aLLoyM: A large language model for alloy phase diagram prediction

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

[*]Equal contribution [1][Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba, 277-8561, Japan] [2][University Grenoble Alpes, CNRS, Grenoble INP, SIMaP, Grenoble, F-38000, France] [3][Research Center for Structural Materials, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047, Japan] [4][Center for Basic Research on Materials, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki, 305-0044, Japan] [5][RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan]

Correspondence to: Ryo Tamura tamura.ryo@nims.go.jp; Koji Tsuda tsuda@k.u-tokyo.ac.jp

Large Language Models are general-purpose tools with wide-ranging applications, including in materials science. In this work, we introduce aLLoyM, a Mistral-Nemo-Instruct-2407-bnb-4bit based model fine-tuned specifically 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) [1] and assessments based on CALPHAD (CALculation of PHAse Diagrams) [2]. Excerpts of the three types of Q&As are represented in Table 1. The dataset was split into training and test sets using an 8:2 ratio under two strategies: an interpolation split, where data points from all systems were randomly distributed, and an extrapolation split, where entire systems in the test set were excluded from training to assess generalization to unseen systems. We compared aLLoyM, which was LoRA finetuned with the training dataset, with the baseline (unfinetuned) Mistral model using a multiple-choice benchmark. Each question required selecting the correct answer from four options, with three distractors sampled from the same alloy system. aLLoyM consistently outperformed the baseline model, demonstrating its ability to generalize to unseen data (Fig 2).

We further evaluated aLLoyM using short-answer tasks, allowing free-form generation without multiple-choice constraints. Model outputs were compared to ground truth using task-specific scoring metrics (see Appendix A). aLLoyM successfully generated plausible phase names and experimental conditions, with performance varying across scenarios (Fig 3, 4).

The short-answer model is capable of generating novel phase diagrams from elemental inputs alone. As a representative example, we examined the ternary system tungsten–tantalum–osmium (W–Ta–Os), whose exceptionally high melting points make experimental investigation challenging. Although no ternary phase diagram has been established for this system, aLLoyM made it possible to generate the ternary phase diagram at 800 K and predicted a three phase coexistence region at intermediate compositions. Notably, the model predicts phases labelled with “WOLF” nomenclature that do not appear in the training data, potentially reflecting latent knowledge from the pre-trained Mistral model.

To promote further research and adoption, we have publicly released the short-answer fine-tuned version of aLLoyM, (https://huggingface.co/Playingyoyo/aLLoyM) along with the complete benchmarking Q&A dataset (https://huggingface.co/datasets/Playingyoyo/aLLoyM-dataset), on Hugging Face.

3.1 Related work

Preliminary investigations have explored LLM applications in phase diagram analysis, including system specific training on Mg-Al-Zn data [3] and experimental diagram annotation [4], suggesting substantial potential for phase diagram analysis.

3.2 Figures and tables

Table 1: Excerpts of three types of Q&As

Task	Full phase information	Phase name	Experimental condition
Instruction	You are an expert in phase diagrams, thermodynamics, and materials science, specializing in alloy systems.		
Question (Input)	What phases form when Silver (34%) + Yttrium (66%) are mixed at 500 K?	What phases form when Silver (34%) + Yttrium (66%) are mixed at 500 K? Answer phase names only.	Under what condition do Silver + Yttrium form HCP_A3 + SOLID?
Answer (Output)	32% HCP_A3 + 68% SOLID with composition ratio Silver : Yttrium = 1 : 1.	HCP_A3 + SOLID.	Silver (34%) + Yttrium (66%) mixed at 500 K.

