npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-026-01966-6

aLLoyM: a large language model for alloy phase diagram prediction

Yuna Oikawa[1], Guillaume Deffrennes[2], Rintaro Shimayoshi[1], Taichi Abe[3], Ryo Tamura[1,4] ✉ & Koji Tsuda[1,4,5] ✉

Large language models (LLMs) are general-purpose tools with wide-ranging applications, including in materials science. In this work, we introduce aLLoyM, a fine-tuned LLM specifically trained on alloy compositions, temperatures, and their corresponding phase information. To develop aLLoyM, we curated question-and-answer (Q&A) pairs for binary and ternary phase diagrams using the open-source Computational Phase Diagram Database (CPDDB) and assessments based on CALPHAD (CALculation of PHAse Diagrams). We fine-tuned Mistral, an open-source pre-trained LLM, for two distinct Q&A formats: multiple-choice and short-answer. Benchmark evaluations demonstrate that fine-tuning substantially enhances performance on multiple-choice phase diagram questions. Moreover, the short-answer model of aLLoyM can generate novel phase diagrams from its components alone, suggesting that it may aid the discovery of new materials systems. To promote further research and adoption, we have publicly released the short-answer fine-tuned version of aLLoyM, along with the complete benchmarking Q&A dataset, on Hugging Face.

Phase diagrams serve as fundamental roadmaps in materials science, providing critical insights into material behavior across varying thermodynamic conditions. The ability to accurately predict and interpret these diagrams represents a cornerstone of efficient materials design, with experienced practitioners often relying on accumulated expertise to anticipate phase relationships. While large experimental databases[1–4] and computational repositories[5–7] have established valuable reference collections, the experimental determination of phase diagrams remains resource-intensive and prohibitively time-consuming for comprehensive materials exploration.

Recent advances in machine learning methodologies have demonstrated promising capabilities for phase diagram prediction, with conventional approaches including neural networks, support vector machines, random forests, and label propagation algorithms showing measurable success[8–15]. Concurrently, the emergence of large language models (LLMs) such as GPT-4, LLaMA, and Mistral has opened novel avenues for materials science applications[16–20]. Unlike specialized machine learning models that operate on isolated datasets, LLMs represent general-purpose architectures capable of leveraging broader scientific knowledge, such as thermodynamic principles and elementary properties encoded during pre-training, into phase diagram predictions. Preliminary investigations have explored LLM applications in phase diagram analysis, including system-specific training on Mg-Al-Zn data[21] and experimental diagram annotation[22], suggesting substantial potential for phase diagram analysis.

In this study, we introduce aLLoyM, an LLM fine-tuned for phase diagram generation (Fig. 1). Due to computational resource constraints, we adopted low-rank adaptation (LoRA) instead of full fine-tuning. The efficacy of LoRA in domain-specific fine-tuning scenarios has been substantiated in prior studies[23–26]. Our approach leverages the Computational Phase Diagram Database (CPDDB)[5], a comprehensive open-source repository published by the National Institute for Materials Science (NIMS), as the primary training corpus. From the CPDDB, thermodynamic database (TDB) files for 389 binary and 38 ternary phase diagrams were obtained, with the distribution of constituent elements illustrated in Figs. S1 and S2. Each TDB file contains Gibbs free energy functions for individual phases, enabling the construction of phase diagrams through CALPHAD assessments. Phase diagram calculations were performed across systematic compositional and temperature grids using Pandat software[27]. For compositional variables, elemental fractions were sampled from 0% to 100% in 2% increments. For binaries, temperature was varied from 200 K to 5000 K in 50 K intervals, while for ternaries, the temperature was fixed at 800 K due to the computational cost. There are two main reasons why the temperature was fixed at 800 K for the ternary systems: (1) it is close to typical annealing temperatures for high-entropy alloys, and (2) it is comparable to annealing

[1]Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan. [2]University Grenoble Alpes, CNRS, Grenoble INP, SIMaP, Grenoble, France. [3]Research Center for Structural Materials, National Institute for Materials Science, Tsukuba, Ibaraki, Japan. [4]Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki, Japan. [5]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. ✉e-mail: tamura.ryo@nims.go.jp; tsuda@k.u-tokyo.ac.jp

