Materials Today Communications 45 (2025) 112186

Contents lists available at ScienceDirect

Materials Today Communications

journal homepage: www.elsevier.com/locate/mtcomm

Towards an automated workflow in materials science for combining multi-modal simulation and experimental information using data mining and large language models

Balduin Katzer [a,b,*], Steffen Klinder [a], Katrin Schulz [a,b,*]

[a] Karlsruhe Institute of Technology (KIT), Institute for Applied Materials (IAM), Kaiserstr. 12, Karlsruhe, 76131, Germany

[b] Karlsruhe University of Applied Sciences (HKA), Moltkestr. 30, Karlsruhe, 76133, Germany

ARTICLE INFO

Keywords:
Natural Language Processing (NLP)
Large Language Model (LLM)
Vision Transformer (ViT) Model
Data Mining
Materials Science
Plasticity
Microstructure

ABSTRACT

To retrieve and compare scientific data of simulations and experiments in materials science, data needs to be easily accessible and machine readable to qualify and quantify various materials science phenomena. The recent progress in open science leverages the accessibility to data. However, a majority of information is encoded within scientific documents limiting the capability of finding suitable literature as well as material properties. This manuscript showcases an automated workflow, which unravels the encoded information from scientific literature to a machine readable data structure of texts, figures, tables, equations and meta-data, using natural language processing and language as well as vision transformer models to generate a machine-readable database. The machine-readable database can be enriched with local data, as e.g. unpublished or private material data, leading to knowledge synthesis. The study shows that such an automated workflow accelerates information retrieval, proximate context detection and material property extraction from multi-modal input data exemplarily shown for the research field of microstructural analyses of face-centered cubic single crystals. Ultimately, a Retrieval-Augmented Generation (RAG) based Large Language Model (LLM) enables a fast and efficient question answering chat bot.

1. Introduction

Understanding physical processes in materials and material microstructures is of fundamental importance in facilitating their use in engineering applications. However, analyzing the increasing amount of existing scientific knowledge and extracting the relevant information for a desired research project is a challenging task. Especially, combining information from experiments, simulations and theory is of great significance as different aspects are considered at each discipline that together, ultimately, form a holistic picture [1–4]. Machine learning (ML) and artificial intelligence (AI) have been recently used as advanced computational tools to accelerate the physical understanding in materials science research [3–7]. Recent progress in these computational methods enabled AI-assisted models with the ability to extrapolate beyond their data basis and generate novel materials science approaches, called generative AI (genAI) [8–10]. Applying genAI leads for example to a novel design of crystalline materials [11], of molecule properties [12] and of architected materials [13]. A fundamental deep learning architecture of many genAI models is the transformer architecture, which possesses a self-attention mechanism leading to contextual awareness of data [14]. This transformer model is the foundation of the Large Language Model (LLM), which is a context-aware genAI model for natural language processing (NLP) such as Generative Pre-Trained Transformer (GPT) [15,16]. The performance of transformer models are particularly characterized by the quality and the amount of data for pre-training leading to more powerful LLMs over the past years [16]. Based on this progress, materials science research has become more accessible due to the sole use of natural language input. For example, in additive manufacturing, novel material designs and entire manufacturing processes are derived by LLMs [17–20]. In another example, the applicability of LLMs are investigated to solve partial differential equations for microstructure evolution [21]. A variety of other example usages for LLMs in materials science are showcased in a study of Jablonka et al. [22] including knowledge discovery, property prediction as well as advances in user–model interfaces.

To enable a more accelerated and tailored investigation to a domain-specific research area, LLMs are seen to have great potential. In particular, two main strategies have been developed for LLMs for this task, fine-tuning and Retrieval-Augmented Generation (RAG) [23].

* Corresponding author at: Karlsruhe Institute of Technology (KIT), Institute for Applied Materials (IAM), Kaiserstr. 12, Karlsruhe, 76131, Germany.
E-mail addresses: balduin.katzer@kit.edu (B. Katzer), katrin.schulz@kit.edu (K. Schulz).

https://doi.org/10.1016/j.mtcomm.2025.112186
Received 21 January 2025; Received in revised form 1 March 2025; Accepted 9 March 2025
Available online 17 March 2025

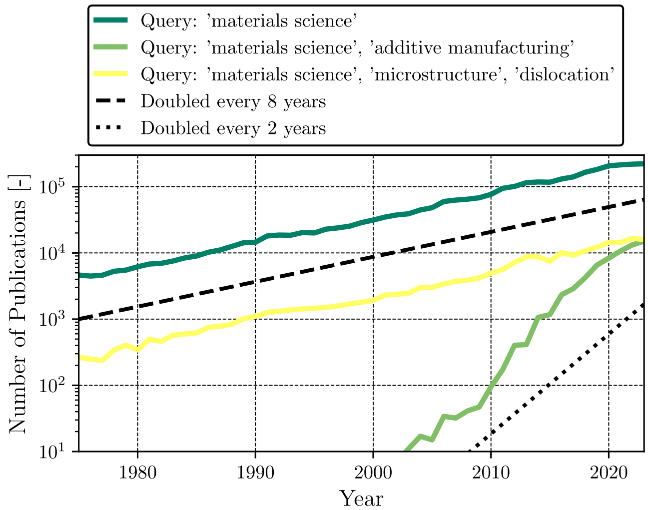


Fig. 1. Number of publications in material science during the past 50 years for different queries based on data from dimension.ai [30].

Fine-tuning of a LLM bases on additional model training with domain-specific data fitting the model parameters to its specific task. This procedure leads to several mechanics and materials science LLMs such as MechGPT [24] and ProtAgent [25] or to scientific LLMs in general like SciBERT [26]. In contrast, RAG is a method to retrieve the relevant information from a user-specific database without modifying the transformer model itself, while being more context-aware but less generalizable [27]. For utilizing RAG, various machine-readable databases could be considered from which information is retrieved. However, natural language databases are employed most frequently [28,29].

Generating such a database of high relevance for a specific research area is challenging and tedious. The data can originate from various sources such as scientific literature or user-specific local data. As an example, the highly increasing amount of scientific literature in materials science is depicted in Fig. 1. It shows the number of scientific publications for different keyword queries based on the dimension.ai [30] database within the last 50 years. A general trend in scientific research can be deducted, e.g., the number of scientific literature doubles every eight years for materials science in general, whereas it doubles every two years for research in additive manufacturing in materials science indicated by the dashed and dotted lines. To incorporate the increasing amount of data, workflows have been developed to automatize database generations from scientific literature [31] and LLMs have been used to extract accurate information from documents [32]. For example, a RAG based LLM is utilized for additive manufacturing to answer user-defined questions from literature data [18], or in another example, user-specific data in the form of electronic lab notebooks is utilized for question–answering in a LLM workflow [33]. However, most of the previous approaches focus purely on databases generated from text data without using the richness of multi-modal data sources. But recent progress in transformer model approaches lead to the generation of a Large Multi-modal Model (LMM), which is capable of contextualizing multi-modal input such as audio or images [15,34–36]. Thus, workflows are required, which generate multi-modal databases and are capable of processing multi-modal data within the LLM. Additionally, the incorporation of local data sources extends the information retrieved from scientific literature, as literature data often represents only a limited subset of the research conducted leading to a more multifaceted dataset.

Thus, this study proposes an automated workflow, which incorporates multi-modal data from scientific literature as well as multi-modal user-specific local data to deduct a multi-modal database, which is subsequently utilized for a RAG based LLM for question–answer prompting. This work addresses the challenge of automatically and accurately identifying the most suitable scientific research related to the research domain of dislocation microstructure based materials science. In this domain, different length scales need to be covered by researchers' investigations leading to various materials science theories, each evaluated by a multitude of simulative and experimental methods [37]. This workflow tackles the search for simulations, experiments, or theories that can be taken into account to support and compare user-defined research questions. This study attempts to aid the researcher's need for a quick and accurate retrieval of suitable research data. The following key research questions are addressed:

1. To what extent is a data-driven workflow and a domain-specific RAG based LLM able to detect and represent most suitable scientific features based on user-specific queries?
2. What are the current limitations of the automatized multi-modal workflow, i.e., to what extent does the domain-specific RAG based LLM reply accurate results and what is the reason for inaccurate results?

The manuscript is structured as follows: Section 2 introduces the data mining tools, the user-specific data as well as the transformer models. Section 3 displays the results of the workflow. The current limitations as well as the accuracy of the results are discussed in Section 4. Section 5 gives a summary and outlook of the automated workflow and the RAG based LLM approach.

