Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d5dd00521c

ComProScanner: a multi-agent based framework for composition-property structured data extraction from scientific literature

Aritra Roy, *[ab] Enrico Grisan, [c] John Buckeridge *[ab] and Chiara Gattinoni *[d]

Modern materials discovery using data-driven techniques relies heavily on large and structured databases of material compositions and properties; however, the majority of information regarding experimentally synthesised materials lies buried within millions of scientific articles. Large language models and agents have now made it possible to extract structured knowledge from scientific text, but, despite several approaches designed for this aim, no highly accurate approach focused on composition and property extraction—the bare minimum for data-driven methods—to create machine learning-ready databases without the need for human assistance has been developed. We therefore developed ComProScanner, an autonomous multi-agent platform that facilitates the extraction, validation, classification and visualisation of machine-readable chemical compositions and properties for comprehensive database creation. ComProScanner is a publisher-to-database framework which incorporates publisher APIs bypassing the need to manually upload papers into the framework and it is capable of scanning thousands of papers without human intervention. We evaluated our framework using 100 journal articles against 10 different LLMs, including both open-source and proprietary models, to extract highly complex compositions associated with ceramic piezoelectric materials and corresponding piezoelectric strain coefficients (d_{33}), motivated by the lack of a large dataset for such materials. DeepSeek-V3-0324 outperformed all models with a significant overall accuracy of 0.82. Even with this small journal sample, the vast majority of the piezoelectric materials we extracted are not included in commonly available databases and we identified one system with a significantly high piezoelectric coefficient. This framework provides a simple, user-friendly, readily usable package for extracting highly complex experimental data buried in the literature to build machine learning or deep learning datasets.

Received 24th November 2025
Accepted 19th March 2026

DOI: 10.1039/d5dd00521c

rsc.li/digitaldiscovery

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

1 Introduction

Contemporary data-driven materials design heavily relies on high-fidelity datasets in machine-readable formats, as the effectiveness of machine learning (ML) and deep learning (DL) methodologies hinges on structured and computationally accessible data containing, at minimum, material compositions and their corresponding physical properties. Over the past decade and a half, the establishment of computational databases of high-throughput screened materials based on Density Functional Theory (DFT) calculations, such as the Materials Project (MP),[1] JARVIS-DFT,[2] Alexandria,[3] and Open Quantum Materials Database (OQMD),[4] together with experimental datasets like the Cambridge Crystallographic Data Centre (CCDC)[5] or High Throughput Experimental Materials (HTEM) database[6] and dataset like ChemPile,[7] has shifted research emphasis toward data-driven materials design. Nevertheless, the preponderance of experimental scientific knowledge regarding solid-state materials, analogous to other domains, remains embedded within millions of scientific journal articles. Extracting this wealth of information into the structured, machine-readable formats required for computational analysis presents a significant challenge that necessitates automated approaches.

Natural language processing (NLP) algorithms have demonstrated remarkable advances in materials science applications, from building toolkits and techniques for automated extraction of chemical information from the scientific literature, such as ChemDataExtractor,[8,9] ChemicalTagger,[10] BatteryBERT,[11] and others. These tools and techniques have been implemented to systematically structure the vast corpus of textual knowledge in the field[11–20] leveraging various techniques, including regular expressions,[21] BiLSTM recurrent neural

[a]Energy, Materials and Environment Research Centre, London South Bank University, London SE1 0AA, UK. E-mail: pgr.aritra.roy@lsbu.ac.uk; j.buckeridge@lsbu.ac.uk
[b]School of Engineering and Design, London South Bank University, London SE1 0AA, UK
[c]Bioscience and Bioengineering Research Centre, London South Bank University, London SE1 0AA, UK
[d]Department of Physics, King's College London, London WC2R 2LS, UK. E-mail: chiara.gattinoni@kcl.ac.uk

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

networks,[22] and smaller transformer-based language models like BERT.[23] These approaches have successfully facilitated the extraction of entity information from diverse sources, including battery materials literature[11,14] and chemical synthesis parameters documented in methodology sections of scientific papers.[13] Entity extraction, and in particular named entity recognition (NER), has dominated these research efforts. Researchers have applied domain-specific labels such as "material" or "property" to specific textual elements, but require an additional post-processing step to construct the relations between these entities, relations that prove essential for training effective machine learning or deep learning models. To exemplify, discrete entities such as "Cu_2O" or "2.1 eV" were targeted rather than establishing the relational connections between them (for example, "2.1 eV" represents the measurement of the band gap for "Cu_2O"), *i.e.*, they do not implement relation extraction (RE) techniques.

In the early 2020s, several end-to-end methods were developed that use a single machine learning model integrating both named entity recognition and relation extraction (NERRE).[24–26] These methodologies demonstrate efficacy in relation extraction tasks; however, they remain fundamentally limited to *n*-ary relation extraction frameworks that are complex in architectural structure and struggle to extract all information if the interconnection between various entities are too high. Following the widespread adoption of various large language models (LLMs), researchers have employed them successfully to extract information from journal articles, replacing traditional sequence-to-sequence approaches with more sophisticated NERRE methods. Approaches ranging from pre-training[27] and fine-tuning LLMs[28–33] to prompt-engineering,[29,32–37] zero-shot[29,30,32,33,38] and few-shot prompting,[29,33,37,38] as well as Retrieval-Augmented Generation (RAG) methods[39,40] have enhanced NERRE-level text extraction from materials science literature. Concurrently, LLM-powered agents have been utilised for various chemistry and material science tasks, including extracting relevant information from journal articles,[41–45] predicting new molecules or materials or their properties,[33,41] automating data handling,[33,43,45–47] enhancing reasoning and computational capabilities of LLMs,[41,43,44,46–48] proposing novel hypotheses,[33] and even semi-automating experiments[47] by integrating expert tools. Several notable implementations have emerged in this domain, such as Eunomia by Ansari *et al.*,[42] an AI agent chemist for developing materials datasets by accessing computational databases and research papers, and, very recently the multi-agent system nanoMINER,[49] which combines LLMs and multimodal analysis to extract information, though it is specifically limited to nanomaterials. However, both Eunomia and nanoMINER lack the capability to integrate Text and Data Mining (TDM) API keys† through the package, requiring users to provide the articles in PDF format by manually downloading them, which represents a labour-intensive and time-consuming process when dealing with large-scale datasets. Additionally, enumerating all explicit chemical formulas from variable compositions (*e.g.*, $Pb_{1-x}K_xNb_2O_6$ where $x = 0.1, 0.2$ *etc.*) into distinct compounds remains beyond the scope of these agentic systems. Recently, Wilhelmi *et al.* published a comprehensive tutorial on using LLMs to extract chemical data as structured output *via* various methods, including prompting, RAG and agentic systems.[50] Nevertheless, an easily configurable automated workflow that enables end users to build, evaluate and visualise datasets through information extraction from journal articles has been lacking.

In this work, we present an autonomous multi-agent agile framework, ComProScanner, for end users to extract, evaluate, categorise and visualise machine-readable structured chemical compositions and properties, combined with synthesis information from journal articles to create extensive databases. When a research article contains chemical composition along with the enquired property value either in full article text or tables, the framework extracts structured JSON data[51] containing both agent-extracted relevant information and journal article metadata obtained *via* APIs. The agent-extracted relevant information comprises the chemical composition of the material and the property value as key–value pairs, property unit, material family, synthesis method, precursors used, brief synthesis steps highlighting the key synthesis conditions and steps used and characterisation techniques employed. Our system combines LLM agents with powerful tools, including RAG and a custom deep learning model for extracting chemical compositions and properties only when property values are available in articles. The workflow supports Elsevier, Springer Nature, IOP Publishing and Wiley articles *via* publishers' TDM APIs or PDFs from local folders. ComProScanner enhances text-mining accuracy by providing flexible contextual parameters to agents while maintaining cost-effectiveness through preliminary article filtering *via* keyword matching. The system supports multiple configurable LLMs for both extraction agents and RAG implementations. ComProScanner can be implemented with fewer than 20 lines of Python code to extract pre-defined structured data, provided that users have access to the TDM APIs of the publishers. We evaluated the extraction performance of ten LLMs using 100 articles containing piezoelectric coefficient d_{33} values, achieving overall accuracy exceeding 80% across various models. Detailed evaluation methods and metrics are presented in the Results and discussion sections.

2 Methods

ComProScanner is a highly configurable multi-agent-based Python package developed using CrewAI,[52] a production-grade framework for orchestrating AI agent workflows, supplemented with custom Python scripts. Custom scripts have been strategically implemented throughout the system to enhance cost-effectiveness and accessibility for researchers engaged in data-driven materials discovery.

