# The Tokenization Bottleneck: How Vocabulary Extension Improves Chemistry Representation Learning in Pretrained Language Models

**Prathamesh Kalamkar**[*†], **Ned Letcher**[*], **Meissane Chami**[*], **Sahger Lad**[*],
**Shayan Mohanty**, **Prasanna Pendse**
Thoughtworks

## Abstract

The application of large language models (LLMs) to chemistry is frequently hampered by a "tokenization bottleneck", where tokenizers tuned on general-domain text tend to fragment chemical representations—such as SMILES—into semantically uninformative sub-tokens. This paper introduces a principled methodology to resolve this bottleneck by unifying the representation of natural language and molecular structures within a single model. Our approach involves targeted vocabulary extension—augmenting a pretrained LLM's vocabulary with chemically salient tokens, followed by continued pretraining on chemistry-domain text to integrate this new knowledge. We provide an empirical demonstration of the effectiveness of this strategy, showing that our methodology leads to superior performance on a range of downstream chemical tasks.

## 1 Introduction

Despite being powerful, LLMs typically lack the specialized domain-specific knowledge required to excel in scientific fields such as chemistry. The unique vocabulary, syntax, and concepts of chemistry necessitate a dedicated adaptation process to transform these generalist models into domain specialists. Continued pretraining (CPT) [Gururangan et al., 2020] involves continuing the self-supervised pretraining objective of a base model on a new, domain-specific corpus, allowing it to acquire specialized knowledge while retaining its pre-existing capabilities. A significant challenge in continued pretraining is the phenomenon of catastrophic forgetting [French, 1999], where a model's performance on general-domain tasks degrades as it continues learning from specialized-domain data.

In this paper, we introduce a principled methodology to jointly model molecular structure (SMILES) and general-purpose text by applying continued pretraining together with vocabulary extension. We also provide an empirical demonstration of its effectiveness, showing that this methodology leads to more efficient learning and superior performance on downstream tasks in the chemistry domain.

## 2 Related Work

The primary methods for representing molecules are string-based notations such as the Simplified Molecular Input Line Entry System (SMILES) or the more robust but less common SELFIES (SELF-referencing embedded strings). While LLMs can handle molecular text and general text [Sadeghi et al., 2024], a core challenge is that standard NLP tokenizers are ill-suited for capturing the formal grammar of these representations, leading to the semantic fragmentation that characterizes the tokenization bottleneck.

[*] Authors contributed equally
[†] correspondence: prathamk@thoughtworks.com

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

A dominant architectural paradigm addresses this challenge by treating text and molecular data as two separate modalities, enabling the use of distinct tokenizers, with approaches such as SMILES Pair Encoding handling molecular representations [Li and Fourches, 2021]. This approach uses a dual-encoder architecture where one encoder processes textual descriptions and another processes molecular structures. The representations from these two encoders are then aligned into a shared embedding space using techniques like contrastive learning. Prominent examples include CLAMP [Seidl et al., 2023], MolFM [Luo et al., 2023], and PubChemSTM [Liu et al., 2023]. This approach, however, is computationally expensive and requires large, high-quality paired datasets. Another strategy is instruction tuning, where a pretrained model is fine-tuned on curated chemical-instruction datasets, such as SMolInstruct [Yu et al., 2024] or Mol-Instructions [Fang et al., 2024]. This injects task-specific knowledge after pretraining, but it does not fix the underlying representational flaws learned during the initial pretraining stages. Some models, such as ChemDFM [Zhao et al., 2024], have also attempted to integrate molecular knowledge directly into a unified text-based LLM during pretraining.

Vocabulary extension during CPT [Tai et al., 2020] is a domain-adaptation technique in which the vocabulary of a pretrained model is augmented with new specialized tokens before continued pretraining is performed, enabling the model to learn more semantically robust representations for domain-specific concepts. More direct, efficient, and foundational strategies for targeted vocabulary extension within existing LLM frameworks remain comparatively underexplored.