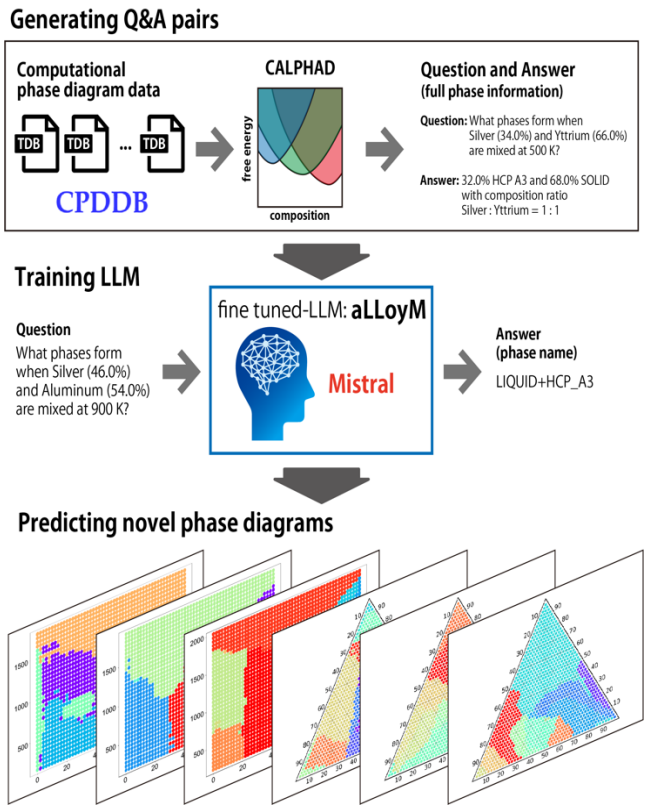

Fig. 1: Schematic of aLLoyM

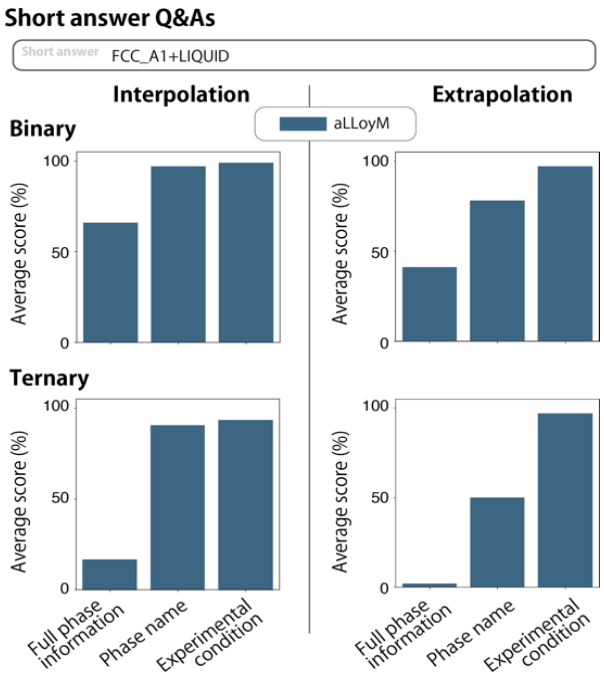

Fig. 3: Average scores of the fine-tuned models (aLLoyM) for short answer Q&As

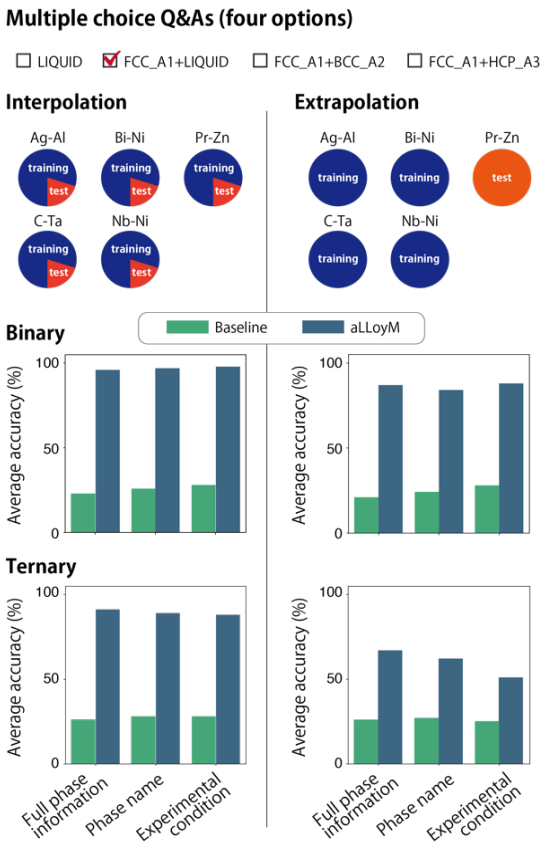

Fig. 2: Accuracies of the baseline model (Mistral-Nemo-Instruct-2407-bnb-4bit) and the fine-tuned model (aLLoyM) on multiple-choice Q&As