ComProScanner's workflow architecture comprises four distinct operational phases: (a) metadata retrieval, (b) article collection, (c) information extraction and (d) evaluation, post-

† TDM agreements differ from standard academic subscriptions granted to institutional libraries, as they specifically govern the scraping and downloading of large volumes of content, which could potentially impact the operational performance of publishers' servers.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

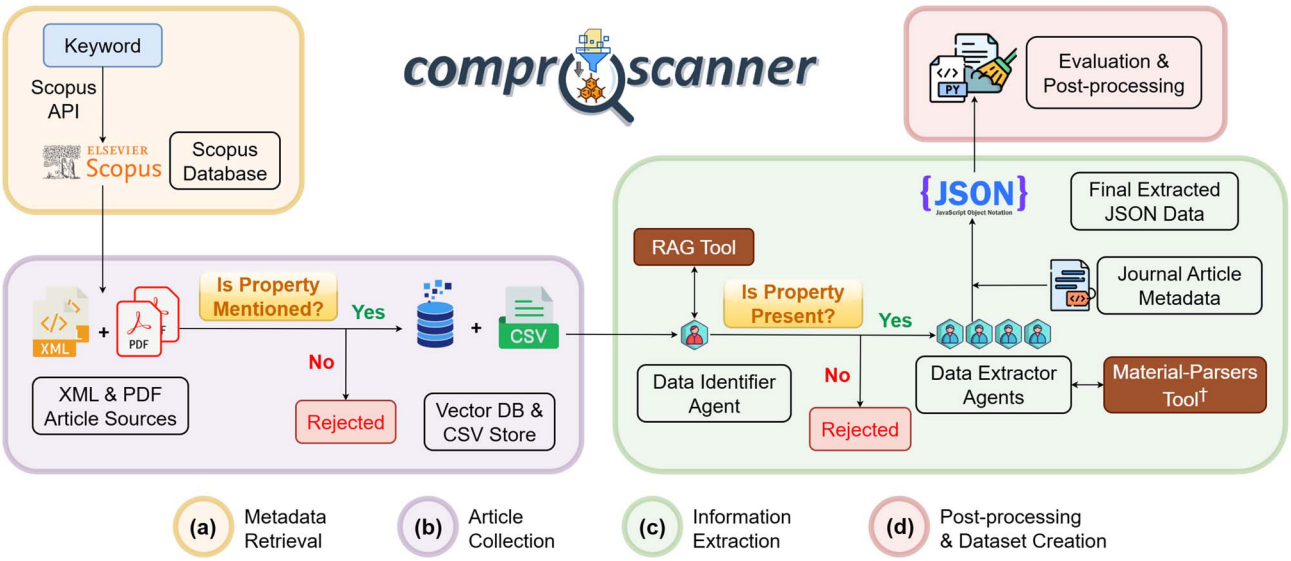

Fig. 1 Overall workflow diagram of ComProScanner framework, separated in four distinct operational phases, distinguished with four different colour regions: (a) metadata retrieval (yellow), (b) article collection (purple), (c) information extraction (green) and (d) evaluation, post-processing and dataset creation (brown).

processing and dataset creation (see Fig. 1). We describe each phase in turn below.

2.1 Metadata retrieval

In this phase, ComProScanner finds metadata for relevant articles associated with the enquired property, including DOI, publication name, ISSN, Scopus ID, article title, article type and publisher name. This section of the program implements property-related article metadata retrieval scripts that function as a Python wrapper for the Scopus Search API.[53] The wrapper enables users to specify primary keyword(s) for relevant metadata search while providing the flexibility to incorporate additional keywords in combination with the primary terms. In alignment with the objectives of the ComProScanner package, this module filters the document formats to include only *Articles* and *Letters*, thereby eliminating other document types such as *Reviews* or *Conference Papers* that could potentially introduce duplicate compositions or properties into the dataset.

2.2 Article collection

This section of ComProScanner accesses full-text articles through publisher-provided TDM APIs. The system currently supports automated extraction of articles from four major publishers *via* their TDM API and manually downloaded PDF articles from all publishers (for further details, see Section S1 of the SI). The system implements preliminary keyword-based filtration for the entire text of the article through Python regular expressions (Python RegEx)[21] to identify relevant articles mentioning the property, thereby optimising data management by avoiding text extraction from irrelevant sources that would unnecessarily inflate the database size. Articles in which the required property is mentioned are organised and stored in CSV format with dedicated columns corresponding to specific article sections (abstract, introduction, experimental methods, computational methods, results and discussion and conclusion), with optional MySQL database[54] integration. Vector databases are generated, along with CSV files, using the open-source ChromaDB[55] package when any of the specified relevant keywords are detected within an article, facilitating future RAG queries.

2.3 Information extraction

The information extraction phase represented in detail in Fig. 2, incorporates five specialised AI agents (Fig. 2), beginning with a property identifier (the materials data identifier) that utilises RAG technology under the RAG Crew. This initial filtering significantly reduces API costs (or computational resource usage for locally hosted LLMs) by eliminating articles that merely mention the required property without containing actual property values. The four remaining agents are organised into two functional subgroups: one dedicated to extracting composition-related data (the composition crew set) and another focused on collecting synthesis information (the synthesis crew set). Each subgroup employs two sequentially ordered agents; the first extracts raw data, while the second formats it. By default, the package uses the provided keyword with some pre-set rules and instruction for the agents. However, *Notes* can be appended to both agents and tasks across all five agent components to provide supplementary instructions to the agents as additional context. For complex compositions with multiple fractions denoted as variables *e.g.*, $Na_{(1-x)}Li_xTiO_3$ where x= 0.1, 0.3, 0.4, the system employs material-parsers, a deep learning model developed by Foppiano *et al.*,[20] as an agent tool that resolves the example into three distinct compositions: $Na(0.9)Li(0.1)TiO_3$, $Na(0.7)Li(0.3)TiO_3$ and $Na(0.6)Li(0.4)TiO_3$. All extracted data are compiled into

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

a unified JSON format, which is subsequently integrated with the corresponding article metadata‡.

2.4 Evaluation, post-processing and dataset creation

The total extracted JSON data comprise two main segments: (i) agent-extracted composition-property and synthesis data, (ii) journal article metadata. A detailed description of each type of agent-extracted data can be found in Section S2 of the SI along with an example of complete extracted JSON data (Fig. S1). ComProScanner offers a built-in comprehensive evaluation framework, both agent-based and semantic-based, designed to assess and visualise the extraction performance of LLM agents with scientific rigour. The framework implements three distinct categories of evaluation metrics: (a) custom weight-based accuracy metrics, (b) conventional classification metrics and (c) normalised classification metrics. The custom weight-based accuracy enables users to assign differential priority values to specific extracted parameters; for instance, users can allocate greater emphasis to composition-property key–value pairs (weight of 0.3) compared to synthesis steps (weight of 0.1) [see Section S2 in the SI for all possible extracted parameters and their weight-based emphasis], where the total weight for all extracted parameters is 1. If the weight for one parameter (*e.g.*, composition-property) is set to 1, keeping all other parameters 0, the accuracy will be determined purely by the composition-property extraction performance. Standard classification metrics include precision, recall and F1-score, which are defined relative to the concepts of true positive (TP), false positive (FP) and false negative (FN). True positive can be defined as a correct value extracted by the agent, false positive when the extracted value does not match the ground truth (the actual text to be extracted and false negative as a value that is expected but has not been extracted by the agent. Both precision and recall can be represented as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

From precision and recall, another metric, F1, can be calculated using eqn (2),

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

Beyond standard classification metrics calculated using the aggregate number of items across all evaluation articles, we have developed normalised classification metrics that consider each article as a single evaluative unit, wherein each extracted item within an article contributes a fractional importance to that article's overall evaluation score. These normalised evaluation metrics were specifically designed to ensure an equitable comparison between articles with significant disparities in the quantity of extractable information. The normalised metrics for all papers are calculated using the modified Precision, Recall and F1-score,

$$\text{normalised precision} = \frac{\sum_{i=1}^{N} \text{TP}_i/n_i}{\sum_{i=1}^{N} \text{TP}_i/n_i + \sum_{i=1}^{N} \text{FP}_i/n_i} \tag{3}$$

$$\text{normalised recall} = \frac{\sum_{i=1}^{n} \text{TP}_i/n_i}{\sum_{i=1}^{n} \text{TP}_i/n_i + \sum_{i=1}^{n} \text{FN}_i/n_i} \tag{4}$$

$$\text{Normalised F1} = \frac{2 \times \text{normalised precision} \times \text{normalised recall}}{\text{normalised precision} + \text{normalised recall}} \tag{5}$$

where, TP_i, FP_i, FN_i = true positives, false positives, false negatives for paper i, n_i = total number of items in paper which can be different for different papers i and N = total number of papers.

Weight-based accuracy metrics, classification metrics and normalised classification metrics all provide the flexibility to use both semantic and agentic approaches for evaluation. The semantic similarity method is used to match ground truth and ComProScanner-extracted information for the semantic approach, whereas LLM agents are instructed to match the ground truth and ComProScanner-extracted information for the agentic approach. Although the evaluation accuracy is expected to be higher for the agentic approach, given that LLM agents will have better comparison ability than semantic comparison between two sentences, the agentic evaluation can take more time and require significantly large numbers of tokens if reasoning models are used for better performance.

ComProScanner provides extensive visualisation capabilities of the evaluation through a diverse array of graphical representations, including bar charts, radar plots, heat maps, histograms and violin charts, all readily accessible within the framework. Additionally, the system offers pie charts and histogram plotting functionalities to facilitate the analysis of data distribution across composition families, precursors and characterisation techniques.