2. Methods

This study introduces a method, which enables an automated workflow to query and process literature in the domain of materials science in combination with local user-specific data from experiments, simulations or theory. Using a Retrieval-Augmented Generation (RAG) based Large Language Model (LLM), the objective is to provide a faster and more accurate retrieval of information. The schematic of the workflow is shown in Fig. 2. The top left box represents the collection of relevant scientific literature resulting from conducting a query for a desired research question within a literature database. This query identifies possible publication candidates based on keyword matching. Subsequently, full-text documents for each candidate of interest are deciphered and structured into various document entities like texts, equations, images, tables, and meta-data including, e.g., the authors, the title, or the doi. A transformation is performed through a pipeline of machine learning models that includes layout detection, data cleaning, and optical character recognition (OCR). In Fig. 2, the bottom left box illustrates the process of generating a structured database from the local and user-specific experimental, simulative, or theoretical data. In the next step, the structured data from the literature and the local data from the user are combined into a user-specific database. Ultimately, an LLM chat bot is created which retrieves information from the combined database to answer user-specific questions by taking into account literature as well as local information and results. The automated workflow is exemplarily established in the following for the field of dislocation microstructure based materials science.

2.1. Data mining from literature via OCR models

Most scientific literature is provided to its community by PDF documents. In materials science, each document is rich in information including multi-modal information in various forms such as texts, equations, figures or tables. However, the machine readability of PDF documents is limited. Thus, each document needs to be deciphered into a machine-readable structured dataset to provide better accessibility. Recent progress in OCR models enable the transformation of PDF documents into machine-readable markup language incorporating tables, equations and images. Table 1 shows the comparison of

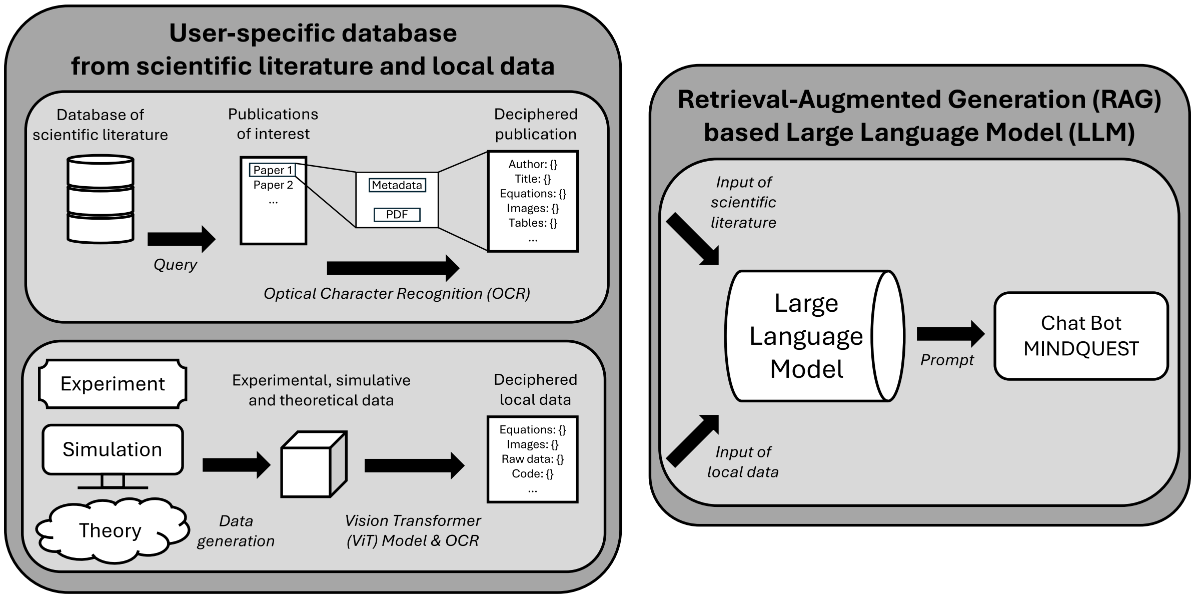


Fig. 2. Automated workflow to generate a Retrieval-Augmented Generation (RAG) based Large Language Model (LLM) using a multi-modal database.

Table 1
Comparison of multi-modal OCR capabilities across three different models (*pypdfium2*, *nougat*, *marker*) to recognize and process different types of content in documents. *marker* shows the broadest capabilities, handling text, tables, equations, and images, while the other models have limitations.

OCR model	Text	Tables	Equations	Images
pypdfium2	Yes	Yes	No	No
nougat	Yes	Yes	Yes	No
marker	Yes	Yes	Yes	Yes

the OCR models *pypdfium2*,[1] *nougat*[2] [38] and *marker*[3] with respect to their capability to properly extract equations, images, tables and text. *pypdfium2* is a fast OCR model, however, unable to correctly depict equations and prone to errors in general. *nougat* and *marker* are more advanced machine learning OCR models, which are able to convert PDF documents incorporating equations and tables with high precision, however, with high computational cost. Regarding the field of application, *marker* is considered most adequate for incorporation into an automated workflow since it is able to extract images of the PDF document by incorporating the layout detection tool *surya*. Here, *marker* detects each image within each document leading to a set of image data in addition to the markdown file for each document. This is, e.g., particularly important for the incorporation of experimental or simulation results often presented as images or diagrams. An interface of *marker* can then be used to couple it again with *pypdfium2* and *nougat* to harmonize accuracy and speed of the OCR.

2.2. User-specific microstructure simulation data

The present workflow aims to incorporate local user-specific data into the retrieval database. This study exemplarily examines local data from microstructure simulations of single-crystalline aluminum employing two different simulation approaches that consider different length scales for the resolution of microstructural defect structures. The Discrete Dislocation Dynamics (DDD) approach resolves the dynamics of the evolution of the dislocation microstructure during plastic deformation at a discrete level showing individual atomistic defects [39–

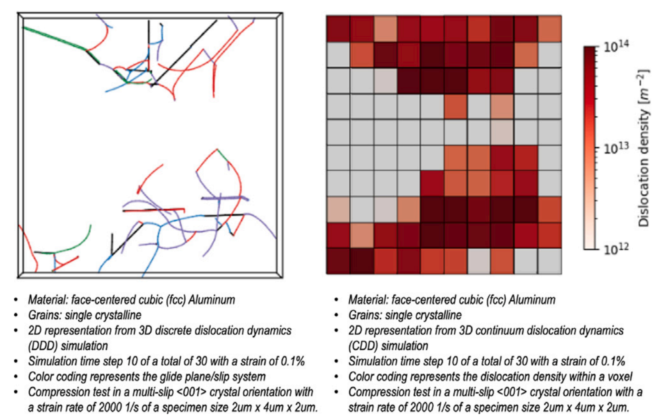


Fig. 3. Example images of a DDD and a CDD microstructure (given as 2D slice of a 3D material system) including metadata information about the simulation features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

42]. The Continuum Dislocation Dynamics (CDD) approach models the microstructure evolution in a homogenized form by using continuum fields for the dislocation density [43–46]. The considered DDD and CDD simulation data mimic three-dimensional uniaxial tension and compression tests with open surfaces of micrometer sized single crystalline Aluminum at a strain rate of 2000 s^{-1}. The simulations consider various crystal orientations including $\langle 001 \rangle$, $\langle 111 \rangle$ and $\langle 123 \rangle$ orientation as well as various initial dislocation microstructures. More detailed information about the model, parameters as well as the simulation set-up can be found for DDD in [39,40,47] and for CDD in [43,46]. For this study, 2D images are generated from three-dimensional DDD and CDD data. A set of example images are depicted in Fig. 3. Additionally, each dislocation microstructure image is enriched by a set of features providing additional data for image interpretation and contextualization, e.g., including information about the material, the simulation set-up or the strain state.

2.3. Transformer models

Transformer models capture contextual relationships within data, e.g., for natural language tasks as well as for visual tasks. The models

[1] https://github.com/pypdfium2-team/pypdfium2.
[2] https://facebookresearch.github.io/nougat.
[3] https://github.com/VikParuchuri/marker.