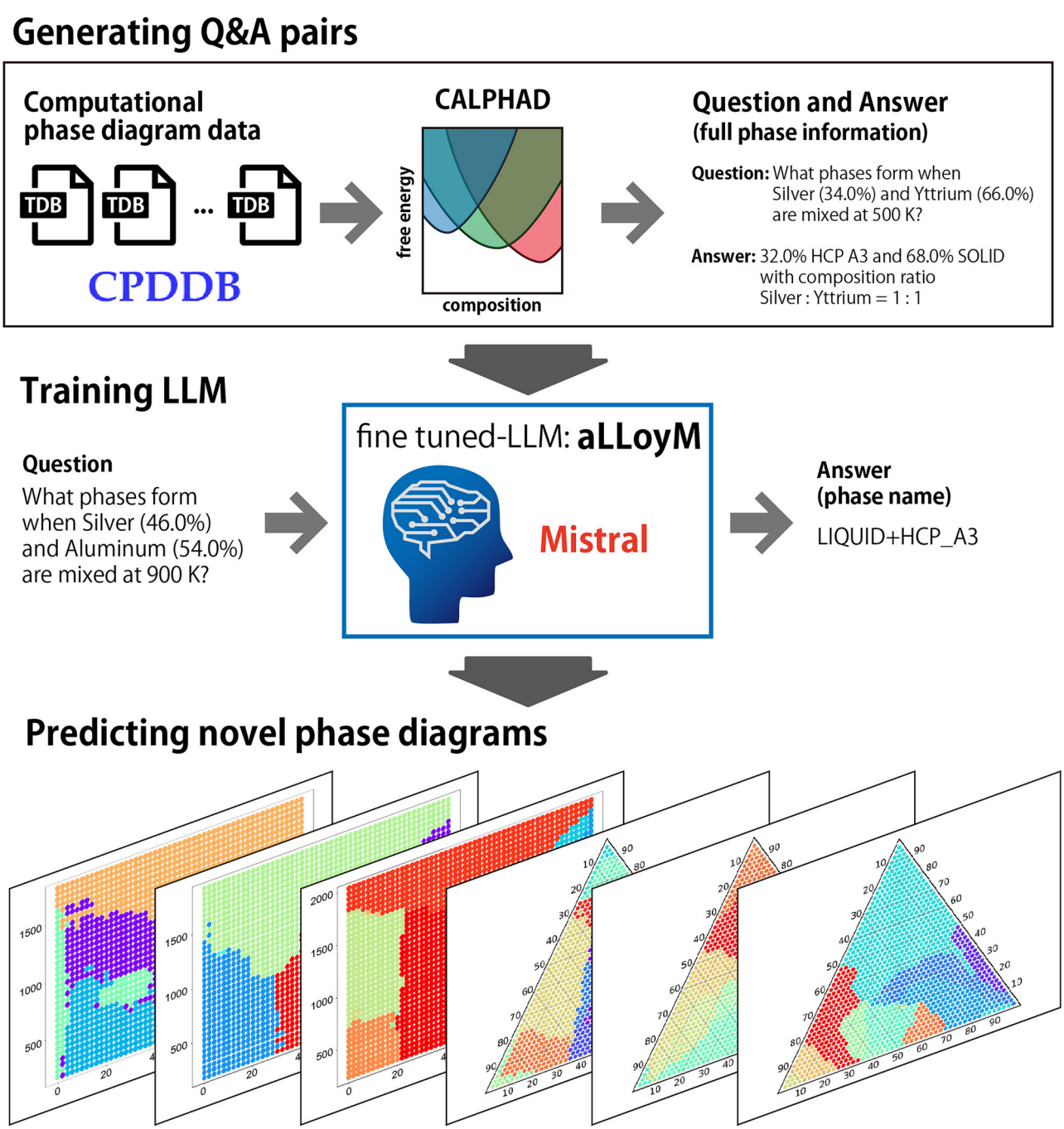

Fig. 1 | Schematic of fine-tuned LLM for phase diagram generation: aLLoyM. Q&As were generated from CPDDB using CALPHAD assessments, and Mistral was fine-tuned on these pairs.

temperatures for steel as well as the aging treatment of Ni superalloys. Consequently, phase diagrams at this temperature are of particular interest from the perspective of phase diagram determination. This systematic sampling approach generated 837,475 data points, each defining the relationship between elemental composition, temperature, and corresponding phase names. From these data points, we constructed question-and-answer (Q&A) pairs. For example, a question might include information about the composition and temperature, and the answer would be the associated phase name. One of the important features of LLMs is their ability to handle multiple tasks within a single model. Thus, in this study, we developed a model capable of performing three different Q&A tasks using a unified architecture. We then fine-tuned Mistral, an open-source pre-trained LLM, on these Q&As to incorporate domain-specific knowledge through selective parameter updates.

The aLLoyM model was comprehensively benchmarked using two distinct Q&A formats: multiple-choice and short-answer. The multiple-choice Q&As facilitated direct comparative analysis between baseline and fine-tuned model performance, with results demonstrating that fine-tuning yielded substantial improvements in predictive accuracy relative to the baseline LLM. In contrast, the short-answer Q&As operate independently of multiple-choice constraints, rendering it particularly suitable for predicting previously unexplored phase diagrams without requiring additional domain knowledge. The implementation of short-answer Q&As with aLLoyM can be employed to generate novel phase diagrams, as exemplified in Fig. 1, and facilitated the generation of illustrative examples. The aLLoyM model optimized for short-answer applications is publicly accessible through the Hugging Face platform (https://huggingface.co/Playingyoyo/aLLoyM).