3 Results

Although the materials project[1] contains the largest database of piezoelectric materials,[59] approximately 700 materials therein have non-zero d_{33} coefficients, which quantify the electric field generated when a piezoelectric material is subjected to applied strain. More critically, fewer than 50 materials are present in the database with d_{33} values exceeding 10 pC N^{-1}, where the highest value reaches up to 738.47 pC N^{-1}. However, tens of thousands of previous works, predominantly experimental, have been

‡ This new article metadata is collected for each specific article containing agent-extracted information, differing from the previously collected metadata that contained limited information for all related articles associated with the property keyword used for metadata collection. This new comprehensive metadata includes a wide range of information: DOI, article title, journal name, year of publication, open access information, author list with their institutional details, and article keywords. These data are obtained either *via* Elsevier's ScienceDirect Article Metadata API[56] (optional) or the Open Access Button's free metadata API[57] developed by OA.Works.[58]

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

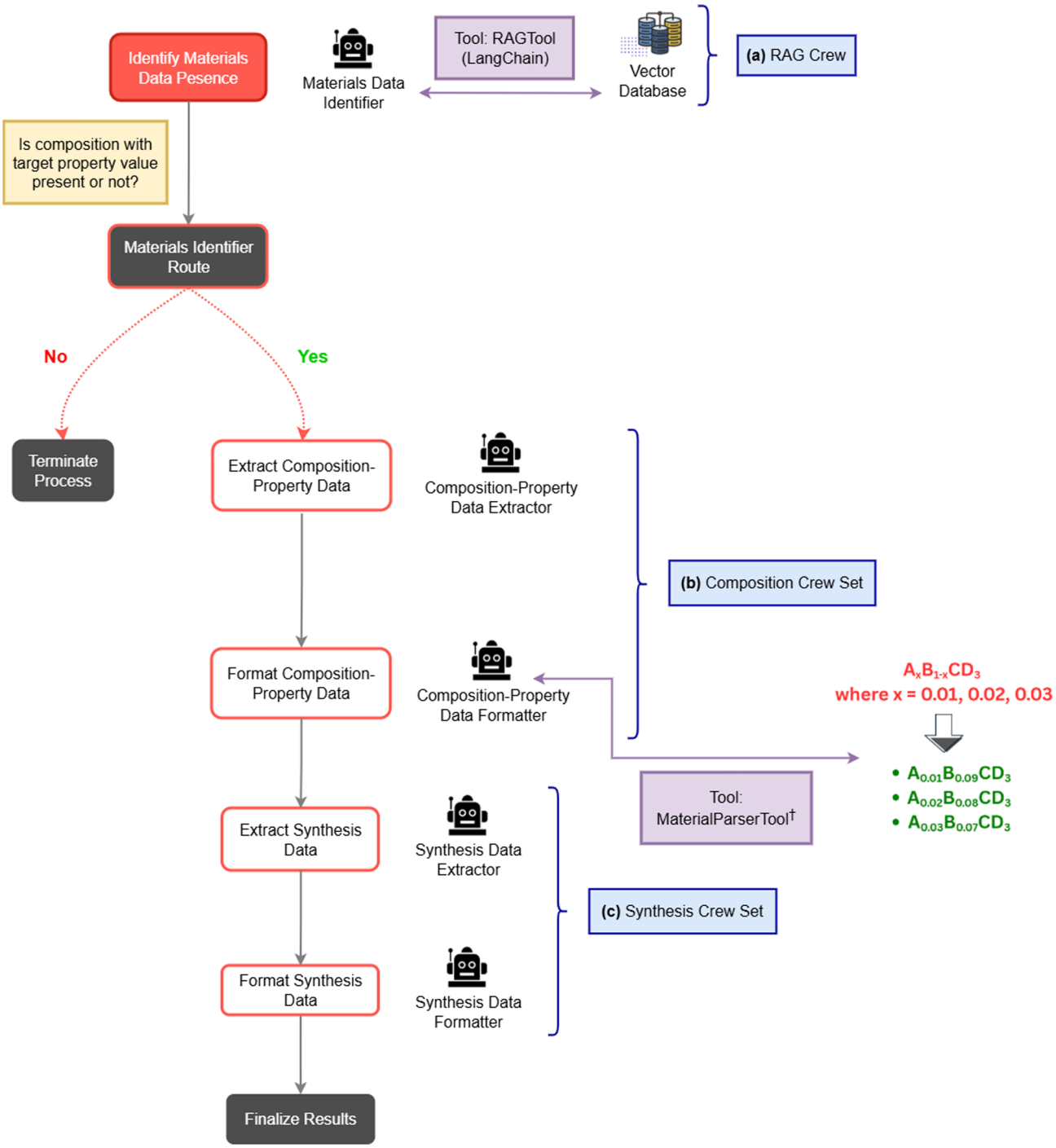

Fig. 2 Comprehensive workflow diagram of the CrewAI-based extraction system in ComProScanner, comprising five specialised agents. The process begins with a property identifier agent ((a) RAG Crew) that leverages Retrieval-Augmented Generation (RAG) technology to filter relevant articles. The remaining four agents are strategically organised into two parallel functional subgroups: one dedicated to composition data extraction ((b) composition crew set) and the other focused on synthesis information collection ((c) synthesis crew set). Each subgroup implements a sequential two-agent architecture—the first agent extracts raw data while the second performs formatting and standardisation. The workflow integrates two essential tools: RAGTool for discriminating between mere property mentions and actual quantitative property values and MaterialParserTool for accurate processing of complex chemical formulations.

already conducted to identify materials with larger d_{33} coefficients through doping or other methods, yet these findings remain in the literature in unstructured, non-machine readable formats. Thus, considering the challenge of extracting composition–property relationships from unstructured literature data as one of the most significant tests for assessing ComProScanner's ability, along with synthesis data, we

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

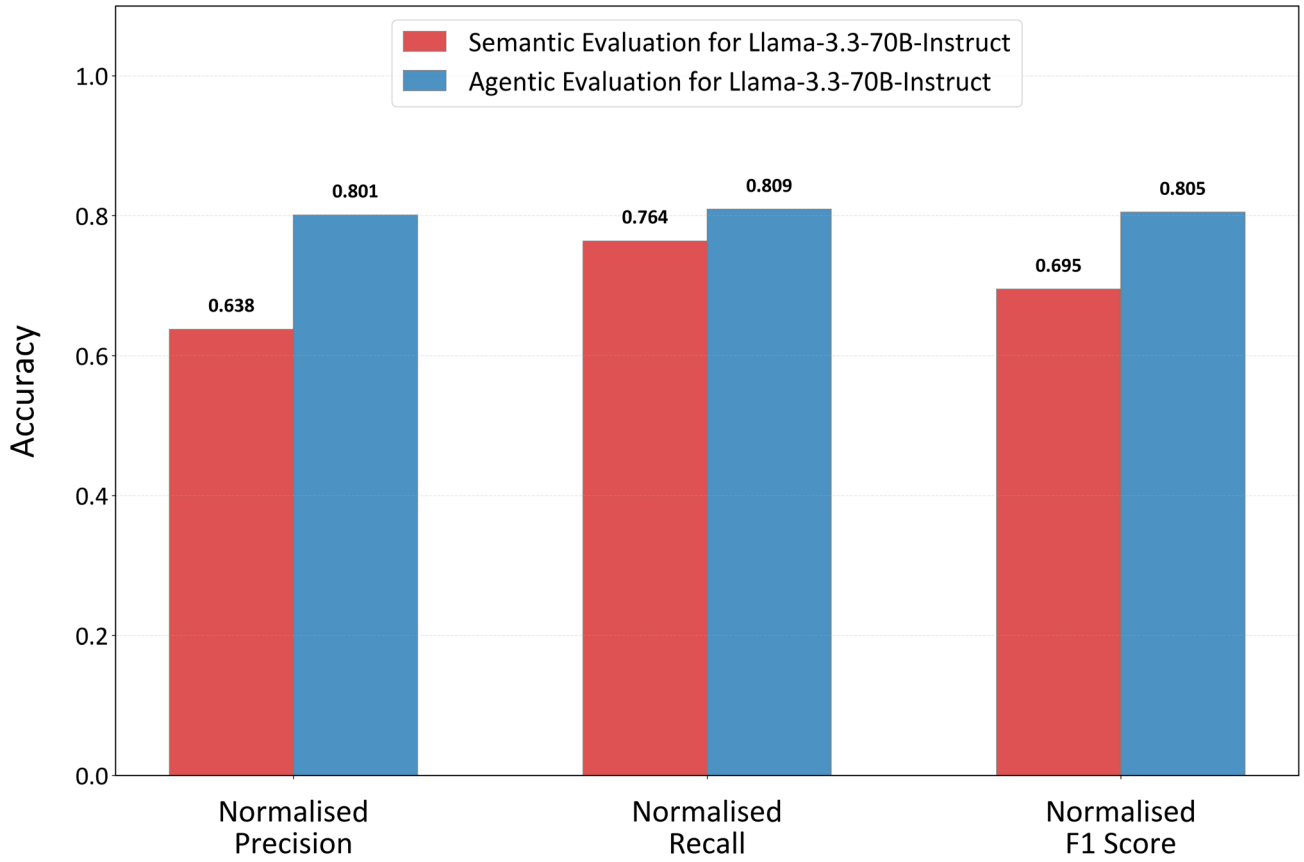

Fig. 3 Normalised classification metrics (precision, recall and F1-score) for model Llama-3.3-70B-Instruct (best performing model considering only normalised metrics) to showcase the performance of ComProScanner's performance capability for the considered models using: semantic evaluation through the PhysBERT model (red bars) and agentic evaluation through the Gemini-2.5-Pro reasoning model (blue bars).

evaluated ceramic piezoelectric materials and their corresponding piezoelectric d_{33} coefficient values, across ten different LLMs.