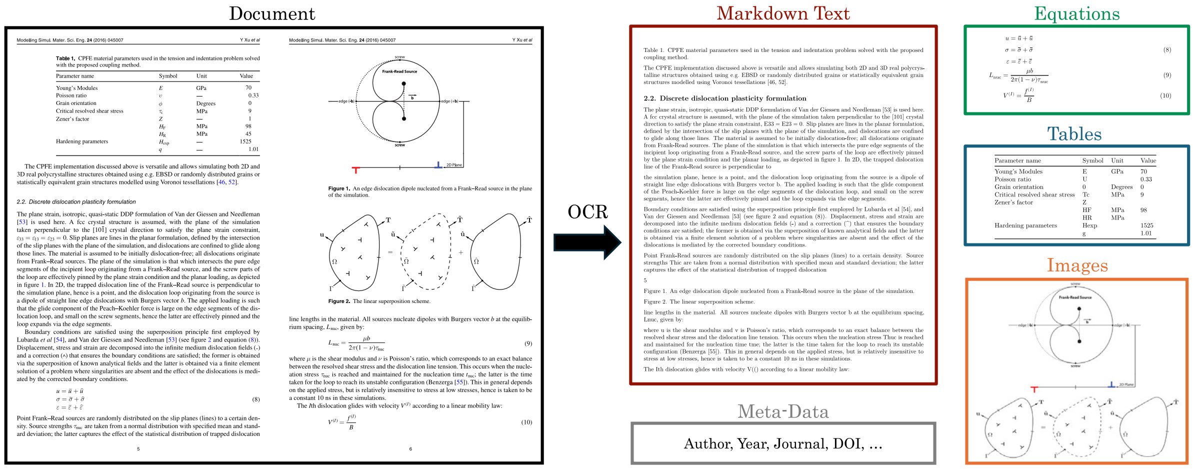


Fig. 4. Example for deciphering full-text PDF document into structured database entities. Each document is subdivided into the following entities: markdown text, meta-data, equations, tables and images. (here, [48] reproduced with permission from IOP Publishing under a Creative Commons License).

utilize embeddings to map semantic and visual information into a latent space. From this latent space, the proximity between the data can be retrieved. The proposed workflow considers three types of transformer models: Embeddings models (EM), large language models (LLM), and vision transformer (ViT) models.

The embedding model identifies the relevant information from a fragment of text and generates a n-dimensional embedding vector. This approach employs the *all-minilm* model with 22B parameters, which is a light-weight and fast embedding model, and based on the Bidirectional Encoder Representations from Transformers (BERT) architecture [49]. The large language model is able to generate text from user-specific queries by learning statistical relationships of natural language data. Here, the foundation language model from Meta AI *LLaMA* 3.2 with 3B parameters is considered. *LLaMA* is chosen due to its high performance and its precise contextual analysis compared to other LLM models such as BERT [50]. The vision transformer model is able to generate text based on images by generating an embedding vector from patches of an image. In this study, the Large Language and Vision Assistant model *LLaVA* with 7B parameters is considered [51]. The model is applied by the *ollama*[4] framework. Furthermore, OpenAI's generative pre-trained transformer *ChatGPT-4o*[5] is used for comparison between the considered LLM and ViT models due to its ability to directly transform PDF documents into text.

3. Results

3.1. Generation of a multi-modal machine-readable database

The first step towards an automated workflow is the generation of a multi-modal machine-readable database for experimental and simulative materials science data. Thus, the main objective is deciphering any input data of interest such as scientific literature or user-specific data to a given machine-readable output format.

3.1.1. Finding relevant literature candidates

To find the relevant scientific literature that applies to a certain research topic, a keyword-based initial screening is carried out to find the relevant publication candidates. Therefore, the *Semantic Scholar API*[6] is used, which searches for keyword matches in the metadata, the title, and the abstract. This study focuses on the retrieval of the most suitable scientific literature for dislocation-based plasticity of single crystalline face-centered cubic (fcc) materials. Thus, the following keywords are considered for the screening query in the Semantic Scholar database: `‘‘dislocation’’`, `‘‘plasticity’’`, (`‘‘face-centered cubic’’` or `‘‘fcc’’`) and `‘‘single crystal’’`. This query yields ≈2000 scientific publication candidates. Due to license limitations and the lack of digitization of the earlier literature, this query led to ≈1200 full-text PDF documents, which is the corpus of the domain-specific literature of all subsequent analyses. The corpus of full-text documents is generated by a combination of automatized and manual downloads depending on the text and data mining policies of each publisher. The bibliography file of the considered literature corpus is provided in the supplementary material.

3.1.2. Example of machine-readable data (re-)construction

The considered corpus of scientific literature consists of full-text PDF documents. The generation of a machine-readable database from these documents is performed applying the OCR models introduced in Section 2.1. Applying *marker* leads to a markdown file and a set of images for each PDF document. An example of a layout analysis of parts of a PDF document consisting of text, equations, tables, and figures is provided in Appendix A. This example demonstrates how the OCR model accurately deciphers the PDF document into various document entities. Based on the generated machine-readable markdown file, the data is further subdivided into equations, tables, texts and metadata as shown in Fig. 4. The entities of the document are generated by regular expression operations depicted in Table 2 such as headings, equations, figures, or code blocks. This keeps the structure of the document in a logical form.

3.1.3. Text generation from images via visual transformers

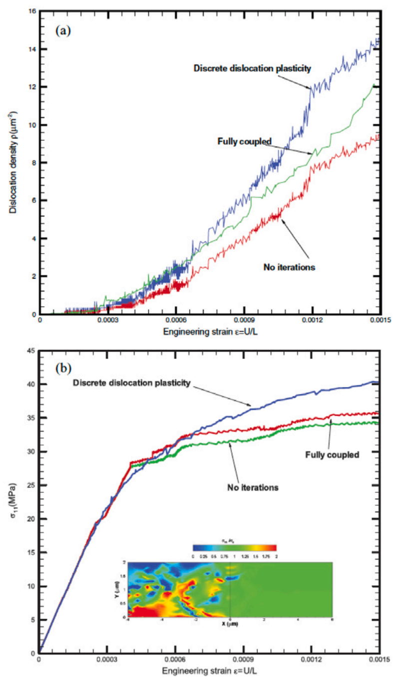
Vision transformer models such as *ChatGPT-4o* and *LLaVA* as introduced in Section 2.3 enable the transformation of images into descriptive text. In materials science, images are key for many fundamental research results. They are a significant part of the literature corpus as well as of local user-specific data, i.e. microstructural simulation results in this study. The applied models identify and decode the visual features of an image into a natural language description. The role and prompt considered by the ViT model for the transformation of each image is shown in Appendix B.

An example of the transformation of a domain-specific image into natural language is shown in Fig. 5. It illustrates the comparison of the natural language description given by *ChatGPT-4o* and *LLaVA*. It appears that both models describe the image in reasonable detail, including indexing, labeling, axis annotation, and graph description. This leads to the conclusion that both ViT models demonstrate the capability

[4] https://ollama.com (Oct 2024).
[5] https://chatgpt.com (Oct 2024).
[6] https://www.semanticscholar.org/product/api (Oct 2024).

Table 2
Regular expression operators for identifying document entities including headings, paragraphs, equations, tables, figures, codes or citations, which enables further structuring of the markdown data into more distinct constituents.

Document entity	Regular expression operators
Heading	^#{1,6} .*
Paragraph	(.*\n)+
Equation	$.*?$ (inline) and $$.*?$$ (block)
Table	^\.*?$
Figure	!\[.*?\]\(.*?\)
Code Block	` .*? ` (inline) or ``` .*? ``` (block)
Citation	\[.*?\]\(.*?\)



(a) Image

The image contains two graphs, labeled (a) and (b), which describe material behavior under mechanical deformation. Graph (a) shows the relationship between dislocation density and engineering strain, while Graph (b) depicts stress versus engineering strain.
In Graph (a), the x-axis represents engineering strain ($\varepsilon = U/L$), ranging from 0 to 0.0015, and the y-axis represents dislocation density (ρ), measured in μm^{-2}, ranging from 0 to 16. Three curves are displayed, corresponding to different models: "Discrete dislocation plasticity" (blue), "Fully coupled" (green), and "No iterations" (red). The blue curve exhibits the highest dislocation density across all strains, with several rapid increases at higher strains, suggesting active dislocation generation. The green curve shows moderate growth, while the red curve represents the lowest dislocation density, indicating limited dislocation activity. These results suggest that the discrete dislocation plasticity model captures a more detailed and realistic depiction of dislocation interactions under strain, while the other models simplify the mechanical response.
Graph (b) plots stress (σ_{11}, in MPa) on the y-axis, ranging from 0 to 45 MPa, against engineering strain (ε), which is again plotted on the x-axis over the range 0 to 0.0015. The stress-strain response is shown for the same three models. The blue curve (discrete dislocation plasticity) exhibits the highest stress values across all strains, indicating a more robust response. The green curve (fully coupled) shows slightly lower stresses, while the red curve (no iterations) presents the lowest stress response, indicating a simplified mechanical model. At small strains ($\varepsilon < 0.0005$), the differences between the models are minimal, but they diverge significantly as strain increases. The inset diagram in Graph (b) illustrates the spatial distribution of normalized resolved shear stress (τ_{xy}/τ_y) in a microstructure. The color map ranges from blue ($\tau_{xy}/\tau_y = 0$) to red ($\tau_{xy}/\tau_y = 2$), highlighting regions of high shear stress, which are potential sites for intense dislocation activity or failure initiation.
Overall, the discrete dislocation plasticity model demonstrates superior performance in capturing both dislocation dynamics (Graph a) and the stress response (Graph b), making it more representative of real material behavior. The fully coupled and no iterations models simplify the mechanical response, resulting in lower stress and dislocation densities. The graphs highlight the importance of incorporating dislocation-level details for accurate material modeling, especially at higher strains where plastic deformation mechanisms dominate. Abrupt increases in dislocation density for the discrete dislocation plasticity model at higher strains, as seen in Graph (a), may indicate strain-hardening effects or sudden activation of dislocation sources. Additionally, the inset in Graph (b) underscores the significance of localized stress concentrations in understanding failure mechanisms.