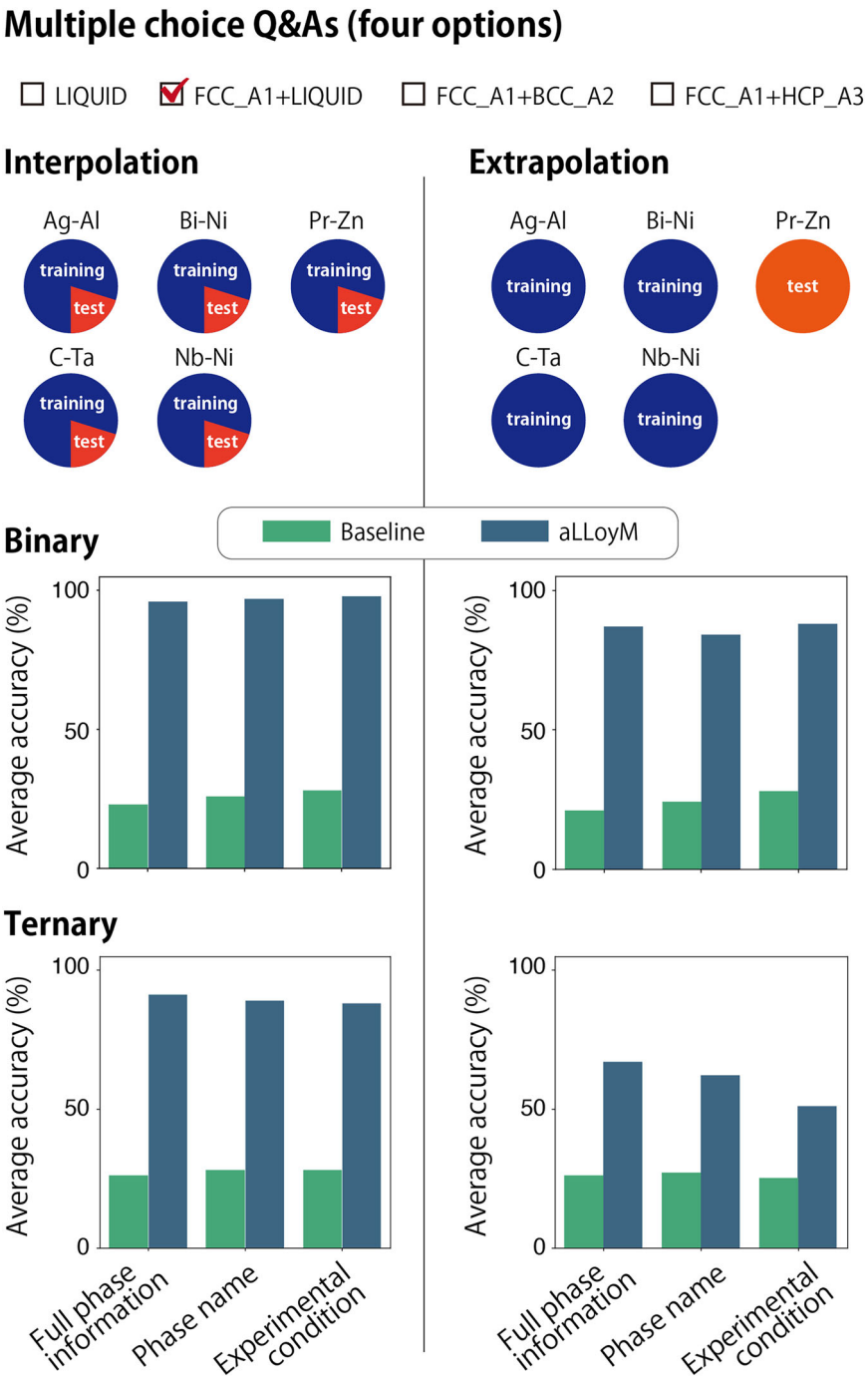

Fig. 2 | Accuracies of the baseline model (Mistral) and the fine-tuned model (aLLoyM) on multiple-choice Q&As. Results are reported separately for interpolation and extrapolation settings, and cover all three Q&A task types: full phase information inference, phase name prediction, and experimental condition inference. Performance is also distinguished between binary and ternary systems.

Results

Multiple choice Q&As

We conducted a benchmark evaluation using multiple-choice questions to compare the performance of aLLoyM against a baseline LLM. Each question required the model to choose the correct answer from four options, where three distractors were randomly selected from answers related to the same systems (Fig. 2). To assess the model's generalization capability, the dataset, comprising binary and ternary systems, was split into training and test sets using an 8:2 ratio. Two distinct data splitting strategies were implemented to evaluate model performance under different generalization scenarios (see Fig. 2). *Interpolation split*: data points were randomly distributed across all available systems, allowing assessment of model performance on familiar systems with varying compositional and thermal conditions. *Extrapolation split*: systems in the test set were completely excluded from the training set, enabling evaluation of the model's ability to generalize to previously unseen systems.

We considered three types of Q&A tasks. *Full phase information*: given the input composition and temperature, the model predicts the complete phase information, including phase names and their corresponding fractions and compositions. An example of this Q&A task is presented in Fig. 1. *Phase name*: the model predicts only the phase names based on the input composition and temperature. The output is the phase domain without specifying phase fractions or compositions for full phase information. *Experimental condition*: given the constitutive elements and a specific phase domain, the model predicts a possible composition and temperature. This task serves as the inverse of the phase name prediction. The examples of each Q&A are summarized in Table S1. These three Q&A tasks were trained within a single LLM.

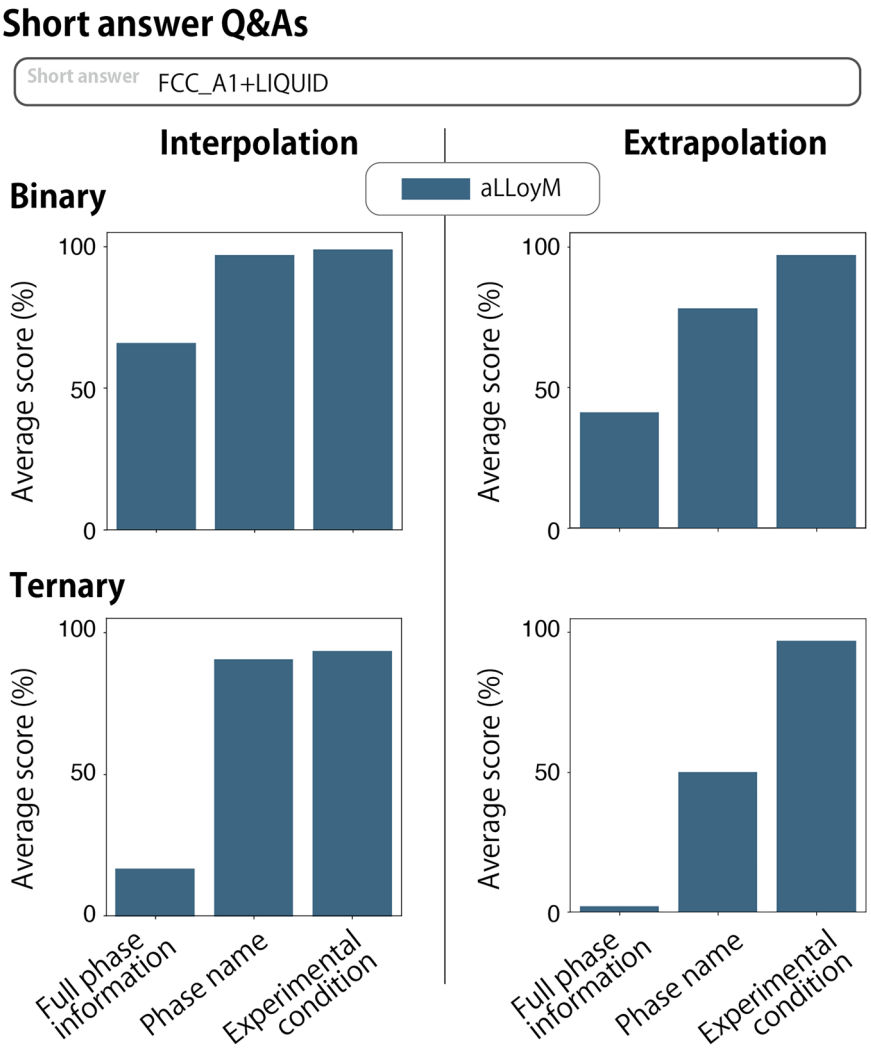

Fig. 3 | Average scores of the fine-tuned models (aLLoyM) for short answer Q&As. Results are presented for interpolation and extrapolation configurations, with individual tasks (full phase information, phase name, and experimental condition). Binary and ternary systems were evaluated separately.

The accuracies of all Q&A tasks in the multiple-choice are shown in Fig. 2, with performance evaluated separately for binary and ternary systems across the three task types. As a baseline, we employed the Mistral-Nemo-Instruct-2407-bnb-4bit model using Hugging Face's causal language modeling interface[28]. The baseline model's performance remained close to random guessing in both interpolation and extrapolation settings, with accuracy only slightly above the level expected by chance. These findings indicate that the baseline language model struggled to produce correct answers to phase diagram questions. Ideally, predictions obtained from conventional machine learning methods should also have been considered as an additional baseline. However, constructing prediction models with such approaches presents a major challenge in numerically encoding phase labels. The phase names appearing in different phase diagrams vary considerably, making it difficult to standardize them or to convert them into numerical representations. Consequently, applying conventional machine learning methods to the present dataset is not straightforward. In contrast, a key advantage of LLMs lies in their ability to handle phase names directly.