# 3 Methodology

Our methodology is predicated on treating SMILES not as a distinct modality requiring a separate encoder but as a structured chemical notation that can be unified with natural language within a single representational space. This approach is motivated by the objective of creating a single cohesive model that learns a joint representation of molecular structures and natural text.

We propose a novel approach for adapting a pretrained LLM to the chemistry domain by extending its vocabulary to include chemically significant substructures before performing continued pretraining against a dataset composed of both chemistry-domain data and general-domain data. The new tokens added to the vocabulary are derived from a data-driven analysis of all SMILES strings occurring in the pretraining data, as described in Section 3.2.

## 3.1 Pretraining Data

To enable continued pretraining, we curated a data blend from the following sources.

1. **USPTO**: We selected chemistry-related US patents from the USPTO dataset. Using the PatentChem library [Subramanian et al., 2023], XML files were parsed into plain text, and the corresponding SMILES strings were inserted at their appropriate locations. To help the model distinguish between text and molecular modalities, all SMILES strings were enclosed within `<SMILES>` and `</SMILES>` tags.

2. **Chemistry Papers**: A high-quality corpus of English-language academic papers focused on chemistry was compiled from the Semantic Scholar Open Research Corpus (S2ORC) [Lo et al., 2020].

3. **SMolInstruct**: This is a large-scale instruction-tuning dataset for chemistry, centered on small molecules [Yu et al., 2024]. It features 14 distinct tasks and over three million curated samples. Given the short length of individual instruction samples, we concatenated multiple samples, separating them with `<EOS>` tokens to help the model recognize record boundaries.

4. **L+M-24**: This dataset provides molecule-text pairs designed to train and evaluate a model's understanding of key natural language concepts in molecule design, such as compositionality, functionality, and abstraction [Edwards et al., 2024].

5. **FineWeb**: To mitigate catastrophic forgetting of general knowledge—a common issue in domain adaptation—we incorporated a 10B token sample of the FineWeb dataset for data replay [Penedo et al., 2024].

The final data blend was constructed by assigning weights to each dataset, as detailed in Table 1. Higher weights were assigned to high-quality, domain-specific sources, including the chemistry papers, SMolInstruct, and L+M-24, to prioritize the learning of specialized chemical knowledge.

Table 1: Pretraining data sizes and weights

| **Data Source** | **Data %** | **Data Tokens (Bn)** | **Training Data %** | **Training Tokens (Bn)** |
|---|---|---|---|---|
| FineWeb | 61% | 10.00 | 50% | 8.18 |
| USPTO | 31% | 5.00 | 35% | 5.73 |
| L+M-24 & SMolInstruct | 3% | 0.43 | 10% | 1.64 |
| Chemistry Papers | 6% | 0.94 | 5% | 0.82 |
| Total | 100% | 16.37 | 100% | 16.37 |

### 3.2 Vocabulary Extension

A core component of our methodology is the extension of the base LLM's vocabulary to incorporate important tokens for the chemistry domain. The additional tokens are derived through two distinct processes that correspond to the two modalities of text and molecular data.

1. **Text modality**: The top 1,000 most frequently occurring out-of-vocabulary tokens produced by the Llama 3 tokenizer after tokenizing the blended dataset, with SMILES strings excluded.
2. **Molecular modality**: Commonly occurring SMILES substructures extracted using the process described below, resulting in 16,795 new tokens.

Our process for extracting new molecular tokens is based on the SMILES Pair Encoding approach, introduced by Li and Fourches [2021], which is analogous to the Byte-Pair Encoding methodology commonly used for tokenization in LLMs [Sennrich et al., 2016].

1. First, all SMILES strings from the USPTO, SMolInstruct, and L+M-24 datasets were extracted and filtered to remove all invalid SMILES using RDKit [Landrum et al., 2006].
2. One round of data augmentation was performed by using RDKit to generate a single random alternative SMILES representation for each SMILES string.
3. SMILES strings were pre-tokenized into atom-level units using a regex-based tokenizer, establishing a foundation of chemically valid candidate tokens.
4. An iterative process was then applied, merging the most frequent pairs in the token sequences. This merging continued until no candidate pair exceeded a frequency threshold of three.

