Accelerating Material Discovery for Metal Organic Frameworks using Large Language Models

Sultan Alrowili
IBM Research
Riyadh, Saudi Arabia
sultan.alrowili@ibm.com

Mathan K. Eswaran
IBM Research
Riyadh, Saudi Arabia
Mathan.K.Eswaran@ibm.com

Shantanu Godbole
IBM Research
Abu Dhabi, UAE
shantanugodbole@in.ibm.com

Abstract

Recent advancements in Machine Learning (ML) have substantially accelerated the material discovery field, yet the utilization of Large Language Models (LLMs) in the Metal-Organic Frameworks (MOFs) research has received limited attention. This work leverages LLMs to build a new set of models that accelerate MOF material discovery. Our strategy relies on pre-training the Granite model using a single H100 GPU on a combination of selective chemical journals and structural data from the PubChem database. Our evaluation demonstrates that this pre-training strategy significantly enhances the performance of LLMs in predicting MOF properties, especially in limited-resource task scenarios. We hope this work can motivate future research to explore the potential of LLMs in enhancing material discovery to build robust and efficient Metal-Organic Frameworks models.

1 Introduction

Metal Organic Frameworks (MOFs) represent a class of porous materials formed through the coordination of metal ions or clusters referred to as secondary building units (SBUs) linked with organic ligands to form extended structures via coordination bonds [1] [2]. This material exhibits high porosity, thermal stability, and a wide range of applications in the field of catalysis [3], water treatment [4] [5] [6], and gas storage [7] [8] [9]. MOFs derive their structural flexibility from the combinations of metal nodes, organic linkers, and network topologies [8] [10] [11]. This tunability produces adjustable physicochemical properties and surface functionalities, offering massive potential to tailor materials for diverse applications. However, the ability to design MOFs with absolute precision and rational architecture is a challenging task.

Previous studies have employed the Crystal Graph Convolutional Neural Network (CGCNN) [12] to predict methane adsorption in MOFs, leveraging its architecture specifically considered for crystalline materials. While CGCNN offers the best performance in property prediction, it depends heavily on accurate 3D structural information, which is a prerequisite for the model input. This requirement poses a challenge, as many MOFs contain hundreds or even thousands of atoms, making their crystal graph representations computationally intensive and memory-inefficient, especially when scaling to large datasets. In addition, combining the 3D structure with the text representation of MOFs will increase the latency and decrease the efficiency of MOF-based models.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat-NeurIPS-2025.

High-throughput calculations and traditional density functional theory calculations have advanced MOF research, but at the same time, they face limitations due to their expensive computational time. Since MOFs are generally composed of metal nodes, organic linkers, and topologies, MOFid offers a compact string format that encodes both chemical and structural information [13]. It combines the Simplified Molecular Input Line Entry System (SMILES) representations [14] for the building blocks with topology and catenation codes from the Reticular Chemistry Structure Resource (RCSR) database [15].

This representation of MOF structure in text-based format (MOFid) enables the Machine Learning (ML) models, which mostly work with text data, to build more efficient MOF models that can predict MOF properties. This also became possible with the introduction of several large-scale MOFid-based databases that have been made publicly available, including CoRE MOF 2019 [16], hypothetical MOF (hMOF) [9], and QMOF [17] [18]. These resources offer detailed atomic structures of MOFs along with computed properties like CO adsorption capacities and electronic band gaps. This allows the research community to utilize the Transformer model [19] in building MOF models, as illustrated in the MOFormer model [20].

In recent years, Generative AI has rapidly evolved with the rise of Large Language Models (LLMs). Although Large Language Models (LLMs) show great success across diverse fields such as biomedical [21] and chemistry [22], the material science field, particularly in the domain of Metal-Organic Frameworks (MOFs), remains relatively understudied. This limitation comes from two challenges. Primarily, the structural complexity of materials like metal–organic frameworks (MOFs) makes it difficult to develop text-compatible input representations that accurately capture their intricate properties. Secondly, there are few available material-specific training data in the field of MOFs. Thus, most LLMs are overly dependent on a limited set of MOF datasets—namely CoRE MOF 2019, QMOF, and hMOF for pre-training. Addressing these challenges is crucial for unlocking the full potential of LLMs in accelerating MOF material discovery.