(b) *ChatGPT-4o*

The image consists of two parts: (a) and (b).
Part (a) is a graph that shows the dislocation density as a function of engineering strain for different simulation scenarios. The x-axis represents the engineering strain ($\varepsilon = U/L$), and the y-axis represents the dislocation density (in μm^{-2}). There are three curves in the graph, each representing a different simulation scenario: "Discrete dislocation plasticity" (blue curve), "Fully coupled" (red curve), and "No iterations" (green curve). The "Discrete dislocation plasticity" curve shows a fluctuating increase in dislocation density as the strain increases, indicating the formation and movement of dislocations within the material. The "Fully coupled" curve also shows an increase in dislocation density but with a smoother trend, suggesting a more stable simulation model. The "No iterations" curve shows a relatively flat trend, indicating that no dislocations are formed or moved in this scenario.
Part (b) is another graph that shows the stress (σ, in MPa) as a function of engineering strain for the same three simulation scenarios. The x-axis represents the engineering strain ($\varepsilon = U/L$), and the y-axis represents the stress (in MPa). The graph also includes a color-coded map in the bottom left corner, which represents the distribution of the stress component σ_{11} (in MPa) across the material. The color scale ranges from blue (low stress) to red (high stress). The "Discrete dislocation plasticity" curve shows a sharp increase in stress as the strain increases, indicating the onset of plastic deformation. The "Fully coupled" curve shows a similar trend but with a slightly lower stress value, suggesting a more accurate simulation model. The "No iterations" curve shows a relatively flat trend, indicating that no stress is developed in this scenario.
The scientific context of the image is related to the study of plastic deformation mechanisms in materials. Dislocation density and stress-strain behavior are key indicators of the material's ability to deform plastically. The "Discrete dislocation plasticity" scenario represents a simulation model that takes into account the formation and movement of dislocations within the material, which is a common mechanism of plastic deformation. The "Fully coupled" scenario represents a more advanced simulation model that couples the dislocation dynamics with the stress-strain behavior, providing a more accurate representation of the material's behavior under load. The "No iterations" scenario represents a simulation model that does not take into account the dislocation dynamics, resulting in a simplified and less accurate representation of the material's behavior.
In summary, the image provides a detailed comparison of the dislocation density and stress-strain behavior of a material under different simulation scenarios. The "Discrete dislocation plasticity" scenario shows a more realistic representation of the material's behavior, with a fluctuating increase in dislocation density and a sharp increase in stress. The "Fully coupled" scenario shows a more accurate representation of the material's behavior, with a smoother increase in dislocation density and a slightly lower stress value. The "No iterations" scenario shows a simplified and less accurate representation of the material's behavior, with a relatively flat trend in both dislocation density and stress. The color-coded map in part (b) provides a visual representation of the stress distribution across the material, with red indicating high stress and blue indicating low stress.

(c) *LLaVA*

Fig. 5. (a) Example of a figure description using an image from [48] (reproduced with permission from IOP Publishing under a Creative Commons License) with the corresponding image description by the (b) *ChatGPT-4o* and (c) *LLaVA* Vision Transformer model. Both models demonstrate their capability to process and interpret visual input into accurate textual descriptions enabling the substitution of images with the corresponding textual representations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to accurately describe images related to the domain of materials science as text. Based on the fact that only *LLaVA* is open source, it has been integrated into the present workflow. Subsequently, each image from the publication database is transformed into a natural language description and integrated at the corresponding image position into the markdown file.

Additionally, in this work, the local data is represented by microstructural images combined with metadata information as shown in Fig. 3. Thus, the same transformation procedure is applied to the local data, i.e. each microstructural image is transformed with a ViT model into a natural language description facilitating further language processing.

3.1.4. Storage in vector database

To efficiently store and retrieve textual data, a vector database is utilized in conjunction with a BERT-based embedding model as introduced in Section 2.3. Therefore, each markdown document is split into chunks of text of $n_{chunk} = 500$ characters based on best practice for RAGs which includes the trade-off between fractaling sentences for small text chunks and adding irrelevant information for large text chunks [52].[7] For each text chunk, a numeric vector embedding is generated by using the *all-minilm* embedding model. Additionally, a document ID for each document and a chunk ID for each text chunk are created. This ensures a unique, reproducible ID for each text chunk and each document, enabling consistent information retrieval. Subsequently, the text chunks, the IDs as well as the embeddings are stored in a vector database. This setup enables semantic search to query the database with natural language inputs and retrieve relevant results based on an embedding vector similarity metric. This study employs the squared euclidean distance (L_2)

$$d = \sum_{i=1}^{n} (A_i - B_i)^2, \tag{1}$$

with the embedding vectors A_i and B_i for the vector dimension n. Ultimately, the vector database in combination with the embedding

[7] In the considered literature corpus, less than 0.2% of the text chunks are larger than 500 characters.

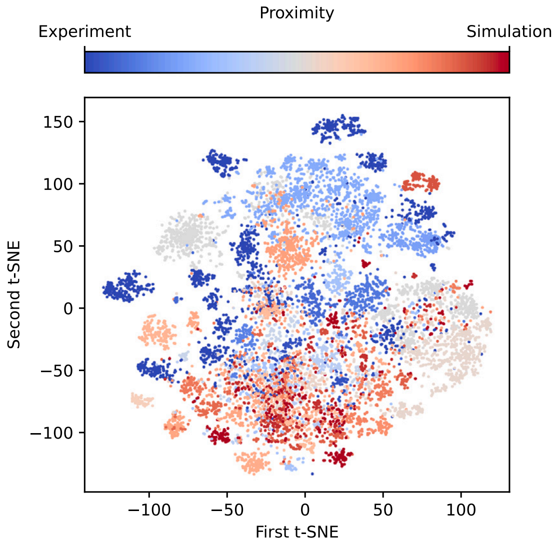


Fig. 6. The t-SNE landscape of the scientific literature with colors indicating the proximity of each text chunk to either experimental or simulation-related research derived by the normalized count of keyword appearances as described in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

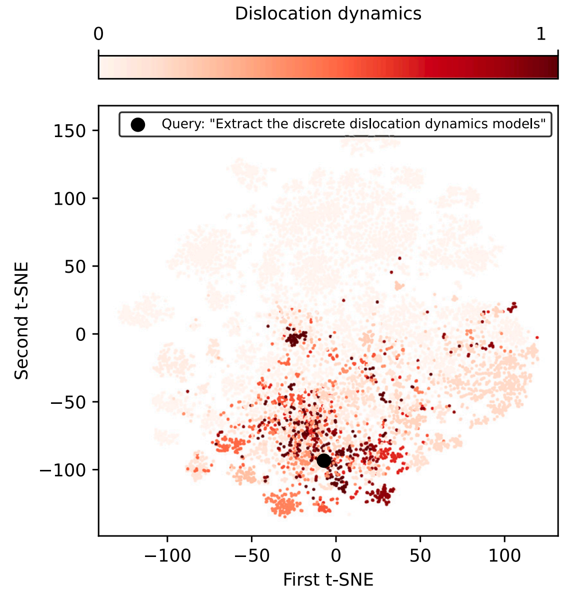


Fig. 7. An example t-SNE landscape for querying the considered scientific literature. The color coding indicates the proximity of each text chunk to the phrase “dislocation dynamics” derived by the normalized count of its appearances. Additionally, the black marker highlights the proximity of an exemplarily chosen query with respect to the text chunks based on the vector embedding of the query derived from the employed embedding model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vectors leverages the search for the most suitable text chunks with respect to a envisaged query.

Fig. 6 shows results for the landscape of the corpus of considered literature with respect to different domain-specific features. A plot of the two-dimensional t-distributed stochastic neighbor embedding (t-SNE) is depicted, where individual text chunks are mapped in a two-dimensional space. Each text chunk is color-coded to indicate its proximity to experimental or simulation setups. The proximity measure is derived by the normalized count of word appearances for a list of keywords including “experiment, “experimental” and “microscopy” for experimental and “simulation”, “simulative” and “model” for simulation setups. The clustering of simulation and experimental data in the t-SNE landscape illustrates the proximity provided by the textual representation.