Metadata of articles related to piezoelectric materials were collected based on *piezoelectric*, *piezoelectricity*, *pyroelectric*, *pyroelectricity*, *ferroelectric* and *ferroelectricity* as the main base keywords. After collecting metadata with only base queries, combinations of base queries and 18 additional keywords such as, *advancements*, *applications*, *ceramics*, *characterization*, *composites*, *crystals*, *etc.*, were used to collect a larger set of metadata that could contain potential piezoelectric materials along with their corresponding d_{33} coefficient values. The complete list of the additional keywords can be found in the SI[60] (see the test_example.py script). Although metadata were collected for all articles published between 1st of January 2019 and 17th of March 2025, only Elsevier papers were considered for the evaluation process, where only 3916 papers mentioned d_{33}, accounting for potential differences in formatting.

Subsequently, 100 test DOIs were selected, based on the presence of the composition-property data using the RAG agent, whilst randomising the metadata order. For RAG and other NLP tasks, text embedding plays a crucial part in ensuring the efficiency of the models. The PhysBERT[61] model has demonstrated superior accuracy compared to various sentence transformer and BERT models in identifying various physics and materials science specific vocabulary. However, to ensure that PhysBERT would perform better than the leading sentence transformer model, all-mpnet-base-v2,[62] in our specific domain, the *thellert/physbert_cased* model from Hugging Face was evaluated against sentence-transformer's all-mpnet-base-v2 model using 12 domain-specific synonyms based on abbreviated forms or chemical formulae and their corresponding full names or trivial names, which are summarised in Table S1 in the SI. The PhysBERT model outperformed all-mpnet-base-v2 in all cases, with remarkable performance differences ranging from highly significant improvements for terms such as DOS (density of states) with a cosine similarity§ difference of 0.8338, to modest improvements for common terms such as PVC (polyvinyl chloride) with a difference of 0.0566. This satisfactory performance of PhysBERT encouraged us to adopt this model as the default embedding model for storing article text data in the ChromaDB vector database for use in the RAG tool illustrated in Fig. 2. For fair evaluation across various models, the RAG environment

§ Cosine similarity measures how similar two text embeddings (vectors) are by calculating the cosine of the angle between them. This provides a score between −1 (completely dissimilar) and 1 (identical), which is useful for tasks such as text or document clustering.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

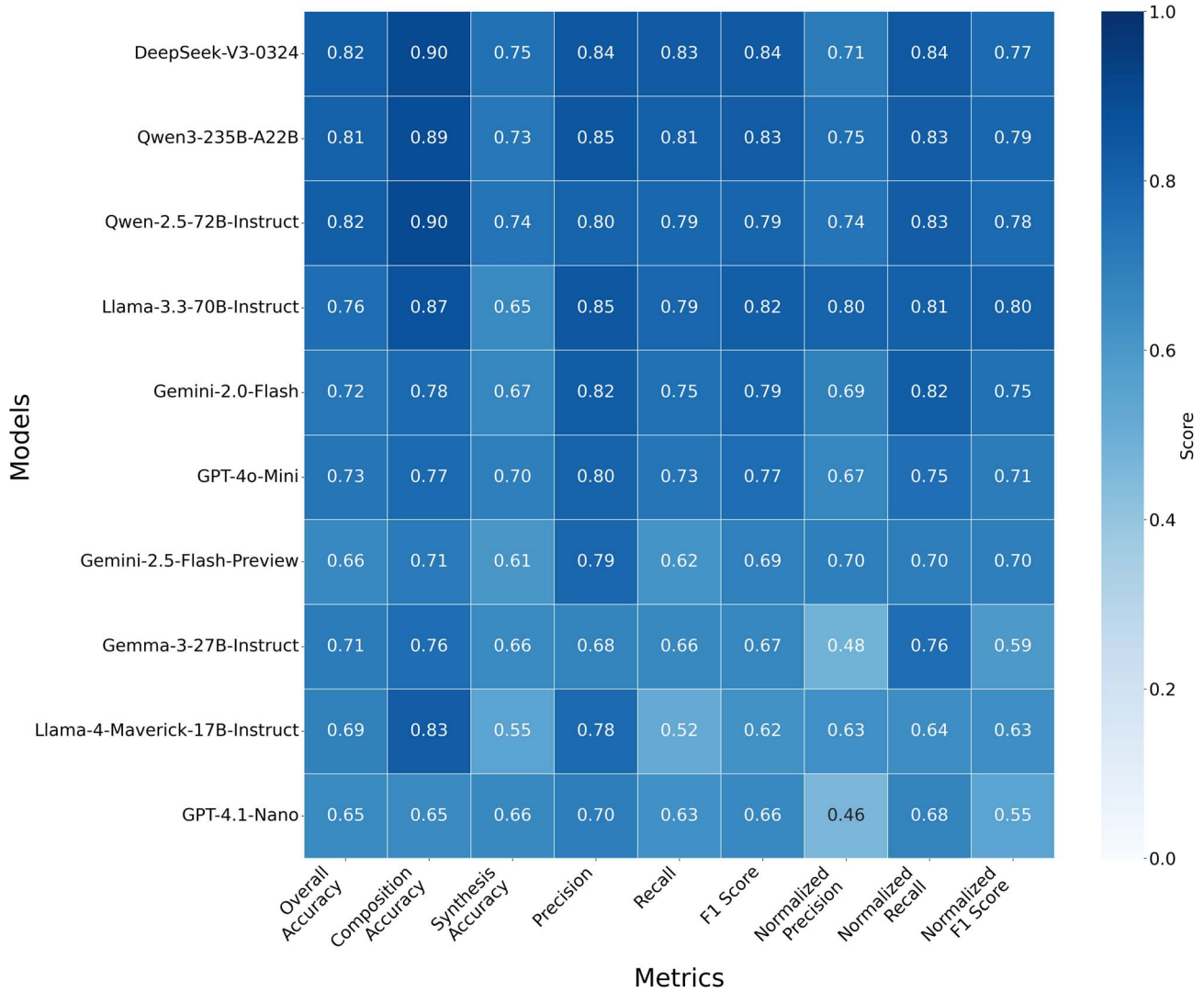

Fig. 4 Confusion matrix from agentic evaluation, showcasing all 9 evaluation parameters, such as weight-based overall accuracy (average of composition accuracy and synthesis accuracy), weight-based composition accuracy and weight-based synthesis accuracy, classification metrics (precision, recall and F1-score) and normalised classification metrics (normalised precision, normalised recall and normalised F1-score), across 10 different LLMs used in this study.

was maintained consistently, as described in detail in Section S3 of the SI.

Finally, extraction agents were used to extract information for the piezoelectric materials only for the test DOIs. We selected LLMs to enable a comparison between open-source models (Google's Gemma-3-27B-Instruct,[63] DeepSeek's DeepSeek-V3-0324,[64] Meta's Llama-3.3-70B-Instruct, Llama-4-Maverick-17B-Instruct,[65,66] and Alibaba's Qwen3-235-A22B,[67] Qwen-2.5-72B-Instruct[68]) and proprietary models (Google's Gemini-2.0-Flash,[69] Gemini-2.5-Flash-Preview,[70] and OpenAI's GPT-4o-mini,[71] GPT-4.1-nano[72]) at similar price points, after analysing the cost-versus-accuracy ratio from the Chatbot Arena LLM Leaderboard,[73] where models had Arena scores exceeding 1250 and output costs below $1/1 M tokens (for more details, see Section S4 and Fig. S2 in the SI). Additional instructions were passed to the agents for better extraction performance specific to piezoelectric materials and d_{33} coefficients (see the test_example.py script from SI[60]). Temperature and maximum output tokens were set to 0.1 and 2048 respectively, which are the default values for ComProScanner's data extraction function.

The normalised classification metrics (precision, recall and F1-score) of different models, as described in the evaluation, post-processing and dataset creation sub-section above, for both semantic and agentic approaches, are represented as grouped bar charts for the model Llama-3.3-70B-Instruct (the best-performing model when considering only normalised metrics) in Fig. 3. Semantic and agentic comparisons based on normalised metrics for all other models can be found in section S5 along with associated Fig. S3 of the SI. We used PhysBERT model for semantic evaluation, while the Gemini-2.5-Pro reasoning model[70] was employed for agentic evaluation. Although normalised classification metrics for both semantic and agentic evaluation show similar trends, the agentic

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 1 Comparison of performance between material-parsers developed by Foppiano *et al.*[20] and ComProScanner regarding variable substitution in material compositions