Finally, in addition to textual and molecular tokens, we added a set of special tokens required for our data format. These included tokens such as `<EOS>`, `<SMILES>`, `</SMILES>`, `<MOLFORMULA>`, and `</MOLFORMULA>`. The top ten SMILES tokens added to the vocabulary by frequency are shown in Table 4. After combining these two sets of tokens, we added 17,795 new tokens to the original Llama 3 vocabulary of 128k [Grattafiori et al., 2024]. See Appendix A for the top ten molecular and textual tokens added.

### 3.3 Continued Pretraining Setup

To study the impact of our approach, we selected the Llama3-8B model as our base architecture. This choice was motivated by the prevalence of English in chemical literature and the demonstrated handling of catastrophic forgetting by continued pretraining of dense models using data replay, learning rate rewarming and redecaying [Ibrahim et al., 2024, Parmar et al., 2024, Gupta et al., 2023].

The addition of new tokens necessitated resizing the model's token embedding matrix. To ensure a smooth integration of the new chemical vocabulary, the embeddings for these new tokens were initialized with the mean of all existing token embeddings in the original Llama 3 model. During training, the learning rate was managed with a cosine decay schedule, following an initial warmup period, with a maximum learning rate of 3e-4. See Appendix B for all hyperparameters used.

### 3.4 Experiments

To assess the efficacy of our proposed methodology, we conducted a series of experiments centered on the continued pretraining of the Llama3-8B model. The primary comparison was between two training configurations: one with our proposed vocabulary extension and a baseline model trained without it. All models were trained for a single epoch on the curated pretraining data blend using a 32-GPU cluster composed of four nodes with eight NVIDIA H100s each.

### 3.5 Model Evaluations

To validate our central objective of developing a model that learns a robust joint representation of natural language and molecular structures, we evaluated our models on a suite of downstream tasks requiring a nuanced understanding of both modalities. We selected the SMolInstruct benchmark [Yu et al., 2024], a comprehensive, high-quality instruction-tuning dataset designed specifically for chemistry tasks involving small molecules, using the held-out benchmark split excluded from our pretraining. The tasks in this benchmark include forward synthesis (predicting reaction products), retrosynthesis (identifying reactants), molecule captioning (generating textual descriptions), name conversion (translating between molecular representations like SMILES and IUPAC), and property prediction (water solubility, octanol/water distribution coefficient, blood-brain barrier permeability, toxicity, HIV replication inhibition, and side effects of drugs).

We provided few-shot examples in the input prompt, which instruct the model to enclose its output in appropriate tags, such as `<SMILES>` and `</SMILES>`, to facilitate information extraction. An example prompt is shown in Appendix C. These predictions were then parsed and fed into corresponding deterministic evaluation metrics for each task.

## 4 Results and Discussion

In this section, we present the results of our evaluations as well as a brief analysis of the learned SMILES tokens that were added to the vocabulary of the customized model.

### 4.1 Evaluation Results

Table 2: Evaluation results for base Llama 3 and our CPT models on five task categories from the SMolInstruct dataset.

| Task | Samples | Metric | Llama3-8B (base) | Llama3-8B CPT | Llama3-8B CPT+ vocab ext |
|---|---|---|---|---|---|
| Forward Synthesis | 4062 | # Invalid | 612 | 60 | **8** |
| | | # Exact Match | 2 | 201 | **2507** |
| | | Morgan FPS | 0.44 | 0.44 | **0.84** |
| Retrosynthesis | 4156 | # Invalid | 1586 | 107 | **16** |
| | | # Exact Match | 0 | 22 | **1366** |
| | | Morgan FPS | 0.40 | 0.31 | **0.69** |
| Molecule Captioning | 2538 | METEOR | 0.19 | 0.16 | **0.22** |
| NC-I2F | 2993 | # Exact Match | 166 | 127 | **289** |
| NC-I2S | 2993 | # Exact Match | 1 | 139 | **1695** |
| NC-S2F | 2993 | # Exact Match | 21 | **357** | 77 |
| NC-S2I | 2993 | # Exact Match | 0 | 7 | **171** |
| Property ESOL | 112 | RMSE | **1.95** | 5.42 | 5.68 |
| Property LIPO | 420 | RMSE | **8.27** | 59.93 | 31.13 |
| Property BBBP | 197 | F1 Score | 0.11 | 0.79 | **0.88** |
| Property TOX | 144 | F1 Score | 0.13 | 0.00 | **0.14** |
| Property HIV | 4107 | F1 Score | 0.06 | **0.07** | 0.06 |
| Property SIDER | 2860 | F1 Score | 0.25 | 0.31 | **0.66** |