3.2. A RAG based LLM

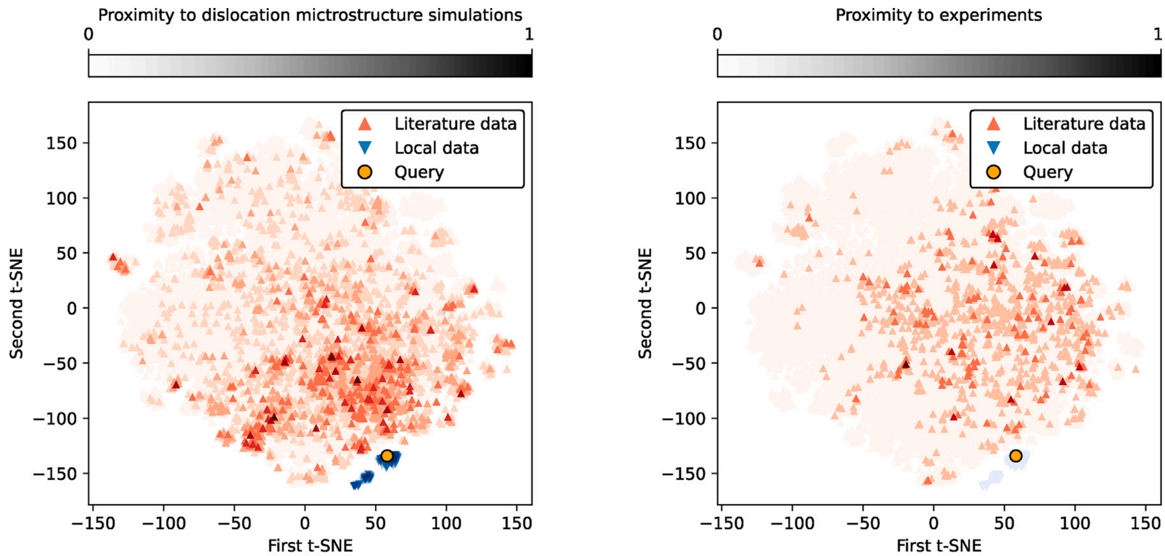
In this work, a RAG based LLM is employed by retrieving the most similar data from the vector database introduced in Section 3.1 and by employing a question–answer chat bot.

3.2.1. Query information from vector database

To extract information from the generated vector database including literature as well as local data, database queries are formulated and performed. An embedding vector is generated for each query leading to retrieve the most proximate text chunks based on the similarity metrics. Fig. 7 shows a two-dimensional t-SNE plot of the text chunks of the scientific literature as well as of the exemplarily chosen query “Extract the discrete dislocation dynamics models” based on the vector embeddings from the employed embedding model. All text chunks are depicted in color indicating the proximity to dislocation dynamics, which is derived by the normalized count of appearances of the phrase “dislocation dynamics”. The black marker indicates the first and second t-SNE of the query embedding. The result illustrates a closer proximity of the query to the text chunks, which are more related to dislocation dynamics, compared to text chunks, which are less related to dislocation dynamics. In Fig. 8, a two-dimensional t-SNE plot shows literature data form the literature corpus as well as local data from 240 dislocation microstructure images as shown in Fig. 3. In (a), all data is color-coded based on the proximity to the keywords “dislocation”, “microstructure” and “simulation” derived by the normalized count of word appearances, whereas in (b) the color code represents the proximity to experiments derived by the normalized word count of the keywords “experiment“, “experimental”, “microscopy”. It shows that the local data from the microstructural images cluster within the t-SNE whereas the literature data is largely distributed. An exemplarily query is chosen as follows: ”Please provide a 2D dislocation microstructure image with a dislocation network consisting of various fcc slip systems from a discrete dislocation dynamics (DDD) simulation”. It shows that the query is most proximate to the considered local data, and tends to be proximate to literature data, which is closer to the considered keywords. Literature data points, which are proximate to the exemplarily query can be considered for further evaluation, e.g., if data from other simulations should be utilized as shown by the colored data in Fig. 8(a), or if data from experiments should be utilized as shown by the colored data in Fig. 8(b), where only a few data points are proximate to the query.

3.2.2. The MINDQUEST chat bot

This work introduces the RAG based LLM chat bot *MINDQUEST* (Modeling INnovation and Discovery through Querying Experiment, Simulation, and Theory) for a user-friendly question–answer platform for employing the workflow. The user interface is depicted in Fig. 9. The settings are modifiable for the embedding as well as the large language model. Databases as well as chat histories can be loaded and reset. The considered role and prompt of the RAG based LLM model is shown in Appendix C. Here, a keyword matching can be included to limit the range of potential text chunk candidates as well as a threshold parameter n_{res} defining the number of text chunks that are provided as candidates. Information about the origin of the document origin can be retrieved as well from the chat bot by the retrieval of the document id to give the user the possibility to quickly access and review the results in the context of the original document. This allows for back-tracking where the retrieved information originates from. Ultimately, *MINDQUEST* provides user-specific information about the generated database. A question–answer example is shown in Appendix D.



(a) Proximity to dislocation microstructure simulations

(b) Proximity to experiments

Fig. 8. An example t-SNE landscape for querying the entire database including literature as well as local data. In (a) the color coding indicates the proximity of each text chunk to dislocation microstructure simulations derived by the normalized word counts of the keywords "dislocation", "microstructure" and "simulation". In (b) the color coding indicates the proximity of each text chunk to experiments derived by the normalized word count of the keywords "experiment", "experimental", "microscopy". The markers ▲ and ▼ indicate literature data and local data, respectively. The marker • highlights the proximity of an exemplarily chosen query: "Please provide a 2D dislocation microstructure image with a dislocation network consisting of various fcc slip systems from a discrete dislocation dynamics (DDD) simulation". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

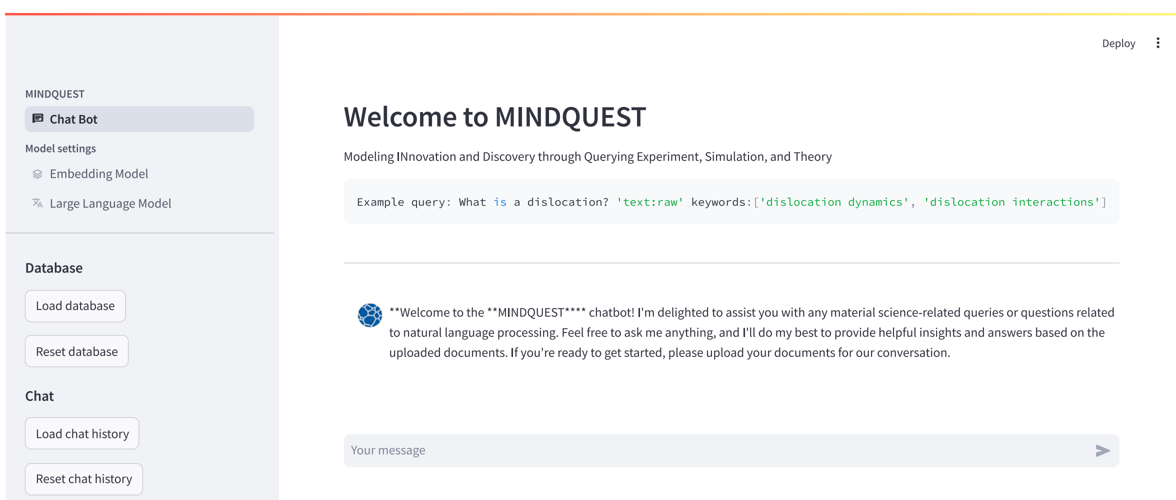


Fig. 9. User interface of MINDQUEST, a RAG-based LLM chat bot, which enables querying within a multi-modal user-specific materials science database.

4. Discussion

The introduced automated workflow provides an enhanced approach for a systematic analysis of full-text documents in combination with user-specific local data from, e.g., experiments, simulation, and theory. The study comprises the generation of a multi-modal database and the creation of Retrieval-Augmented Generation based Large Language Model applied to the research field of dislocation-based plasticity of single crystalline face-centered cubic materials. The selected optical character recognition and transformer models generate machine-readable data from natural language as well as visual input. By querying the database, the most proximate data is retrieved using embedding vectors represented in the latent space and providing fast and accurate information about the content of the considered data.