DOI	Item	Details
10.1016/j.jallcom.2024.176609	Text	The 0.12Pb($Ni_{1/3}Ta_{2/3}$)O_{3-x}PbZr$O_{3-(0.88-x)}$PbTiO_3 piezoelectric ceramics with 2 mol% MnO_2 (abbreviated as PNT-*x*PZ-PT-Mn, $x = 0.41, 0.42, 0.43, 0.44$) were fabricated by the conventional solid-state reaction method
	Material-parsers	1. 0.12Pb 2. $Ni_{1/3}Ta_{2/3}O_{2.59}PbZrO_3$-(0.87.59)$PbTiO_3$ 3. $Ni_{1/3}Ta_{2/3}O_{2.58}PbZrO_3$-(0.87.58)$PbTiO_3$ 4. $Ni_{1/3}Ta_{2/3}O_{2.57}PbZrO_3$-(0.87.57)$PbTiO_3$ 5. $Ni_{1/3}Ta_{2/3}O_{2.56}PbZrO_3$-(0.87.56)$PbTiO_3$
	ComProScanner	1. 0.12Pb($Ni_{1/3}Ta_{2/3}$)O_3-0.41PbZrO_3-0.47PbTiO_3 + 2% MnO_2 2. 0.12Pb($Ni_{1/3}Ta_{2/3}$)O_3-0.42PbZrO_3-0.46PbTiO_3 + 2% MnO_2 3. 0.12Pb($Ni_{1/3}Ta_{2/3}$)O_3-0.43PbZrO_3-0.45PbTiO_3 + 2% MnO_2 4. 0.12Pb($Ni_{1/3}Ta_{2/3}$)O_3-0.44PbZrO_3-0.44PbTiO_3 + 2% MnO_2
10.1016/j.jeurceramsoc.2025.117193	Text	In this study, dense $Pb_{(1-x)}K_{2x}[Nb_{0.96}Ta_{0.04}]_2O_6$ (PK_xNT, $x = 0.05, 0.10, 0.15, 0.20$) ceramics were prepared *via* the solid-state reaction method
	Material-parsers	1. In 2. $Pb_{(0.95)}K_{20.05}[Nb_{0.96}Ta_{0.04}]_2O_6$ 3. $Pb_{(0.9)}K_{20.10}[Nb_{0.96}Ta_{0.04}]_2O_6$ 4. $Pb_{(0.85)}K_{20.15}[Nb_{0.96}Ta_{0.04}]_2O_6$ 5. $Pb_{(0.8)}K_{20.20}[Nb_{0.96}Ta_{0.04}]_2O_6$
	ComProScanner	1. $Pb_{0.95}K_{0.1}[Nb_{0.96}Ta_{0.04}]_2O_6$ 2. $Pb_{0.9}K_{0.2}[Nb_{0.96}Ta_{0.04}]_2O_6$ 3. $Pb_{0.85}K_{0.3}[Nb_{0.96}Ta_{0.04}]_2O_6$ 4. $Pb_{0.8}K_{0.4}[Nb_{0.96}Ta_{0.04}]_2O_6$
10.1016/j.ceramint.2024.09.282	Text	$BaCO_3$ (99.8%, Aladdin), TiO_2 (99.0%, McLean, Shanghai, China), SnO_2 (99.9%, Aladdin), $CaCO_3$ (99.0%, Sinopharm), Bi_2O_3 (99.9%, McLean), Fe_2O_3 (99.0%, Sinopharm) are used as raw materials, which were accurately weighed according to a composition of $(1-x)$ ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)O_{3-x}BiFeO_3 (BCTSO-*x*BFO, $x = 0, 0.1, 0.5, 0.9$ mol%) and milled with ethanol for 16 h
	Material-parsers	1. $BaCO_3$ 2. TiO_2 3. SnO_2 (99.9%, Aladdin 4. $CaCO_3$ (99.0%, Sinopharm 5. Bi_2O_3 (99.9%, McLean), Fe2O3 6. (1.0) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{3.0}$BiFeO_3 7. (0.9) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{2.9}$BiFeO_3 8. (0.5) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{2.5}$BiFeO_3 9. (1.0) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{3.0}$BiFeO_3 10. (0.9) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{2.9}$BiFeO_3 11. (0.5) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)$O_{2.5}$BiFeO_3 12. (0.1) (Ba0.95Ca0.05) (Ti0.89Sn0.11)O2.1BiFeO_3 13. (−15.0) ($Ba_{0.95}Ca_{0.05}$) ($Ti_{0.89}Sn_{0.11}$)O-13.0BiFeO_3
	ComProScanner	1. ($Ba_{0.95}Ca_{0.05}$)($Ti_{0.89}Sn_{0.11}$)O_3 2. ($Ba_{0.95}Ca_{0.05}$)($Ti_{0.89}Sn_{0.11}$)O_3-(0.1)BiFeO_3 3. ($Ba_{0.95}Ca_{0.05}$)($Ti_{0.89}Sn_{0.11}$)O_3-(0.5)BiFeO_3 4. ($Ba_{0.95}Ca_{0.05}$)($Ti_{0.89}Sn_{0.11}$)O_3-(0.9)BiFeO_3
10.1016/j.ceramint.2024.10.314	Text	Lead-free piezoelectric ceramics with the formula $Ba_{1-x}Sr_xTi_{0.92}Zr_{0.08}O_3$ [$x = 0$, 0.04, 0.08, 0.12, 0.16, 0.20 (mol)] were prepared using the solid-state reaction technique
	Material-parsers & ComProScanner	1. $Ba_{1.0}Sr_0Ti_{0.92}Zr_{0.08}O_3$ 2. $Ba_{0.96}Sr_{0.04}Ti_{0.92}Zr_{0.08}O_3$ 3. $Ba_{0.92}Sr_{0.08}Ti_{0.92}Zr_{0.08}O_3$ 4. $Ba_{0.88}Sr_{0.12}Ti_{0.92}Zr_{0.08}O_3$ 5. $Ba_{0.84}Sr_{0.16}Ti_{0.92}Zr_{0.08}O_3$ 6. $Ba_{0.80}Sr_{0.20}Ti_{0.92}Zr_{0.08}O_3$
10.1016/j.jeurceramsoc.2024.117065	Text	Pure $CaBi_2Nb_2O_9$ and rare-earth thulium-substituted $CaBi_2Nb_2O_9$ powders with nominal compositions of $Ca_{1-x}Tm_xBi_2Nb_2O_9$ (CBN-100xTm) were prepared through a solid-phase reaction method. To characterize the phase transition in detail, a composition range of $x = 0.01$–0.05 was selected
	Material-parsers	1. $CaBi_2Nb_2O9$ 2. $CaBi_2Nb_2O_9$ 3. $Ca_{1-x}Tm_xBi_2Nb_2O_9$

© 2026 The Author(s). Published by the Royal Society of Chemistry

Table 1 (*Contd.*)

DOI	Item	Details
	ComProScanner	1. $CaBi_2Nb_2O_9$-1Tm 2. $CaBi_2Nb_2O_9$-2Tm 3. $CaBi_2Nb_2O_9$-3Tm 4. $CaBi_2Nb_2O_9$-4Tm 5. $CaBi_2Nb_2O_9$-5Tm
	Actual resolved compositions	1. $Ca_{0.99}Tm_{0.01}Bi_2Nb_2O_9$ 2. $Ca_{0.98}Tm_{0.02}Bi_2Nb_2O_9$ 3. $Ca_{0.97}Tm_{0.03}Bi_2Nb_2O_9$ 4. $Ca_{0.96}Tm_{0.04}Bi_2Nb_2O_9$ 5. $Ca_{0.95}Tm_{0.05}Bi_2Nb_2O_9$

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

evaluation demonstrates superior performance accuracy compared to semantic evaluation, which is understandable given that reasoning models such as Gemini-2.5-Pro possess greater capability to compare sentence structures with equivalent meanings. Given the superior accuracy demonstrated by agentic evaluation, we focus on these results to identify the best-performing models for practical implementation. As mentioned earlier, Llama-3.3-70B-Instruct outperforms all other models in normalised classification metrics with a Precision value of 0.80, Recall value of 0.81 and F1-score of 0.80 (Fig. 3).

The confusion matrix (Fig. 4) reveals distinct performance patterns across the evaluated models for piezoelectric materials extraction taking into account all performance metrics. DeepSeek-V3-0324 emerged as the top-performing model for data extraction, demonstrating consistently high scores across all metrics, with particularly strong performance in composition accuracy (0.90), precision (0.84), recall (0.83) and F1-score (0.84). This model showed balanced performance with an overall accuracy of 0.82 and robust synthesis accuracy of 0.75. The Qwen model family demonstrated competitive performance, with both Qwen3-235B-A22B and Qwen-2.5-72B-Instruct achieving comparable results. Notably, both models excelled in composition accuracy (0.89–0.90) and maintained consistent performance across precision, recall and F1-score metrics (0.79–0.85). Llama-3.3-70B-Instruct showed strong overall performance with an accuracy of 0.76 and exceptional composition accuracy (0.87). Google's Gemini models presented mixed results. While Gemini-2.0-Flash achieved moderate performance with balanced metrics, Gemini-2.5-Flash-Preview unexpectedly underperformed compared to its predecessor, showing lower scores across most metrics (0.61–0.71). Llama-4-Maverick-17B-Instruct demonstrated notable strengths in specific areas despite its overall lower performance, achieving commendable composition accuracy (0.83) and precision (0.78). However, the model struggled significantly with synthesis accuracy (0.55) and normalised precision (0.63). The most concerning performance was observed with GPT-4.1-nano, which consistently scored lowest across all metrics, particularly struggling with normalised Precision (0.46). Similarly, Gemma-3-27B-Instruct showed suboptimal performance, with notable weaknesses in synthesis accuracy and normalised precision. We have also taken into consideration the incorrect and hallucinated information extracted by the best-performing model, *i.e.*, DeepSeek-V3-0324. Although incorrect extractions occasionally occur, the total number of hallucinated extractions is significantly low (see Table S2 in the SI). Interestingly, although only compositions specific to the relevant work were instructed to be extracted, compositions mentioned in that specific article as part of a literature study were occasionally also extracted. However, these nuances can be fine-tuned through robust prompt engineering for each specific task.

Furthermore, to compare ComProScanner's variable parsing ability with the original material-parsers tool developed by Foppiano *et al.*,[20] we tested several examples from the test dataset by processing them directly through material-parsers, with results summarised in Table 1. Whilst ComProScanner outperformed material-parsers in most cases (first three examples), both tools successfully resolved the chemical formulae for relatively straightforward compositions (fourth example) and both occasionally failed, as demonstrated in the fifth example. Throughout the entire test set, ComProScanner demonstrated superior performance in most instances and equivalent performance in others when compared to material-parsers, thereby validating ComProScanner's capabilities. Furthermore, we compared our framework with similar existing frameworks, Eunomia[42] and the extraction agent by CMEG-IITR.[48] ComProScanner outperformed both frameworks in all metrics by significant values. More details about the comparison can be found in section S6 of the ESI.

