

Automated Extraction of Materials System Charts using a Large Language Model Framework

Authors: Quanliang Liu¹, Maciej P Polak¹, MD Al Amin Shuvo¹, Hrishikesh Shridhar Deodhar¹, Jeongsoo Han¹, Dane Morgan¹, Hyunseok Oh^{1*}

¹Department of Materials Science & Engineering, University of Wisconsin-Madison; Madison, WI, USA, 53706

*Corresponding author. Hyunseok Oh (hyunseok.oh@wisc.edu)

Keywords

Materials system chart; Metallurgy; Processing-Structure-Property relationships; Large language model; Prompt engineering

Abstract

A framework leveraging large language models (LLMs) is developed to systematically extract and organize Processing-Mechanism-Structure-Mechanism-Property (P-M-S-M-P) relationships from materials science and engineering literature, with a particular focus on metallurgy. Using multi-stage prompts, our method identifies key properties, microstructures, processing methods, and associated mechanisms, then integrates them to generate comprehensive materials system charts. Additionally, the framework refines the extracted system charts for visualization, enabling the creation of informative diagrams that capture essential insights from each paper. Evaluated across 70 papers spanning multiple alloy systems and research types, the approach achieves 94% accuracy in mechanism extraction, 87% in information source labeling, and 97% in the human-machine readability index for processing, structure, and property entities. The prompts and codes are provided alongside guidelines for researchers unfamiliar with coding. This framework offers an effective methodology for knowledge extraction in materials science using the P-M-S-M-P framework.

Materials science continues to advance at a remarkable pace [1]. A prime example of this growth is high-entropy alloys (HEAs): since the first paper was published in 2004, the Web of Science database (keyword: “High entropy alloys” has recorded 3,206 articles in 2024. While the expanding literature fosters cross-fertilization, it also poses challenges for researchers to efficiently track and synthesize emerging knowledge.

Recently, large language models (LLMs) have demonstrated impressive capabilities in text retrieval, generation, and reasoning for complex language tasks [2–5]. Many applications exist within materials science, ranging from extracting numerical [6–9] and textual [10–13] data from the literature to planning and guiding autonomous experiments [14,15]. In particular, LLMs have an extraordinary capacity to generate responses based on instructions provided in the input prompt, without updating its internal parameters [16,17]. This “in-context learning” capability makes them ideally suited to perform the knowledge extraction task through prompt engineering. Although these capabilities have been applied to literature-based data mining in general, most work focuses on extracting simple numerical or textual data, rather than structuring materials science knowledge, which remains challenging due to the complex and hierarchical interplay of factors governing materials behavior.

We were inspired by the observation that materials science knowledge can be systematically organized using the Processing-Mechanism-(Micro)Structure-Mechanism-Property (P-M-S-M-P) framework [18,19]. The materials system chart—embodying this interconnected P-M-S-M-P framework—has been proposed as a fundamental tool [20–25]. These charts organize hierarchical entities within each processing, structure, and property domains while elucidating mechanistic links, providing a linearized representation of both deductive cause-and-effect scientific information flow (e.g., how a processing method influences microstructure) and inductive goal-means engineering empirical information flow (e.g., what processing methods should be chosen to achieve a target microstructure) [18,26–28]. Consequently, materials system charts enable researchers to efficiently review findings, identify knowledge gaps, and transfer insights across disciplines. Moreover, structured representations of materials knowledge can be readily integrated into computation tools and AI models, accelerating predictive modeling and materials discover.

In this paper, we introduce an LLM-assisted framework that extracts materials system charts from individual metallurgy papers (**Fig. 1**). Our approach guides the model through multiple engineered prompts to identify relevant properties, microstructures, processing methods, and mechanisms, consolidating the extracted information into a single structured table. Manual evaluation by the authors confirms high extraction accuracy. A separate set of prompts optimizes the table for visualization, by categorizing microstructures and reducing redundancy.

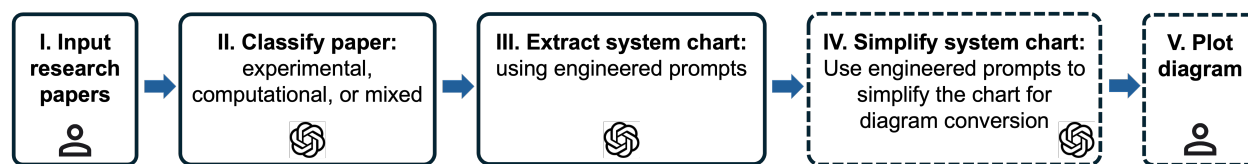


Fig. 1. Overview of the system chart extraction and visualization workflow using a conversational LLM.