In this paper, we build a new set of small LLM models to accelerate material discovery in metal–organic frameworks (MOFs) by continually pre-training the Granite model on combinations of journal and chemical structured datasets. Our strategy relies on studying the overlooked impact of chemical corpora selection to enhance the performance of LLM on MOF tasks. The evaluation of our models shows that we outperform existing Transformer-based models on several MOF tasks. Moreover, our models show a comparable performance compared to existing multi-models that incorporate the 3D MOF structure representation. Furthermore, our study also shows the significant impact of our adapted strategy in enhancing the performance of LLMs in scenarios where we have a limited supervised finetuning SFT dataset in the MOF domain.

2 Proposed Model

To investigate the impact of continual pre-training of LLMs on the MOF domain, our method consists of two phases. In the first phase, we have continually pre-train the Granite 3.3 2B instruct model [1] on different chemical corpora setups. In the second phase, we have fine-tuned our pre-trained models on the training set of both QMOF and hMOF datasets in Supervised Finetuning (SFT) format.

2.1 PreTraining Phase

To address the limited MOF contextual representation in the Granite model (e.g, SMILES, MOFid), we performed continual pre-training on a variety of chemical corpora. These chemical corpora include the PubChem structural dataset [23] as well as MOF-related articles and abstracts sourced from the PMC Open Access Set [24]. Table 1 shows the different combinations of setups that we used in the pre-training phase. These diverse setups are designed to uncover the underexplored influence of corpora choice on enhancing the performance of LLMs on MOF-related tasks.

PMC Open Access Subset The PMC Open Access Subset includes 3.4 million journal articles and preprints, which have a more permissive license than the regular PMC articles. To select articles related to the MOF domain, we use a set of keywords such as mof, metal-organic framework, pore size, CO2 adsorption, CO2/N2 selectivity, CO2/H2O selectivity, CO2/CH4 selectivity, CH4 adsorption, and CO2 uptake. The final MOF subset consists of 28,525 articles and abstracts. To prepare these

[1] `https://huggingface.co/ibm-granite/granite-3.3-2b-instruct`

Table 1: Details of the chemical corpora setups used in the continual pre-training stage. The size of the corpora is measured in bytes rather than token count, as the general LLMs tokenizer is not trained on SMILES representation, which will inflate token count for the PubChem corpora compared to the Open Access PMC corpora.

Setup	Pre-Training Corpora Setup	Size
1	PubChem (500K)	254M
2	Open Access PMC (MOF) + PubChem (500K)	1.13GB
3	Open Access PMC (MOF) + PubChem (2M)	1.87GB

articles for the pre-training phase, we use the NLTK sentence tokenizer [25] to split each sentence into a new line.

PubChem PubChem is the world's largest collection of freely accessible chemical information. It includes more than 122M chemical compounds that each have information such as IUPAC name, SMILES representation, molecular formula, and chemical properties. Our hypothesis assumes that including the PubChem dataset, which consists of MOF-related properties, such as SMILES name, could help improve the performance on MOF-related tasks. We prepare the PubChem dataset by appending related properties, such as SMILES structure, chemical formula, and molecular weight, to each compound name in one line, as this will help add context to each property.

2.2 FineTuning Phase

We follow the continual pre-training phase with a supervised fine-tuning (SFT) phase, which uses the training set of QMOF and hMOF datasets in the prompt and output style. The prompt in this case is represented by the MOFid, where the output is the targeted property.

hMOF: The hypothetical MOFs (hMOF) is a dataset that consists of 137,652 MOFs (102,858 with MOFid), which capture the gas adsorption properties of CO2 and CH4 in mol kg1 at 0.05, 0.5, and 2.5 bar of pressure. The hMOF consists of 72,000, 15,428, and 15,428 samples in train, validation, and test splits, respectively [20]. Each sample has a MOFid, and 6 properties including: CO2 and CH4 adsorption at 0.05, 0.5, and 2.5 bar of pressure. We prepare our hMOF SFT dataset by having the MOFid in the prompt and all six properties in the target sequence (output), where the title and value for each property are placed on a new line.

