvent pairs identified are ($MAPbI_3$-DMF), ($MAPbBr_3$-DMF), and ($MAPbI_3$-DMSO), highlighting their dominant role in perovskite synthesis.

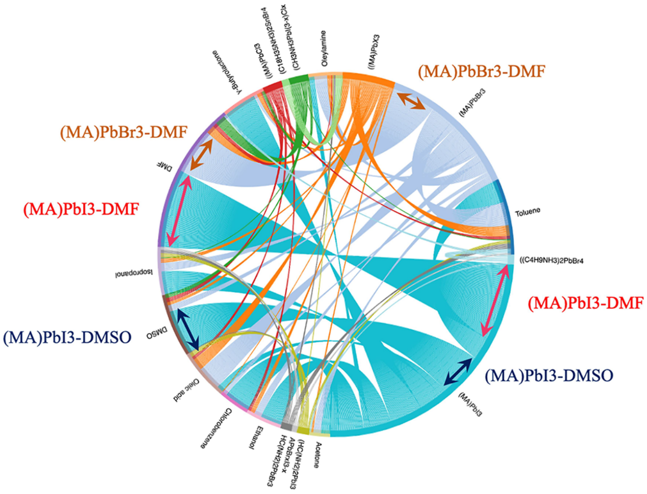


Figure 9. Chord diagram illustrating the associations between the top 10 most frequently reported perovskites and top 10 solvents used in their synthesis. The width of each chord represents the frequency of co-occurrence in the literature, with perovskites (source nodes) linked to solvents (target nodes). This visualization highlights dominant perovskite-solvent pairs, such as ($MAPbI_3$-DMF), ($MAPbBr_3$-DMF), and ($MAPbI_3$-DMSO).

Figure 10 illustrates the marginal distribution of the most commonly occurring solvents among the 25 unanimously identified solvents, focusing on those cumulatively accounting for at least 85% of the total occurrence. DMF dominates the distribution with a share of 23.4%, followed by Toluene (14.5%) and DMSO (11.8%). DMF is commonly used for the dissolution of lead and Methylammonium (MA) salts,[58,59] and hence, it is no surprise that it appears at the top of the list.

Having established a comprehensive analysis of the solvents identified by the two methods and their associations with perovskite synthesis, we now focus on the marginal distribution of the perovskites that appear in conjunction with the 25 identified solvents. Understanding the prevalence of specific perovskite types and their mutual associations with solvents provides critical insights into the broader synthesis trends and highlights the dominant perovskite-solvent pairs driving research and development in the field. We identified all the organic perovskites mentioned in the synthesis paragraphs that we extracted. We were able to acquire more than 350 uniquely mentioned organic perovskites, most of which are MA-based (>40%), while Formamidinium (FA) and Butylammonium (BA) based perovskites constitute around 10% each. A list of the most occurring 73 perovskites, along with the associated 25 solvents, is given in a spreadsheet titled 'top_solvent_perovskite.xlsx' in the Supporting Information. As solvents are required for different activities during perovskite synthesis,[60,61] we looked up their mutual distribution in the analyzed papers (see Figure 11). Our study reveals that most solvents are reported in conjunction with MA lead halide perovskites. This is unsurprising given that the MA-based perovskites have been attractive due to higher efficiency and better stability.[62,63] We further looked into the distribution of these organic perovskites based on their frequency of mutual occurrences with the solvents and plotted the chart shown in Figure 11. This chart shows that out of all the associations between organic perovskites and solvents, more than 3/4th involve MA lead halide perovskites. This reflects the scale of the study conducted on these perovskites so far. FA and BA-based perovskites seem to offer alternative choices, but their number is dwarfed by the MA-based ones. The perovskite (MA)PbI3 accounts for 42.7% of the distribution, making it the most frequently occurring perovskite in the data set. The second most common perovskite, (MA)PbBr3, accounts for 22.0% of the distribution.

4. CASE STUDY: UNCERTAINTY-INFORMED ENDOCRINE DISRUPTION NATURE OF SOLVENTS

Understanding the endocrine-disrupting (ED) nature of industrial solvents is a critical area of research, given the potential health implications associated with exposure to these chemicals. The EPA's Endocrine Disruptor Screening Program (EDSP) is a critical initiative aimed at assessing the potential endocrine activity of various chemicals, that includes the use of ML models to predict estrogen receptor (ER) activity efficiently.[64] However, such predictions alone are not sufficient. Including epistemic uncertainty arising from limitations in the training data sets of these machine learning models is critical. It allows the identification of areas where predictions are less reliable, ensuring that resources can focus on resolving ambiguities and refining data. In this section, we assess if a given solvent associated with perovskite synthesis is an endocrine-disrupting chemical using a pretrained classification model developed in our earlier work.[65]

We have used a multioutput binary classification model[65] designed to predict whether a chemical, such as a solvent, has