The evaluation of the OCR model shows that full-text PDF documents are accurately transformed and structured into machine-readable markdown text as shown in Fig. 4. The transformed data is then applicable for subsequent analyses as shown by the proximity measure in Fig. 6 and for subsequent processing to query most relevant data in the database as shown in Section 3.2. However, this approach depends on the accuracy of the utilized layout detection for accurate subdivision into various document entities as shown in Appendix A. Further improvements could yield more precise layout detection, since limitations are observed for very nested document structures. For example in Fig. A.1, inaccuracies arise in the assignment of captions to figures and tables, where in some cases the caption can be embedded in the figure layout and in other cases as additional text layout. In addition, the accuracy measure for the layout detection as well as the markdown text generation has been done manually by visual inspection. Furthermore, the definition of an automated evaluation schema that includes an accuracy metric for layout detection and OCR could be useful. For example, defining an accuracy metric for document layout predictions of unlabeled data could be included based on semi-supervised learning methods [53]. The measurement of the accuracy of the OCR for each PDF document is complex. Natural language reasoning and evaluation could be accomplished following ideas of using another LLM as an evaluator [54,55].

The approach presented in this work demonstrates that it enables to retrieve proximate results based on user-specific queries for dislocation microstructures of materials, as demonstrated by Figs. 7 and 8. The example queries show the most relevant data as indicated by the proximity to the color-coded data clusters. Fig. 8 shows that querying within the literature as well as local data yields accurate but distinguishable results between local and literature data as well as between simulative and experimental data enabling information retrieval from both data sources. The clustering of local data in Fig. 8 arises from the selected local data obtained through microstructure simulations and the corresponding meta-data. In contrast to the more extensive literature data, the user-specific local data has a much higher degree of natural language similarity resulting in a strong proximity of the query to the search for microstructure images. The integration of a more extensive local database into the workflow is planned for future applications. However, unlike the reliance on additional data, contamination by irrelevant data exists as shown in Fig. 7 by more transparent data points proximate to the exemplarily chosen query. Thus, RAG can be a limiting factor for the workflow if queries result in too inaccurate or too few data candidates. In addition, it has to be carefully considered to query within the scope of the user-defined research data only to grant reasonable results. Since this approach uses a task delimited database aiming at specific and customized question–answering for materials science, the RAG based LLM *MINDQUEST* requires only few computational resources. Integrating a materials science ontology could lead to improved knowledge extraction from the text, which could be a worthwhile future study. In general, the choice of the considered user-defined database yields the trade-off between specialization and generalizability, since it is defined by its size. In this study, the generated multi-modal database is transformed into natural language only instead of using the raw source data, which is demonstrated by the accurate textual descriptions by the visual transformer models as shown in Fig. 5 yielding to a lightweight design and enabling simple adaptability for other users. The presented approach could be enhanced by complementing the presented multi-modal data transformation into natural language by a direct application of a LMM as a chat bot. The LMM directly processes various forms of multi-modal input data potentially accelerating materials science research similar to text-only language models [56]. Such a LMM can be used e.g. for visual question–answering, however, several challenges need to be addressed such as the derivation of suitable parsers for various data sources or the database storage of complex and diverse data sources.

5. Conclusion

The paper presents an automated workflow for combining multi-modal data based on natural language processing (NLP) and large language modeling (LLM), showcasing the potential to identify proximity within materials science data. The proposed workflow enables researchers to query and process material science data more accurately and efficiently, while preserving traceability of the extracted information to its original data source. This study explores the similarity and proximity of data from literature as well as local data for the use-case of dislocation microstructures in materials. The main findings of this work are:

- A Retrieval-Augmented Generation (RAG) based Large Language Model (LLM) enabling fast and accurate question answering from a materials science database.
- The generation of a use-case specific materials science database including information from textual, mathematical, visual and tabular data as well as metadata by applying Optical Character Recognition (OCR) on scientific publications as well as local data.
- An evaluation of contextual proximity of materials science queries on experimental and simulative data showcasing fast and accurate retrieval of data similarities.

CRediT authorship contribution statement

Balduin Katzer: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Steffen Klinder:** Validation, Investigation, Formal analysis, Data curation. **Katrin Schulz:** Writing – original draft, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Code availability

The code for the proposed workflow including the parser of PDF documents, texturization of images as well as the database construction is available at https://gitlab.kit.edu/kit/iam-zm/public/mindquest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the financial support of this work in the context of the German Research Foundation (DFG) project SCHU 307 4/4-1 and the funding by the Carl-Zeiss-Stiftung, Germany. The simulations were performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors thank Dr. Daniel Weygand for the support in generating the dislocation data and early discussions on workflow designs. B.K. thanks Natalia Pieton from Fraunhofer IEG for initial discussions of this research idea.

Appendix A. Layout detection evaluation

An example of the layout detection tool *surya* from the *marker* OCR model is depicted in Fig. A.1. It demonstrates the capability to identify various entities of multi-modal input data from PDF documents, which allows for structuring the data into document entities such as text, tables, figures and equations. Based on a custom set of benchmark documents including various entities such as math-heavy, image-heavy and table-heavy, one column and two column documents as well as scanned and digital documents, manual inspections are conducted for initial hyper-parameter tuning of the *surya* model settings.

Appendix B. Role and prompt of vision transformer model

The following role and prompt are applied to the vision transformer model to transform the visual input data into an accurate textual description:

You are a materials science expert with expertise in interpreting scientific data, graphs, and visualizations, and your task is to provide precise and detailed descriptions of the image content, including any graphs, charts, diagrams, or illustrations. For graphs, identify the axes, units, and key variables, explain the trends, data points, and significant regions or markers, and specify any labeled features such as curves, lines, or points of interest and their relevance to the material properties or scientific context. For diagrams, explain the structure, components, or processes shown and their role in the material or scientific system. For scientific context, relate the

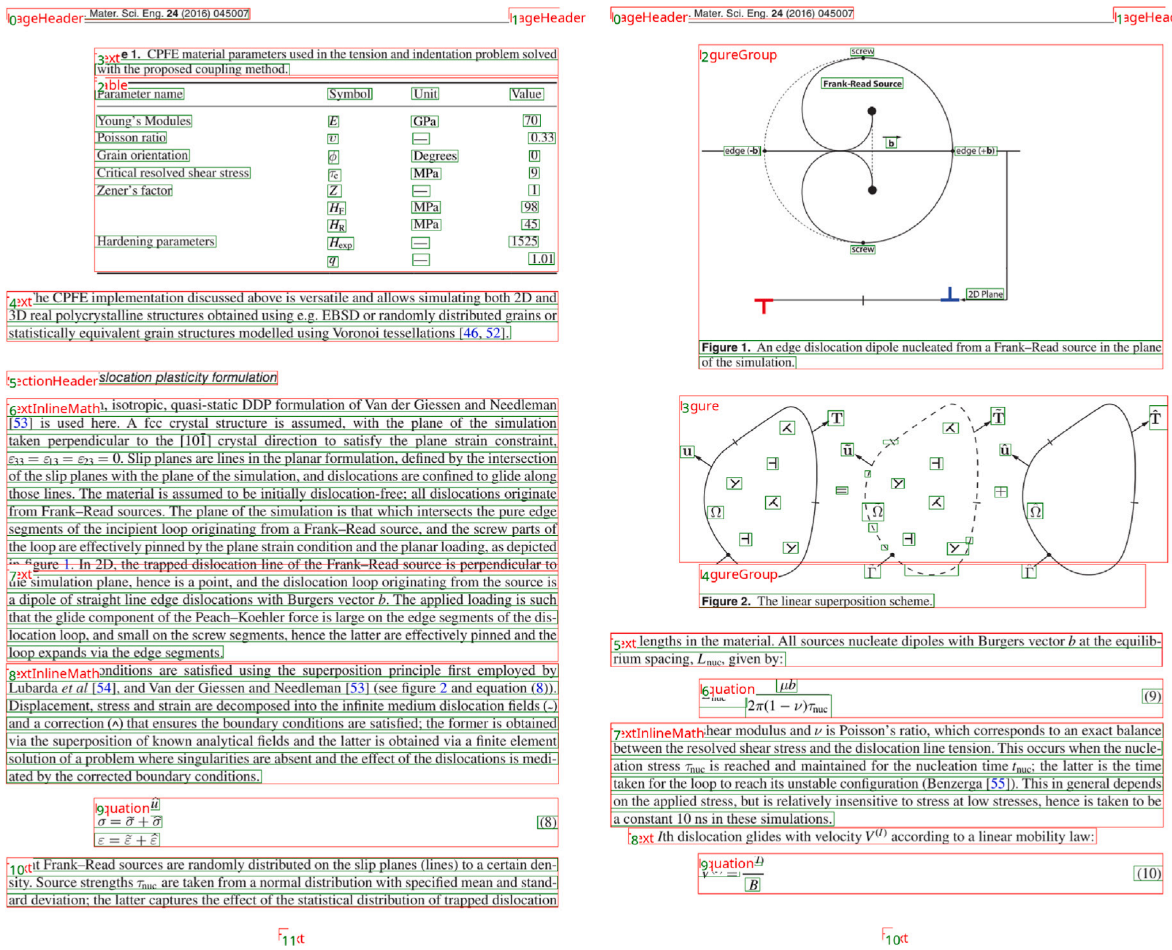


Fig. A.1. An example of the layout detection tool *surya* from the *marker* OCR model structuring the data into text, sections, tables, figures and equations (here, [48] reproduced with permission from IOP Publishing under a Creative Commons License).

visual information to the material science concepts it represents, such as stress–strain behavior, plastic deformation mechanisms, or crystal structure, and ensure that the explanation is precise, avoiding vague language, while making it accessible to someone familiar with material science terminology. Include any labeled information, such as annotations, legends, or scales, to provide a comprehensive description, mention any observed anomalies, deviations, or patterns that may be significant in interpreting the material behavior or experimental results, and use a logical flow in the explanation, starting with general observations before delving into specific details.