In contrast to the baseline, the fine-tuned models exhibited substantial performance improvements across all tasks. For both interpolation and extrapolation settings, individual models were fine-tuned on the complete ensemble of three Q&A tasks. In all cases, the fine-tuned models outperformed the baseline. As anticipated, performance was generally higher on interpolation tasks compared to extrapolation tasks. Furthermore, predictions for ternary systems proved more challenging than those for binary systems, while performance differences among the three Q&A tasks were relatively minor. These results demonstrate that, when provided with suitable training data, LLMs are capable of accurately predicting phase diagram information. Notably, the model's success on extrapolation setting suggests an ability to generalize knowledge from known systems to make informed predictions for previously unseen combinations.

We evaluated the average accuracy for ternary systems as a function of the number of constituent binary pairs included in the training dataset. The results of phase name prediction for the extrapolation split are shown in Fig. S3. Although the average accuracy generally improved with an increasing number of constituent binary pairs, the variability remained substantial. The training was performed using a mixture of binary and ternary data, albeit with an imbalanced distribution. To assess the impact of this imbalance, we constructed a model trained exclusively on ternary data and compared its accuracy (see Fig. S4). The results demonstrated that excluding binary data reduced the prediction accuracy for ternary phase diagrams. These findings suggest that, in the present case, distributional imbalance does not inherently impair the prediction performance of LLMs.

Short answer Q&As

Adopting the short-answer questions allows the model to generate responses without relying on predefined multiple-choice options. Consistent with the multiple-choice Q&As, the fine-tuned models were trained using the full data corresponding to all three Q&A tasks. To evaluate the alignment between the ground-truth answers and those generated by aLLoyM, we introduced a scoring metric described in the "Methods" section depending on the Q&A task. The score ranges from 0 to 100%, with higher values indicating greater agreement between the generated and ground-truth answers.

Figure 3 presents the average scores for each task. As anticipated, performance was superior on interpolation settings relative to extrapolation

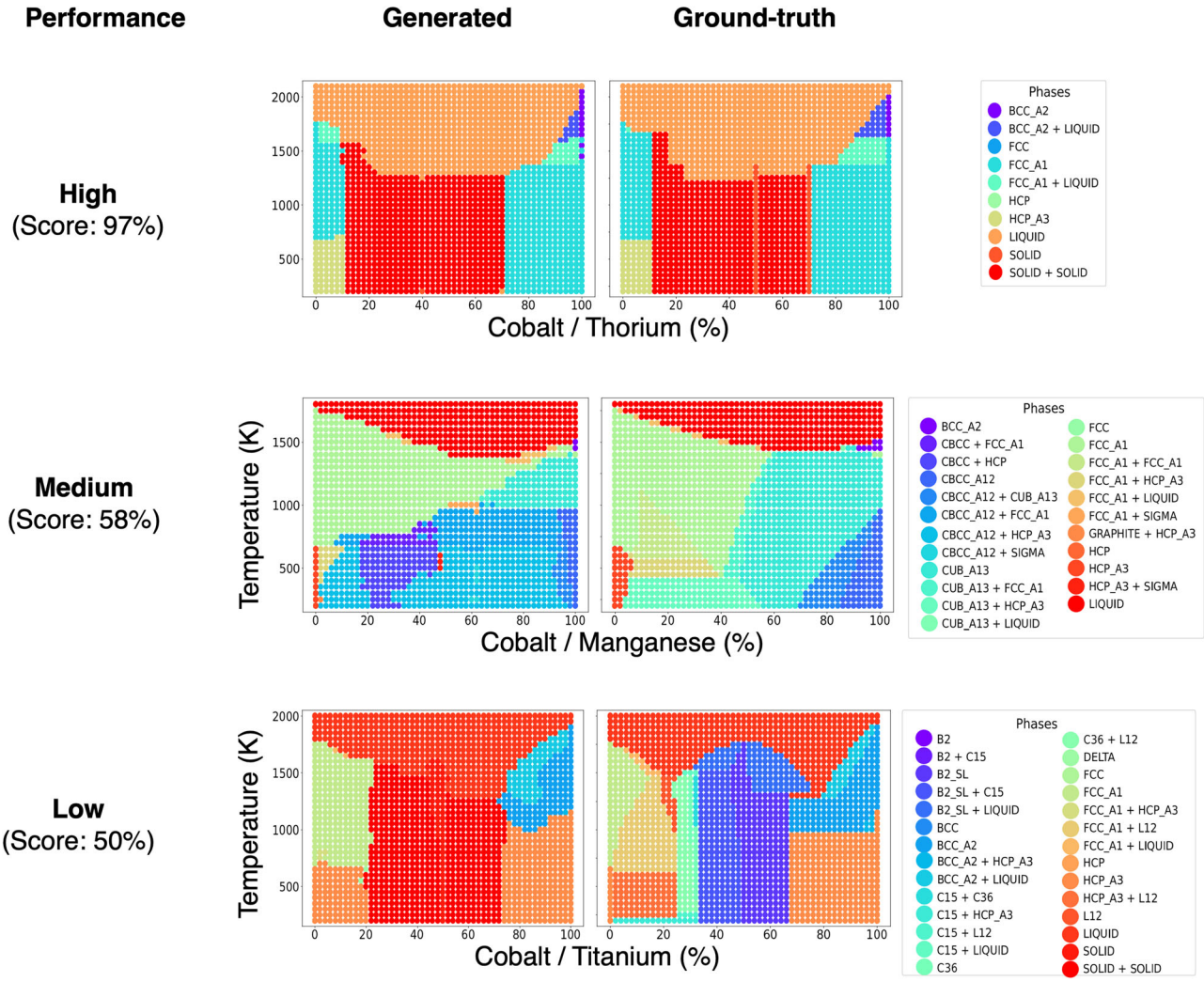

Fig. 4 | Representative binary phase diagrams exhibiting varying predictive performance, as generated by aLLoyM for the phase name prediction task. The ground-truth phase diagrams are also shown. Lower scores denote greater discrepancies between generated and ground-truth phase diagrams.

settings. Among the three Q&A task categories, predicting complete phase information proved most challenging. Nevertheless, the model demonstrated robust performance in predicting phase names, even under extrapolation conditions. Furthermore, it successfully generated appropriate experimental conditions from specified phase information in extrapolation settings, suggesting that when a target phase is designated, aLLoyM possesses the capacity to reliably propose suitable experimental parameters. Across all tasks, predictions for ternary systems were consistently more challenging than those for binary systems. Note that Supplementary Note A was prepared to analyze the sources of prediction errors, such as missing phases or inaccurate temperatures. In addition, the effect of jointly fine-tuning all three Q&A tasks, as opposed to fine-tuning them separately, is of particular relevance for evaluating the capabilities of LLMs. In the binary prediction task, we compared these two training strategies and observed that the resulting accuracies were nearly identical (see Fig. S5). These results demonstrate that LLMs are capable of effectively learning multiple tasks within a single model.