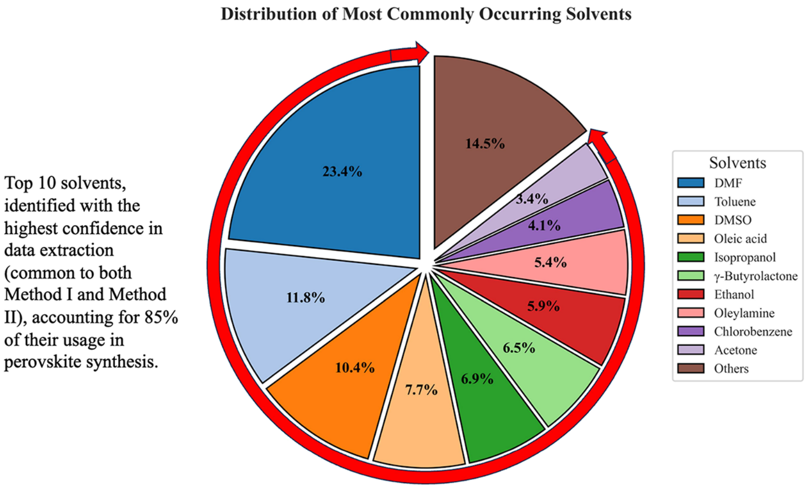


Figure 10. Distribution of the most commonly occurring solvents in perovskite synthesis, with DMF accounting for the largest share (23.4%), followed by Toluene (14.5%) and DMSO (11.8%). The chart highlights the dominant role of these solvents in synthesis practices based on high-confidence data extracted from both methods.

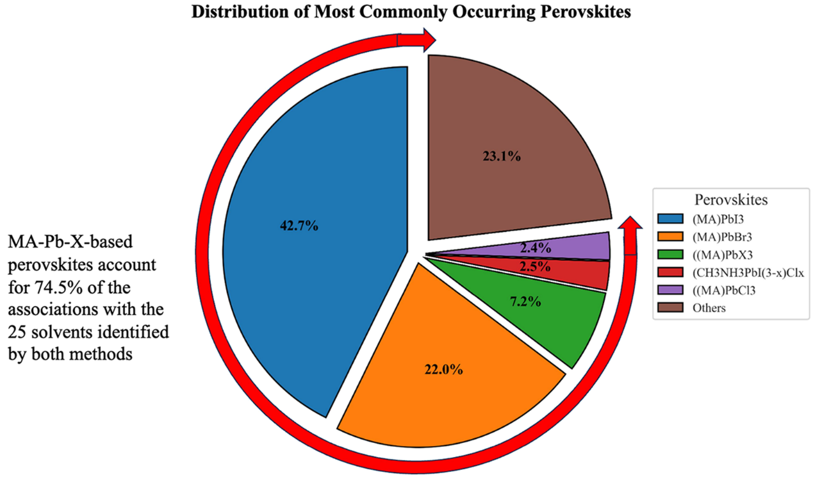


Figure 11. Pie chart shows distribution of organic perovskites based on solvent-perovskite mutual occurrences. About 75% of solvent-perovskite association was found in the literature with methylammonium (MA) lead halide perovskites.

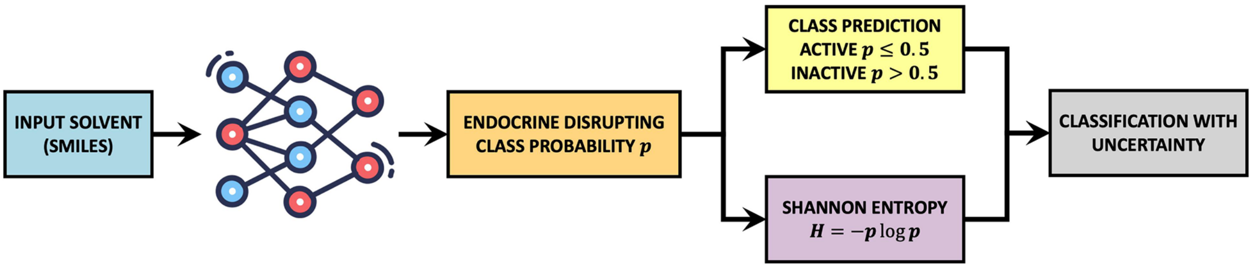


Figure 12. Workflow for assessing the prediction uncertainty of endocrine-disrupting chemicals. The process begins with the input of solvent data in the form of SMILES codes, which are processed by a deep neural network model to generate the class probability. p of a solvent being endocrine-disrupting. This probability is then used for class prediction (active/inactive) and for calculating Shannon entropy. $H = - p \log p$ To assess the uncertainty of the classification. The final output is the classification with an associated uncertainty measure.