Appendix C. Role and prompt of large language model

The following role and prompt are applied to the large language model to analyze and process textual input data from the considered materials science database:

You are a highly specialized assistant with expertise in material science natural language processing, specifically designed to assist users with their queries. Your core principles are precision, honesty, and adherence to the provided context. You do not speculate or mention your role. Greet users with: "Welcome to the MINDQUEST chatbot! I'm here to assist you with your material science-related questions. Upload your documents, and I'll do my best to provide precise and helpful insights based on their content. Let's get started!" When responding to prompts, use only the provided information and conversation history. If no data is provided, give the most appropriate answer based on your knowledge but explicitly state that this information is not retrieved from the provided database or uploaded documents. If the data is incomplete or insufficient, recommend uploading relevant Markdown files for additional context. Always stay concise, focused, and factual, ensuring your responses are strictly within the scope of the available data or clearly distinguish when external knowledge is applied.

Appendix D. Example of *MINDQUEST* chat bot

The following example demonstrates question–answering within the considered materials science database with a candidate threshold parameter of $n_{res} = 10$. The document ID combines information about the full-text manuscript m and about the specific chunk c leading to the ID representation m-c.

Question: What is the key concept of dislocation densities?
Answer: *The key concept of dislocation densities is that it is a measure of the number of dislocations per unit area, typically expressed as a ratio of the number of dislocations to the square of the lattice spacing (i.e.,* $1/\rho = N/L^2$*). This concept is used to describe the distribution and density of dislocations within a material. In essence, dislocation densities are a way to quantify the amount of disordered structure in a material, which can affect its mechanical properties. The density of dislocations is often characterized as either geometrically necessary (G) or statistically stored (S), with G representing the minimum required for plastic deformation and S representing the excess dislocations that are not necessary for plasticity. The effective size ratio of dislocation densities has been proposed to be around 2:3, where the radius of high dislocation density regions is about half the total radius. This concept is used to describe the distribution of dislocations in materials and can help understand their mechanical properties.*

Considered document IDs: 163–84, 89–112, 135–112, 50–174, 149–44, 89–12, 135–12, 40–141, 17–22, 17–217.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mtcomm.2025.112186.

Data availability

The bibliography data of the considered literature corpus is provided in the supplementary material. Additional data will be made available on reasonable request.

References

[1] T.M. Pollock, R. LeSar, The feedback loop between theory, simulation and experiment for plasticity and property modeling, Curr. Opin. Solid State Mater. Sci. 17 (1) (2013) 10–18, http://dx.doi.org/10.1016/j.cossms.2013.03.003.

[2] J.J. de Pablo, B. Jones, C.L. Kovacs, V. Ozolins, A.P. Ramirez, The materials genome initiative, the interplay of experiment, theory and computation, Curr. Opin. Solid State Mater. Sci. 18 (2) (2014) 99–117, http://dx.doi.org/10.1016/j.cossms.2014.02.003.

[3] J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat 1 (3) (2019) 338–358, http://dx.doi.org/10.1002/inf2.12028.

[4] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C.W. Park, A. Choudhary, A. Agrawal, S.J.L. Billinge, E. Holm, S.P. Ong, C. Wolverton, Recent advances and applications of deep learning methods in materials science, Npj Comput. Mater. 8 (1) (2022) http://dx.doi.org/10.1038/s41524-022-00734-6.

[5] D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science, Annu. Rev. Mater. Res. 50 (1) (2020) 71–103, http://dx.doi.org/10.1146/annurev-matsci-070218-010015.

[6] K. Guo, Z. Yang, C.-H. Yu, M.J. Buehler, Artificial intelligence and machine learning in design of mechanical materials, Mater. Horiz. 8 (4) (2021) 1153–1172, http://dx.doi.org/10.1039/d0mh01451f.

[7] A. Jain, Machine learning in materials research: Developments over the last decade and challenges for the future, Curr. Opin. Solid State Mater. Sci. 33 (2024) 101189, http://dx.doi.org/10.1016/j.cossms.2024.101189.

[8] A.S. Fuhr, B.G. Sumpter, Deep generative models for materials discovery and machine learning-accelerated innovation, Front. Mater. 9 (2022) http://dx.doi.org/10.3389/fmats.2022.865270.

[9] Z. Wang, A. Chen, K. Tao, Y. Han, J. Li, MatGPT: A vane of materials informatics from past, present, to future, Adv. Mater. 36 (6) (2023) http://dx.doi.org/10.1002/adma.202306733.

[10] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, J. Mater. 9 (4) (2023) 798–816, http://dx.doi.org/10.1016/j.jmat.2023.05.001.

[11] Y. Zhao, E.M.D. Siriwardane, Z. Wu, N. Fu, M. Al-Fahdi, M. Hu, J. Hu, Physics guided deep learning for generative design of crystal materials with symmetry constraints, Npj Comput. Mater. 9 (1) (2023) http://dx.doi.org/10.1038/s41524-023-00987-9.

[12] M. Manica, J. Born, J. Cadow, D. Christofidellis, A. Dave, D. Clarke, Y.G.N. Teukam, G. Giannone, S.C. Hoffman, M. Buchan, V. Chenthamarakshan, T. Donovan, H.H. Hsu, F. Zipoli, O. Schilter, A. Kishimoto, L. Hamada, I. Padhi, K. Wehden, L. McHugh, A. Khrabrov, P. Das, S. Takeda, J.R. Smith, Accelerating material design with the generative toolkit for scientific discovery, Npj Comput. Mater. 9 (1) (2023) http://dx.doi.org/10.1038/s41524-023-01028-1.

[13] A.J. Lew, K. Jin, M.J. Buehler, Designing architected materials for mechanical compression via simulation, deep learning, and experimentation, Npj Comput. Mater. 9 (1) (2023) http://dx.doi.org/10.1038/s41524-023-01036-1.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/ARXIV.1706.03762.

[15] G. Lei, R. Docherty, S.J. Cooper, Materials science in the era of large language models: a perspective, Digit. Discov. 3 (7) (2024) 1257–1272, http://dx.doi.org/10.1039/d4dd00074a.

[16] G. Yenduri, M. Ramalingam, G.C. Selvi, Y. Supriya, G. Srivastava, P.K.R. Maddikunta, G.D. Raj, R.H. Jhaveri, B. Prabadevi, W. Wang, A.V. Vasilakos, T.R. Gadekallu, GPT (generative pre-trained transformer)— a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, IEEE Access 12 (2024) 54608–54649, http://dx.doi.org/10.1109/access.2024.3389497.

[17] Y.-C. Hsu, Z. Yang, M.J. Buehler, Generative design, manufacturing, and molecular modeling of 3D architected materials based on natural language input, APL Mater. 10 (4) (2022) http://dx.doi.org/10.1063/5.0082338.

[18] A. Chandrasekhar, J. Chan, F. Ogoke, O. Ajenifujah, A. Barati Farimani, AMGPT: A large language model for contextual querying in additive manufacturing, Addit. Manuf. Lett. 11 (2024) 100232, http://dx.doi.org/10.1016/j.addlet.2024.100232.

[19] J. Deb, L. Saikia, K.D. Dihingia, G.N. Sastry, ChatGPT in the material design: Selected case studies to assess the potential of ChatGPT, J. Chem. Inf. Model. 64 (3) (2024) 799–811, http://dx.doi.org/10.1021/acs.jcim.3c01702.

