Knowlab's Submission to L+M Shared Task: All you need is continued pretraining of chemistry texts even for molecule captioning

Yunsoo Kim
University College London
yunsoo.kim.23@ucl.ac.uk

Honghan Wu
University College London
honghan.wu@ucl.ac.uk

Abstract

This paper presents our submission to the L+M-24 shared task, focused on translating molecular structures into natural language descriptions, known as the molecule captioning task. We selected a small language model (SLM), Phi-3-mini-4k, to evaluate the impact of continued pretraining and instruction tuning for domain-specific chemical knowledge. The Phi-3 model was continued pretrained with 90M chemistry textbooks and abstracts, followed by instruction tuning on 150K question answering sets of SMILES and general chemistry knowledge. Despite the continued pretraining phase not including direct exposure to SMILES representations, it significantly enhanced the Phi-3 model's performance, a 300% increase for the BLEU scores, in the molecule captioning task. The code and model are released at `https://github.com/bluesky333/Phi3KnowChem` to facilitate research in chemical small language modeling.

1 Introduction

The intersection of natural language processing (NLP) and chemistry began with drug discovery and biochemistry but recently moved to the other fields of chemistry such as electrochemistry for battery and rheology for chemical property prediction (Krallinger et al., 2015; Li et al., 2016; Huang and Cole, 2022; Kim et al., 2023). With the recent advancement of large language models (LLMs), the language model for chemistry domain knowledge started to cover molecule representation such as the simplified molecular-input line-entry system (SMILES) and 3D structure of molecules (Edwards et al., 2022; Taylor et al., 2022; Fang et al., 2023; Zhang et al., 2024a,b). As the research in LLMs has been facilitated by the benchmark datasets for evaluating the model's understanding of domain knowledge, there has been a pressing need for a benchmark specifically for molecule and language models (Hendrycks et al., 2020).

To address this gap, the L+M-24 shared task was introduced as one of the first competitions focused on translating between language and molecule representations (Edwards et al., 2024). This task involves generating captions based on input molecules represented in SMILES, pushing the boundaries of molecule captioning by leveraging language models. The task covers four key applications within chemical knowledge: biochemistry, electrochemistry, organoleptics, and agricultural chemistry. Progress in these specific is essential for building foundational models applicable to small molecule applications.

Traditionally, models designed for such tasks require extensive domain-specific pretraining and fine-tuning with molecule representation to understand and generate chemistry-related text effectively (Edwards et al., 2022; Taylor et al., 2022; Fang et al., 2023; Zhang et al., 2024a). This process is often resource-intensive and requires large, specialized datasets.

In this work, we explore the efficacy of continued pretraining and instruction tuning on a small language model (SLM), specifically the 3.8B parameter model, Phi-3-mini-4k, for the molecule captioning task.

Our approach involves two primary stages:

1. Continued Pretraining: We further pretrain the Phi-3-mini-4k model using a corpus of 90 million chemistry textbooks and abstracts. This step aimed to infuse the model with a broad and deep understanding of chemical language and concepts.

2. Instruction Tuning: We further refined the model with 150,000 instruction tuning tasks focused on SMILES question answering and general chemistry knowledge question answering. This step was designed to enhance the model's ability to handle SMILES representation and chemical queries.

Proceedings of the 1st Workshop on Language + Molecules (L+M 2024), pages 92–97

The resulting models underwent fine-tuning with the shared task's training data for 1 epoch, and our best-performing model surpassed the performance of the MolT5-base model, which was trained with 100 million SMILES strings (Edwards et al., 2022).

The contributions of this paper are as follows:

1. **Pretraining without molecule still helps.** Our study demonstrates that continued pre-training using a chemical text corpus significantly enhances performance in molecule captioning tasks, even without direct exposure to molecular representations. We saw an almost 300% increase for BLEU scores, about a 67% increase for the ROUGE score, and a 19% increase for Meteor with the continued pretraining.

2. **Phi-3KnowChem model.** We introduce the Phi-3KnowChem model, a small language model (SLM) based on the Phi-3 architecture, pretrained with a chemical corpus and instruction-tuning datasets. To the best of our knowledge, this is the first Phi-3 model specifically trained for the chemical domain.