endocrine-disrupting (ED) potential by analyzing its molecular structure. It utilizes the Simplified Molecular Input Line Entry System (SMILES) representation to encode chemical structures into machine-readable strings. The SMILES strings are first numerically encoded using a bag-of-words approach, where each character in the SMILES vocabulary is assigned a unique integer. Then, the encoded sequences are padded with zeros at the beginning to achieve a fixed length of 130 characters. The encoded and padded SMILES string serves as the input to the classification model. The deep neural network classification model is a stack of ten convolutions and two LSTM layers, followed by two dense layers. The convolution layers progressively extract the spatially correlated local features from the SMILES, while the LSTM layers are used for sequential data processing. The final dense layer has two diverging sigmoid layers that output the probabilities indicating whether the chemical is active or inactive regarding endocrine receptor interaction: 'Agonist' and 'Binding.' The multioutput binary classification model was trained on 3,236 chemicals from the Tox21 data set and 4,492 chemicals from the CERAPP data set, achieving testing accuracies of 90.7% for agonist activity and 89.6% for binding activity on a combined evaluation data set. Our proposed UQ specifically evaluates the epistemic uncertainty of the model's predictions concerning this training and evaluation data. Details of model architecture and accuracies are reported in the Supporting Information.

Figure 12 demonstrates the use of Shannon entropy to estimate epistemic uncertainty in the classification of solvents for endocrine-disrupting (ED) potential. After processing the input solvent's SMILES representation through the model, the output is a class probability p, where $p < 0.5$ indicates active and $p > 0.5$ indicates inactive. The prediction probability density function (or mass function for discrete output) conditioned on the model structure is given as

$$p_i = p(y_i) = p_F(y_i|x, D) \tag{1}$$

The class probability using the last sigmoid layer of the deep learning model given in Figure 12 can be written as

$$y_i = \sigma_i(F(x)) \quad i = 1, 2 \tag{2}$$

$$\sigma_i = \frac{1}{1 + e^{-\beta_i F(x)}}$$

Where $F(x)$ represents the input to the sigmoid function from the preceding layers of the neural network. This function maps the input features of a solvent to a probability p_i indicating the likelihood of the solvent being an EDC. Also, $i = 1{,}2$ determines the class of EDC (Agonist or Binding), and σ is the sigmoid function. Given an organic molecule x_j, $j = 1$ to N belonging to the list of solvents given in Table 2, the prediction probabilities p_{ij} are given by the function $p_{ij} = \sigma_i(F(x_j))$, $p_{ij} \in [0{,}1]$. The relationship between uncertainty and output probability is not linear. The classification model can have low activation values in all the remaining neurons, but still can have high sigmoid values. Thus, using only the sigmoid output as a measure of model uncertainty can be misleading. Shannon entropy removes this drawback by weighing the prediction probability p_{ij} with the logarithm of the reciprocal of p_{ij} and thereby used to measure the information content of each prediction. The basic intuition behind such a formulation is that the unlikely event will be more informative, while the likely events have little information, and the extreme case events should have no information. The self-information or Shannon information function is the information content associated with a single prediction and is defined as

$$I(p_i) = -\log p_i \tag{3}$$

The Shannon entropy for the j^{th} solvent for the i^{th} class is measured as

$$H_{ij} = -p_{ij}\log p_{ij} - (1 - p_{ij})\log(1 - p_{ij}) \tag{4}$$

This calculation effectively captures the uncertainty of the prediction by considering both the probability of the event occurring and not occurring. This measure reaches its maximum when $p = 0.5$, indicating maximum uncertainty, and is minimal (zero) when p is 0 or 1. The maximum entropy or the total uncertainty for the whole list of solvents for j^{th}class (Agonist or Binding) is $S_j = \sum H_{ij}$. The uncertainty associated with each i^{th} solvent for the j^{th} class of EDC is estimated as the ratio of the prediction entropy H_{ij} and the maximum entropy S_j, providing a normalized measure of the uncertainty across all solvents in a class.

Shannon entropy, using the class probabilities provided by the sigmoid layers, provides a postprediction uncertainty analysis[66,67] that assesses the precision of the data-driven model by quantifying the uncertainty associated with the predictions. It is important to note that this epistemic uncertainty reflects the confidence of the trained model in its predictions and arises from the inherent complexity or limitations of the model's learned representations. The ML model[65] trained on the list of EDCs from the ToxCast and Tox21 needs to be representative of organic molecules in general to obtain an interpretable prediction to accurately classify a solvent as either active or inactive for each class. This does not involve any uncertainty related to the data extracted from automated data curation using either Method I or Method II. Instead, it is an intrinsic measure of the probabilistic output of the model, quantifying ambiguity in decision-making based solely on the learned patterns from the training data.