Based on the phase names predicted by aLLoyM, we reconstructed the phase diagrams for the element sets in the extrapolated test set. Figures 4 and 5 present representative binary and ternary phase diagrams exhibiting varying levels of predictive performance. The scores represent averages across each complete phase diagram, with corresponding ground-truth phase diagrams provided for comparison. Across all cases, predictive performance remains consistently higher in regions proximate to pure elements and diminishes progressively as compositions approach intermediate regions. When the intermediate compositional range exhibits relatively simple phase behavior, the generated phase diagrams demonstrate greater accuracy, yielding elevated scores as observed in the Co-Th and Mg-Si-Cu systems. Conversely, systems characterized by more complex intermediate phase behavior frequently produce lower scores, as exemplified by the Co-Ti and Cr-Ni-Al systems. These findings indicate that the inherent complexity of intermediate compositional regions is a key factor contributing to the difficulty of phase diagram generation for aLLoyM.

To check the training-testing data division dependence of accuracy, five-fold cross-validation on the full phase information task for the short answer Q&As was performed (see Fig. S6). It was confirmed that the accuracy fluctuates significantly across each fold, suggesting that the accuracy of the answers varies depending on the chemical distance reflected in the data split.

Novel phase diagram generation

aLLoyM enables the generation of entirely novel phase diagrams, including those that are currently unknown or extremely difficult to construct experimentally. Figure 6 presents examples of such phase diagrams for both binary and ternary systems, generated using aLLoyM with the short-answer Q&A format. We first examine the results for binary systems. Phase diagrams were generated for the Th-Ac (thorium-actinium) and U-Nh (uranium-nihonium) systems. In the case of Th-Ac, pure thorium was

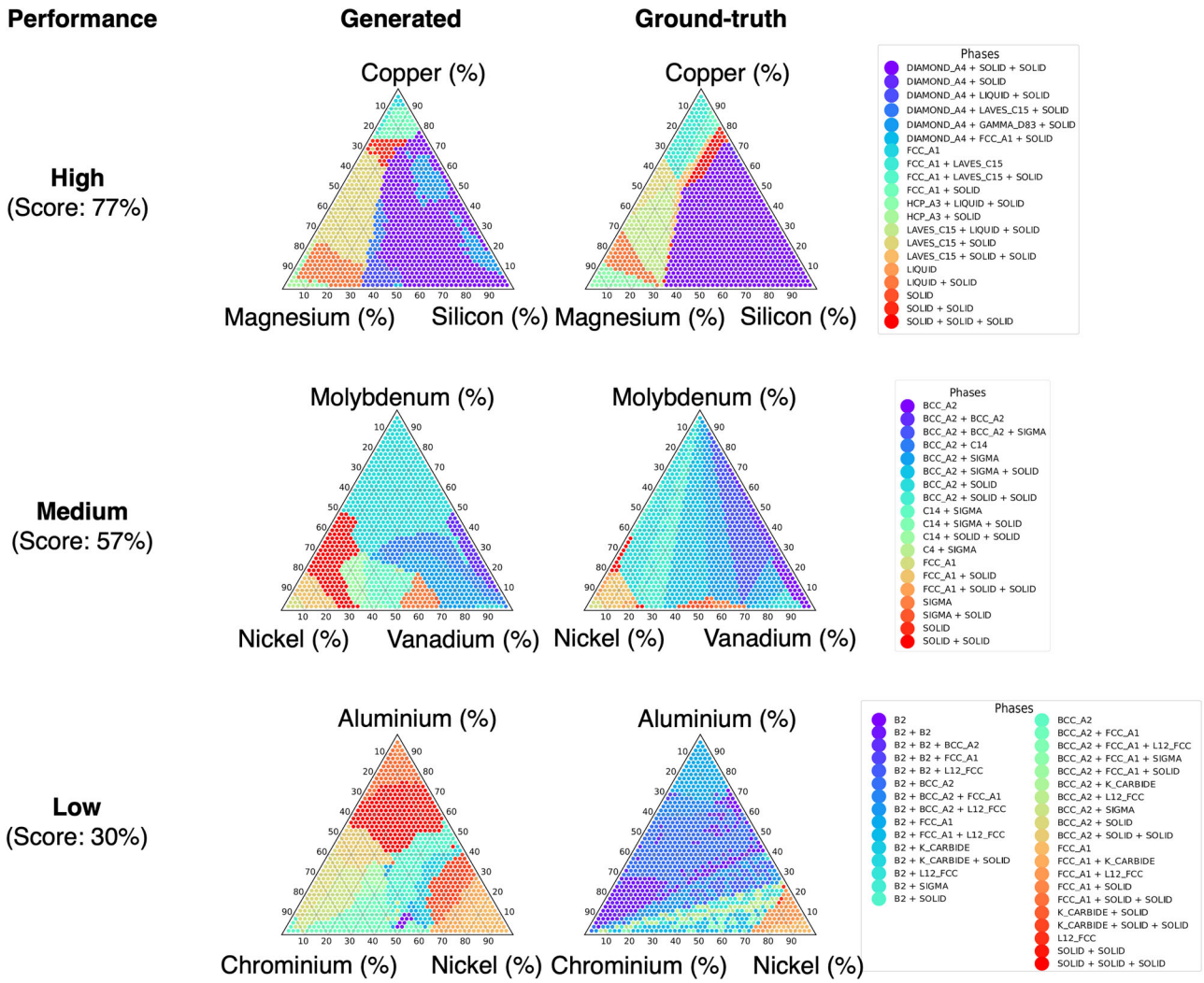

Fig. 5 | Representative 800K ternary isothermal sections exhibiting varying predictive performance, as generated by aLLoyM for the phase name prediction task. The ground-truth phase diagrams are also shown. Lower scores denote greater discrepancies between generated and ground-truth phase diagrams.