The OpenAI ChatGPT API is implemented using Python 3.10.8 and version 1.54.3 of the openai package. We employ the gpt-4-1106-preview snapshot model with a temperature setting of 0.0 to produce deterministic, focused responses, ensuring consistent and reliable outputs. The maximum output is capped

at 4,000 tokens. The system message is configured as “You are an expert in the alloy field of Materials Science and Engineering.” All other parameters retain their default values: frequency_penalty = 0.0, presence_penalty = 0.0, top_p = 1.0, logprobs = False, stream = False, n = 1, logit_bias = null, and stop = null. For comparing results across different models, we use the up-to-date reasoning model o3-mini-2025-01-31, set to a reasoning_effort of medium, without a temperature parameter, and with all other defaults matching those listed above.

This section describes the step-by-step process for extracting materials system charts from metallurgy papers in PDF format (**Fig. 2**). Detailed prompt structures are available in **Note S1** of the **SI**. To illustrate the workflow, we present a case study based on the paper “*Ultrastrong Steel via Minimal Lattice Misfit and High-Density Nanoprecipitation*” [29].

The first step classifies the paper as experimental, computational, or mixed research to improve extraction accuracy. We observe that for computational papers, the model often misidentifies simulation tools—such as density functional theory (DFT) calculations—as processing methods. To prevent such errors, processing methods are not extracted from computational papers. While this step enhances precision, future improvements in LLM reasoning capabilities [30–32] may eliminate the need for this step.

Following classification, the system sequentially extracts key properties, structural characteristics, and associated mechanisms linking structure and properties, forming Sub-table 1. It then extracts key processing methods and mechanisms linking processing and structure, forming Sub-table 2. Finally, the two sub-tables are merged into a single comprehensive table, using the ‘structure’ column as the primary link. To optimize extraction, several key considerations are applied:

- **Property extraction:** General mechanical properties (e.g., strength, ductility) are provided as examples, allowing adaptability across different material domains. The prompt emphasizes measurement environments (e.g., cryogenic temperatures) to contextualize extracted data.
- **Structure extraction:** The model focuses on key structural characteristics influencing the extracted mechanical properties, avoiding trivial details. Typical structures are provided as examples, such as precipitate shape; matrix phase; slip pattern; dislocations; defect networks; interfaces; distribution of atoms, among the others. In **Fig. 2**, minimal lattice misfit between the precipitate and the matrix, a crucial factor for high yield strength and ductility, is successfully extracted.
- **Mechanism (Structure-Property) extraction:** If a mechanism is directly quoted or closely paraphrased from the source text, it is labeled “(From text)”; if inferred or generated by the LLM from its pretrained knowledge, it is labeled “(From knowledge base).” In **Fig. 2**, the extracted mechanism associated with low lattice misfit is taken directly from the paper, and the model correctly labels it as “(From text)”.
- **Sub-table 1 construction:** Extracted microstructure-mechanism-property relationships are organized into Sub-table 1, where each row represents a separate microstructure-mechanism-property pair.
- **Processing methods and composition extraction:** The entire processing chain—from raw materials to final heat treatment—can involve many steps. The model is repeatedly verified to extract only the most relevant processing methods affecting the microstructures identified in Sub-table 1. Each process or element is paired with a specific mechanism explaining its impact on microstructure. For example, in **Fig. 2**, aging is extracted not only for its role in precipitate formation but also for its effect on dislocation density and interface coherence. Similarly, Mo addition is highlighted for its role in stabilizing B2 precipitates and reducing lattice misfit.

- **Final merging:** Sub-tables 1 and 2 are merged, using the structure column serves as the primary merging criterion, since it serves as the central link between processing and properties. This integration ensures logical P-M-S-M-P connections. The final materials system chart is provided in **Fig. 3**.
- **Computational and Mixed papers:** For computational papers, only compositions and key element additions are extracted; processing methods are not included. For mixed papers, content is divided into three sections: A) Computational model and theoretical analysis of the material, B) Material synthesis and manufacturing, and C) Material's performance testing. Processing extraction is limited to Section B to maintain accuracy.
- **Error mitigation:** The model occasionally misclassifies terms, such as identifying “precipitate” and “dislocation” as properties or including trivial terms like “polish” for processing methods. While the advanced models may mitigate these issues, we define three sets of prohibited terms—one for properties, one for microstructures, and one for processing methods to address this issue—filtering responses via Python word matching (**Note S1** of the **SI**).
- **Paper Indicator:** After merging, a paper indicator is appended to the system chart to ensure traceability and prevent referencing errors, a known LLM weakness [33,34].