In our analysis, we have categorized the organic solvents in perovskite synthesis, obtained from both methods of automated data extraction, into two subclasses of endocrine disruptors (EDs)—'Agonist' and 'Binding'—as shown in Table 2. We have used the deep learning model to make our predictions. The studies that substantiate our data are cited in the table's last column, reinforcing the reliability of our classifications. For example, DMF is listed as a potential endocrine disruptor in a study of chemicals used in natural gas extraction.[68] In a study conducted on workers exposed to DMF in the synthetic leather industry, it has been found to have adverse effects on sperm function.[69] A European analysis of birth weight and length of gestation due to occupational exposure to endocrine-disrupting chemicals has listed Toluene as an endocrine-disrupting solvent.[70] Such a nature of Toluene has also been established in research that studied low-dose effects and nonmonotonic dose responses of hormones and endocrine-disrupting chemicals.[71] Alterations in enzyme activities were reported in rat liver due to *n*-Octane administration.[77] While these studies reinforce our classifications, there are also some conflicting reports. Our classification of Acetone as an inactive endocrine-disrupting solvent is confirmed in the EPA's report,[73] but we also came across an article that says the opposite.[72] Similarly, *n*-Hexane was reported as a potential EDC in one study[75] but was ruled out in the other.[76] Simply put, for some solvents in our study, there is data to back up their screening as EDC, while for some, there is vague information in the literature, and for the rest, the information is hard to find. However, using a deep learning model that has 90% accuracy, we have given a tool to the scientific community to screen out the potential EDCs when we do not have relevant data on the chemicals. That means our work puts a red flag on these chemicals, so that careful consideration is given before using them. In other words, our work can act as a guide in safer solvent selection for perovskite synthesis. For example, almost all solvents have been used in the synthesis of MA lead halide perovskites, but by using this work, one can easily opt for a solvent that is not an active EDC. Both DMF and DMSO are polar solvents and are excellent at dissolving perovskite precursors. However, DMF is an EDC chemical, while DMSO is not. Hence, one can

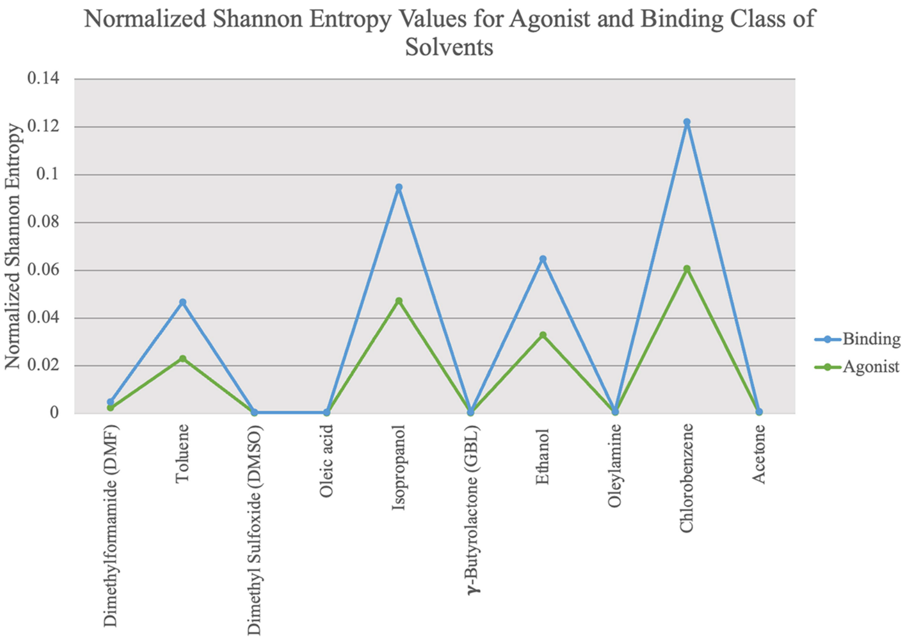


Figure 13. Uncertainties associated with predicting the solvents into agonist and binding classes calculated using Shannon Entropy. A lower value of uncertainty indicates higher confidence in the corresponding prediction. Higher entropy values indicate greater uncertainty in the classification, emphasizing the need for careful consideration and further validation of these results. The green and blue lines, representing 'Agonist' and 'Binding' classes, respectively, have overlapped, indicating similar levels

immediately choose to substitute DMF for DMSO in the synthesis of MA lead halide perovskites. Solvents such as Toluene, Isopropanol, and Chlorobenzene are antisolvents and are used to wash/rinse the solvents to get precursor precipitates.[78] However, Toluene and Chlorobenzene are active EDCs and, hence, are advised to be replaced by Isopropanol or some other antisolvents with matching properties.

Figure 13 shows the uncertainty computed using the Shannon entropy formula for the ten most frequently appearing solvents used in the synthesis of common perovskites. The figure shows nonoverlapping lines for normalized Shannon entropy values of 'Agonist' (green) and 'Binding' (blue) classes, indicating different uncertainty levels in the classification of the solvents across the two classes. From the figure, Chlorobenzene and Isopropanol exhibit higher entropy values, suggesting a lower degree of confidence in their classification, while DMF, DMSO, Oleic acid, Oleylamine, and Acetone indicate a more confident classification. Our classification model, as explained before, which uses SMILES notation as input, processes these representations through convolutional layers followed by LSTM layers and fully connected layers. As mentioned earlier, the convolution layers extract spatially correlated local features or critical substructures within the molecule, and the LSTM layer maps the sequential dependencies or the order and arrangement of atoms and substructures identified by the convolution layers. Thus, high uncertainty for certain solvents, such as Chlorobenzene and Isopropanol, may indicate that the chemical substructures within the molecule and their arrangements are difficult for our classification model to identify. The specific structure and/or the substructure may not be well represented in the training data set.