QMOF: The QMOF data set contains quantum-chemical properties for metal–organic frameworks (MOFs), which contains 20,375 (7,466 with MOFid) along with the label of a density functional-based tight-binding (DFTB) [26] [27], which calculates the band gap in Electron Volts (eV). The QMOF dataset consists of 5,226, 1,119, and 1,119 samples in train, validation, and test splits, respectively [20]. In alignment with the hMOF dataset configuration, we set the MOFid as our prompt and the band gap value as our targeted output.

2.3 Baseline Models and Evaluation Metrics

Our baseline models include the following state-of-the-art MOF models: Crystal Graph Convolutional Neural Network (CGCNN) [12], Smooth Overlap of Atomic Positions (SOAP) [28] [29] [30], and MOFormer [20]. These three models are the top-performing models on both QMOF and hMOF tasks as reported by [20]. In addition, to study the impact of our pre-training strategies where other design factors (e.g, architecture, base-scale) are fixed in the experimental setup, we evaluate our pre-trained models against Granite 3.3 2B, the foundation model that we use during the pre-training phase.

For the evaluation metrics, we adopted the same standards in the literature [20] by using the mean absolute error (MAE) to evaluate both QMOF and hMOF. We use the validation set of QMOF and hMOF to find the best hyperparameters (e.g., batch size, learning rate) during the fine-tuning phase. Then, we adopted these hyperparameters to report our results on the test set. Table 3 shows more details about our hyperparameter choices for both the pre-training and fine-tuning phases.

Table 2: The Mean Absolute Error (MAE) results of our pre-trained models against our baseline models on the test set of QMOF (eV) and hMOF (mol/kg) datasets. We use the reported results by [20] for our baseline models. PubChem: 500K chemical compounds, PubChem+: 2M chemical compounds.

		QMOF	CO2 bar (hMOF)			CH4 bar (hMOF)		
Model	SFT	eV	0.05	0.5	2.5	0.05	0.5	2.5
CGCNN	Full	**0.256**	**0.110**	**0.330**	0.645	0.025	**0.099**	0.258
SOAP	Full	0.424	<u>0.115</u>	<u>0.339</u>	0.666	<u>0.022</u>	0.106	**0.239**
MOFormer	Full	0.367	0.158	0.545	0.982	0.033	0.161	0.384
Granite 3.3 2B Instruct	Full	0.338	0.119	0.366	0.640	**0.019**	0.102	<u>0.243</u>
- PMC + PubChem	Full	<u>0.308</u>	0.120	0.366	**0.635**	**0.019**	0.102	0.244
- PMC + PubChem+	Full	0.314	0.127	0.401	0.694	0.035	0.174	0.272
- PubChem	Full	0.329	0.119	0.364	<u>0.638</u>	**0.019**	<u>0.100</u>	<u>0.243</u>
Granite 3.3 2B	1K	0.513	0.285	1.145	2.397	0.043	0.283	0.778
- PMC + PubChem	1K	0.430	0.158	0.494	0.900	0.025	0.137	0.333
Granite 3.3 2B	3K	0.396	0.199	0.616	1.254	0.029	0.162	0.410
- PMC + PubChem	3K	0.348	0.150	0.465	0.850	0.024	0.130	0.315
Granite 3.3 2B	10K	-	0.156	0.480	0.878	0.024	0.130	0.317
- PMC + PubChem	10K	-	0.142	0.484	0.784	0.023	0.146	0.293

3 Results and Discussions

The first section of Table 2 shows the results of our models against our baseline models on QMOF and hMOF tasks. In the following sections of the table, we show the evaluation of our pre-trained models against the Granite model, where we use different subsets of the SFT training set (e.g, 1K, 3K, 10K) instead of using the full training set of QMOF and hMOF. This setup will help us to understand the impact of the pre-training phase in scenarios where the SFT dataset is limited in size.