Input: Paper “Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation”		
I. Paper classification	Q	Is this paper experimental, computational, or a mix of both?
	A	Experimental
II. Sub-table 1	Property Microstructure Mechanism	
	Q	Example Prompt (Property extraction): Identify the alloy's significant mechanical properties discussed in the text, focusing on mechanical (e.g., strength, ductility) and physical aspects under certain environments (e.g., temperature), if any...
	A	Yield strength of $1,947 \pm 23$ MPa; ultimate tensile strength of $2,197 \pm 33$ MPa; total elongation (ductility) of $8.2\% \pm 0.7\%$
	Q	Example Prompt (Microstructure extraction): Comprehensively summarize structural characteristics relevant to the alloy's mechanical properties or deformation mechanisms as discussed in the paper...
	A	Ultrafine spherical precipitates; lath martensite matrix;...; minimal lattice misfit
	Q	Example Prompt (Mechanism extraction): Construct a table linking structure, associated deformation mechanisms, and resultant mechanical properties...
	A	Low lattice misfit decreases the nucleation barrier for precipitation, enabling the stabilization of nanoprecipitates and maintaining ductility (From text)...
III. Sub-table 2	Microstructure Processing Mechanism	
	Q	Example Prompt (Processing extraction): List the key processing or manufacturing techniques explicitly stated in the text provided for this alloy...
	A	Homogenization under Ar atmosphere; cold rolling; annealing; water quenching; ageing
IV. Final merged table	Q	Merge the two extracted sub-tables into a single comprehensive table, using the 'Microstructure' column as the primary criterion for merging.
	A	Processing Mechanism Microstructure Mechanism Property

Fig. 1. Detailed materials system chart extraction from individual papers. The example is from the paper titled “Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation” [29]

Processing	Mechanism	Microstructure	Mechanism	Property
Ageing Mo additions	Ageing at 500 °C stabilizes ultrafine spherical precipitates via controlled growth and coarsening; Mo retards coarsening of B2 precipitates (From text).	Ultrafine spherical precipitates	High-density nanoprecipitation with minimal lattice misfit leads to a high number density of precipitates, which strengthen the alloy by impeding dislocation movement (From text).	Yield strength of $1,947 \pm 23$ MPa; ultimate tensile strength of $2,197 \pm 33$ MPa
Ageing	Ageing promotes a high dislocation density by enabling dislocation interactions with precipitates (From text).	High dislocation density	Dislocations provide obstacles to slip, increasing the strength and work-hardening capacity, which contributes to both strength and ductility (From text).	Yield strength of $1,947 \pm 23$ MPa; total elongation (ductility) of $8.2\% \pm 0.7\%$
Ageing Ni(Al,Fe) precipitates	Ageing facilitates coherent precipitate/matrix interfaces through minimal lattice misfit and uniform precipitate distribution; Ni(Al,Fe) precipitates enhance coherency (From text).	Coherent precipitate/matrix interfaces	Coherency reduces stress concentrations at interfaces, preventing crack initiation and contributing to ductility (From text).	Total elongation (ductility) of $8.2\% \pm 0.7\%$
Ageing	Ageing leads to uniform precipitate distribution by homogenizing the precipitate nucleation sites (From text).	Uniform precipitate distribution	Even distribution of precipitates avoids stress concentration and ensures uniform deformation, enhancing ductility (From text).	Total elongation (ductility) of $8.2\% \pm 0.7\%$
Ageing Ni(Al,Fe) precipitates	Ageing results in the formation of a B2 ordered structure by driving the ordering transformation at elevated temperatures; Ni(Al,Fe) precipitates form the B2 structure (From text).	B2 ordered structure	The chemical ordering effect creates backstresses when precipitates are cut by dislocations, contributing to the strength (From text).	Yield strength of $1,947 \pm 23$ MPa
Ageing Mo additions	Ageing maintains minimal lattice misfit by controlling the diffusion and partitioning of elements; Mo reduces the lattice mismatch (From text).	Minimal lattice misfit	Low lattice misfit decreases the nucleation barrier for precipitation, enabling the stabilization of nanoprecipitates and maintaining ductility (From text).	Yield strength of $1,947 \pm 23$ MPa; total elongation (ductility) of $8.2\% \pm 0.7\%$
N/A	N/A	Lath martensite matrix	The martensitic structure provides a high dislocation density that contributes to the strength through dislocation interactions and work hardening (From text).	Yield strength of $1,947 \pm 23$ MPa

Fig. 3. Exemplary system chart for the paper titled “Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation” [29]. The yellow-highlighted block represents a PN mechanism. All ‘N/A’ entries are excluded from the evaluation.

a1

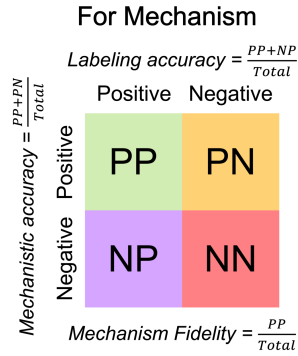
For Processing, Microstructure, and Property