5. CONCLUSIONS

This study presents an ensemble approach for addressing the challenges of sparse and unstructured data in scientific literature, specifically within the niche domain of perovskite synthesis, by juxtaposing deterministic outputs from smaller contextual language models (e.g., BERT, ELMo) with the broader contextual capabilities of large language models (e.g., GPT-3.5). This ensemble methodology addresses the strengths of combining multiple models to overcome the lack of benchmarking and mitigate challenges such as hallucination and overgeneralization in data extraction. Our work compares two methods for hierarchical data extraction, identifying 35 solvents using Method I and 54 solvents using Method II, with 25 solvents unanimously identified by both. Among these, DMF, Toluene, and DMSO dominate, collectively accounting for over 50% of occurrences. Further analysis of organic perovskites reveals that over 40% are MA-based, with FA and BA-based perovskites each contributing around 10%. Most solvent-perovskite associations involve MA lead halide perovskites, reflecting their popularity due to efficiency and stability. The most frequent perovskites are (MA)PbI3 (42.7%) and (MA)PbBr3 (22.0%). This information is crucial as it highlights the specific solvent-perovskite combinations that optimize device performance and manufacturing efficiency in perovskite-based solar cells.

While our evaluation originally treated Method I and Method II as independent pipelines, our findings support a shift toward a conjunctive strategy. Rather than using both methods in parallel, a guided pipeline—where high-confidence extractions from Method I are used to inform and constrain GPT-based generation—could yield more robust and contextually accurate results. Although the current intersection approach reduces hallucination significantly, it suffers from compounded omission, as the conservative outputs of Method I bleed into the final set.

Furthermore, this work demonstrates how structured data sets extracted via language models can feed into predictive models, enabling actionable insights for safer and more sustainable solvent choices. We apply the curated data set in a case study that explores the endocrine-disrupting potential of solvents using a deep learning-based uncertainty quantification (UQ) framework. Notably, the UQ is applied to the predictions generated by the deep learning classification model, not the LLM-derived results. The UQ framework specifically addresses epistemic uncertainty or uncertainty arising from limited or incomplete training data by quantifying the variability in class probabilities for each prediction using Shannon entropy. Results show high confidence in prediction for solvents like DMF and DMSO, and lower confidence for Toluene and Diethyl ether, requiring further investigation and consideration for expansion of training data. By leveraging Shannon entropy to assess prediction confidence, the approach highlights areas of low confidence, providing a clear measure of the reliability of toxicity predictions and offering potential pathways for evaluating alternative solvents in future toxicological studies.

This study also recognizes the broader ethical and technological implications of applying generative AI in scientific discovery. While LLMs offer powerful capabilities for data extraction, they can also introduce hidden risks in safety-critical domains such as toxicology and materials synthesis. Therefore, we have verified a sample of our output through manual data extraction and evaluation for the integration of structured verification pipelines, such as knowledge graph grounding and hallucination detection, as essential safeguards to enhance trust, transparency, and accountability in automated systems. This verification method includes a set of urgent measures: (1) the development of benchmark data sets with ground-truth annotations for materials synthesis; (2) the coupling of LLMs with domain-specific knowledge-graphs and structured postprocessing; and (3) the establishment of uncertainty-aware decision protocols that contextualize model confidence when applied to real-world screening scenarios. These steps are critical for enabling responsible, trustworthy, and scientifically valid deployment of LLM-based systems in materials research.

■ ASSOCIATED CONTENT

Data Availability Statement

The software codes for data extraction using language models and the associated configuration file have been attached in a zip file in the Supporting Information. The results used to plot different figures of this article have been attached in the form of spreadsheets in the Supporting Information. The Deep Learning model for predicting the EDC nature of solvents is available at https://github.com/MatInfoUB/VisualTox. All other data are available upon reasonable request from the authors. The software packages used in this study include Transfomers, OpenAI, TensorFlow, ChemDataExtractor, Scikit-learn and RDkit.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00612.

Supporting Information document detailing: (i) Self-Attention Mechanism in NLP Models (ii) BERT, ELMo, and GPT-3.5 methodologies for literature mining. (iii) Example results for Method I and Method II of data extraction. (iv) Deep Learning model architecture for endocrine disruptor prediction and (v) Chord diagram visualization insights for solvent usage trends (PDF)

Spreadsheet containing the paragraphs extracted from scientific literature using BERT and CDQA, along with the Named Entity Recognition (NER) results from ELMo. (i) Sheet 1 (CDQA): Extracted paragraphs relevant to perovskite synthesis, responses from BERT, and confidence scores. (ii) Sheet 2 (NER): Parsed text with word-level classifications, including solvents, precursors, perovskites, and other relevant entities (XLSX)

Results of responses generated by GPT-3.5 for 400 research papers related to perovskite. The spreadsheet contains extracted solvent names, synthesis processes, and precursor details using a structured multilevel prompting approach (XLSX)