The results of the different chemistry tasks used for evaluation are summarized in Table 2. These results confirm that continued pretraining (CPT) substantially improves the base model's performance on downstream chemistry tasks. Crucially, the Llama3-8B model with our proposed vocabulary extension consistently outperformed the model that underwent standard CPT, validating the benefits of integrating chemically aware tokens into the model's vocabulary. Clear improvements over the baseline can be observed for all the tasks except the quantitative prediction tasks, such as lipophilicity (LIPO) and ESOL. This performance gap indicates that while the pretrained model successfully learns abstract features for SMILES understanding and classification, it struggles to learn fine-grained structural nuances required for precise numerical prediction.

### 4.2 Analysis of Chemical Tokens

Here we provide a high-level analysis of the SMILES tokens that were learned via the SMILES Pair Encoding process described in Section 3.2.

Comparing the tokens produced by the two tokenizers over all SMILES in our corpus, the base tokenizer yields a median of 41 tokens per string, whereas the extended vocabulary yields 10. This lower fertility in the extended tokenizer is desirable, as it shortens sequences—reducing compute and fragmentation. Consistent with the goal of allocating vocabulary to frequent, reusable motifs and representing rarer forms via recomposable subunits, the extended tokenizer captures many common molecules as single tokens: 4% of SMILES are represented by one token. See Appendix D for a comparison of the full token-length distributions.

Table 3: A selection of SMILES from the dataset, comparing tokenization with the base Llama 3 tokenizer and the extended tokenizer.

|  | SMILES string | Llama 3 tokenizer | Extended tokenizer |
|---|---|---|---|
| 1. | `N[C@@H](CCC(=O)O)C(=O)O` | `N, [C, @@, H, ](, CCC, (= O, ), O, ), C, (=, O, ), O` | `N[C@@H](C, CC(=O)O), C(=O)O` |
| 2. | `[1*]NC(=O)N[2*]` | `[, 1, *, ], NC, (=, O, ), N [, 2, *, ]` | `[1*]N, C(=O)N, [, 2, *, ]` |
| 3. | `COC(=O)c1ccc(C)cc1` | `C, OC, (=, O, )c, 1, ccc (C, ), cc, 1` | `COC(=O), c1ccc(C)cc1` |
| 4. | `Cc1ccccc1` | `Cc, 1, cc, ccc, 1` | `Cc1ccccc1` |

Table 3 shows example SMILES from the corpus and their tokenization under both base and extended tokenizers. Beyond reduced token fertility, the extended tokenizer captures chemically meaningful substructures, often corresponding to complete or near-complete functional groups with characteristic reactivity. In the first two SMILES, we see `C(=O)O` and `C(=O)N`, corresponding to the carboxylic acid and carboxamide functional groups. The third is decomposed into `COC(=O)` and `c1ccc(C)cc1`, a methyl ester functional group and a para-methylphenyl aromatic ring. Finally, `Cc1ccccc1`, the most frequent SMILES in our corpus (the molecule toluene), is captured as a single token.

## 5 Conclusion

This work demonstrates that the tokenization bottleneck in chemistry representation learning can be mitigated through targeted vocabulary extension and continued pretraining. Our findings establish this vocabulary-centric approach as a powerful strategy for creating unified foundation models with superior performance on downstream tasks in the chemistry domain.