Human Machine-Readability Index (HMI)	
Assessment Categories	Score
Incorrect	0
Partially correct	0.5
Correct	1

a2

Human Machine-Readability Index (HMI)	
Assessment Categories	Occurrences
Incorrect	0
Partially correct	0
Correct	20
Key Takeaway Missed?	No
Final HMI Score	100% (20/20)

b1



b2

Mechanism scores	
Assessment Categories	Occurrences
PP	12
PN	1
NP	0
NN	0
Mechanism Fidelity	92.3% (12/13)
Labeling Accuracy	92.3% (12/13)
Mechanistic Accuracy	100% (13/13)

c

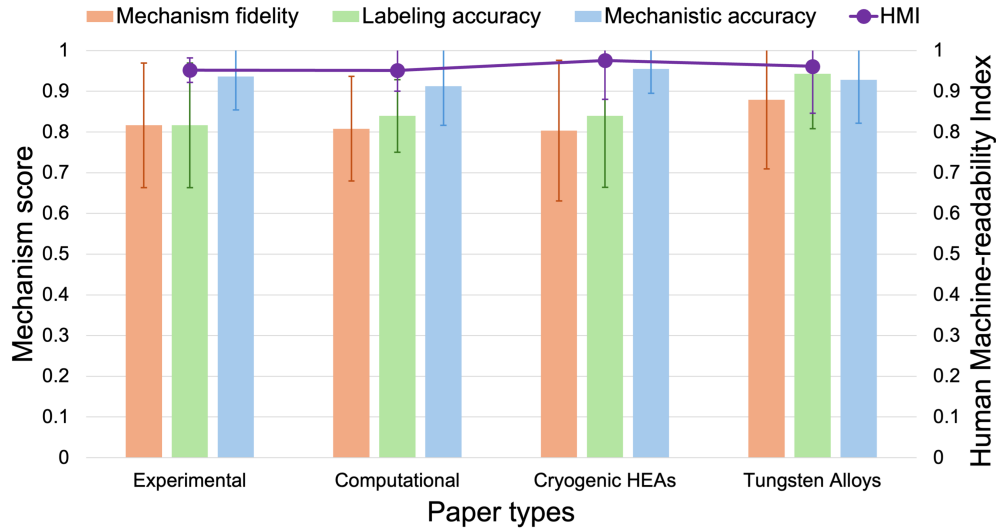


Fig. 4. Evaluation criteria and results for the materials system charts. (a1) Evaluation of the processing, microstructure, and property columns using the Human Machine-readability Index (HMI), categorizing responses as Correct, Partially Correct, and Incorrect. (a2) HMI score for the paper “Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation” [29]. If a paper’s key takeaway is missing from the table, 20% is deducted from the final HMI score. (b1) Evaluation of the mechanism columns involves categorizing individual mechanisms as PP, PN, NP, or NN. For instance, PN indicates an accurate mechanism but inaccurate labeling, while NP indicates accurate labeling but an inaccurate mechanism. Three scores are derived from this categorization: Mechanism Fidelity, Labeling Accuracy, and Mechanistic Accuracy. (b2) Mechanism scores for the same paper: mechanism fidelity (92.3%), labeling accuracy (92.3%), and mechanistic accuracy (100%). (c) HMI and mechanism scores across different paper types: experimental, computational, cryogenic HEAs, and tungsten alloys. Columns represent mechanism scores, and points connected by a line give the HMI scores. Error bars represent two standard deviations across each evaluated set of papers.

We select 70 papers across four datasets to optimize (20) and evaluate (50) the extraction process. First, the initial prompt is built using ten experimental papers from high-impact journals, as they typically present information in a well-organized manner, using concise language and emphasizing critical experimental details and underlying mechanisms. Second, ten computational papers from high-impact journals are included to extend the prompt’s capabilities to simulation-based studies. We use these two sets to iteratively refine the prompts by initializing a first version, testing it on the 20 papers to assess overall performance, and identifying any system charts that failed to capture key takeaways. If failures occur, we modify the prompt and evaluate (1) whether the specific system chart improved and (2) whether other system charts worsened. If condition (1) is met and condition (2) is not, we retain the modification and continue iterating; otherwise, we revert the change and adjust the prompt differently. We stop modifying once the average HMI and average mechanism score exceeds 0.8, as explained in the next section. Third, we select 30 cryogenic HEA papers—recognized for their rigorous science—to test our extraction process for varied paper lengths and research focuses. Lastly, 20 papers on tungsten alloys for fusion applications are selected to evaluate performance across various experimental conditions (e.g., irradiation and high-temperature conditions), target properties (e.g., sputtering resistance, radiation resistance, and high-temperature strength), and processing routes (e.g., sintering and ion implantation).

The materials system charts are evaluated using two metrics: the human machine-readability index (HMI) [35] and a mechanism score specifically defined for this study (**Fig. 4a1 and 4b1**). HMI involves manual evaluation of the accuracy and relevance of content in the processing, structure, and property columns. The mechanism score assesses the mechanistic accuracy of the extracted mechanisms and source labeling accuracy in the Mechanism columns. Mechanism fidelity counts mechanisms with both accurate mechanistic explanations and correct labeling (PP in the figure). Detailed explanations are provided below.