Spreadsheet containing the endocrine-disrupting chemical (EDC) activity predictions for solvents identified in perovskite synthesis. The spreadsheet includes Solvents identified by each method, associated SMILES codes and Shannon entropy values to quantify uncertainty in toxicity classification. Two key columns: Agonist Activity and Binding Activity, representing ML model predictions for potential endocrine disruption (XLSX)

Most frequently occurring solvents in perovskite synthesis based on literature extraction. The spreadsheet provides a ranked list of solvents and their association with different perovskite formulations that is used for chord diagram visualization in the main text (XLSX)

Python codes for extracting textual data from scientific articles using language models and an associated configuration file. The code for Method II is capable of using more recent models such as GPT 4.0. However, in this work, we have used GPT 3.5 Turbo for generating all the results (ZIP)

A manually curated data set containing paper-specific knowledge graphs for 50 additional articles with the following fields for each entry: DOI, Title, Perovskite, Solvent, and Synthesis Description. These entries serve as ground truth references for evaluating extraction accuracy and were used to construct per-paper evaluation benchmarks (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Krishna Rajan − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260−1660, United States;* orcid.org/0000-0001-9303-2797; Email: krajan3@buffalo.edu

Authors

Arpan Mukherjee − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260−1660, United States;* orcid.org/0000-0001-5698-6268

Deepesh Giri − *Laurel Ridge Community College, Middletown, Virginia 22645, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00612

Author Contributions

A.M: Investigation, methodology, code writing and maintenance, formal analysis, writing-original draft and editing. D.G: Investigation, methodology, formal analysis, writing-original draft. K.R.: Conceptualization, resources, writing-review and editing, supervision, funding acquisition.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge support from NSF Award No. 2315307: NSF Engines Development Award and the Collaboratory for a Regenerative Economy (CoRE center) in the Department of Materials Design and Innovation - University at Buffalo.

■ ADDITIONAL NOTE

†https://platform.openai.com/playground/prompts?models=gpt-4o-mini.

■ REFERENCES

(1) 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.

(2) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C.; et al. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater* **2022**, *8*, 59.

(3) Schilling-Wilhelmi, M.; et al. *From Text to Insight: Large Language Models for Materials Science Data Extraction.* **2025**, *54*, 1125.

(4) Kim, E.; et al. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

(5) Olivetti, E. A.; et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **2020**, 7, No. 041317.

(6) Kim, E.; et al. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf Model* **2020**, *60*, 1194−1201.

(7) Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8*, 102.

(8) Li, B.et al. Deliberate then Generate: Enhanced Prompting Framework for Text Generation. *arXiv preprint* 2023.

(9) Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A.; et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418.

(10) Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, *15*, 1569.

(11) Chen, Z.-Y.; et al. MatChat: A large language model and application service platform for materials science. *Chinese Physics B* **2023**, *32*, 118104.

(12) Wang, H.; et al. Evaluating the Performance and Robustness of LLMs in Materials Science Q&A and Property Predictions. *Digital Discovery* **2024**, 1612.

(13) Ethayarajh, K.How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 55−65 (Association for Computational Linguistics: Stroudsburg, PA, USA, 2019). doi: .

(14) Huang, S.; Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf Model* **2022**, *62*, 6365−6377.

(15) Zhao, J.; Huang, S.; Cole, J. M. OpticalBERT and OpticalTable-SQA: Text- and Table-Based Language Models for the Optical-Materials Domain. *J. Chem. Inf Model* **2023**, *63*, 1961−1981.

(16) Devlin, J.; Chang, M.-W.; Lee, K., Google, K. T. & Language, A. I*BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.* https://github.com/tensorflow/tensor2tensor (2019).

(17) Buehler, M. J. MechGPT, a Language-Based Strategy for Mechanics and Materials Modeling That Connects Knowledge Across Scales, Disciplines, and Modalities. *Appl. Mech Rev.* **2024**, *76*, No. 021001.

(18) Turhan, G. D.Life Cycle Assessment for the Unconventional Construction Materials in Collaboration with a Large Language Model.*Proceedings of the International Conference on Education and Research in Computer Aided Architectural Design in Europe*; Education and research in Computer Aided Architectural Design in Europe 39−48 (2023). doi: .

(19) Guerreiro, N. M.; et al. Hallucinations in Large Multilingual Translation Models. *Trans Assoc Comput. Linguist* **2023**, *11*, 1500−1517.

(20) McKenna, N.; et al. Sources of Hallucination by Large Language Models on Inference Tasks. *arXiv preprint* **2023**.

(21) Radanliev, P.; Santos, O.; Brandon-Jones, A.; Joinson, A. Ethics and responsible AI deployment. *Front. Artif. Intell.* **2024**, *7*, No. 1377011.

(22) Radanliev, P. AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Appl. Artif. Intell.* **2025**, *39*, No. 2463722.

(23) Binette, O.; Steorts, R. C. (Almost) all of entity resolution. *Sci. Adv.* **2022**, *8*, No. eabi8021.

(24) Yazdani, A.; Rouhizadeh, H.; Bornet, A.; Teodoro, D. CONORM: Context-Aware Entity Normalization for Adverse Drug Event Detection. *medRxiv* **2023**, 2023-09.

