


# Are frontier models superhuman chemists?

Nawaf Alampara<sup>1</sup>, Adrian Mirza<sup>1</sup>, Sreekanth Kunchapu<sup>1</sup>, Aswanth Krishnan<sup>1</sup>,  
Mara Willhelmi<sup>1</sup>, Macjonathan Oreke<sup>1</sup>, Benedict Emoekabu<sup>1</sup>, Tanya Gupta<sup>5</sup>,  
Philippe Schwaller<sup>5</sup>, Michael Pieler<sup>1</sup>, and Kevin Maik Jablonka <sup>1,2,3</sup>

<sup>1</sup>Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich  
Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany

<sup>2</sup>Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich  
Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany

<sup>3</sup>Helmholtz Institute for Polymers in Energy Applications (HIPOLE),  
Philosophenweg 7a, 07743 Jena, Germany

<sup>1</sup>mail@kjablonka.com

March 5, 2024

**Abstract**

## 1 Introduction

Large language models (LLMs) are frontier machine learning (ML) models trained on massive amounts of text to complete sentences. While some see in them “sparks of artificial general intelligence (AGI)”,<sup>?</sup> others consider them as “stochastic parrots”—i.e., systems that only regurgitate what they have been trained on.<sup>?</sup>

Chemists and materials scientists have quickly caught on the mounting attention given to LLMs, some even suggesting that “the future of chemistry is language”.<sup>1</sup> This statement is motivated by a growing number of reports that use LLMs to properties of molecules or materials,<sup>???</sup> to optimize reactions<sup>??</sup> or generate materials,<sup>???</sup> extract information,<sup>?????</sup> or to even build “autonomous” systems that can physically perform reactions provided a command in natural language.<sup>???</sup> The rapid increase in capabilities led to concerns about the potential for dual use of these technologies, e.g. for the design of chemical weapons.<sup>2–4</sup> Moreover, the use of these models is now also widespread among students<sup>?</sup> as well as research groups. For some users, misleading information—especially about safety-related aspects—might lead to harmful outcomes. Unfortunately, apart from anecdotal reports there is little evidence on how LLMs perform compared to experts.

Thus, to better understand what LLMs can do for chemistry and materials science, and where they might be improved with further developments, a comprehensive analysis is needed. For the development of LLMs, such evaluation is currently mostly performed via standardized benchmarks such as BigBench<sup>5</sup> or the LM Eval Harness.<sup>6</sup> The former contains, among 204 tasks, only two tasks classified as “chemistry related” whereas the latter contains no specific chemistry tasks. Due to the lack of widely excepted standard benchmarks, the developers of chemical language models<sup>7?–9</sup> frequently utilize language-interfaced<sup>10</sup> tabular datasets such as the ones reported in MoleculeNet,<sup>11</sup> Therapeutic Data Commons<sup>12</sup> or MatBench.<sup>13</sup> While those evaluations can measure how well models can make predictions for very specific tasks, they only give a poor measure of how useful those models might be as a chemical assistant.

While some benchmark based on university entrance exams<sup>14,15</sup> or automatic text mining<sup>16,17</sup> have been proposed, also those do not satisfy the following basic criteria chemistry benchmarks should satisfy:

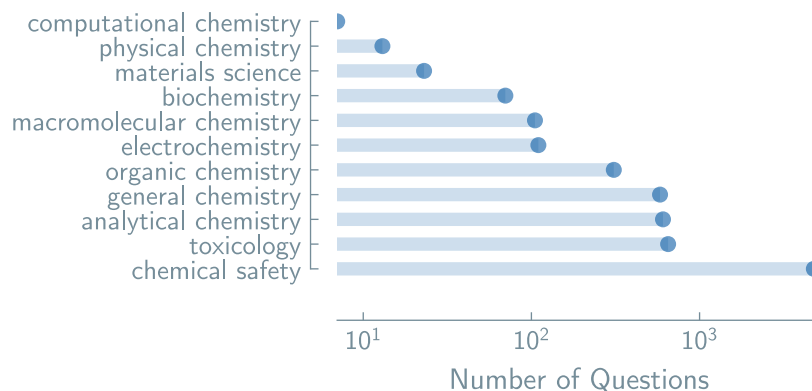
- *End-to-end automation.* For model development, the evaluations will need to be run many times (e.g., on regular intervals of a training run). Approaches that rely on humans scoring the answers of a system<sup>??</sup> can thus not be used.

