Are large language models superhuman chemists?

Adrian Mirza [1,★], Nawaf Alampara [1,★], Sreekanth Kunchapu [1,★], Benedict Emoekabu , Aswanth Krishnan [2], Tanya Gupta [3,4], Mara Wilhelmi [1], Macjonathan Okereke [1], Mehrdad Asgari [5], Juliane Eberhardt [6], Amir Mohammad Elahi [7], Christina Glaubitz , Maximilian Greiner[1], Caroline T. Holick[1], Tim Hoffmann [1], Lea C. Klepsch [1], Yannik Köster [1], Fabian Alexander Kreth [8, 9], Jakob Meyer[1], Santiago Miret [10], Jan Matthias Peschel [1], Michael Ringleb [1], Nicole Roesner [1, 11], Johanna Schreiber [1, 10], Ulrich S. Schubert [1,8,11, 12], Leanne M. Stafast [1, 11], Dinga Wonanke [13], Michael Pieler [14,15], Philippe Schwaller [3,4], and Kevin Maik Jablonka [1,8,11, 12, ✉]

[1]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany
[2]QpiVolta Technologies Pvt Ltd
[3]Laboratory of Artificial Chemical Intelligence (LIAC), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[4]National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[5]Department of Chemical Engineering & Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom
[6]Macromolecular Chemistry, University of Bayreuth, 95447 Bayreuth, Germany
[7]Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland
[8]Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[9]Institute for Technical Chemistry and Environmental Chemistry (ITUC), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[10]Intel Labs
[11]Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
[12]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany
[13]Theoretical Chemistry, Technische Universität Dresden, Dresden 01062, Germany
[14]OpenBioML.org
[15]Stability.AI
[✉]mail@kjablonka.com
[★]These authors contributed equally.

March 30, 2024

Abstract

Large language models (LLMs) have gained widespread interest due to their ability to process human language and perform tasks on which they have not been explicitly trained. This is relevant for the chemical sciences, which face the problem of small and diverse datasets that are frequently in the form of text. LLMs have shown promise in addressing these issues and are increasingly being harnessed to predict chemical properties, optimize reactions, and even design and conduct experiments autonomously.

However, we still have only a very limited systematic understanding of the chemical reasoning capabilities of LLMs, which would be required to improve models and mitigate potential harms. Here, we introduce "ChemBench," an automated framework designed to rigorously evaluate the chemical knowledge and reasoning abilities of state-of-the-art LLMs against the expertise of human chemists.

We curated more than 7,000 question-answer pairs for a wide array of subfields of the chemical sciences, evaluated leading open and closed-source LLMs, and found that the best models outperformed the best human chemists in our study on average. The models, however, struggle with some chemical reasoning tasks that are easy for human experts and provide overconfident, misleading predictions, such as about chemicals' safety profiles.

These findings underscore the dual reality that, although LLMs demonstrate remarkable proficiency in chemical tasks, further research is critical to enhancing their safety and utility in chemical sciences. Our findings also indicate a need for adaptations to chemistry curricula and highlight the importance of continuing to develop evaluation frameworks to improve safe and useful LLMs.

1 Introduction

Large language models (LLMs) are machine learning (ML) models trained on massive amounts of text to complete sentences. Aggressive scaling of these models has led to a rapid increase in their capabilities,[brown2020language, zhong2024benchmarking] with the leading models now being able to pass in some evaluations the United States Medical Licensing Examination.[kung2023performance] They also have been shown to design and autonomously perform chemical reactions when augmented with external tools such as web search and synthesis planners.[openai2024gpt4, Boiko_2023, bran2023chemcrow] While some see "sparks of artificial general intelligence (AGI)" in them,[bubeck2023sparks] others consider them as "stochastic parrots"—i.e., systems that only regurgitate what they have been trained on.[bender2021dangers] Nevertheless, the promise of these models is that they have shown the ability to solve a wide variety of tasks they have not been explicitly trained on.[bommasani2021opportunities, anderljung2023frontier] This has led to tremendous economic interest and investment in such generative models, with an expected market of more than \$1.3 trillion (almost \$30 billion for drug discovery applications) by 2032.[bloomberg]