(25) Gozzi, M.; Di Maio, F. Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. *Multitask Prompts. Electronics (Basel)* **2024**, *13*, 4712.

(26) Chang, K.; et al. Efficient Prompting Methods for Large Language Models: A Survey. *arXiv preprint* **2024**.

(27) Basyal, L.; Sanghvi, M. Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. *arXiv preprint* **2023**.

(28) Wang, J.; Huang, J. X.; Tu, X.; Wang, J.; Huang, A. J.; Laskar, M. T. R.; Bhuiyan, A.; et al. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. *ACM Comput. Surv.* **2024**, *56*, 1−33.

(29) Foppiano, L.; Lambard, G.; Amagasa, T.; Ishii, M. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods* **2024**, *4*, No. 2356506.

(30) Chen, J.et al. A Combined Encoder and Transformer Approach for Coherent and High-Quality Text Generation; *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*; IEEE 2024.

(31) Liu, S.; Wen, T.; Pattamatta, A. S. L. S.; Srolovitz, D. J. A prompt-engineered large language model, deep learning workflow for materials classification. *Mater. Today* **2024**, *80*, 240−249.

(32) Insuasti, J.; Roa, F.; Zapata-Jaramillo, C. M. Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models. *Big Data and Cognitive Computing* **2023**, *7*, 182.

(33) Hu, Y.; Buehler, M. J. Deep language models for interpretative and predictive materials science. *APL Mach. Learning* **2023**, *1*, No. 010901.

(34) Hu, E. J.et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint* 2021.

(35) Zhou, G.; Zhang, M.; Ji, D.; Zhu, Q. Hierarchical learning strategy in semantic relation extraction. *Inf Process Manag* **2008**, *44*, 1008−1021.

(36) Hendricks, G.; Tkaczyk, D.; Lin, J.; Feeney, P. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies* **2020**, *1*, 414−427.

(37) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* **2013**.

(38) Radford, A.et al. *Language Models Are Unsupervised Multitask Learners.* https://github.com/codelucas/newspaper.

(39) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint* **2014**.

(40) Sun, H.; et al. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. *arXiv preprint* **2018**.

(41) Sukhbaatar, S.; Arthur, S.; Jason, W.; Rob, F. Weakly supervised memory networks. *arXiv preprint* 2015.

(42) Cho, K.; et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint* **2014**.

(43) Sutskever Google, I.; Vinyals Google, O.; Le Google, Q. V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* 2014.

(44) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint* 2014.

(45) Giri, D.; Mukherjee, A.; Rajan, K.; Lazou, A.; Daehn, K.; Fleuriault, C.; Gökelma, M.; Olivetti, E.; Meskers, C.; Giri, D.; Mukherjee, A.; Rajan, K.Informatics Driven Materials Innovation for a Regenerative Economy: Harnessing NLP for Safer Chemistry in Manufacturing of Solar Cells. *REWAS 2022: Developing Tomorrow's Technical Cycles (Volume I)*; Springer International Publishing 11192022DOI: .

(46) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J Cheminform* **2011**, *3*, 17.

(47) Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J Cheminform* **2011**, *3*, 41.

(48) Eltyeb, S.; Salim, N. Chemical named entities recognition: a review on approaches and applications. *J. Cheminform* **2014**, *6*, 17.

(49) Lozano, A.; Fleming, S. L.; Chiang, C.-C.; Shah, N.Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature. in *Biocomputing 2024 8−23* (WORLD SCIENTIFIC, 2023). doi: .

(50) Buehler, M. J. Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design. *ACS Engineering Au* **2024**, *4*, 241−277.

(51) Wu, S.; et al. Retrieval-Augmented Generation for Natural Language Processing: A Survey. *arXiv preprint* **2024**.

(52) Miao, J.; et al. Chain of Thought Utilization in Large Language Models and Application in Nephrology. *Medicina (B Aires)* **2024**, *60*, 148.

(53) Xia, Y.; et al. Beyond Chain-of-Thought: A Survey of Chain-of-X Paradigms for LLMs. *arXiv preprint* **2024**.

(54) He, B.; et al. Retrieving, Rethinking and Revising: The Chain-of-Verification Can. Improve Retrieval Augmented Generation. *arXiv preprint* **2024**.

(55) Kouemo Ngassom, S.; Moradi Dakhel, A.; Tambon, F.; Khomh, F.Chain of Targeted Verification Questions to Improve the Reliability of Code Generated by LLMs. in *Proceedings of the 1st ACM International Conference on AI-Powered Software* 122−130 (ACM: New York, NY, USA, 2024). doi: .

(56) Gill, J.; Chetty, M.; Lim, S.; Hallinan, J. Knowledge-Based Intelligent Text Simplification for Biological Relation Extraction. *Informatics* **2023**, *10*, 89.

(57) Zhang, X.; et al. Hybrid Perovskite Light-Emitting Diodes Based on Perovskite Nanocrystals with Organic−Inorganic Mixed Cations. *Adv. Mater.* **2017**, *29*, No. 1606405.