ComProScanner also offers built-in data distribution visualisation functions to represent various material families, synthesis precursors and characterisation techniques as either histograms or pie-charts through a semantic clustering mechanism. Fig. 5 shows these data distributions, where similarity thresholds of 0.8 (default in ComProScanner) were applied for material families and precursors, while 0.78 was found to be best for characterisation techniques during semantic clustering. The resulting distributions reveal the prevalence of different components in piezoelectric materials research across the evaluated 100 articles. In terms of material families, $BaTiO_3$ dominates at 39.0%, followed by KNN (16.0%) and PZT (14.0%), with various other compositions including CaBi2Nb2O9 (9.0%) and BNT-based materials (3.0%) comprising the remainder. For synthesis precursors, Bi2O3 is most frequently used (18.9%),

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

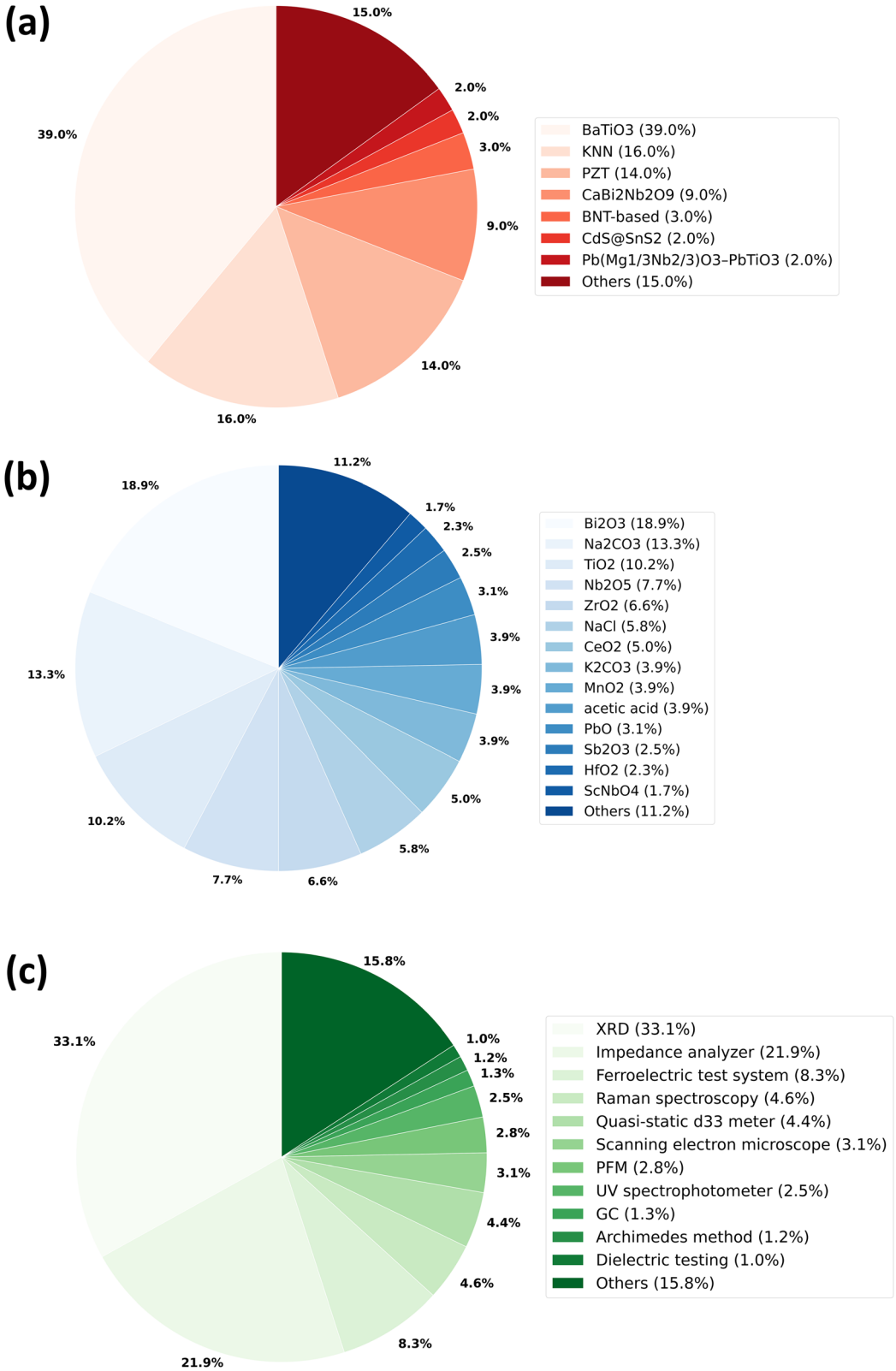

Fig. 5 Distribution of various data types across the evaluated 100 articles: (a) piezoelectric material families, (b) synthesis precursors and (c) characterisation techniques. Similarity thresholds of 0.8 were applied for families and precursors, whilst 0.78 was used for characterisation techniques to group semantically similar items.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

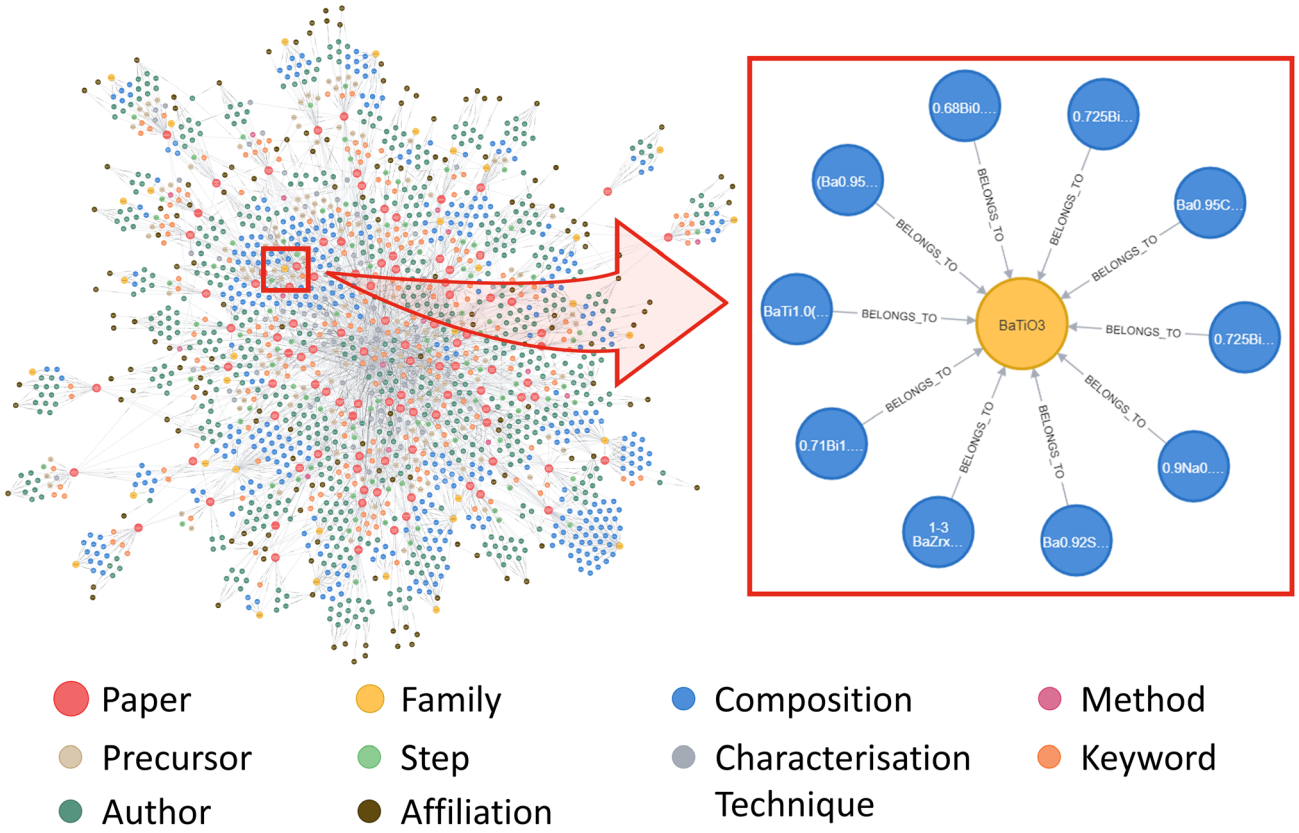

Fig. 6 Custom schema-based relationship graph generated and visualised using *neo4j* from composition-property and synthesis data as well as metadata information extracted from 100 articles using the ComProScanner package. The inset shows 10 randomly chosen items from the 79 compositions associated with the $BaTiO_3$ family node using cypher query.

followed by Na2CO3 (13.3%) and TiO2 (10.2%), with a diverse range of other precursors including various carbonates, oxides and acids distributed across smaller percentages. The characterisation techniques show XRD as the predominant method (33.1%), which is expected for crystalline phase analysis, followed by impedance analysers (21.9%) for electrical property measurements and ferroelectric test systems (8.3%) for specific piezoelectric characterisation, with various other analytical techniques contributing to comprehensive materials evaluation.