As shown in the first section of Table 2, our model (PMC + PubChem), which was continually pre-trained on a collection of MOF-related articles from the PMC Open Access set and 500k chemical compounds from PubChem data, achieves superior performance on the QMOF task, outperforming all the text-based models, including Granite, MOFormer, and SOAP. Our model (PMC + PubChem) also outperforms the CGCNN model on several hMOF tasks, even though the CGCNN incorporates the 3D MOF structure. The results also show that the gap in performance between our model (PMC + PubChem) and other text-based models is larger on the QMOF than the hMOF task. This performance gap is because the QMOF dataset consists of only 5.2K samples in the train set against 72K samples for hMOF. These findings highlight that our continual pre-training strategies are particularly effective in scenarios where the SFT dataset is limited in scale.

Our comparative analysis of the three pre-trained models reveals that the combination of PMC articles and the PubChem dataset (500K) serves as the most effective pre-training setup, consistently outperforming the alternatives. Using the PubChem dataset alone decreases performance significantly on QMOF. Additionally, our results demonstrate that scaling the PubChem dataset to 2M chemical compounds (PubChem+) leads to a noticeable decline in performance on the QMOF task. The decline in performance is attributed to the fact that adding more structured datasets, such as PubChem, to the pre-training corpora decreased the impact of the Open PMC dataset on the contextual representation, as it has a lower ratio in the pre-training dataset. This conclusion is further validated by the pre-training results obtained when using the PubChem dataset alone. Moreover, our results indicate that changes in pre-training corpora have minimal influence on the hMOF task. These results show that in scenarios where we have a large SFT dataset, the continual pre-training stage may not be needed. However, in scenarios where the SFT dataset is limited in size, as in the case of QMOF, pre-training LLMs on a combination of PMC and PubChem datasets helps address this limitation in the SFT dataset.

To further evaluate this hypothesis, Table 2 presents a comparative analysis of our pretrained models and the Granite model across SFT training subsets of 1K, 3K, and 10K samples, instead of using the full QMOF (5.2K) and hMOF (72K) dataset. The results with this evaluation setup confirm our early

hypothesis, which we concluded on the QMOF task. The results show a significant gap in margin between Granite and our pre-trained models across all tasks. However, this margin decreases when we use 10K examples from the hMOF training set. These results suggest that having a moderate <5K samples in the training set of SFT tasks could be better to evaluate the MOF contextual representation in LLMs. These findings could influence the research community's decision when building future benchmarks for the material discovery field.

4 Conclusion

This study presents a targeted strategy for constructing small-scale language models for MOF applications, emphasizing the role of curated pre-training datasets. The results show that we outperform existing state-of-the-art MOF frameworks across multiple tasks in both the QMOF and hMOF benchmarks, which shows the potential of LLMs in addressing the MOF field. In addition, focusing more on enhancing the quality of the pre-training dataset could eventually increase the generalization of the LLMs across various tasks in the material discovery and MOF fields, especially in cases where the SFT dataset is limited. For future work, we are planning to increase the scale of our pre-trained models and investigate the impact of including the 3D structure of MOFs with Multi-Model LLMs.

References

[1] Stuart James. Metal-organic frameworks. *Chemical Society reviews*, 32:276–88, 10 2003.

[2] Hong-Cai Zhou, Jeffrey Long, and Omar Yaghi. Introduction to metal-organic frameworks. *Chemical reviews*, 112:673–4, 02 2012.

[3] Anastasiya Bavykina, Nikita Kolobov, Il Son Khan, Jeremy Bau, Adrian Ramirez, and Jorge Gascon. Metal–organic frameworks in heterogeneous catalysis: Recent progress, new trends, and future perspectives. *Chemical Reviews*, 120, 03 2020.

[4] Husam Almassad, Rada Abaza, Lama Siwwan, Bassem Al-maythalony, and Kyle Cordova. Environmentally adaptive mof-based device enables continuous self-optimizing atmospheric water harvesting. *Nature Communications*, 13:4873, 08 2022.

[5] Zhonglin Cao, Vincent Liu, and Amir Barati Farimani. Water desalination with two-dimensional metal-organic framework membranes. *Nano Letters*, 2019, 10 2019.

[6] Nikita Hanikel, Mathieu S. Prévot, and Omar M. Yaghi. Mof water harvesters. *Nature Nanotechnology*, 15(5):348–355, May 2020.