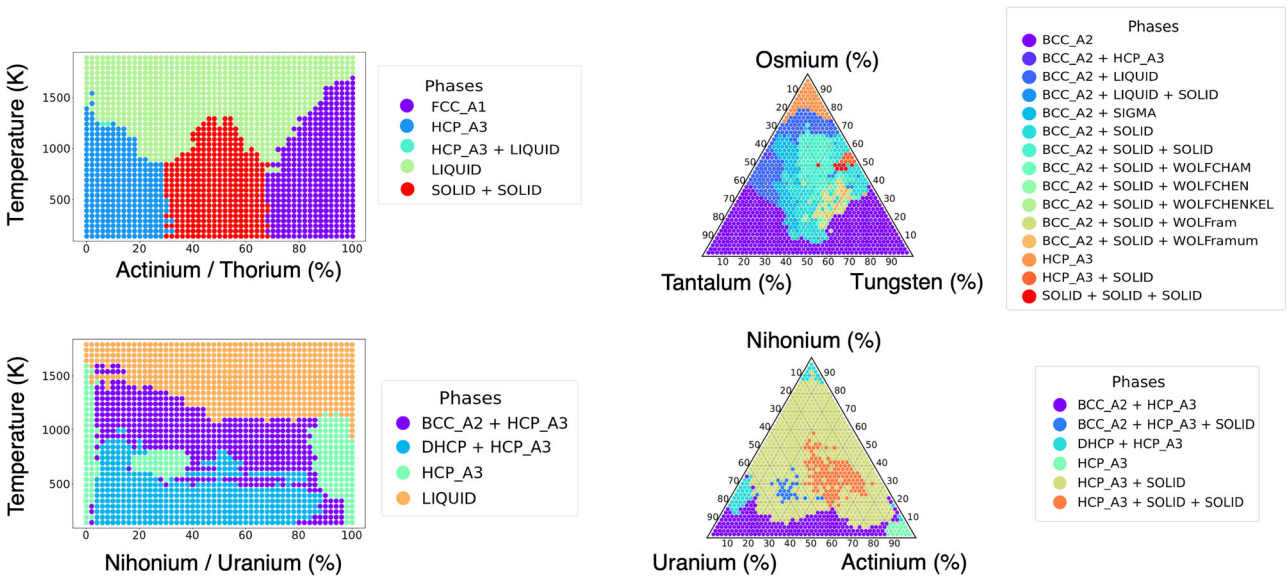

Fig. 6 | Examples of unknown phase diagram generations. Binary phase diagrams and 800K ternary isothermal sections were inferred by aLLoyM in the phase name prediction task.

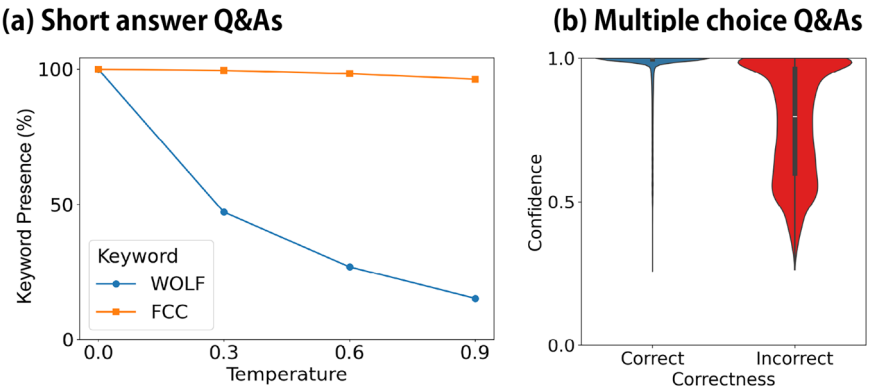

Fig. 7 | Reliability analysis of phase diagram generations. a Proportion of generated answers that included the specific keyword depending on the sampling temperature. 100 responses are generated for the questions where the answers including the "WOLF" (novel prediction) and FCC (correct). **b** Confidence distributions between correct and incorrect predictions for the full phase prediction task of multiple choice Q&As.

incorporated within the training dataset, whereas actinium was omitted owing to its short half-life. aLLoyM predicted the melting point of actinium to be approximately 1400 °C, which is consistent with the experimental value of approximately 1050 °C[29]. However, while the stable crystal structure of actinium is known to be face-centered cubic (FCC)[30], the model incorrectly predicted it as hexagonal close-packed (HCP). For the U-Nh system, neither uranium nor nihonium was included in the training data, making this an entirely extrapolative prediction. The predicted melting point for uranium was approximately 900 °C, compared to the known value of 1135 °C[31], indicating only a moderate deviation. However, aLLoyM erroneously predicted HCP as the stable structure, while uranium's actual stable structure is body-centered cubic (BCC)[32]. For nihonium, no experimental data on melting point or crystal structure are currently available. Nevertheless, the model was able to generate phase diagram outputs, illustrating its potential to make predictions in domains where experimental data are scarce or nonexistent. It should be noted that the pre-trained Mistral model correctly predicted the stable low-temperature structures of actinium and uranium. However, after fine-tuning, aLLoyM generated incorrect crystal structures, indicating that catastrophic forgetting occurred even when RoLA fine-tuning was applied.

We subsequently examine the results for ternary systems. tungsten (W), tantalum (Ta), and osmium (Os) are all elements characterized by exceptionally high melting points, rendering experimental investigation of their ternary phase diagram particularly challenging. To date, no ternary phase diagrams have been established for this system, although all three constituent elements are present in the training data for binary systems. Using aLLoyM, we reconstructed the ternary phase diagram for this system at 800 K. In the intermediate compositional region, the model predicts the emergence of three-phase coexistence. Notably, aLLoyM also predicts the existence of phases designated with "WOLF" nomenclature that are absent from the training data. These may reflect latent knowledge embedded within the pre-trained Mistral model. Finally, we generated a ternary phase diagram for nihonium, uranium, and actinium at 800 K, representing an entirely hypothetical system that cannot be experimentally realized. Here as well, the model predicts three-phase coexistence in the intermediate compositional region as well as the W-Ta-Os system.

Since the phase diagrams generated above remain beyond experimental validation, they should be regarded only as illustrative examples. As the next step of this research, it is important to predict realistic novel phase diagrams that can be experimentally tested and to carry out their experimental evaluation.

Reliability assessment

Evaluating the reliability of generated answers is crucial for demonstrating the accuracy of the predictions. Here, we address methods for assessing the reliability of aLLoyM's output using both temperature-based evaluation and confidence-based evaluation.