To visualise the relationships between the distribution of all data types mentioned in the Methods section for the evaluated dataset, a custom schema-based relationship graph has been constructed and visualised using the *neo4j*[74] library within the ComProScanner package (Fig. 6). The schema defines hierarchical relationships between extracted compositional data, material properties, synthesis parameters and metadata. The produced *neo4j* relationship graph from 100 test articles contains a total of 1825 graph nodes, which are summarised in Table S2 of the SI. Cypher[75] queries can be utilised to retrieve relational information for specific nodes, for example, the inset of Fig. 6 represents 10 random items among 79 compositions associated with the $BaTiO_3$ family across 100 test articles. Detailed information about all nodes associated with the test data can be found in Table S3 in the SI.

4 Discussion

Although d_{33} was mentioned in 3916 papers published within the considered time period, only data from the 100 test papers were extracted for evaluation, as our aim here is to introduce the robust data-extraction framework rather than providing a dataset. However, with this limited dataset of 100 test samples, we identified a piezoelectric composition $Pb(In_{1/2}Nb_{1/2})O_3$-$Pb(Mg_{1/3}Nb_{2/3})O_3$-$PbTiO_3$ achieving 2090 pC N^{-1}, which demonstrates the significance of this framework. On top of this, over 99% of the extracted piezoelectric materials are not in the Materials Project piezoelectric database, which emphasises the importance of ComProScanner for creating datasets from materials information buried within the materials science literature. When using ComProScanner to extract data for a use case different to the piezoelectric materials described here, substituting the piezoelectric material and d_{33} coefficient-related keywords with one's own choice of property keywords in the additional information may be sufficient; however, one may need to introduce further prompt engineering for better extraction performance. During evaluation with one's own data, despite being superior in evaluation, the agentic approach can result in substantial costs, as the pricing of reasoning models is considerably higher than that of chat models. In such cases, to determine the suitable model for data extraction, semantic evaluation can serve as a cost-efficient approach.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

For the piezoelectric materials considered, balanced performance in each metric with an overall accuracy of 0.82 indicates that DeepSeek-V3-0324 possesses the most reliable extraction capabilities for complex piezoelectric material data. The consistency in various metrics for both Qwen models suggests these models are also well-suited for systematic materials data extraction tasks. Llama-3.3-70B-Instruct's results makes it particularly valuable for applications requiring high Precision in materials identification. However, its synthesis accuracy (0.65) is relatively lower, indicating potential challenges in extracting complex synthesis information. The counter-intuitive results from two Gemini models suggests that model updates do not always guarantee improved performance for domain-specific tasks. The results for Llama-4-Maverick-17B-Instruct suggest it may be more suitable for composition-focused extraction tasks rather than comprehensive materials informatics applications. Poor performances from GPT-4.1-Nano and Gemma-3-27B-Instruct highlight the importance of model selection for materials informatics applications, where domain-specific performance can vary significantly from general language tasks. The comparison between the original material-parsers tool and ComProScanner demonstrates that our package performs significantly more efficiently than material-parsers when resolving complex chemical compositions containing variables.

Though LLM agents attempt to ensure consistent results across runs, the underlying LLMs are nondeterministic by nature, which forms the core limitation of any type of LLM-based approach; consequently, results may vary slightly between runs. For the *Materials Data identifier* agent, the RAG question, chunk size, chunk overlap, top k value and RAG chat model may require adjustment and testing according to the specific use case. Although manual evaluation would serve as the optimal evaluation technique compared to semantic and agentic approaches, it is not practical for large dataset evaluation. With this consideration, semantic and agentic approaches are incorporated into the framework and depending on the chosen reasoning model, evaluation results can vary slightly.

ComProScanner establishes the essential foundation for the next generation of AI in material science, creating a pathway to develop extensive text-mined datasets from journal articles. The framework we have developed enables a seamless, user-friendly automated data extraction pipeline. However, OCR technology or VLMs could be integrated with the framework in the future to extract information from graphs or other image formats. Additionally, flexibility to modify the structure of the extracted JSON data could be incorporated into the framework to extract multiple material properties.

5 Conclusions

Although researchers have attempted to automate the extraction of structured information from journal articles, a user-friendly, ready-to-use framework was lacking. Here, we have introduced a multi-agent framework, ComProScanner, to accomplish this task. We assessed our framework using 100 scientific articles across 10 LLMs for highly complex ceramic piezoelectric material compositions and corresponding d_{33} coefficient values. We found ComProScanner could extract extremely complex piezoelectric material compositions with DeepSeek-V3-0324 achieving an overall accuracy of 0.82 and compositional accuracy of 0.90 under its optimal settings. Both considered Qwen models (Qwen3-235B-A22B and Qwen-2.5-72B-Instruct) and Llama-3.3-70B-Instruct also demonstrated competitive performance with an average composition-property accuracy ranging from 0.87–0.90. Surprisingly Gemini-2.5-Flash-Preview underperformed compared to its predecessor in most of the evaluation metrics. Correct assessment of these complex textual data is challenging and must be performed manually which is practically impossible for extensive datasets. However, both semantic and agentic evaluation suggest the potential application of ComProScanner for creating vast datasets of complex material compositions and associated properties, along with synthesis information. As demonstrated by ComProScanner's performance, LLM-based multi-agent frameworks represent a promising approach for automated scientific data extraction, potentially accelerating materials discovery and database construction for data-driven research.

Author contributions

AR: conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, writing – original draft. EG: resources, supervision. JB: conceptualisation, formal analysis, funding acquisition, investigation, resources, validation, writing – original draft, writing – review & editing, supervision. CG: conceptualisation, formal analysis, funding acquisition, investigation, resources, validation, writing – original draft, writing – review & editing, supervision.

Conflicts of interest

There are no conflicts to declare.

Data availability

ComProScanner code is available at https://github.com/slimeslab/ComProScanner for reuse and modification under the MIT licence. To support long-term preservation and reproducibility, a static snapshot of the repository corresponding to the manuscript has been archived in Zenodo (DOI: https://doi.org/10.5281/zenodo.19033754), alongside the archived version of the latest software release. All data pertaining to the evaluation process can be found in the *examples* folder. The Python package is hosted on the Python Package Index (PyPI) at https://pypi.org/project/comproscanner/ for straightforward installation *via* pip. Comprehensive documentation detailing package usage with custom configurations and all available functions with their accepted arguments is available at https://slimeslab.github.io/ComProScanner.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00521c.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Acknowledgements

AR and JB thank London South Bank University for financial and legal support to obtain Elsevier, Wiley and Springer Nature publisher's TDM licences. CG thanks King's College London for legal support in obtaining IOP Publishing's TDM licence. CG was supported by the EPSRC through a New Investigator Award [grant number UKRI132].

References

1 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, *APL Mater.*, 2013, **1**, 011002.
2 K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, *et al.*, *npj Comput. Mater.*, 2020, **6**, 173.
3 M. M. Ghahremanpour, P. J. Van Maaren and D. Van Der Spoel, *Sci. Data*, 2018, **5**, 1–10.
4 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *Jom*, 2013, **65**, 1501–1509.
5 F. H. Allen, *Sci. Data*, 2002, **58**, 380–388.
6 A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas and C. Phillips, *Sci. Data*, 2018, **5**, 1–12.
7 A. Mirza, N. Alampara, M. Ríos-García, M. Abdelalim, J. Butler, B. Connolly, T. Dogan, M. Nezhurina, B. Şen, S. Tirunagari, M. Worrall, A. Young, P. Schwaller, M. Pieler and K. M. Jablonka, ChemPile: A 250GB Diverse and Curated Dataset for Chemical Foundation Models, *arXiv*, 2025, preprint, arXiv:2505.12534, DOI: 10.48550/arXiv.2505.12534, https://arxiv.org/abs/2505.12534.
8 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
9 D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 41.
10 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 1–13.
11 S. Huang and J. M. Cole, *J. Chem. Inf. Model.*, 2022, **62**, 6365–6377.
12 E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder and E. Olivetti, *Chem. Mater.*, 2017, **29**, 9436–9444.
13 O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan and G. Ceder, *Sci. Data*, 2019, **6**, 203.
14 S. Huang and J. M. Cole, *Sci. Data*, 2020, **7**, 260.
15 O. Sierepeklis and J. M. Cole, *Sci. Data*, 2022, **9**, 648.
16 Q. Dong and J. M. Cole, *Sci. Data*, 2022, **9**, 193.
17 K. Cruse, A. Trewartha, S. Lee, Z. Wang, H. Huo, T. He, O. Kononova, A. Jain and G. Ceder, *Sci. Data*, 2022, **9**, 234.
18 E. J. Beard and J. M. Cole, *Sci. Data*, 2022, **9**, 329.
19 A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder and A. Jain, *Patterns*, 2022, **3**, 100488.
20 L. Foppiano, P. B. Castro, P. Ortiz Suarez, K. Terashima, Y. Takano and M. Ishii, *Sci. Technol. Adv. Mater.:Methods*, 2023, **3**, 2153633.
21 re — Regular expression operations, https://docs.python.org/3/library/re.html, 2025, Accessed: 2025-03-19.
22 A. A. Sharfuddin, M. N. Tihami and M. S. Islam, *2018 International conference on Bangla speech and language processing (ICBSLP)*, 2018, pp. , pp. 1–4.
23 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv*, 2019, preprint, arXiv:1810.04805, DOI: 10.48550/arXiv.1810.04805.
24 J. Giorgi, G. D. Bader and B. Wang, *A Sequence-to-Sequence Approach for Document-Level Relation Extraction*, 2022.
25 P.-L. H. Cabot and R. Navigli, *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2370–2381.
26 B. Townsend, E. Ito-Fisher, L. Zhang and M. May, Doc2Dict: Information Extraction as Text Generation, *arXiv*, 2021, preprint, arXiv:2105.07510, DOI: 10.48550/arXiv.2105.07510, https://arxiv.org/abs/2105.07510.
27 V. Mishra, S. Singh, D. Ahlawat, M. Zaki, V. Bihani, H. S. Grover, B. Mishra, S. Miret, Mausam and N. M. A. Krishnan, Foundational Large Language Models for Materials Research, *arXiv*, 2025, preprint, arXiv:2412.09560, DOI: 10.48550/arXiv.2412.09560, https://arxiv.org/abs/2412.09560.
28 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, *Nat. Commun.*, 2024, **15**, 1418.
29 L. Foppiano, G. Lambard, T. Amagasa and M. Ishii, *Sci. Technol. Adv. Mater.:Methods*, 2024, **4**, 2356506.
30 M. P. Polak, S. Modi, A. Latosinska, J. Zhang, C.-W. Wang, S. Wang, A. D. Hazra and D. Morgan, *Digital Discovery*, 2024, **3**, 1221–1235.
31 Y. Ye, J. Ren, S. Wang, Y. Wan, H. Wang, I. Razzak, B. Hoex, T. Xie and W. Zhang, Construction and Application of Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model, *arXiv*, 2024, preprint, arXiv:2404.03080, DOI: 10.48550/arXiv.2404.03080, https://arxiv.org/abs/2404.03080.
32 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, *et al.*, *Digital Discovery*, 2023, **2**, 1233–1250.
33 Y. Zimmermann, A. Bazgir and Z. Afzal, Reflections from the 2024 Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry, *arXiv*, 2025, preprint, arXiv:2411.15221, DOI: 10.48550/arXiv.2411.15221, https://arxiv.org/abs/2411.15221.
34 M. P. Polak and D. Morgan, *Nat. Commun.*, 2024, **15**, 1569.
35 N. Alampara, M. Schilling-Wilhelmi, M. Ríos-García, I. Mandal, P. Khetarpal, H. S. Grover, N. M. A. Krishnan and K. M. Jablonka, Probing the limitations of multimodal language models for chemistry and materials research, *arXiv*, 2025, preprint, arXiv:2411.16955, DOI: 10.48550/arXiv.2411.16955, https://arxiv.org/abs/2411.16955.
36 D. Prasad, M. Pimpude and A. Alankar, Towards Development of Automated Knowledge Maps and Databases for Materials Engineering using Large Language

