LLaMat: Large Language Models for Materials Science

Vaibhav Mishra[1], **Somaditya Singh**[1], **Mohd Zaki**[2], **Hargun Singh Grover**[3], **Santiago Miret**[4], **Mausam**[1,3], **N. M. Anoop Krishnan**[2,3]
[1]Department of Computer Science and Engineering, [2]Department of Civil Engineering
[3]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi
[4]Intel labs
{cs1200448, cs1200389, cez198233, aiz218326}@iitd.ac.in
{santiago.miret}@intel.com, {mausam, krishnan}@iitd.ac.in

Abstract

Large language models are versatile tools that have been recently used in the materials science domain for tasks ranging from information extraction to acting as scientific assistants in materials discovery. It is believed that using domain-specific large language models will help improve performance on such tasks. In this work, we address the challenge of efficiently accessing and utilizing vast textual knowledge in materials science using continued pre-training of Meta's LLaMA-2-7B on curated materials science texts, enhancing its domain-specific capabilities. We also developed LLaMat-Chat, an instruction fine-tuned variant of LLaMat that is tailored through a dataset of one million instruction-output pairs, enabling interactive applications and proficient performance in natural language processing tasks within materials science. We show that LLaMat achieves state-of-the-art performance on several information extraction tasks from materials science text. Since the pre-training corpus also included crystallographic information files, it will be interesting in future to evaluate the materials discovery applications of LLaMat.

1 Introduction

Knowledge about materials has been reported in the form of text, which includes books, research papers, patents, and technical reports, to name a few. It is humanly intractable for humans to go through a large amount of text and find answers to specific questions related to different materials science aspectsHira et al. (2024); Miret & Krishnan (2024). Dissemination of textual information in a natural language is an important aspect of democratising access to knowledge about materials science. However, developing a model capable of performing different types of tasks with high accuracy is a challenging task, which has been taken up by several researchers trying to address it by developing foundational models. Large language models are one of them.

Large language models have started revolutionizing both scientific and non-scientific domains. Due to their capability to perform a variety of tasks by understanding input through human language, they are also called foundational models. Recently, several researchers have attempted to develop and understand the capabilities of foundational models for chemistry (Zhang et al. (2024); Mirza et al. (2024)) and the medical domain(Chen et al. (2023)) or use general-purpose foundational models for domain-specific tasks either directly or after finetuning(Dagdelen et al. (2024); Polak & Morgan (2024); M. Bran et al. (2024); Boiko et al. (2023); Song et al. (2023b)). The benefits of domain adaptation of foundational models are well documented. Considering the wide variety of sub-domains for the part of materials science, a foundational model understanding the broad domain of materials science will enable the researchers to get the answers to highly specialised questions.

In response to the growing need for a foundational language model tailored to the domain of materials science, we propose LLaMat (Large Language Model for Materials Science). This model builds upon the architecture of LLaMA-2-7BTouvron et al. (2023), undergoing further pretraining on a carefully

curated corpus of high-quality materials science texts. This extended pretraining aims to enhance the model's domain-specific knowledge and performance.

To endow LLaMat with robust conversational abilities, we introduce LLaMat-Chat. This variant has been instruction fine-tuned using a dataset comprising approximately one million instruction-output pairs. The instruction fine-tuning process equips the model with the capability to understand and generate responses based on given instructions, thus facilitating interactive and user-friendly applications. This advanced model is proficient in performing classical natural language processing (NLP) tasks such as Named Entity Recognition, Abstract Classification, Relation Extraction, and Event Extraction within materials science datasets. In addition to these tasks, LLaMat-Chat can provide succinct and detailed answers to questions related to materials science tailored to the user's requirements. Fig. 1 shows the pipeline of development of LLaMat and LLaMat-Chat models.

2 R2CID - PRETRAIN DATASET

For pretraining LLaMat, we consider the text from **R**esearch papers, a subset of **R**edpajama dataset, **Ci**f (crystallographic information files) files **D**ataset. We call our training corpus the R2CID database. The details of each part are provided as follows.