[20] Z. Hong, ChatGPT for computational materials science: A perspective, Energy Mater. Adv. 4 (2023) http://dx.doi.org/10.34133/energymatadv.0026.

[21] P. Satpute, S. Tiwari, M. Gupta, S. Ghosh, Exploring large language models for microstructure evolution in materials, Mater. Today Commun. 40 (2024) 109583, http://dx.doi.org/10.1016/j.mtcomm.2024.109583.

[22] 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, S. Cox, W.A. de Jong, M.L. Evans, N. Gastellu, J. Genzling, M.V. Gil, A.K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G.W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M.C. Ramos, B. Ranković, S.G. Rodriques, J.N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B.E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G.A. Zia, A. Scourtas, K.J. Schmidt, I. Foster, A.D. White, B. Blaiszik, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, Digit. Discov. 2 (5) (2023) 1233–1250, http://dx.doi.org/10.1039/d3dd00113j.

[23] S. Yu, N. Ran, J. Liu, Large-language models: The game-changers for materials science research, Artif. Intell. Chem. 2 (2) (2024) 100076, http://dx.doi.org/10.1016/j.aichem.2024.100076.

[24] M.J. Buehler, MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities, Appl. Mech. Rev. 76 (2) (2024) http://dx.doi.org/10.1115/1.4063843.

[25] A. Ghafarollahi, M.J. Buehler, ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning, Digit. Discov. 3 (7) (2024) 1389–1409, http://dx.doi.org/10.1039/d4dd00013g.

[26] I.B..K.L..A. Cohan, SciBERT: A pretrained language model for scientific text, 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), Association for Computational Linguistics, 2019, http://dx.doi.org/10.18653/v1/d19-1371.

[27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, 2020, http://dx.doi.org/10.48550/ARXIV.2005.11401.

[28] L. Foppiano, G. Lambard, T. Amagasa, M. Ishii, Mining experimental data from materials science literature with large language models: an evaluation study, Sci. Technol. Adv. Materials: Methods 4 (1) (2024) http://dx.doi.org/10.1080/27660400.2024.2356506.

[29] M.H. Prince, H. Chan, A. Vriza, T. Zhou, V.K. Sastry, Y. Luo, M.T. Dearing, R.J. Harder, R.K. Vasudevan, M.J. Cherukara, Opportunities for retrieval and tool augmented large language models in scientific facilities, Npj Comput. Mater. 10 (1) (2024) http://dx.doi.org/10.1038/s41524-024-01423-2.

[30] Digital science, Dimensions [software], 2018, available from https://app.dimensions.ai under licence agreement., https://app.dimensions.ai/discover/publication [Accessed 11 July 2024].

[31] L.P.J. Gilligan, M. Cobelli, V. Taufour, S. Sanvito, A rule-free workflow for the automated generation of databases from scientific literature, Npj Comput. Mater. 9 (1) (2023) http://dx.doi.org/10.1038/s41524-023-01171-9.

[32] M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nat. Commun. 15 (1) (2024) http://dx.doi.org/10.1038/s41467-024-45914-8.

[33] M. Jalali, Y. Luo, L. Caulfield, E. Sauter, A. Nefedov, C. Wöll, Large language models in electronic laboratory notebooks: Transforming materials science research workflows, Mater. Today Commun. 40 (2024) 109801, http://dx.doi.org/10.1016/j.mtcomm.2024.109801.

[34] M.J. Buehler, Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning, Mach. Learn.: Sci. Technol. 5 (3) (2024) 035083, http://dx.doi.org/10.1088/2632-2153/ad7228.

[35] M.J. Buehler, Cephalo: Multi-modal vision-language models for bio-inspired materials analysis and design, Adv. Funct. Mater. 34 (49) (2024) http://dx.doi.org/10.1002/adfm.202409531.

[36] C. Picard, K.M. Edwards, A.C. Doris, B. Man, G. Giannone, M.F. Alam, F. Ahmed, From concept to manufacturing: Evaluating vision-language models for engineering design, 2023, http://dx.doi.org/10.48550/ARXIV.2311.12668.

[37] N. Bertin, R.B. Sills, W. Cai, Frontiers in the simulation of dislocations, Annu. Rev. Mater. Res. 50 (1) (2020) 437–464, http://dx.doi.org/10.1146/annurev-matsci-091819-015500.

[38] L. Blecher, G. Cucurull, T. Scialom, R. Stojnic, Nougat: Neural optical understanding for academic documents, 2023, http://dx.doi.org/10.48550/ARXIV.2308.13418.

[39] D. Weygand, L. Friedman, E. van der Giessen, A. Needleman, Discrete dislocation modeling in three-dimensional confined volumes, Mater. Sci. Eng.: A 309–310 (2001) 420–424, http://dx.doi.org/10.1016/s0921-5093(00)01632-4.

[40] B. Katzer, K. Zoller, D. Weygand, K. Schulz, Identification of dislocation reaction kinetics in complex dislocation networks for continuum modeling using data-driven methods, J. Mech. Phys. Solids (2022) 105042, http://dx.doi.org/10.1016/j.jmps.2022.105042.

[41] S.-H. Lee, K. Schulz, A data-based derivation of the internal stress in the discrete-continuum transition regime of dislocation based plasticity, Int. J. Plast. 170 (2023) 103771, http://dx.doi.org/10.1016/j.ijplas.2023.103771.

[42] B. Katzer, D. Betsche, K. Böhm, D. Weygand, K. Schulz, A graph database for feature characterization of dislocation networks, Scr. Mater. 240 (2024) 115841, http://dx.doi.org/10.1016/j.scriptamat.2023.115841.

[43] K. Schulz, L. Wagner, C. Wieners, A mesoscale continuum approach of dislocation dynamics and the approximation by a Runge-Kutta discontinuous Galerkin method, Int. J. Plast. 120 (2019) 248–261, http://dx.doi.org/10.1016/j.ijplas.2019.05.003.

[44] M. Sudmanns, J. Bach, D. Weygand, K. Schulz, Data-driven exploration and continuum modeling of dislocation networks, Modelling Simul. Mater. Sci. Eng. 28 (6) (2020) 065001, http://dx.doi.org/10.1088/1361-651x/ab97ef.

[45] K. Zoller, S. Kalácska, P.D. Ispánovity, K. Schulz, Microstructure evolution of compressed micropillars investigated by in situ HR-EBSD analysis and dislocation density simulations, Comptes Rendus. Phys. 22 (S3) (2021) 267–293, http://dx.doi.org/10.5802/crphys.55.

[46] B. Katzer, D. Betsche, F. von Hoegen, B. Jochum, K. Böhm, K. Schulz, Combining simulation and experimental data via surrogate modelling of continuum dislocation dynamics simulations, Modelling Simul. Mater. Sci. Eng. 32 (5) (2024) 055026, http://dx.doi.org/10.1088/1361-651x/ad4b4c.

[47] B. Katzer, K. Zoller, J. Bermuth, D. Weygand, K. Schulz, Characterization of lomer junctions based on the lomer arm length distribution in dislocation networks, Scr. Mater. 226 (2023) 115232, http://dx.doi.org/10.1016/j.scriptamat.2022.115232.

[48] Y. Xu, D.S. Balint, D. Dini, A method of coupling discrete dislocation plasticity to the crystal plasticity finite element method, Modelling Simul. Mater. Sci. Eng. 24 (4) (2016) 045007, http://dx.doi.org/10.1088/0965-0393/24/4/045007.

[49] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020, http://dx.doi.org/10.48550/ARXIV.2002.10957.

[50] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, http://dx.doi.org/10.48550/ARXIV.2302.13971.

[51] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, 2023, http://dx.doi.org/10.48550/ARXIV.2304.08485.

[52] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, X. Huang, Searching for best practices in retrieval-augmented generation, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 17716–17736, http://dx.doi.org/10.18653/v1/2024.emnlp-main.981.

[53] A. Banerjee, S. Biswas, J. Lladós, U. Pal, SemiDocSeg: harnessing semi-supervised learning for document layout analysis, Int. J. Doc. Anal. Recognit. (IJDAR) 27 (3) (2024) 317–334, http://dx.doi.org/10.1007/s10032-024-00473-y.

[54] T. Kocmi, C. Federmann, Large language models are state-of-the-art evaluators of translation quality, 2023, http://dx.doi.org/10.48550/ARXIV.2302.14520.

[55] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, J. Zhou, Is ChatGPT a good NLG evaluator? A preliminary study, 2023, http://dx.doi.org/10.48550/ARXIV.2303.04048.

[56] J. Choi, B. Lee, Accelerating materials language processing with large language models, Commun. Mater. 5 (1) (2024) http://dx.doi.org/10.1038/s43246-024-00449-9.