- *Careful validation by experts.* Manual curation is needed to minimize number of incorrect or unanswerable questions.<sup>18</sup> This is motivated by the observation that many widely used benchmarks are plagued by noisiness.<sup>19,20</sup>
- *Usable with models that support special treatment of molecules.* Some models such as Galactica<sup>21</sup> use special tokenization or encoding procedures for molecules or equations. To support this, the benchmark system must encode the semantic meaning of various parts of the question or answer.
- *Usable with black box systems.* Many relevant systems do not provide access to model weights or even just the raw logits. This might be the case because the systems are proprietary or because they involve not only LLMs but also external tools such as search application programming interfaces (APIs) or code executors.<sup>???</sup> Thus, a benchmark should not assume access to the raw model outputs but be able to operate on text completions.
- *Probing capabilities beyond answering of multiple-choice questions (MCQs).* In real world chemistry as well as higher-level university education multiple choice question are seldom utilized. Yet, most benchmarking frameworks focus on the MCQ setting because of the ease of evaluation. Realistic evaluations must measure capabilities beyond the answering of MCQ.
- *Cover a diverse set of topics.* Chemistry, as the “central science”, bridges multiple disciplines.<sup>22</sup> To even just approximate “chemistry capabilities” the topics covered by a chemistry benchmark must be very diverse.

In this work, we report a novel benchmarking framework, chembench, and use it to reveal limitations of current frontier models for the use in the chemical sciences. Our benchmark consists more than 6000 question answer pairs manually or semi-automatically compiled from diverse sources. It covers a large fraction of the topics taught in undergraduate chemistry curricula at various skill levels and can be used with any system that can return text (i.e., also tool-augmented systems).

To contextualize the scores, we also surveyed more than XX experts in chemistry on a subset of the benchmark corpus to be able to compare the performance of current frontier models with the one of humans. Our results indicate that current frontier models perform “superhuman” on some aspects of chemistry but in many cases, included safety-related ones, might be very misleading.

## 2 Results



**Figure 1: Number of questions for different topics.** The topics have been assigned using a combination of a rule-based system (mostly based on the source the question has been sampled from) as well as a classifier operating on a word-embedding of the question. The figure shows that not all aspects of chemistry are equally represented in our corpus. The corpus currently focuses on safety-related aspects.

## 2.1 Benchmark dataset

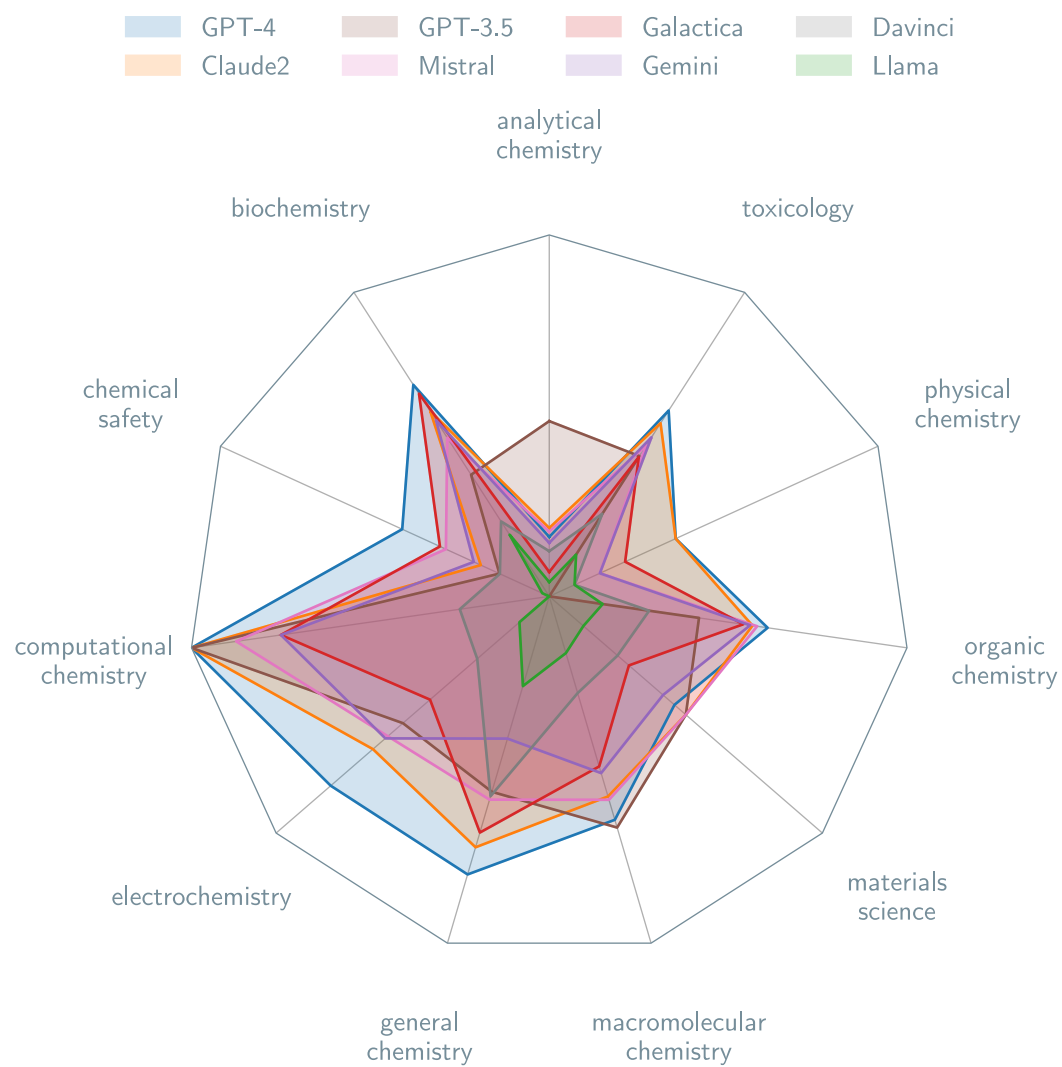
To compile our benchmark corpus we utilized a broad list of sources (Section 4.1), ranging from university exams to semi-automatically generated questions based on curated subsets of data in chemical databases. To ensure maximal interoperability, we curated the data in an extended form of the BigBench format. This also implies that future baselines can be built on top of our infrastructure as long as they are saved in the same format. For quality assurance, all questions have been reviewed by at least one scientist in addition to the original curator.