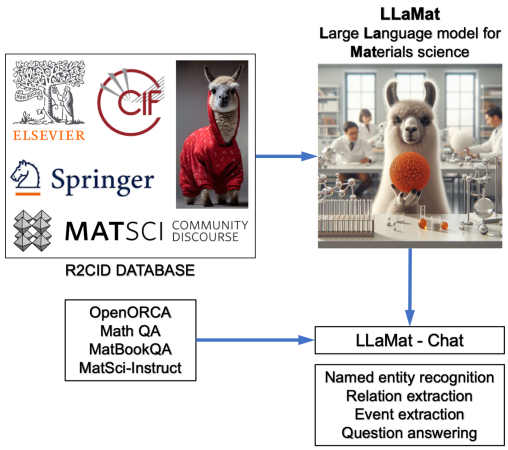

Figure 1: LLaMat and LLaMat-Chat development pipeline

2.1 RESEARCH PAPERS

We sourced research papers from around 500 Elsevierels (a) journals and 300 Springerspr journals to compile a comprehensive and high-quality dataset. The inclusion criteria required full-text availability in XML format for Elsevier papers and HTML format for Springer papers, ensuring compatibility with our data processing pipeline. The choice of Elsevier and Springer journals was influenced by the constraints of our institution's subscription contract, which provided access to a wide range of journals from these publishers. This contractual limitation shaped the scope of our dataset. The selected research papers' Digital Object Identifiers (DOIs) were retrieved using the CrossRef APIFarley. After obtaining the DOIs, the full texts of the research papers were downloaded using the publisher specific APIsels (b); spr. These APIs facilitated access to the papers in the specified formats (XML for Elsevier and HTML for Springer), which were then incorporated into the R2CID corpus.

2.2 REDPAJAMA SAMPLE

The RedPajama datasetred (2024) was employed as the foundational corpus for the initial training phase of the LLaMA-2 model. We systematically extracted approximately 700 million tokens from this corpus to ensure a representative sample. The primary objective of incorporating this subset into R2CID is to address the issue of catastrophic forgetting, thereby preserving the model's comprehension and utility derived from its original, general-purpose training corpus. This ensures the model retains its foundational knowledge while effectively assimilating new information.

2.3 CRYSTALLOGRAPHIC INFORMATION FILES

Material structures are best obtained through diffraction studies and are reported as Crystallography Information Files. These are standardized text files used for storing and exchanging crystallographic data. These files contain unit cell parameters like the lengths of cell edges and angles between them. They also include symmetry information, such as the space group and symmetry operations, and atomic coordinates that specify the positions of atoms within the unit cell. To allow an increased understanding of CIF files, we considered a total of 470k CIF files and obtained their description in

natural language using RoboCrystallographerGanose & Jain (2019). The R2CID consists of these CIF files and their descriptions from the Materials ProjectJain et al. (2020), Google GNoMEMerchant et al. (2023), and AMSCSD databaseamc.

Merging the components to form R2CID: To enhance the effectiveness of model training and mitigate catastrophic forgetting, research papers were periodically interleaved with text from the RedPajama corpus. The periodic interleaving strategy was refined through a series of empirical evaluations. The selected interleaving period of 100 million research-related tokens with 2.3 million RedPajama tokens provided a balance that enhanced the model's ability to generalize and retain relevant information from both datasets. The CIF files were included in the posterior 10% of the corpus and interleaved with research papers.

3 PRETRAINING METHODOLOGY

LLaMat was initialised with weights of LLaMA-2-7B and pretrained on R2CID for one epoch. The learning rate was initialised at 0, increased to 3×10^{-4} and then adhered a cosine decay schedule to stop at 3×10^{-5}. The pretraining was done using the Megatron-LLM library introduced by Shoeybi et al. (2019) and extended to LLaMA-2-7B by Chen et al. (2023), that utilises 3D model parallelism for efficient training of LLMs. The pretraining was done on 16 A100 NVIDIA GPUs for approximately 9 days. The loss curve can be seen in the Appendix (Fig. 2).