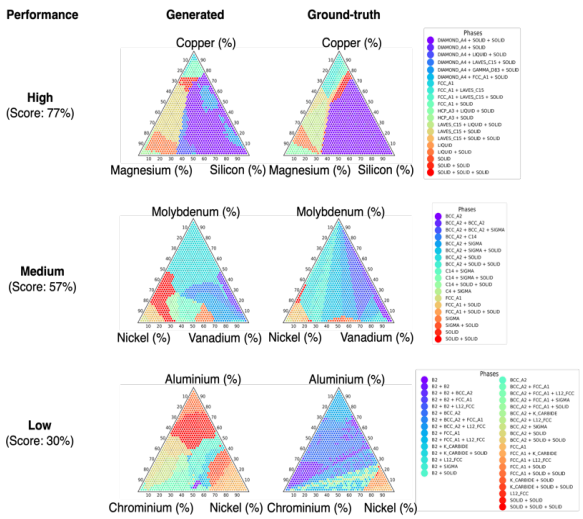

Fig 4: Representative 800K ternary isothermal sections exhibiting varying predictive performance, as generated by aLLoyM for the phase name prediction task

Acknowledgments

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.

References

[1] Computational phase diagram database (CPDDB). URL https://cpddb.nims.go. jp/.
[2] Jung, I.-H. & Van Ende, M.-A. Computational thermodynamic calculations: FactSage from CALPHAD thermodynamic database to virtual process simulation. Metallurgical and Materials Transactions B 51, 1851-1874 (2020). URL http://dx.doi.org/10.1007/s11663-020-01908-7
[3] Yan, Z. et al. PDGPT: A large language model for acquiring phase diagram information in magnesium alloys. Materials Genome Engineering Advances 2, e77 (2024). URL http://dx.doi.org/10.1002/mgea.77
[4] 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). URL http://dx.doi.org/10.3390/math12193141.

Appendix A. Scoring criteria for the free-form generated answers

The scoring criteria for generated answers depends 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.

Full phase information: This requires an exact match between the generated answer and the ground-truth answer. The score is defined as

$$\text{Score} = \begin{cases} 100\% & : \text{if the generated answer exactly matches the ground truth} \\ 0\% & : \text{otherwise} \end{cases}.$$

Phase name: We first construct lists of all phase domains in the generated answer and the ground truth, denoted as A and B, respectively. Let k and l be the number of phase domains in the generated answer and the ground truth, respectively. The resulting lists are as follows:

$$A = [\text{phase name 1},\ \text{phase name 2},\ \ldots,\ \text{phase name } k],$$

$$B = [\text{phase name 1},\ \text{phase name 2},\ \ldots,\ \text{phase name } l].$$

The score is then evaluated using the Jaccard similarity of perfectly matching phase names, defined as:

$$\text{Score} = \begin{cases} |A \cap B| / |A \cup B| \times 100\% & : \text{if } |A \cup B| > 0 \\ 0\% & : \text{otherwise} \end{cases}.$$

With this definition, scores can be evaluated even when the number of phase domains differs between the generated answer and the ground truth.

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. Composition Accuracy measures how closely the generated composition matches the ground-truth composition. The compositions of the generated answer and the ground-truth are defined as

$$\{x_n\}_{n=1,\ldots,N} \text{ and } \{x_n^*\}_{n=1,\ldots,N}$$

where N is the number of elements in the target system. The composition accuracy is defined as

$$A_\mathrm{c} = \frac{1}{N} \sum_{n=1}^{N} \left(1 - \frac{|x_n^* - x_n|}{100} \right).$$

To standardize the values, the each composition difference

$$|x_n^* - x_n|$$

is divided by 100 which is the maximum difference. Temperature Accuracy measures how closely the generated temperature T matches the ground-truth temperature T ∗ . It normalizes the absolute difference ΔT by the full range of temperatures available in the target elemental system. The temperature accuracy is defined as

$$A_\mathrm{t} = 1 - \frac{|T^* - T|}{\Delta T}.$$

Using these accuracies, the score for experimental condition is defined as

$$\text{Score} = \left(\frac{1}{2} A_\mathrm{c} + \frac{1}{2} A_\mathrm{t} \right) \times 100\%.$$

This score is 100% when both the compositions and the temperature in the generated answer exactly match with any of the ground truths. Note that in this case, multiple ground-truth answers may exist. Therefore, the above score is computed for each ground-truth answer, and the maximum value is adopted.