Chemists and materials scientists have quickly caught on to the mounting attention given to LLMs, with some voices even suggesting that "the future of chemistry is language".[White_2023] This statement is motivated by a growing number of reports that use LLMs to predict properties of molecules or materials,[jablonka202314, jablonka2024leveraging, xie2024fine, liao2024words, z optimize reactions[ramos2023bayesian, kristiadi2024sober] generate materials,[rubungo2023llm, flam2023language, gruver2024fine] extract information,[Patiny_2023, Dagdelen_2024, Zheng_2024, lala2023paperqa, caufield2023structured, gupta2022discomat] or to even prototype systems that can autonomously perform reactions in the physical world based on commands provided in natural language.[bran2023chemcrow, Boiko_2023, darvish2024organa]

In addition, since a lot—if not most—of the information about chemistry is currently stored and communicated in text, there is a strong reason to believe that there is still a lot of untapped potential in LLMs for chemistry and materials science.[miret2024llms] For instance, most insights in chemical research do not directly originate from data stored in databases but rather from the scientists and their ability to interpret data. Many of these insights are in the form of text in scientific publications. Thus, operating on such texts might be our best way of unlocking these insights and learning from them. This might ultimately lead to general copilot systems for chemists that can provide answers to questions or even suggest new experiments based on vastly more information than a human could ever read. Such a usage mode is, in particular, interesting in the face of recent advances in autonomous laboratories.[Boiko_2023, bran2023chemcrow, darvish2024organa, granda2018controlling,

However, the rapid increase in capabilities of chemical ML models led (even before the recent interest in LLMs) to concerns about the potential for dual use of these technologies, e.g., for the design of chemical weapons.[gopal2023releasing, ganguli2022red, Urbina_2022, campbell2023censoring, mou To some extent, this is not surprising as any technology that, for instance, is used to design non-toxic molecules can also be used inversely to predict toxic ones. Yet, it is important to recognize that while LLMs can facilitate access to information and tools, this information about controlled chemicals and task-specific tools has been available for years and requires access to controlled facilities such as laboratories. Still, it is essential to realize that such models' user base is wider than that of chemistry and materials science experts who critically reflect on every output such models produce. For example, many students frequently consult these tools—perhaps even to prepare chemical experiments.[Intelligent.com_2023] This also applies to users from the general public, who might consider using LLMs to answer questions about the safety of chemicals. Thus, for some users, misleading information—especially about safety-related aspects—might lead to harmful outcomes. However, even for experts, chemical understanding and reasoning capabilities are essential as they will determine the capabilities and limitations of their models in their work, e.g., in copilot systems for chemists. Unfortunately, apart from anecdotal reports, there is little evidence on how LLMs perform compared to expert (human) chemists.

Thus, to better understand what LLMs can do for the chemical sciences and where they might be improved with further developments, evaluation frameworks are needed to allow us to measure progress and mitigate potential harms systematically. For the development of LLMs, evaluation is currently primarily performed via standardized benchmark suites such as BigBench[srivastava2022beyond] or the LM Eval Harness.[eval-harness] Among 204 tasks (such as linguistic puzzles), the former contains only two tasks classified as "chemistry related", whereas the latter contains no specific chemistry tasks. Due to the lack of widely accepted standard benchmarks, the developers of chemical language models[jablonka2024leveraging, guo2023large, ahmad2022chemberta2, Cai_2024, frey2023neural] frequently utilize language-interfaced[dinh2022lift] tabular datasets such as the ones reported in MoleculeNet,[wu2018moleculenet] Therapeutic Data Commons[huang2021therapeutics] or MatBench.[Dunn_2020] In these cases, the models are evaluated on predicting very specific properties of molecules (e.g., solubility, toxicity, melting temperature or reactivity) or on predicting the outcome of specific chemical reactions. This, however, only gives a very limited view of the general chemical capabilities of the models.