4 INSTRUCTION FINE-TUNE METHODOLOGY

LLaMat-Chat was initialized with the weights of LLaMat. The instruction fine-tuning process was conducted in three distinct stages:

- **Stage 1:** LLaMat-Chat was first fine-tuned on the OpenOrca dataset for one epoch. The objective of this stage was to enable the pretrained model to learn how to follow common English instructions.
- **Stage 2:** The model was further fine-tuned on a dataset of mathematical questions for three epochs. This stage aimed to enhance the mathematical reasoning capabilities of LLaMat-Chat. Due to the relatively small size of this dataset, we observed a decrease in validation loss over the three epochs.
- **Stage 3:** In the final stage, LLaMat-Chat was fine-tuned on a combined dataset constructed from MatSciInstruct, MatSciNLP, MatBookQA, and MaScQA (for one epoch).

The fine-tuning process utilized the Megatron-LLM library. The learning rate for each stage was initialized at 2×10^{-6} and increased to 2×10^{-5} over the first 10% of the total iterations. Following this initial increase, the learning rate adhered to a cosine decay schedule.

5 RESULTS

5.1 DOWNSTREAM TASKS

To continuously evaluate the improved understanding of Materials Science principles gained by pretraining on **R2CID** as well as to measure any potential degradation in understanding conversational or informal English, we curated a dataset consisting of Materials Science and English Comprehension tasks. Table 1 shows the list of different tasks, datasets, and the number of samples in training and validation sets. The dataset has the following tasks: **sc**: sentence classification, **re**: relation extraction, **ner**: named entity extraction, **sar**: synthesis action retrieval, a type of classfification task, **pc**: paragraph classification, **ee**: entity extraction, **sf**: slot filling, **qna**: question answering, and **mcq** : multiple choice question answering. The details of these task can be found in Song et al. (2023a). The samples from the training set were used to fine-tune the models before evaluation on the validation set to ensure that the models learned to follow the instructions. The performance of different models on these datasets is shown in Table 2. The Macro and Micro F1 scores were averaged over all the tasks.

Table 1: Details of downstream datasets

Task	Dataset	Train	Val
sc	sofc_sent	1893	1889
re	structured_re	1788	1786
ner	matscholar	1062	1061
ner	sc_comics	937	936
sar	synthesis_actions	565	569
re	sc_comics	376	373
pc	glass_non_glass	300	299
ee	sc_comics	287	288
ner	sofc_token	175	177
sf	sofc_token	175	179
qna	squad	1042	1042
mcq	hellaswag	981	980
mcq	boolqa	500	499

Table 2: Performance on validation set for downstream tasks

Model	Macro F1	Micro F1
LLaMA-2-7B	77.745	84.239
LLaMat	82.26	87.85
LLaMat-Chat	**84.66**	**89.51**

5.2 MATSCI-NLP

To benchmark and compare the performance of LLaMat-Chat against other state-of-the-art models within the materials science domain, we utilized the MatSci-NLP dataset, a comprehensive benchmark for materials science NLP tasks Song et al. (2023a). The evaluation was conducted in a zero-shot manner. The substantial improvement in performance indicates that our pretraining corpus effectively imparts knowledge of various materials science principles to LLaMat-Chat, while our fine-tuning process enhances its instruction-following capabilities. Table **??** shows the performance of various models on the MatSci-NLP. Performance numbers for other models have been adapted from Song et al. (2023c) while ensuring identical experimental settings. The scores are Macro-F1(Top) and Micro-F1(Bottom).

Table 3: Zero-shot performance of LLMs based on MatSci-NLP by Song et al. (2023a).