3. **Open source.** To foster research in chemical small language modeling, we will release the model weights and codes to prompote reproducibility and collaboration in the field.

2 Methods

2.1 Language Model

Phi-3 (Abdin et al., 2024) We use Phi-3-mini-4K model, which we refer to as Phi-3 model throughout our paper. Phi-3 has 3.8B parameters and is trained on an augmented textbook corpus and high-quality web data consisting of 3.3 trillion tokens. While specific training details are not disclosed, Phi-3 building blocks are reported to be a similar structure to the Llama-2 model (Touvron et al., 2023). Phi-3 showed an outstanding performance in MMLU which includes high school chemistry and college chemistry subjects (Hendrycks et al., 2020).

2.2 Pretrain Data

The pretraining corpus comprised 8 million tokens sourced from chemistry textbooks and an additional 82 million tokens extracted from chemical journal abstracts published by the Royal Society of Chemistry (Chen et al., 2020). The textbook data was acquired from the HuggingFace repository [1]. This diverse corpus provides a rich source of chemical language and concepts, enabling the model to develop a comprehensive understanding of the domain. Examples of the pretraining corpus are provided in Table 1.

Textbook
To discuss the electronic states of atoms we need a system of notation for multi-electron wavefunctions. As we saw in Chapter 8, the assignment of electrons to orbitals is called the electron configuration of the atom. One creates an electronic configuration representing the electronic structure of a multi-electron atom or ion in its ground or lowest-energy state as follows.
RSC Abstract
Rhenium, the non - noble metal with an acceptable price, was found to be a good additive that largely improved Pt / WO3 / ZrO2 catalysis for glycerol hydrogenolysis. Compared with conventionally employed Pt / WO3 / ZrO2, the Re - promoted catalyst led to almost quantitative glycerol conversion (> 99% vs. 57.7%), giving useful C3 alcohols in excellent total selectivity (> 95%) under reduced reaction pressure (2.5 MPa). The addition of Re led to such an impressive enhancement of the catalyst activity that even the reaction performed under atmospheric H2 pressure (0.1 MPa) afforded 96.8% glycerol conversion and a good selectivity of C3 compounds at 95.2%. Further XRD, Raman, BET, CO chemisorption, TEM, H2-TPR, XPS, NH3-TPD, 1H MAS NMR and Py-IR studies indicated that introduction of Re greatly improved the dispersion of Pt and catalyst acidity, and resulted in this largely enhanced catalyst activity.

Table 1: Examples of Pretrain Corpus.

2.3 Instruction Tuning Data

The instruction tuning phase utilized two primary datasets:

1. **SMILES Question Answering Pairs:** Consisting of 129,699 question-answer pairs from ChEMBL, this dataset contains 3,892 unique

[1] Hack90/libre_chem_textbooks

SMILES representations. The dataset was acquired from the HuggingFace repository. [2] Each pair involves a question related to the chemical properties of the molecule, along with the corresponding answer.

2. **CAMEL-AI Chemistry Dataset:** Chemistry instruction tuning dataset which was generated using GPT-4 was used to cover a broad range of chemical queries and tasks (Li et al., 2024). The dataset is made up of 20,000 instruction pairs in 25 topics and 25 subtopics for each topic. By incorporating this diverse dataset into the instruction tuning process, we aimed to further refine the model's ability to comprehend and respond to various chemical queries beyond SMILES representation.

The examples of the instruction tuning data are shown in Table 2.

SMILES QA
SMILES: CC(=O)/N=c1/sc(S(N)(=O)=O)nn1C Question: How many rotatable bonds does this compound have? Answer: 1
CAMEL-AI Chemistry QA
What are the expected products of the reaction between sodium chloride (NaCl) and silver nitrate (AgNO3)? Provide the balanced chemical equation and identify the type of reaction that occurs. Answer: When sodium chloride (NaCl) reacts with silver nitrate (AgNO3), a double displacement reaction occurs, also known as a metathesis reaction. The expected products of this reaction are silver chloride (AgCl) and sodium nitrate (NaNO3). The balanced chemical equation for this reaction is: NaCl (aq) + AgNO3 (aq) $\rightarrow$ AgCl (s) + NaNO3 (aq) In this reaction, the sodium (Na) and silver (Ag) ions switch places, forming new compounds. Silver chloride (AgCl) is an insoluble solid that precipitates out of the solution, while sodium nitrate (NaNO3) remains dissolved in the solution.