Importantly, our large pool of questions encompassed a wide range of topics. This can be seen, for example, in Figure 1 in which we compare the number of questions in different categories (see Section 4.4 for details on how we assigned topics). By design, a focus of our corpus is on safety-related aspects with a (currently) limited sampling of questions from physical or theoretical chemistry, which might be extended in future work.

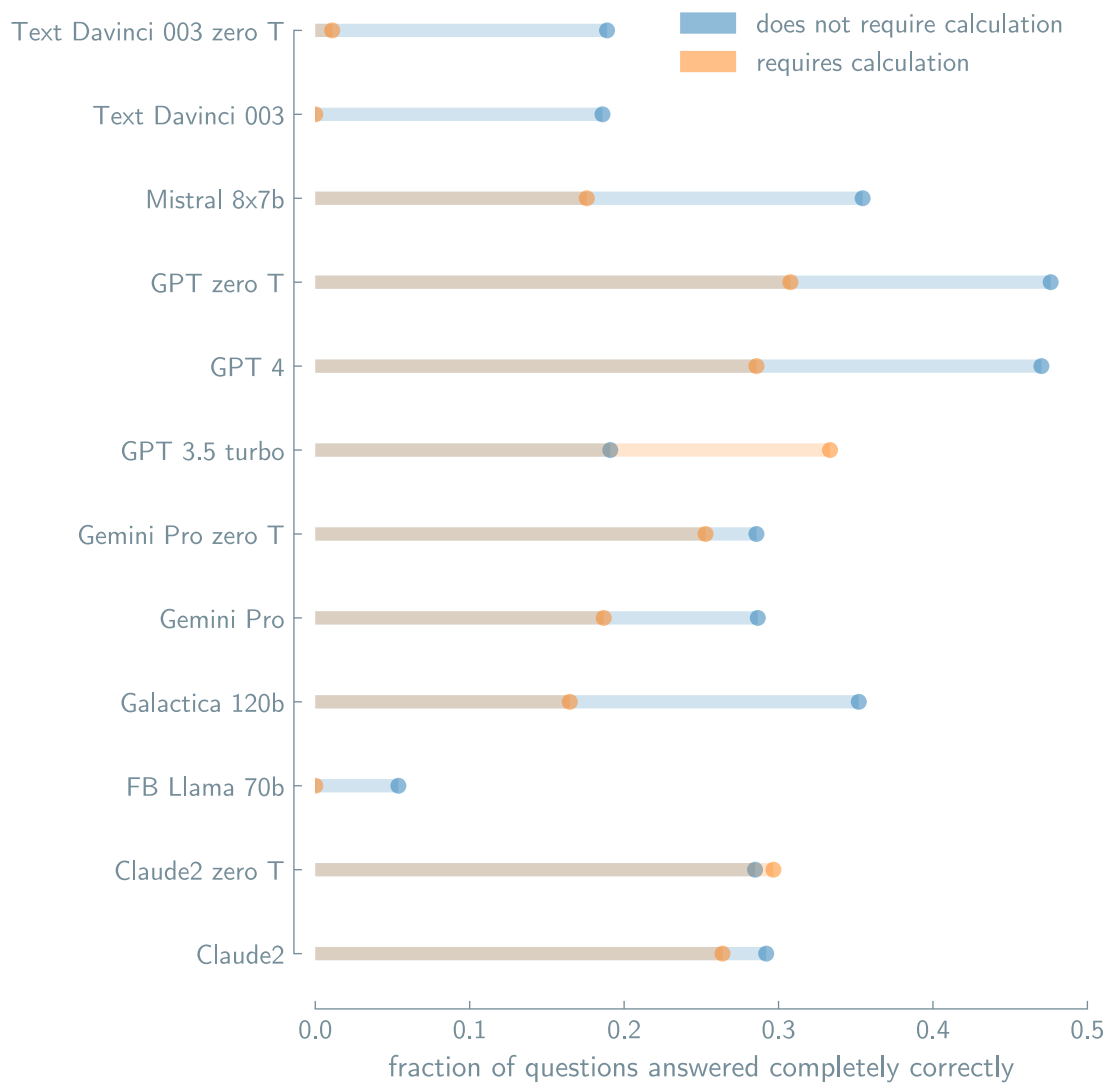
Importantly, the corpus samples both MCQ and open-ended questions in a balanced way (). As one might expect, most questions are difficult to read according to the FleschKincaid readability test (57(18)).<sup>23</sup>