Temperature-based evaluation: aLLoyM occasionally produces novel phase-name predictions in response to short-answer Q&As, such as a phase name containing "WOLF." To assess the reliability of such predictions, we perform a temperature-based evaluation. For each question where a phase name containing "WOLF" was generated at sampling temperature $T = 0$, we generated 100 responses per temperature setting, increasing the temperature in increments of 0.3. The proportion of generated phase names that contained the keyword "WOLF" at each temperature is shown in Fig. 7a. For comparison, we conducted the same analysis for a question whose ground-truth answer is a phase name containing "FCC." We observed that the proportion of FCC-containing predictions remained stable even as the temperature increased, whereas WOLF-containing predictions disappeared rapidly. This indicates that "WOLF" corresponds to a low-reliability prediction. Such analyses are therefore essential whenever a novel phase-name prediction is obtained, to evaluate the robustness of the model's output.

Confidence-based evaluation: we next examined a confidence-based method for assessing prediction reliability. Because confidence can be evaluated at the token level, we focused on multiple-choice Q&As in which all answers share a uniform length. For each question, we estimated confidence by computing the model's log-likelihood of generating each candidate option label (a, b, c, or d) as the next token, applying a softmax over these scores, and taking the probability assigned to the most likely label. We then compared the resulting confidence distributions for correct versus incorrect predictions in the full phase prediction task, as shown in Fig. 7b. The confidence associated with incorrect predictions is clearly lower than that of correct predictions. This indicates that, for multiple-choice Q&As, inspecting the model's confidence provides an effective means of evaluating the reliability of its answers.

Discussion

In this work, we developed aLLoyM, a fine-tuned Large Language Model specialized for relations between alloy compositions, temperatures, and phase information. The model was fine-tuned on Q&As for binary and ternary phase diagrams constructed from the open-source Computational Phase Diagram Database (CPDDB) using CALculation of PHAse Diagrams (CALPHAD) assessments. Our benchmark results demonstrated that fine-tuning significantly improves the accuracy of the model in selecting the correct responses to multiple-choice questions concerning phase diagrams. Furthermore, the short-answer model of aLLoyM can be used to generate phase diagrams for previously unreported systems. These results indicate that aLLoyM provides a potentially useful framework for phase diagram prediction, with some capacity for extrapolation to novel systems. Its ability to infer phase behavior in previously unexplored compositional spaces could facilitate the design and discovery of new materials.

The consistently stronger performance on binary systems compared to ternary systems across all evaluations can be attributed to the relatively limited availability of ternary training data. Moreover, the absence of temperature-dependent training data for ternary systems prevents aLLoyM from making reliable predictions across different temperatures, particularly below 800 K. Future work should therefore prioritize expanding the training data for ternary and higher-order systems with explicit temperature

dependence to enable more robust predictions of multi-component phase diagrams. In addition, the integration of experimental phase diagrams using LLMs represents an important future perspective, where phase diagram annotation techniques based on LLMs can be leveraged.

A key advantage of aLLoyM's natural language framework lies in its ability to utilize the virtually unlimited vocabulary of elements and phase names acquired during pretraining, thus supporting broad generalization to diverse chemical systems. While the current implementation tends to generate phase names seen during training, this limitation opens promising opportunities for improvement through advanced prompt engineering. In particular, incorporating thermodynamics-aware prompts may help guide the model toward applying physically meaningful reasoning during inference, thereby enhancing prediction accuracy. aLLoyM's training utilized a standardized prompt template, which means its prediction quality may be sensitive to variations in how prompts are phrased or input formats are changed. We encourage users to experiment with various prompting approaches and share successful strategies, as the field of prompt engineering is constantly evolving to optimize LLM performance through input design[33,34]. To advance phase diagram prediction, integrating thermodynamic information, particularly Gibbs energy, will be essential. As Gibbs energy data are available in TDB files, future models should be trained to incorporate this information directly. Another current limitation of aLLoyM is the absence of uncertainty quantification, which restricts it to deterministic outputs. Embedding mechanisms for uncertainty estimation will be critical for enabling reliable predictions in materials design. In this study, we examined the generation of phase diagrams through discrete Q&A as a proof of concept. Nevertheless, we recognize the importance of developing approaches that can handle phase boundaries in a continuous manner. One promising direction is the use of Q&A grounded in graph-based representations, along with the development of strategies for constructing Q&A datasets that are better tailored to phase diagram generation. Collectively, these directions represent important pathways for future research toward developing more capable LLMs specifically tailored to phase diagram prediction and materials discovery.

Methods

Fine-tuning

We fine-tuned the Mistral-Nemo-Instruct-2407 model using LoRA (Low-Rank Adaptation) with rank 16 and alpha 16, targeting attention and feed-forward projections. These hyperparameter values correspond to the default LoRA adapters in the pretrained model's official demo https://colab.research.google.com/github/unslothai/studio/blob/main/colabs/mistral_nemo_12b.ipynb. We confirmed that modifying these values does not substantially affect the accuracies (see Fig. S7). Training data was formatted using a structured prompt template with Instruction, Input, and Output sections (see Table S1). The model was trained for 15,000 steps with a learning rate of 2×10^{-4}, batch size of 16 per device, and 4 gradient accumulation steps using the AdamW optimizer with bfloat16 precision. Training of the full-aLLoyM required 32 h on a NVIDIA A100 GPU (80GB PCIe). The training was conducted using Python 3.10.10 in a Linux environment (6.8.0-55-generic). On the same GPU, generation required approximately 1 s per question. We confirmed that general-purpose language tasks can indeed be performed even after LoRA-based fine-tuning, indicating that knowledge retention is maintained.

Scoring criteria for generated answers

For the short answer Q&As, the scoring criteria for generated answers depend on the specific Q&A task, as both the answer format and the target subject vary across tasks. The definition of the scores for each task is shown below. All of the following scores are defined with a maximum value of 100%. The details are summarized in Supplementary Note B. *Full phase information*: the exact match between the generated answer and the ground-truth answer was used. *Phase name*: the score was evaluated using the Jaccard similarity of perfectly matching phase names. *Experimental condition*: the scoring of experimental conditions evaluates how well the element compositions and temperature match one of the ground-truths by comparing composition accuracy and temperature accuracy.

Data availability

All Q&A data used in this study are publicly available at: https://huggingface.co/datasets/Playingyoyo/aLLoyM-dataset. The short-answer version of aLLoyM, fine-tuned on the full dataset, can be accessed at: https://huggingface.co/Playingyoyo/aLLoyM.