## 6 Future Research Directions

Integrating our continually pretrained model with external chemistry tools and reasoning frameworks, as demonstrated by systems like ChemCrow [M. Bran et al., 2024] and CACTUS [McNaughton et al., 2024], presents a promising path forward. We believe that a synergistic approach that combines a semantically sound foundation model with post-training alignment and tool-use capabilities has the potential to unlock state-of-the-art performance on complex chemistry tasks.

## Acknowledgments and Disclosure of Funding

The authors thank NVIDIA for supporting this research.

## A Top Chemistry Tokens

Table 4: The top 10 SMILES and textual tokens added to the new vocabulary.

| | SMILES token | | Textual token |
|---|---|---|---|
| 1. | `1C` | 1. | `invention` |
| 2. | `C2` | 2. | `mmol` |
| 3. | `(=O)N` | 3. | `compounds` |
| 4. | `(=O)C` | 4. | `mixture` |
| 5. | `C)` | 5. | `embodiments` |
| 6. | `C3` | 6. | `preferably` |
| 7. | `CC` | 7. | `described` |
| 8. | `=C1` | 8. | `embodiment` |
| 9. | `c1` | 9. | `herein` |
| 10. | `ccc(` | 10. | `thereof` |

## B Training Hyperparameters

Table 5: Training hyperparameters for continued pretraining of Llama3-8B

| Hyperparameters | Value |
|---|---:|
| Context Window | 8192 |
| Global Batch Size | 16 |
| Optimizer | Distributed Fused Adam |
| Weight Decay | 0.1 |
| Betas | 0.9, 0.95 |
| Scheduler | Cosine Annealing |
| Warmup Steps | 200 |
| Max Learning Rate | 3e-4 |
| Min Learning Rate | 3e-6 |

## C Example Outputs

### Example predictions for forward synthesis task

**Input Prompt**

```
You are a chemistry assistant. For the given reactants and reagents,
predict the resulting product. Always enclose the final SMILES string
in <SMILES> and </SMILES> tags.

Input: <SMILES>CCOC(=N)C1=C(F)C=CC=C1F.O=C=NC1=CC(Cl)=C(OC2=NC=C(C(F)
(F)F)C=C2Cl)C(Cl)=C1</SMILES>
Based on the reactants and reagents given above, suggest a possible
product.

Answer: A possible product can be
<SMILES>CCOC(=NC(=O)NC1=CC(Cl)=C(OC2=NC=C(C(F(F)F)C=C2Cl)C(Cl)=C1)C1
=C(F)C=CC=C1F</SMILES>

Input: Propose a potential product given these reactants and reagents.
<SMILES>C1CSCCN1.CC1=CC=CC=C1.ClCCl.NC1=CNN=C1C1=NC2=CC(CN3CCOCC3)
=CC=C2N1.O=C(Cl)Cl</SMILES>

Answer:
<SMILES>O=C(NC1=CNN=C1C1=NC2=CC(CN3CCOCC3)=CC=C2N1)N1CCSCC1</SMILES>

Input:
<SMILES>COCCOCCOC1=CC=CC2=C1C(=O)N(CC1=CC=C(OC)C=C1)C2=O.[Al+3].
[H-].[Li+]</SMILES>

Answer:
<SMILES>
```