## 2.2 Model evaluation





**Figure 3:** Caption



**Figure 4:** Caption

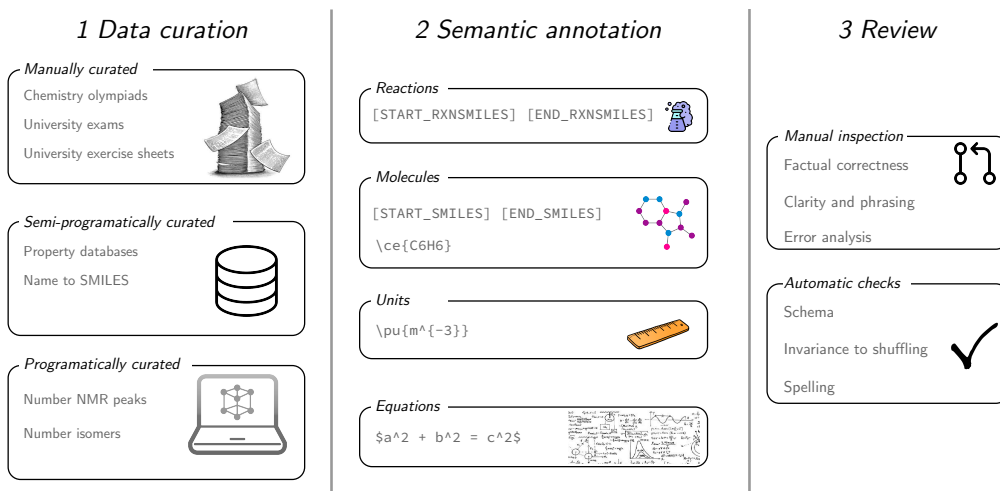


Figure 5: Data curation workflow.

## Manually curated questions

## Semi-programmatically generated questions

Oxidation states

Electron counts

## Programmatically generated questions

**Number of NMR peaks** To generate tasks about the number of NMR peaks, we randomly sampled SMILES from the ZINC dataset and then used OpenChemLib<sup>24</sup> to compute the number of diastereotopically distinct hydrogen atoms.

GHS classification

Hazard statements

Number of isomers

## 4.2 Model evaluation workflow

### Prompting



**Parsing** Our parsing workflow is, by default, multistep and primarily based on regular expressions. In the case of instruction-tuned models we first attempt to identify the [ANSWER][\ANSWER] environment we prompt the model to report the answer in. In the case of completion models, this step is skipped. From there, we attempt to extract the relevant enumeration letters (for multiple-choice questions) or numbers. In the case of numbers, our regular expression was engineered to be able to deal with various forms of scientific notation. As initial tests indicated that models sometimes return integers in the form of words, e.g. “one” instead of “1”, we also implemented a word-to-number conversion. In case these hard-coded parsing steps fail, we fall back to using a LLM, e.g. Claude-2, to parse the completion. The frequency of this fallback being triggered was very different for different models (see XX). Manual verification (see XX) indicates that this LLM-based parsing is not a relevant error source.