For HMI, the criteria include 1) Incorrect actions: Fabricated definitions, inaccuracies, irrelevant information; 2) Partially correct actions: Accurate content miscategorized, mixed correct and incorrect elements; 3) Correct actions: Precise information mentioned in the papers. After counting the number of occurrences of each action, HMI can be calculated as follows:

$$HMI = \frac{I(0)+PC(0.5)+C(1)}{A_T} \times 100\%,$$

where I = incorrect actions PC = partially correct actions C = correct actions A_T = total actions. Moreover, if the table does not include the core ideas of the paper 20% is deducted from the HMI.

For the mechanism score, the criteria include: 1) Labeling accuracy: Whether the mechanism is correctly classified as “From text” or “From knowledge base” based on its source; 2) Mechanistic accuracy: Whether the mechanism is scientifically reasonable and correctly illustrates the relationship between components. For “From text” mechanisms, accuracy is evaluated based on statements in the original research paper, and for “From knowledge base” mechanisms, accuracy is evaluated based on the authors’ knowledge; 3) Mechanism fidelity: The number of mechanisms that are both mechanistically sound and correctly labeled. An exemplary evaluation result is shown in **Fig. 3**. The yellow-highlighted mechanism in the table is scientifically reasonable but not discussed in the paper and is thus classified as a PN case. Exemplary evaluation results from five papers can be found in **Table S1** of the **SI**, while the results for all 70 papers are available in the repository [36].

Fig. 4c presents a consolidated summary of the model performance for 70 papers. On the x-axis, the papers are grouped by category, while the left y-axis corresponds to the Mechanism Score, which is further divided into three specific subscores. The right y-axis indicates the HMI. Each stacked bar represents the average values for Mechanism fidelity (orange), Labeling accuracy (green), and Mechanistic accuracy (blue), while the purple line traces the HMI for each paper category.

Mechanistic accuracy consistently scores above 0.9, indicating that the model can extract scientifically valid mechanisms. Some NP cases encountered during evaluation include: 1. In the system chart of “*The*

Evolution of the Deformation Substructure in a Ni-Co-Cr Equiatomic Solid Solution Alloy [37],” one mechanism states: “cold rolling induces plastic deformation, creating dislocation networks and increasing dislocation density.” This is inaccurate because, in the context of the paper, the relevant dislocation networks are primarily generated during tensile tests, not cold rolling. 2. In the same paper, the model attempts to link the alloy composition (equiatomic NiCoCr) with HCP lamellae, stating, “Equiatomic ratio promotes high entropy, which stabilizes the solid solution and influences defect structures.” This is also inaccurate, as the immediate cause of HCP formation is mechanical strain rather than simply high entropy. Detailed evaluation results for this paper can be found in **Table S1** in **SI**.

Labeling accuracy remains above 0.8 in all cases. Across all PN and NN cases among 70 papers, the model mistakenly regards some mechanisms as being derived from the paper 81 times (for example, the PN mechanism in **Fig. 3**). Conversely, there are 24 instances where the model fails to recognize that certain mechanisms are mentioned in the paper.

Meanwhile, the HMI remains at or above approximately 0.95, underscoring the accuracy in capturing processing, microstructure, and property information. Although not shown in the figure, only two of the 70 papers—in the tungsten alloy category—receive a 20% score deduction due to missing key findings. The first instance appears in the paper “*Development of yttrium-containing self-passivating tungsten alloys for future fusion power plants* [38].” The main thrust of this paper is that W-Cr-Y alloys can self-passivate under accident conditions by forming a stable Cr_2O_3 layer, which drastically reduces the release of radioactive WO_3 . The table, however, focuses primarily on conventional microstructure-property relationships (grain size, precipitates, toughness) and merely includes a generic note on “high-temperature oxidation,” without connecting these points to the optimized Cr/Y balance. The second instance occurs in “*Performance of tungsten nitride compound surfaces to resist sputtering under intense irradiation in nuclear fusion reactors* [39].” The paper’s key message is that although WN_x surfaces are intrinsically sputter-resistant, hydrogen ultimately drives nitrogen depletion by forming ammonia (NH_3), which undermines that resistance. However, the table does not explicitly illustrate this hydrogenation effect. Detailed evaluation results can be found in **Table S1** in **SI**.

Taken as a whole, this evaluation demonstrates the robustness of the present system across diverse scientific contexts. Looking ahead, we can not only integrate features to aid and streamline the human evaluation process but also strive for higher accuracy of extracted system charts. For instance, to more easily identify those subtle NP and PN cases, the LLM could be prompted to attach confidence scores for both its source labeling and scientific plausibility in context when generating each mechanism. Such a feature would allow users to more efficiently target entries requiring closer inspection during review. Additionally, a second LLM could be implemented to cross-check each mechanism against the original paper text, effectively proofreading the correctness of the quoted or paraphrased mechanisms. Separately, to enhance the accuracy of system charts, there are some advanced LLM techniques waiting to be explored, involving causal analysis [40,41], fine tuning [42,43], ontology integration [44,45], which will be explored in future work. Furthermore, with the rapid evolution of LLMs, newer models with larger sizes and improved architectures will likely offer significantly enhanced reasoning capabilities, making them better at capturing nuanced relationships [30–32]. This could potentially improve both Mechanism and HMI scores. To assess these improvements, we use one of the latest models, ‘o3-mini-2025-01-31’ from OpenAI to test the four lowest-performing papers from the evaluation—one from each category—and confirm the improvements (Mechanism Fidelity improved from 63% (gpt-4-1106-preview) to 73% (o3-mini-2025-01-31); and HMI scores increased from 82% to 94%). Detailed evaluation results are provided in **Table S2** and **Figure S1**.