Table 2: Examples of instruction tuning dataset.

[2] alxfgh/ChEMBL_Drug_Instruction_Tuning

3 Experiment and Results

3.1 SLM Training

We conducted continued pretraining and instruction tuning using 8 A6000 GPUs, each with 48 GB GPU memory. We used LLaMaFactory[3] and used Deep3 for efficient training (Rasley et al., 2020). Continued pretraining was done with an epoch of 3 and a learning rate of 1e-5. The instruction tuning was done with the same learning rate as the continued pretraining but only for 1 epoch. The total training took less than a day. We named the resulting model as Phi-3-KnowChem.

For the shared task fine-tuning, we used a different computational resource, 2 A5000 GPUs with a total GPU memory of 48 GB. The finetuning for the shared task was done using low-rank adaptation (LoRA) and deepSpeed zero redundancy optimizer to reduce the GPU memory requirement (Rasley et al., 2020; Hu et al., 2021). We trained the model for the captioning task with a learning rate of 1e-3 and epoch 1. Hyperparameters for LoRA were as follows: rank - 128, alpha 256, and projector learning rate 2e-5. We used a simple prompt for the training and the evaluation: 'Describe the input molecule represented in SMILES. SMILES string'. The whole train dataset for the task was used for the fine-tuning.

3.2 Evaluation

For the evaluation, we used 1 A5000 GPU and used temperature of 1. For the baseline performance, we used the greedy search as the decoding strategy. We change this strategy with the beam search using multinomial sampling. The number of beams was 3. For the evaluation metrics scoring, we submitted the text output to the codabench. The evaluation metrics used include BLEU-2, BLEU-4, Meteor, ROUGE-1, ROUGE-2, and ROUGE-L.

3.3 Evaluation Results

The results in Table 3 for the chemical language training provide several notable trends. Firstly, the baseline Phi-3 model perform very poorly on the task even after 1 epoch of finetuning. This suggests that the model has a limited understanding of chemistry-specific language, as indicated by lowest scores across all metrics (Meteor: 0.5582, ROUGE-1: 0.4225, ROUGE-2: 0.3186, ROUGE-L: 0.3016). These baseline scores highlight the challenges of molecule captioning when using a general language

[3] https://github.com/hiyouga/LLaMA-Factory

Model	BLEU-2	BLEU-4	Meteor	ROUGE-1	ROUGE-2	ROUGE-L
Phi-3	0.2299	0.1678	0.5582	0.4225	0.3186	0.3016
Phi-3+PT	0.6804	0.4922	0.6644	0.7074	0.5308	0.5142
Phi-3+PT+IT	0.6959	0.5023	0.6750	0.7232	0.5373	0.5193
Phi-3+PT+IT+Beam	0.7056	0.5083	0.6901	0.7261	0.5361	0.5201

Table 3: Molecule Captioning Evaluation Results. The result is sorted by the ROUGE-L. PT stands for continued pretrained model. IT stands for instruction-tuned model. Beam stands for the beam search decoding strategy.

Prompt
Describe the input molecule represented in SMILES. CSc1ccc(C(C)N(C)C)cn1
Phi-3
The molecule is a anti viral. The molecule is a hiv inhibitor and is anti viral compound. The molecule is a hiv integrase inhibitor and is anti viral agent. The molecule is a hiv replication inhibitor and is anti viral agent compound. The molecule is a hiv protease inhibitor and is anti viral agent, belonging to the anti viral class of molecules. The molecule is a hiv treatment and is anti viral compound, anti viral activity, and anti viral. The molecule is a hiv integrase inhibitor and is anti viral compound, anti viral activity, and anti viral. The molecule is a hiv inhibitor and aids treatment, belonging to the anti viral compound class of molecules. The molecule is a hiv treatment and aids treatment, impacting both hiv treatment and aids. The molecule is a hiv inhibitor and aids treatment, impacting both hiv replication and cancer
Phi-3+PT
The molecule is a anti inflammatory.
Phi-3+PT+IT
The molecule is a anti inflammatory.
Phi-3+PT+IT+Beam
The molecule is a anti viral and belongs to the anti viral agent class of molecules, with the characteristic of being anti viral.