Code availability

Code for aLLoyM is available at https://github.com/tsudalab/aLLoyM/tree/main.

Received: 29 July 2025; Accepted: 11 January 2026;
Published online: 22 January 2026

References

1. Schlesinger, M. E. & Mueller, E. M. (eds) *ASM Handbook*, Vol. 3 (ASM International, 1983).
2. Massalski, T. B. & Okamoto, H. (eds) *Binary Alloy Phase Diagrams* (ASM International, 1990).
3. Villars, P., Prince, A. & Okamoto, H. *Handbook of Ternary Alloy Phase Diagrams* (ASM International, 1995).
4. Okamoto, H. *Desk Handbook* 2nd edn. ASM Handbooks (ASM International, 2010).
5. Computational phase diagram database (CPDDB). https://cpddb.nims.go.jp/.
6. Jung, I.-H. & Van Ende, M.-A. Computational thermodynamic calculations: FactSage from CALPHAD thermodynamic database to virtual process simulation. *Metall. Mater. Trans. B* **51**, 1851–1874 (2020).
7. Hallstedt, B., Noori, M., Kies, F., Oppermann, F. & Haase, C. Thermodynamic database for multi-principal element alloys within the system Al-Co-Cr-Fe-Mn-Ni-C. *Calphad* **83**, 102644 (2023).
8. Terayama, K. et al. Efficient construction method for phase diagrams using uncertainty sampling. *Phys. Rev. Mater.* **3**, https://doi.org/10.1103/PhysRevMaterials.3.033802 (2019).
9. Aghaaminiha, M., Ghanadian, S. A., Ahmadi, E. & Farnoud, A. M. A machine learning approach to estimation of phase diagrams for three-component lipid mixtures. *Biochim. Biophys. Acta Biomembr.* **1862**, 183350 (2020).
10. Dai, C. & Glotzer, S. C. Efficient phase diagram sampling by active learning. *J. Phys. Chem. B* **124**, 1275–1284 (2020).
11. Lund, J., Wang, H., Braatz, R. D. & García, R. E. Machine learning of phase diagrams. *Mater. Adv.* **3**, 8485–8497 (2022).
12. Zipoli, F., Viterbo, V., Schilter, O., Kahle, L. & Laino, T. Prediction of phase diagrams and associated phase structural properties. *Ind. Eng. Chem. Res.* **61**, 8378–8389 (2022).
13. Tamura, R. et al. Machine-learning-based phase diagram construction for high-throughput batch experiments. *Sci. Technol. Adv. Mater. Methods* **2**, 153–161 (2022).
14. Deffrennes, G., Terayama, K., Abe, T. & Tamura, R. A machine learning-based classification approach for phase diagram prediction. *Mater. Des.* **215**, 110497 (2022).
15. Tamura, R. et al. AIPHAD, an active learning web application for visual understanding of phase diagrams. *Commun. Mater.* **5**, 139 (2024).
16. Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discov.* **2**, 1233–1250 (2023).
17. Liu, Y. et al. Generative artificial intelligence and its applications in materials science: current situation and future perspectives. *J. Materiomics* **9**, 798–816 (2023).
18. Lei, G., Docherty, R. & Cooper, S. J. Materials science in the era of large language models: a perspective. *Digital Discov.* **3**, 1257–1272 (2024).

19. Deb, J., Saikia, L., Dihingia, K. D. & Sastry, G. N. ChatGPT in the material design: selected case studies to assess the potential of ChatGPT. *J. Chem. Inf. Model.* **64**, 799–811 (2024).
20. Jiang, X. et al. Applications of natural language processing and large language models in materials discovery. *npj Comput. Mater.* **11**, 79 (2025).
21. Yan, Z. et al. PDGPT: a large language model for acquiring phase diagram information in magnesium alloys. *Mater. Genome Eng. Adv.* **2**, e77 (2024).
22. Zha, Y., Li, Y. & Lu, X.-G. Enhancing large language model comprehension of material phase diagrams through prompt engineering and benchmark datasets. *Mathematics* **12**, 3141 (2024).
23. Hu, E. J. et al. Lora: low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685 (2021).
24. Gruver, N. et al. Fine-tuned language models generate stable inorganic materials as text. https://arxiv.org/abs/2402.04379 (2025).
25. Harod, H. et al. Effichem: efficient adaptation of chemical language models for molecular property prediction. https://doi.org/10.26434/chemrxiv-2025-2lljt (2025).
26. Gao, B. et al. Lora-chem: modular machine learning for multitask prediction in organic reactions. https://doi.org/10.26434/chemrxiv-2025-p7sxn (2025).
27. Pandat software. https://computherm.com/.
28. Hugging face, mistral-nemo-instruct-2407-bnb-4bit. https://huggingface.co/unsloth/Mistral-Nemo-Instruct-2407-bnb-4bit.
29. Periodic table, actinium. https://periodic-table.rsc.org/element/89/actinium.
30. Farr, J., Giorgi, A., Bowman, M. & Money, R. The crystal structure of actinium metal and actinium hydride. *J. Inorg. Nucl. Chem.* **18**, 42–47 (1961).
31. Periodic table, uranium. https://periodic-table.rsc.org/element/92/uranium.
32. Grenthe, I. et al. *Uranium*, 253–698 (Springer, 2008).
33. Sahoo, P. et al. A systematic survey of prompt engineering in large language models: techniques and applications. https://arxiv.org/abs/2402.07927 (2025).
34. Rodriguez, A. D., Dearstyne, K. R. & Cleland-Huang, J. Prompts matter: insights and strategies for prompt engineering in automated software traceability. https://doi.org/10.1109/REW57809.2023.00087 (2023).

Acknowledgements

The authors would like to thank Etsuko Ogamino for data collection. This study was supported by a project subsidized by JSPS KAKENHI (25K01492 and 25KJ0870), JST-CREST (JPMJCR21O2), and MEXT Program: Data Creation and Utilization Type Material Research and Development Project (JPMXP1122715503 and JPMXP1122712807).

Author contributions

All the authors conceived the original idea. Y.O. and R.S. prepared Q&As from CALPHAD assessment data and developed aLLoyM. G.D., T.A., and R.T. prepared the CALPHAD assessment data of phase diagrams. Y.O., R.T., and K.T. wrote the original manuscript. All the authors discussed the results, commented on the manuscript, and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-026-01966-6.

Correspondence and requests for materials should be addressed to Ryo Tamura or Koji Tsuda.