Finally, the system chart can be converted into a visual diagram—as shown in step IV and V in **Fig. 1**—allowing readers to quickly grasp the paper’s key insights. Before plotting, another set of prompts is used to refine the system chart, as shown in **Fig. 5**. Specifically, after categorizing different microstructures and tagging them by their mechanistic relevance, the chart undergoes property standardization (e.g., removing extraneous parameters while preserving key compositional details). Next, mechanisms are carefully

segregated so that processing-induced phenomena and composition-driven changes remain distinct, and any missing entries under the “Structure” or “Property” columns are filled in if sufficient information is available. Once these pre-plotting refinements are complete, the final, clean system chart is ready to be plotted. **Fig. 5** shows the automatically created visual diagram for the exemplary paper in **Figs. 2, 3**. The plotting-ready version of the system chart for and detailed prompts can be found in **Table S3** and **Note S2** of the **SI**, respectively.

As the P-M-S-M-P framework is central to materials science and engineering, our work naturally complements other LLM-based approaches for materials data extraction and utilization. First, our method significantly enhances downstream applications such as scientific hypotheses generation. The extracted system charts effectively distill key information from literature—often with a token count around ten times smaller than that of a full paper. Researchers can integrate multiple system charts from different studies into a single hypotheses generation prompt, thereby capturing a wider range of ideas while staying within token limits. These hypotheses can then be seamlessly integrated into automated laboratory systems or high-throughput simulations such as CALPHAD, as demonstrated in our another work currently under review [46].

Second, compared to general-purpose summarization LLM tools like “TLDR This,” which produce concise yet relatively unstructured and less controllable abstracts (see **Note S3** in the **SI**), our framework provides targeted and hierarchically structured knowledge extraction with greater interpretability and controllability. Furthermore, while advanced LLM-based methods such as Dagdelen *et al.* [10] and Polak and Morgan [13] enable sophisticated named entity recognition, relation extraction, and precise zero-shot triplet extraction, our structured P-M-S-M-P framework distinctly captures higher-order hierarchical mechanistic relationships. This structured representation is particularly advantageous for integration into systems engineering frameworks, such as Axiomatic Design [47,48], systematically guiding the coupling and decoupling of processing, structure, and property parameters essential for materials design optimization.

Finally, integrating the normalization, error-correction, and entity-validation capabilities from these advanced LLM-based extraction methods ([10] and [13]) into our hierarchical P-M-S-M-P framework could further enhance the consistency, clarity, and interoperability of generated system charts. Such integration would streamline validation and improve the reliability and accuracy of hierarchical mechanistic relationships extracted across diverse literature sources.

In summary, this work presents an LLM-based framework that systematically extracts plottable system charts from metallurgy literature, serving as a basis for materials discovery and knowledge synthesis. Optimized and evaluated on 70 papers, the method achieves 83% fidelity for mechanisms and 97% human machine-readability index for processing, structure, and property components, demonstrating its accuracy and robustness. Looking ahead, further advancements in LLM reasoning, causal inference, and multimodal integration hold the potential to enhance the depth and automation of knowledge extraction, promoting autonomous knowledge-driven materials design.

Input: System chart of the paper “ <i>Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation</i> ”		
I. Microstructure categorization	Q	...Categorize only those microstructures into groups based on their mechanistic relationships in materials science...
	A	1. Precipitation-related microstructures: - B2 ordered structure - Minimal lattice misfit ... 2. Dislocation-related microstructures: - High dislocation density 3. Martensitic transformation-related microstructures: - Lath martensite matrix
II. Microstructure category tagging	Q	In the previous step, all microstructures were categorized. Based on these results, add a tag for each microstructure to represent its category...
III. Property standardization	Q	1. If any row contains specific parameters, simplify them. 2. Do not remove any composition and key elements. ...
IV. Mechanism separation	Q	Separate the mechanisms for processing methods and the mechanisms for composition/key elements into individual rows...
V. Property separation and mapping	Q	1. First, separate each property into different rows, ensuring that the content of the properties remains unchanged... 2. Critically determine the correlation between property and mechanisms...
VI. Missing data filling	Q	Check if there is any 'N/A' in the 'Structure' and 'Property' columns...
VII. Final chart visualization	Q	Transform the system chart into diagram format using a different set of prompts
	A	Mechanisms 10: Mo reduces the lattice mismatch (Ft). 17, 18: Low lattice misfit decreases the nucleation barrier for precipitation, enabling the stabilization of nanoprecipitates and maintaining ductility (Ft).