Model	Named Entity Recognition	Relation Extraction	Event Argument Extraction	Paragraph Classification	Synthesis Action Retrieval	Sentence Classification	Slot Filling	Overall (All Tasks)
Zero-Shot LLM Performance								
LLaMA-7b	0.042	0.094	0.160	0.279	0.052	0.096	0.142	0.208
(Touvron et al., 2023)	0.064	0.013	0.042	0.218	0.013	0.087	0.010	0.064
LLaMA-13b	0.057	0.109	0.042	0.233	0.039	0.079	0.138	0.1
(Touvron et al., 2023)	0.066	0.016	0.054	0.189	0.009	0.074	0.008	0.059
Alpaca-7b	0.031	0.053	0.029	0.375	0.179	0.180	0.139	0.141
(Taori et al., 2023)	0.018	0.037	0.009	0.294	0.129	0.180	0.039	0.101
Alpaca-13b	0.053	0.016	0.111	0.310	0.442	0.375	0.110	0.202
(Taori et al., 2023)	0.046	0.035	0.072	0.237	0.278	0.334	0.015	0.145
Chat-GPT	0.063	0.232	0.204	0.433	0.300	0.320	0.368	0.274
(OpenAI, 2022)	0.052	0.145	0.203	0.450	0.183	0.318	0.280	0.233
Claude	0.063	0.232	0.195	0.442	0.280	0.329	0.393	0.276
(Bai et al., 2022)	0.048	0.143	0.169	0.467	0.177	0.326	0.305	0.234
GPT-4	0.189	0.445	0.453	0.679	0.743	0.788	0.502	0.543
(OpenAI, 2023)	0.121	0.432	0.353	0.522	0.677	0.689	0.483	0.468
LLaMat-Chat	**0.827**	**0.968**	**0.633**	**0.843**	**0.938**	**0.773**	**0.744**	**0.813**
	0.898	**0.952**	**0.836**	**0.871**	**0.962**	**0.917**	**0.839**	**0.894**

6 CONCLUSION AND FUTURE WORK

The results indicate that domain-specific continued pre-training helps improve performance on several downstream tasks, which are useful for materials discovery. Since the training corpus included information about different tasks related to materials discovery, like, research papers, crystallography information files, information extraction tasks, and question-answering pair, it will be interesting to evaluate the effect of each component of corpus on the final performance of the model. Further, Several open-source small and large language models are being released, which also lack materials science domain knowledge. Therefore, in future, it will be interesting to see the effect of continued pre-training on the performance of these models and deploy them in materials discovery pipelines.

REFERENCES

American Mineralogist Crystal Structure Database. URL https://rruff.geo.arizona.edu/AMS/amcsd.php.

ScienceDirect.com | Science, health and medical journals, full text articles and books., a. URL https://www.sciencedirect.com/.

Elsevier Developer Portal, b. URL https://dev.elsevier.com/.

Springer Nature Developer Portal | APIs for Research Papers. URL https://dev.springernature.com/.

togethercomputer/RedPajama-Data-1T · Datasets at Hugging Face, July 2024. URL https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, 2024.

Isaac Farley. Documentation. URL https://www.crossref.org/documentation/.

Alex M. Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019. doi: 10.1557/mrc.2019.94.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Kausik Hira, Mohd Zaki, Dhruvil Sheth, NM Anoop Krishnan, et al. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digital Discovery*, 3(5):1021–1037, 2024.

Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils ER Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pp. 1751–1784, 2020.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

OpenAI. openaiintroducingchatgpt. `https://openai.com/blog/chatgpt`, 2022. [Accessed 22-Jun-2023].

OpenAI. Gpt-4 technical report, 2023.

Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264*, 2023a.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. HoneyBee: Progressive instruction finetuning of large language models for materials science. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5724–5739, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.380. URL `https://aclanthology.org/2023.findings-emnlp.380`.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*, 2023c.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

A Appendix

A.1 Instruction Fine-Tune Dataset

We use various openly available instruction fine-tuning datasets related to Material science and general English question answering. We also construct a dataset for free-form question answering for material science questions by prompting GPT4 with a context and asking it to generate questions. We call this dataset MatBookQA (Material Science Book-based Question Answering dataset). We also introduce another question-answering dataset based on questions asked in the GATE examination in

India, which is taken by undergraduate students to apply for admissions in Masters and PhD programs in premier institutes in India and some foreign institutions of repute. The details of each dataset are provided as follows.