[7] Alauddin Ahmed, Saona Seth, Justin Purewal, Antek Wong-Foy, Michael Veenstra, Adam Matzger, and Donald Siegel. Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nature Communications*, 10, 04 2019.

[8] Peter Boyd, Arunraj Chidambaram, Enrique García Diez, Christopher Ireland, Thomas Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Moosavi, Mercedes Maroto-Valer, Jeffrey Reimer, Jorge Navarro, Tom Woo, Susana García, Kyriakos Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas co2 capture. *Nature*, 576:253–256, 12 2019.

[9] Christopher Wilmer, Michael Leaf, Chang Yeon Lee, Omar Farha, Brad Hauser, Joseph Hupp, and Randall Snurr. Large-scale screening of hypothetical metal-organic frameworks. *Nature chemistry*, 4:83–9, 02 2012.

[10] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, August 2020.

[11] Conor Sharp, Brandon Bukowski, Hongyu Li, Eric Johnson, Stefan Ilic, Amanda Morris, Dilip Gersappe, Randall Snurr, and John Morris. Nanoconfinement and mass transport in metal–organic frameworks. *Chemical Society Reviews*, 50, 10 2021.

[12] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018.

[13] Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*, 19(11):6682–6697, 2019.

[14] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[15] Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The reticular chemistry structure resource (rcsr) database of, and symbols for, crystal nets. *Accounts of Chemical Research*, 41(12):1782–1789, 2008. PMID: 18834152.

[16] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, 2019.

[17] Andrew Rosen, Shaelyn Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin Notestein, and Randall Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4, 04 2021.

[18] Andrew Rosen, Victor Fung, Patrick Huck, Cody O'Donnell, Matthew Horton, Donald Truhlar, Kristin Persson, Justin Notestein, and Randall Snurr. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *npj Computational Materials*, 8:112, 05 2022.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: Self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023. PMID: 36706365.

[21] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023.

[22] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. Chemllm: A chemical large language model, 2024.

[23] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 11 2024.

[24] PMC Open Access Subset. `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`, 2025 Aug 18.

[25] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[26] Pieremanuele Canepa, Calvin A. Arter, Eliot M. Conwill, Daniel H. Johnson, Brian A. Shoemaker, Karim Z. Soliman, and Timo Thonhauser. High-throughput screening of small-molecule adsorption in mof. *J. Mater. Chem. A*, 1:13597–13604, 2013.

[27] Andrew S. Rosen, Justin M. Notestein, and Randall Q. Snurr. Identifying promising metal–organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. *Journal of Computational Chemistry*, 40(12):1305–1318, 2019.

[28] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[29] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12):e1701816, 2017.

[30] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[32] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

[33] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[34] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Table 3: Hyperparameter choices for both pre-training and Supervised Finetuning SFT phases. Packing refers to the feature where multiple sequences are packed in the same batch, which improves the efficiency of the training process.

Setting	Pre-Training Setup	Supervised Finetuning Setup
Batch Size	1	16
Learning Rate	5e-5	3e-5
Max Sequance Length	4096	512
Epoch	1	4
Packing	True	False
Neftune Noise Alpha	5	5
Gradient Checkpointing	True	True

A PreTraining and FineTuning Hyperparameters

The hyperparameter setup for pre-training and fine-tuning is detailed in Table 3. Our environmental setup uses a single H100 80GB GPU for both pre-training and fine-tuning, where we use the Transformers [31], Transformer Reinforcement Learning (TRL) [32], and Flash attention2 [33] [34] libraries. During pre-training, although we maintained a batch size of 1, we extended the maximum sequence length to 4096. This was achieved with the packing feature enabled, a choice made to ensure sequences are fully occupied rather than padded with tokens. In contrast, for the Supervised Fine-Tuning (SFT) phase, the packing feature was disabled. This adjustment was complemented by increasing the batch size to 16 and reducing the maximum sequence length to 512. This strategy was employed considering that the vast majority of instances in our QMOF and HMOF datasets have sequence lengths below 512. We made this choice of disabling the packing feature as we find that adapting this approach improves the performance and stability in the SFT phase.