## Models

### 4.3 Human baseline

**App** To facilitate the collection of responses, we developed a responsive web application using NextJS/React and Postgresql. This app also rendered molecules using XX, wherefore users were not required to parse SMILES.

**Question selection** Since we anticipated that we will not be able to collect enough responses for every question to allow for a meaningful statistical analysis, we decided on showing a relevant subset of all questions to the human scorers. For selecting the subset, we decided on addressing two questions:

- Are the questions for which the models scored poorly just too difficult or unanswerable?
- Are there areas in which the performance of humans is very different from the ones of the models?

To answer the first question we selected X questions that all LLMs (model names) from an initial scoring round did not answer correctly. From those we picked X diverse one using greedy MaxMin sampling on the embeddings on the questions computed using BART (see below).

**Study design** For our initial study we wanted to maximize the response rate given our available resources. For this reason, we did not opt for a highly controlled study setting. That is, while users were prompted to not use external tools other than a calculator and to not consult with other humans, we do not

have any way to verify that the participants complied with those rules. Note that users were also allowed to skip questions.

#### **4.4 Classification of questions into topics**

When curating our dataset we systematically recorded keywords and sources. To allow for analysis of the model performance as a function of the topic, we leverage this information together with sequence classification models. For questions which can easily be assigned to a topic based on the source (e.g., number of NMR peaks, chemical compatibility, toxicology exam questions) we use this information to make the assignment. For the remaining ones, e.g., from chemistry olympiad questions, we use zero-shot sequence classification<sup>25</sup> using the BART model,<sup>26,27</sup> which our preliminary analysis found to be more robust than topic modeling based on embeddings from OpenAI’s ada model or Cohere’s Cohembed-english-v3.0 model.