(58) Doolin, A. J.; et al. Sustainable solvent selection for the manufacture of methylammonium lead triiodide (MAPbI $_3$) perovskite solar cells. *Green Chem.* **2021**, *23*, 2471−2486.

(59) Wang, J.; et al. Highly Efficient Perovskite Solar Cells Using Non-Toxic Industry Compatible Solvent System. *Solar RRL* **2017**, *1*, No. 1700091.

(60) Park, G.; Oh, I. H.; Park, J. M. S.; Jung, J.; You, C. Y.; Kim, J. S.; Kim, Y.; Jung, J. H.; Hur, N.; Kim, Y.; Kim, J. Y.; Hong, C. S.; Kim, K. Y.; et al. Solvent-dependent self-assembly of two dimensional layered perovskite (C6H5CH2CH2NH3)2MCl4 (M = Cu, Mn) thin films in ambient humidity. *Sci Rep* **2018**, *8*, 4661.

(61) Kim, M.; et al. Coordinating Solvent-Assisted Synthesis of Phase-Stable Perovskite Nanocrystals with High Yield Production for Optoelectronic Applications. *Chem. Mater.* **2021**, *33*, 547−553.

(62) Xu, Z.; et al. A Thermodynamically Favored Crystal Orientation in Mixed Formamidinium/Methylammonium Perovskite for Efficient Solar Cells. *Adv. Mater.* **2019**, *31*, No. 1900390.

(63) Saliba, M.; et al. Cesium-containing triple cation perovskite solar cells: improved stability, reproducibility and high efficiency. *Energy Environ. Sci.* **2016**, *9*, 1989−1997.

(64) Mansouri, K.; et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect* **2016**, *124*, 1023−1033.

(65) Mukherjee, A.; Su, A.; Rajan, K. Deep Learning Model for Identifying Critical Structural Motifs in Potential Endocrine Disruptors. *J. Chem. Inf Model* **2021**, *61*, 2187−2197.

(66) Olivier, A.; Shields, M. D.; Graham-Brady, L. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Comput. Methods Appl. Mech Eng.* **2021**, *386*, No. 114079.

(67) Kabir, H. M. D.; Khosravi, A.; Hosen, M. A.; Nahavandi, S. Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications. *IEEE Access* **2018**, *6*, 36218−36234.

(68) Kassotis, C. D.; Tillitt, D. E.; Davis, J. W.; Hormann, A. M.; Nagel, S. C. Estrogen and Androgen Receptor Activities of Hydraulic Fracturing Chemicals and Surface and Ground Water in a Drilling-Dense Region. *Endocrinology* **2014**, *155*, 897−907.

(69) Chang, H.-Y.; Shih, T.-S.; Guo, Y. L.; Tsai, C.-Y.; Hsu, P.-C. Sperm function in workers exposed to N,N-dimethylformamide in the synthetic leather industry. *Fertil Steril* **2004**, *81*, 1589−1594.

(70) Birks, L.; et al. Occupational Exposure to Endocrine-Disrupting Chemicals and Birth Weight and Length of Gestation: A European Meta-Analysis. *Environ. Health Perspect* **2016**, *124*, 1785−1793.

(71) Vandenberg, L. N.; et al. Hormones and Endocrine-Disrupting Chemicals: Low-Dose Effects and Nonmonotonic Dose Responses. *Endocr Rev.* **2012**, *33*, 378−455.

(72) Bolden, A. L.; Schultz, K.; Pelch, K. E.; Kwiatkowski, C. F. Exploring the endocrine activity of air pollutants associated with unconventional oil and gas extraction. *Environmental Health* **2018**, *17*, 26.

(73) Akerman, G.; Trujillo, J.; Blankinship, A.*UNITED STATES ENVIRONMENTAL PROTECTION AGENCY OFFICE OF CHEMICAL SAFETY AND POLLUTION PREVENTION MEMORANDUM THROUGH*. https://www.regulations.gov/document/EPA-HQ-OPP-2009-0634-0252 (2015).

(74) Sepp, K.; et al. The Role of Uron and Chlorobenzene Derivatives, as Potential Endocrine Disrupting Compounds, in the Secretion of ACTH and PRL. *Int. J. Endocrinol* **2018**, *2018*, 1−7.

(75) Harris, M. O.; Corcoran, J. *TOXICOLOGICAL PROFILE FOR N-HEXANE* 1999.

(76) Ruiz-García, L.; et al. Possible role of n-hexane as endocrine disruptor in occupationally exposed women at reproductive age. *Toxicol. Lett.* **2018**, *295*, S233.

(77) Khan, S.; Mukhtar, H.; Pandya, K. P. n-octane and n-nonane induced alterations in xenobiotic metabolising enzyme activities and lipid peroxidation of rat liver. *Toxicology* **1980**, *16*, 239−245.

(78) Kara, K.; et al. Solvent washing with toluene enhances efficiency and increases reproducibility in perovskite solar cells. *RSC Adv.* **2016**, *6*, 26606−26611.