A.1.1 OPENORCA

This dataset comprises $800,000$ high-quality and diverse textual instructions. A model fine-tuned on this dataset may demonstrate enhanced performance in comprehending technical jargon, responding to complex queries, and producing coherent and contextually appropriate text across various domains. Previous research, as detailed in Mukherjee et al. (2023), has demonstrated that large language models (LLMs) fine-tuned on this dataset outperform other models on a range of benchmarks.

A.1.2 MATH

To induce the ability of mathematical problem-solving in our model, we train our model on the MATH dataset introduced by Hendrycks et al. (2021). It consists of 7500 instructions aimed at complex mathematical reasoning.

A.1.3 MATSCI

We utilize openly available instruction fine-tuning datasets for material science, complemented by a curated dataset generated through GPT-4(gpt-4-0613). By prompting GPT-4 with open-source material science textbooks, we elicit contextually complete questions covering various subdomains of material science. This diverse prompting ensures comprehensive coverage of the field.

We incorporate MatSciInstruct, as introduced in Song et al. (2023c). MatSciInstruct generates specialized instruction data through a two-step framework—Generation and Verification. In the Generation step, an instructor model creates domain-specific instruction data focused on materials science. The Verification step involves a separate verifier model for cross-verifying the instruction data for accuracy and relevance. Additionally, we employ the MatSciNLP training dataset and augment it with our MatBookQA dataset, as discussed below.

A.1.4 MATBOOKQA

We use an open-source book on Material Science and prompt GPT4 with one chapter at a time. We ask it to generate both short and long question-answer pairs for each chapter. We first curate a list of ten prompts each (see Appendix) to obtain short and long descriptions. This resulted in 2069 question-answer pairs, of which 1887 are short and 182 are long.

A.1.5 MASCQA

This dataset consists of 1036 and 549 questions from the civil and chemical engineering exams, respectively. The questions in this dataset can be divided into four types based on their structure: multiple-choice questions, matching-type questions, numerical answer questions with multiple choices, and numerical answer-based questions with no options. More details about the question structure can be found in Zaki et al. Zaki et al. (2024) An earlier version of MaScQA reported by Zaki et al.Zaki et al. (2024) also comprises 650 questions from the same materials science-related questions from the GATE exam. These questions come from various subdomains of materials science, like atomic structure, thermodynamics, electrical and magnetic behaviour of materials, materials manufacturing, applications, processing, and testing. Both these datasets cover vast subdomains of materials science, therefore serving as a challenging benchmark for evaluating the performance of large language models.

A.2 PRE-TRAINING LOSS CURVE

The loss curves indicate that performance on RedPajama, the original corpus of LLaMA-2, degrades over time. However, we were able to contain the degradation by our methods of combining the corpus.

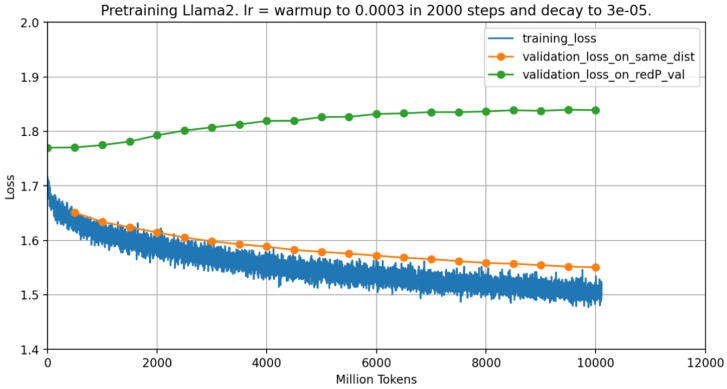

Figure 2: Pre-training and validation loss curve for LLaMat