Fig. 5. Detailed optimization of system charts for automatically plotting from individual papers. The original paper corresponding to the processed system chart will also be included in the workflow. Example is from the paper titled “*Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation*” [29].

Data availability

In this repository [36], we provide a version of the Python code used throughout the framework presented in **Fig. 1**, including the prompts. The “README” file in the repository includes instructions on using the code in the Google Colab environment, which can be completed in about half an hour, even by those unfamiliar with coding.

Acknowledgements

This work was supported by the start-up funds from the University of Wisconsin-Madison provided by the Wisconsin Alumni Research Foundation (WARF) and National Science Foundation EAGER Award No. 2513480. D.M. and M.P.P. acknowledge the support from National Science Foundation Cyberinfrastructure for Sustained Scientific Innovation (CSSI) Award No. 1931298.

CRedit authorship contribution statement

Q.L., H.O., M.P., and D.M. conceptualized and designed the study; Q.L. and H.O. designed the framework and prompts; Q.L. and M.P. developed the code; Q.L., H.D., A.S., J.H., and H.O. conducted the evaluations (HMI, Mechanism score); Q.L. and H.O. wrote the paper and the supplementary material; and all authors discussed the results, revised the manuscript, and approved the final version of the manuscript.

References

- [1] how-important-is-the-impact-factor-to-the-materials-science-community n.d.
<https://www.wiley.com/en-us/network/publishing/research-publishing/choosing-a-journal/how-important-is-the-impact-factor-to-the-materials-science-community> (accessed March 20, 2024).
- [2] Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit Discov* 2023;2:1233–50. <https://doi.org/10.1039/D3DD000113J>.
- [3] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc.; 2020, p. 1877–901.
- [4] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models 2023.
- [5] Lee U, Jung H, Jeon Y, Sohn Y, Hwang W, Moon J, et al. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Educ Inf Technol* 2023. <https://doi.org/10.1007/s10639-023-12249-8>.
- [6] Polak MP, Morgan D. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering 2023.
- [7] Zhang L, Huang Y, Yan L, Ge J, Ma X, Liu Z, et al. Fast Exploring Literature by Language Machine Learning for Perovskite Solar Cell Materials Design. *Adv Intell Syst* 2024;6:2300678. <https://doi.org/10.1002/aisy.202300678>.
- [8] Kumar R, Joshi A, Khan SA, Misra S. Automated extraction of synthesis parameters of pulsed laser-deposited materials from scientific literature. *Digit Discov* 2024;3:944–53. <https://doi.org/10.1039/D4DD00051J>.
- [9] Hira K, Zaki M, Sheth D, Mausam, Krishnan NMA. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digit Discov* 2024;3:1021–37. <https://doi.org/10.1039/D4DD00032C>.
- [10] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, et al. Structured information extraction

from scientific text with large language models. *Nat Commun* 2024;15:1418.

<https://doi.org/10.1038/s41467-024-45563-x>.

[11]Choi J, Lee B. Accelerating materials language processing with large language models. *Commun Mater* 2024;5:13. <https://doi.org/10.1038/s43246-024-00449-9>.

[12]Buehler MJ. Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design. *ACS Eng Au* 2024;4:241–77. <https://doi.org/10.1021/acsengineeringau.3c00058>.

[13]Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat Commun* 2024;15:1569. <https://doi.org/10.1038/s41467-024-45914-8>.

[14]Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. *Nature* 2023;624:570–8. <https://doi.org/10.1038/s41586-023-06792-0>.

[15]M. Bran A, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. *Nat Mach Intell* 2024;6:525–35. <https://doi.org/10.1038/s42256-024-00832-8>.

[16]Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models 2022. <https://doi.org/10.48550/arXiv.2206.07682>.

[17]Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. A Survey on In-context Learning 2024. <https://doi.org/10.48550/arXiv.2301.00234>.

[18]Olson GB. Computational Design of Hierarchically Structured Materials. *Science* 1997;277:1237–42. <https://doi.org/10.1126/science.277.5330.1237>.

[19]Pollock TM, Van Der Ven A. The evolving landscape for alloy design. *MRS Bull* 2019;44:238–46. <https://doi.org/10.1557/mrs.2019.69>.

[20]Olson GB, Liu ZK. Genomic materials design: CALculation of PHase Dynamics. *Calphad* 2023;82:102590. <https://doi.org/10.1016/j.calphad.2023.102590>.

[21]McDowell DL, Olson GB. Concurrent design of hierarchical materials and structures n.d.