Table 4: Response comparison between baseline Phi-3 and trained models. The example was randomly sampled.

model that lacks specialized training in the chemical domain.

Additionally, the effect of continued training is observed. Continued pretraining (PT) on a large corpus of chemistry texts resulted in significant improvements across all evaluation metrics. The continued pretraining effectively infused the model with domain-specific knowledge, enhancing its performance by a substantial margin even without exposure to SMILES string. The performance boost with the continued pretraining was almost 300% increase for BLEU scores, about 67% increase for the ROUGE score, and 19% increase for Meteor. Adding an instruction tuning (IT) phase which contained about 4K unique SMILES representations further improved the model's performance in the molecule captioning task for all the evaluation metrics. This shows that prior exposure to SMILES representation can improve the performance of the related downstream task.

Beam search with multinomial sampling also increased the performance in all the evaluation metrics except ROUGE-2. This suggests that while continued pretraining and instruction tuning lay a strong foundation for chemical language understanding, advanced decoding techniques like beam search can further refine the output quality.

The model responses were compared as shown in Table 4. Rather than giving the right caption for the molecule, the Phi-3 model repeated sentences describing an anti-viral agent. Also, we see grammatical mistakes, using ‘a’ instead of ‘an’. The continued pretrained model and instruction-tuned model both gave the same response, while the beam search strategy yielded a different response that was slightly longer than the other two models.

4 Conclusion

In this paper, we have presented our approach and findings from experimenting with the Phi-3-mini-4k model on the molecule captioning task as part of the L+M-24 shared task. Our experiment focused on evaluating the efficacy of continued pretraining and instruction tuning for enhancing the model's domain-specific chemical knowledge and its ability to generate accurate molecular descriptions.

The results demonstrated that both continued pretraining and instruction tuning play critical roles in enhancing the performance of the Phi-3 model on molecule captioning tasks. Continued pretraining with a chemistry-specific corpus provides a substantial boost by enriching the model's knowledge base, while instruction tuning with targeted question-answer pairs refines its ability to handle specific queries related to chemical structures. The addition of beam search decoding, though providing marginal gains, contributes to producing higher-quality and more accurate descriptions. These findings highlight the importance of domain-specific training and advanced decoding strategies in improving the capabilities of language models for specialized tasks like molecule captioning. In fact, on the leaderboard, our Phi-3-KnowChem outperformed the MolT5-base model.

Nevertheless, while these findings provide valuable insights, further in-depth analysis is warranted to explore the nuances of model performance in the chemical domain fully. The exploration of other tasks such as chemical property prediction can contribute to more accurate and comprehensive assessments of LM performance in real-world chemical applications.

Limitation

The computational constraints restricted the size and complexity of the models that could be feasibly trained and evaluated. There are larger versions of the Phi-3 model, 7B, and 14B models which can potentially perform much better than the version we used in this study.

Broader Impacts and Ethics Statement

Our work does not raise any major ethical concerns regarding the usage of the Phi-3 model as it was used for research purposes only. However, our Phi-3-KnowChem is not rigorously tested for use in real-world chemical applications or scenarios. Thus, they may not be suitable for use in the decision-making process for the chemical industry.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Yifan Chen, Tamara Polajnar, Colin Batchelor, and Simone Teufel. 2020. A corpus of very short scientific summaries. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 153–164.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+m-24: Building a dataset for language+ molecules@ acl 2024. *arXiv preprint arXiv:2403.00791*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shu Huang and Jacqueline M Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *Journal of chemical information and modeling*, 62(24):6365–6377.

Yunsoo Kim, Hyuk Ko, Jane Lee, Hyun Young Heo, Jinyoung Yang, Sungsoo Lee, and Kyu-hwang Lee. 2023. Chemical language understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 404–411.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7:1–17.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

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

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

Kaiwei Zhang, Yange Lin, Guangcheng Wu, Yuxiang Ren, Xuecang Zhang, Xiaoyu Zhang, Weitao Du, et al. 2024b. Sculpting molecules in 3d: A flexible substructure aware framework for text-oriented molecular optimization. *arXiv preprint arXiv:2403.03425*.