### **Data and code availability**

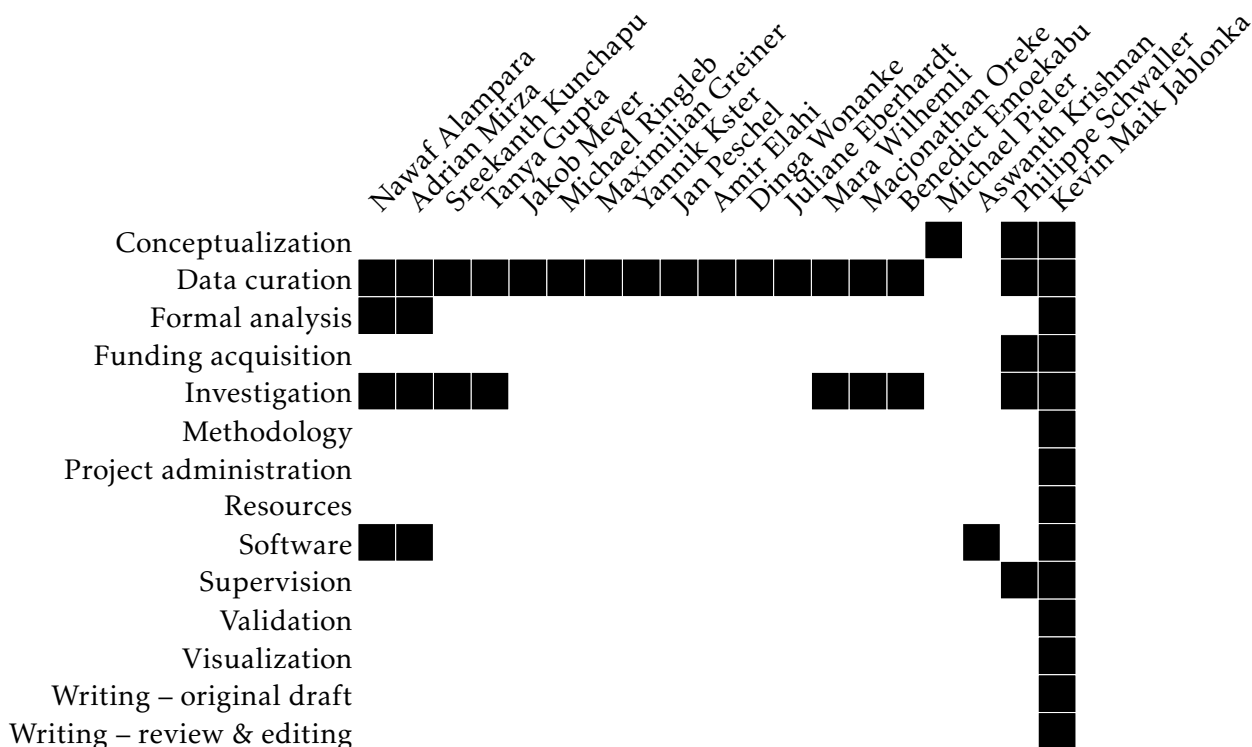
#### **Acknowledgements**

This work was supported by the Carl Zeiss Foundation and a “Talent Fund” of the “Life” profile line of the Friedrich Schiller University Jena. We also want to acknowledge access to the HPC cluster of Stability.AI.

#### **Conflicts of interest**

K.M.J. is a paid consultant for OpenAI. M.P. is employee of Stability.ai and A.M. and N.A. are paid contractors of Stability.ai.

## Author contributions



## References

- [1] White, A. D. *Nature Reviews Chemistry* **2023**, 7, 457458.
- [2] Gopal, A.; Helm-Burger, N.; Justen, L.; Soice, E. H.; Tzeng, T.; Jeyapragasan, G.; Grimm, S.; Mueller, B.; Esvelt, K. M. Will releasing the weights of future large language models grant widespread access to pandemic agents? 2023.
- [3] Ganguli, D. et al. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. 2022.
- [4] Urbina, F.; Lentzos, F.; Invernizzi, C.; Ekins, S. *Nature Machine Intelligence* **2022**, 4, 189191.
- [5] others,, et al. *arXiv preprint arXiv:2206.04615* **2022**,
- [6] Gao, L. et al. A framework for few-shot language model evaluation. 2023; <https://zenodo.org/records/10256836>.

- [7] Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. 2023.
- [8] Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. 2022.
- [9] Cai, X.; Lai, H.; Wang, X.; Wang, L.; Liu, W.; Wang, Y.; Wang, Z.; Cao, D.; Zeng, X. *Methods* **2024**, 222, 133141.
- [10] Dinh, T.; Zeng, Y.; Zhang, R.; Lin, Z.; Gira, M.; Rajput, S.; Sohn, J.-y.; Pappaliopoulos, D.; Lee, K. *Advances in Neural Information Processing Systems* **2022**, 35, 11763–11784.
- [11] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. *Chemical science* **2018**, 9, 513–530.
- [12] Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. *arXiv preprint arXiv:2102.09548* **2021**,
- [13] Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. *npj Computational Materials* **2020**, 6, 138.
- [14] Zaki, M.; Jayadeva,; Mausam,; Krishnan, N. M. A. *Digital Discovery* **2024**,
- [15] Arora, D.; Singh, H. G.; Mausam, Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. 2023.
- [16] Song, Y.; Miret, S.; Zhang, H.; Liu, B. HoneyBee: Progressive Instruction Finetuning of Large Language Models for Materials Science. 2023.
- [17] Wei, Z.; Ji, W.; Geng, X.; Chen, Y.; Chen, B.; Qin, T.; Jiang, D. Chemistry{QA}: A Complex Question Answering Dataset from Chemistry. 2021; <https://openreview.net/forum?id=oeHTRAehiFF>.
- [18] Northcutt, C. G.; Athalye, A.; Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. 2021.
- [19] Frye, C. PubMedQA noisy. 2023; [https://twitter.com/charles\\_irl/status/1731854677711507650](https://twitter.com/charles_irl/status/1731854677711507650).
- [20] Awg, Broken benchmark: MMLU. <https://www.lesswrong.com/posts/rQBafqtqKMfG2uMiWb/broken-benchmark-mmlu>.
- [21] Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. Galactica: A Large Language Model for Science. 2022.

- [22] Aspuru-Guzik, A.; Lindh, R.; Reiher, M. *ACS Central Science* **2018**, *4*, 144152.
- [23] Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; Chissom, B. S. **1975**,
- [24] Actelion, OpenChemLib. <https://github.com/actelion/openchemlib>.
- [25] Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. 2019; <https://arxiv.org/abs/1909.00161>.
- [26] Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 2019; <https://arxiv.org/abs/1910.13461>.
- [27] Facebook, Facebook/Bart-Large-mnli hugging face. <https://huggingface.co/facebook/bart-large-mnli/>.