**LLM Generated Output**

```
COCCOCCOC1=CC=CC2=C1CN(CC1=CC=C(OC)C=C1)C2</SMILES>
```

# D Reduced Token Fertility for SMILES

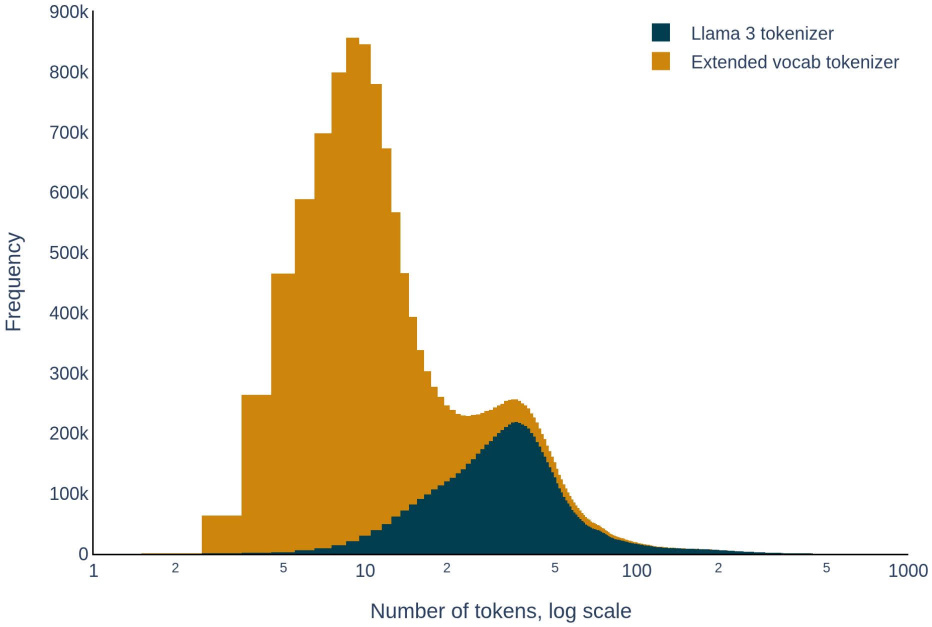

Figure 1: Histogram of the number of tokens per SMILES string

## References

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. L+M-24: Building a dataset for Language+Molecules @ ACL 2024. In Carl Edwards, Qingyun Wang, Manling Li, Lawrence Zhao, Tom Hope, and Heng Ji, editors, *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pages 1–9, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.langmol-1.1. URL `https://aclanthology.org/2024.langmol-1.1/`.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models, 2024. URL `https://arxiv.org/abs/2306.08018`.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3 (4):128–135, 1999.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to (re)warm your model?, 2023. URL `https://arxiv.org/abs/2308.04014`.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th*

*Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740/`.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models, 2024. URL `https://arxiv.org/abs/2403.08763`.

Greg Landrum et al. RDKit: Open-source cheminformatics software. `https://www.rdkit.org`, 2006. [Online; accessed 15-November-2025].

Xinhao Li and Denis Fourches. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. *Journal of chemical information and modeling*, 61(4):1560–1569, 2021.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, 2020.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. MolFM: A multimodal molecular foundation model, 2023. URL `https://arxiv.org/abs/2307.09484`.

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, 2024.

Andrew D McNaughton, Gautham Krishna Sankar Ramalaxmi, Agustin Kruel, Carter R Knutson, Rohith A Varikoti, and Neeraj Kumar. Cactus: Chemistry agent connecting tool usage to science. *ACS omega*, 9(46):46563–46573, 2024.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Reuse, don't retrain: A recipe for continued pretraining of language models, 2024. URL `https://arxiv.org/abs/2407.07263`.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Shaghayegh Sadeghi, Alan Bui, Ali Forooghi, Jianguo Lu, and Alioune Ngom. Can large language models understand molecules? *BMC bioinformatics*, 25(1):225, 2024.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pages 30458–30490. PMLR, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162/`.

Akshay Subramanian, Kevin P. Greenman, Alexis Gervaix, Tzuhsiung Yang, and Rafael Gómez-Bombarelli. Automated patent extraction powers generative modeling in focused chemical spaces. *Digital Discovery*, 2(4):1006–1015, 2023. ISSN 2635-098X. doi: 10.1039/D3DD00041A. URL `https://pubs.rsc.org/en/content/articlelanding/2023/dd/d3dd00041a`.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources.

In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.129. URL https://aclanthology.org/2020.findings-emnlp.129/.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset, 2024. URL https://arxiv.org/abs/2402.09391.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, et al. ChemDFM: A large language foundation model for chemistry. In *NeurIPS 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024.