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Models, *arXiv*, 2024, preprint, arXiv:2402.11323, DOI: 10.48550/arXiv.2402.11323, https://arxiv.org/abs/2402.11323.
37 S. Gupta, A. Mahmood, P. Shetty, A. Adeboye and R. Ramprasad, *Commun. Mater.*, 2024, **5**, 269.
38 C. Ekuma, Dynamic In-context Learning with Conversational Models for Data Extraction and Materials Property Prediction, *arXiv*, 2024, preprint, arXiv:2405.10448, DOI: 10.48550/arXiv.2405.10448.
39 J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques and A. D. White, PaperQA: Retrieval-Augmented Generative Agent for Scientific Research, *arXiv*, 2023, preprint, arXiv:2312.07559, DOI: 10.48550/arXiv.2312.07559.
40 P. R. Maharana, A. Verma and K. Joshi, *J. Phys.*, 2025, **8**, 035006.
41 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, **6**, 525–535.
42 M. Ansari and S. M. Moosavi, *Digital Discovery*, 2024, **3**, 2607–2617.
43 H. Zhang, Y. Song, Z. Hou, S. Miret and B. Liu, HoneyComb: A Flexible LLM-Based Agent System for Materials Science, *arXiv*, 2024, preprint, arXiv:2409.00135, DOI: 10.48550/arXiv.2409.00135, https://arxiv.org/abs/2409.00135.
44 M. D. Skarlinski, S. Cox, J. M. Laurent, J. D. Braza, M. Hinks, M. J. Hammerling, M. Ponnapati, S. G. Rodriques and A. D. White, Language agents achieve superhuman synthesis of scientific knowledge, *arXiv*, 2024, preprint, arXiv:2409.13740, DOI: 10.48550/arXiv.2409.13740, https://arxiv.org/abs/2409.13740.
45 R. Feng, Y. Liang, T. Yin, P. Gao and W. Wang, *Agentic Assistant for Material Scientists*, 2025.
46 Y. Chiang, E. Hsieh, C.-H. Chou and J. Riebesell, LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation, *arXiv*, 2024, preprint, arXiv:2401.17244, DOI: 10.48550/arXiv.2401.17244, https://arxiv.org/abs/2401.17244.
47 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
48 S. Ghosh and A. Tewari, *Comput. Mater. Sci.*, 2026, **265**, 114521.
49 R. Odobesku, K. Romanova, S. Mirzaeva, O. Zagorulko, R. Sim, R. Khakimullin, J. Razlivina, A. Dmitrenko and V. Vinogradov, *npj Comput. Mater.*, 2025, **11**, 194.
50 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chem. Soc. Rev.*, 2025, 1125–1150.
51 JSON: A Lightweight data-interchange format, 2025, https://www.json.org/json-en.html, Accessed: 2025-05-12.
52 crewAI: Framework for orchestrating role-playing, autonomous AI agents, 2025, https://docs.crewai.com/, Accessed: 2025-05-12.
53 Scopus Search API, https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl, 2025, Accessed: 2025-03-02.
54 MySQL, https://www.mysql.com/, 2025, Accessed: March 4, 2025.
55 Chroma, https://www.trychroma.com/, 2025, Accessed: March 4, 2025.
56 ScienceDirect Article Metadata API, https://dev.elsevier.com/documentation/ArticleMetadataAPI.wadl, 2025, Accessed: 2025-03-07.
57 Open Access Button Metadata API, https://openaccessbutton.org/api, 2025, Accessed: 2025-03-07.
58 OA.Works, https://oa.works/, 2025, Accessed: 2025-03-07.
59 M. De Jong, W. Chen, H. Geerlings, M. Asta and K. A. Persson, *Sci. Data*, 2015, **2**, 1–13.
60 A. Roy, ComProScanner, https://github.com/slimeslab/ComProScanner, 2025, Accessed: 2025-10-26.
61 T. Hellert, J. Montenegro and A. Pollastro, PhysBERT: A Text Embedding Model for Physics Scientific Literature, arXiv, 2024, preprint, arXiv:2408.09574, DOI: 10.48550/arXiv.2408.09574, https://arxiv.org/abs/2408.09574.
62 sentence-transformers, sentence-transformers/all-mpnet-base-v2, 2021, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, https://huggingface.co/sentence-transformers/all-mpnet-base-v2, Accessed: 2025-07-22.
63 G. Team, A. Kamath and J. Ferret, Gemma 3 Technical Report, arXiv, 2025, preprint, arXiv:2503.19786, DOI: 10.48550/arXiv.2503.19786, https://arxiv.org/abs/2503.19786.
64 *DeepSeek-V3-0324 Release*, https://api-docs.deepseek.com/news/news250325, 2025, Accessed: 2025-05-16.
65 A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, The Llama 3 Herd of Models, *arXiv*, 2024, preprint, arXiv:2407.21783, DOI: 10.48550/arXiv.2407.21783, https://arxiv.org/abs/2407.21783.
66 The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 2025, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, Accessed: 2025-05-16.
67 Q. Team, Qwen3 Technical Report, *arXiv*, 2025, preprint, arXiv:2505.09388, DOI: 10.48550/arXiv.2505.09388, https://arxiv.org/abs/2505.09388.
68 A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang and Z. Qiu, Qwen2.5 Technical Report, *arXiv*, 2025, preprint, arXiv:2412.15115, DOI: 10.48550/arXiv.2412.15115, https://arxiv.org/abs/2412.15115.
69 Google, Introducing Gemini 2.0: our new AI model for the agentic era, 2024, https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, Accessed: 2025-08-11.
70 G. Comanici, E. Bieber, M. Schaekermann, *et al.*, Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, *arXiv*, 2025, preprint, arXiv:2507.06261, DOI: 10.48550/arXiv.2507.06261, https://arxiv.org/abs/2507.06261.

 © 2026 The Author(s). Published by the Royal Society of Chemistry

71 J. A. OpenAI, S. Adler, S. Agarwal, *et al.*, GPT-4 Technical Report, *arXiv*, 2024, preprint, arXiv:2303.08774, DOI: 10.48550/arXiv.2303.08774, https://arxiv.org/abs/2303.08774.
72 OpenAI, Introducing GPT-4.1 in the API, 2025, https://openai.com/index/gpt-4-1/, Accessed: 22 July 2025.
73 W. L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez and I. Stoica, *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, 2024.
74 Neo4j, Inc., Neo4j, 2025, https://neo4j.com/, https://neo4j.com/, Accessed: 2025-08-11.
75 Neo4j, Inc., Cypher Manual: Introduction, 2025, https://neo4j.com/docs/cypher-manual/current/introduction/, https://neo4j.com/docs/cypher-manual/current/introduction/, Accessed: 2025-08-11.

Open Access Article. Published on 25 March 2026. Downloaded on 4/22/2026 2:57:22 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.