- [22] Ghosh G. Thermodynamic and kinetic modeling of the Cr-Ti-V system. *J Phase Equilibria* 2002;23:310. <https://doi.org/10.1361/105497102770331569>.
- [23] Ghosh G, Olson GB. Integrated design of Nb-based superalloys: Ab initio calculations, computational thermodynamics and kinetics, and experimental results. *Acta Mater* 2007;55:3281–303. <https://doi.org/10.1016/j.actamat.2007.01.036>.
- [24] Mishra RS, Thapliyal S. Design approaches for printability-performance synergy in Al alloys for laser-powder bed additive manufacturing. *Mater Des* 2021;204:109640. <https://doi.org/10.1016/j.matdes.2021.109640>.
- [25] Oh HS, Odbadrakh K, Ikeda Y, Mu S, Körmann F, Sun C-J, et al. Element-resolved local lattice distortion in complex concentrated alloys: An observable signature of electronic effects. *Acta Mater* 2021;216:117135. <https://doi.org/10.1016/j.actamat.2021.117135>.
- [26] Xiong W, Olson GB. Cybermaterials: materials by design and accelerated insertion of materials. *Npj Comput Mater* 2016;2:15009. <https://doi.org/10.1038/npjcompumats.2015.9>.
- [27] McDowell DL, Kalidindi SR. The materials innovation ecosystem: A key enabler for the Materials Genome Initiative. *MRS Bull* 2016;41:326–37. <https://doi.org/10.1557/mrs.2016.61>.
- [28] Cohen M. Metallurgy and the Evolution of Materials Science and Engineering. *Innov. Ultrah.-Strength Steel Technol.*, US Government Printing Office; n.d., p. 67-85. n.d.
- [29] Jiang S, Wang H, Wu Y, Liu X, Chen H, Yao M, et al. Ultrastrong steel via minimal lattice misfit and high-density nanoprecipitation. *Nature* 2017;544:460–4. <https://doi.org/10.1038/nature22032>.
- [30] Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, et al. Scaling Laws for Neural Language Models 2020. <https://doi.org/10.48550/arXiv.2001.08361>.
- [31] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4 2023. <https://doi.org/10.48550/arXiv.2303.12712>.
- [32] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback n.d.

- [33]Huang C, Wu Z, Hu Y, Wang W. Training Language Models to Generate Text with Citations via Fine-grained Rewards 2024. <https://doi.org/10.48550/arXiv.2402.04315>.
- [34]Huang J, Chang KC-C. Citation: A Key to Building Responsible and Accountable Large Language Models 2024. <https://doi.org/10.48550/arXiv.2307.02185>.
- [35]Suvarna M, Vaucher AC, Mitchell S, Laino T, Pérez-Ramírez J. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nat Commun* 2023;14:7964. <https://doi.org/10.1038/s41467-023-43836-5>.
- [36]Liu, Q. et al. Codes to the paper entitled 'A Large Language Model Framework for Automated Extraction of System Charts in Materials Science'. DOI:<https://doi.org/10.6084/m9.figshare.28405409> n.d.
- [37]Miao J, Slone CE, Smith TM, Niu C, Bei H, Ghazisaeidi M, et al. The evolution of the deformation substructure in a Ni-Co-Cr equiatomic solid solution alloy. *Acta Mater* 2017;132:35–48. <https://doi.org/10.1016/j.actamat.2017.04.033>.
- [38]Wegener T, Klein F, Litnovsky A, Rasinski M, Brinkmann J, Koch F, et al. Development of yttrium-containing self-passivating tungsten alloys for future fusion power plants. *Nucl Mater Energy* 2016;9:394–8. <https://doi.org/10.1016/j.nme.2016.07.011>.
- [39]Yu PF, Pan BC. Performance of tungsten nitride compound surfaces to resist sputtering under intense irradiation in nuclear fusion reactors. *Appl Surf Sci* 2022;600:154072. <https://doi.org/10.1016/j.apsusc.2022.154072>.
- [40]Ashwani S, Hegde K, Mannuru NR, Jindal M, Sengar DS, Kathala KCR, et al. Cause and Effect: Can Large Language Models Truly Understand Causality? 2024.
- [41]Ban T, Chen L, Wang X, Chen H. From Query Tools to Causal Architects: Harnessing Large Language Models for Advanced Causal Discovery from Data 2023.
- [42]Fine-Tuning BERT Model for Materials Named Entity Recognition | IEEE Conference Publication | IEEE Xplore n.d. <https://ieeexplore.ieee.org/document/9671697> (accessed March 25, 2024).
- [43]Gruver N, Sriram A, Madotto A, Wilson AG, Zitnick CL, Ulissi Z. Fine-Tuned Language Models

Generate Stable Inorganic Materials as Text n.d.

[44] Ye Y, Ren J, Wang S, Wan Y, Wang H, Razzak I, et al. Construction and Application of Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model 2024.

[45] Venugopal V, Olivetti E. MatKG-2: Unveiling precise material science ontology through autonomous committees of LLMs n.d.

[46] Liu, Q., et al. "Beyond designer's knowledge: Generating materials design hypotheses via a large language model", under review. n.d.

[47] Suh NP. Axiomatic design theory for systems. Res Eng Des 1998;10:189–209.

[48] Suh NP. Axiomatic design and fabrication of composite structures: applications in robots, machine tools, and automobiles. OXFORD university press; 2006.