While some benchmarks based on university entrance exams[Zaki_2024, arora2023llms] or automatic text mining[song2023honeybee, wei2021chemistryqa, song-etal-2023-matsci] have been proposed, none of them have been widely accepted. This is likely because they cannot automatically be used with black box (or tool-augmented) systems, do not cover a wide range of topics, or are not carefully validated by experts. On top of that, the existing benchmarks are not designed to be used with models that support special treatment of molecules or equations and do not provide insights on how the models compare relative to experts.

In this work, we report a novel benchmarking framework (Figure 1), which we call ChemBench, and use it to reveal limitations of current frontier models for use in the chemical sciences. Our benchmark consists of output/total$_{n}umber_{o}f_{q}uestions.txtquestion -$

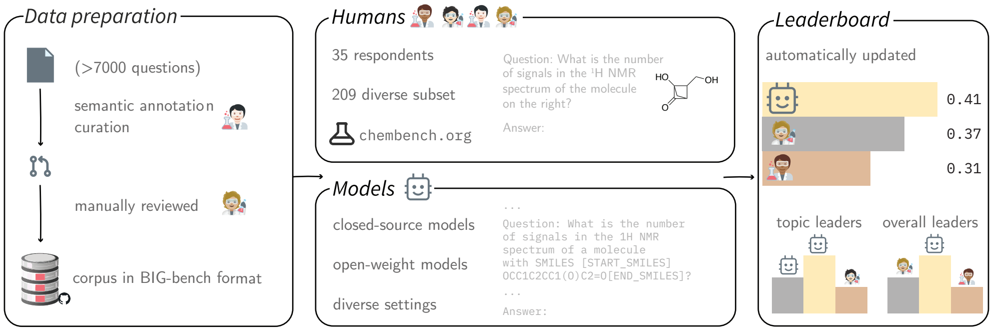
Data preparation
(>7000 questions)
semantic annotation curation
manually reviewed
corpus in BIG-bench format
Humans
35 respondents
209 diverse subset
chembench.org
Question: What is the number of signals in the ¹H NMR spectrum of the molecule on the right?
HO
OH
O
Answer:
Models
closed-source models
open-weight models
diverse settings
...
Question: What is the number of signals in the 1H NMR spectrum of a molecule with SMILES [START_SMILES] OCC1C2CC1(O)C2=O[END_SMILES]?
...
Answer:
Leaderboard
automatically updated
0.41
0.37
0.31
topic leaders
overall leaders

Figure 1: Overview of the ChemBench framework. The figure shows the different components of the ChemBench framework. The framework's foundation is the benchmark corpus that consists of many questions and answers that we manually or semi-automatically compiled from various sources. We then used this corpus to evaluate the performance of various models and tool-augmented systems using a custom framework. To provide a baseline, we built a web application that we used to survey experts in chemistry. The results of the evaluations are then compiled in publicly accessible leaderboards, which we propose as a foundation for evaluating future models.

$answerpairsmanually\,(output/manually_{g}enerated.txt)orsemi-automatically\,(output/automatically_{g}ene$
$augmentedsystems$).

To contextualize the scores, we also surveyed more than output/number$_{e}xperts.txtexpertsinchemistryonas$
$relatedones, mightbeverymisleading.Wefindthattherearestillnumerouslimitationsforcurrentmodelstoo$

2 Results and Discussion

2.1 Benchmark corpus

To compile our benchmark corpus, we utilized a broad list of sources (see **??**), ranging from university exams to semi-automatically generated questions based on curated subsets of data in chemical databases. For quality assurance, all questions have been reviewed by at least one scientist in addition to the original curator and automated checks.

Importantly, our large pool of questions encompasses a wide range of topics. This can be seen, for example, in Figure 2 in which we compare the number of questions in different subfields of the chemical sciences (see **??** for details on how we assigned topics). The distribution of topics is also evident from Figure 3 in which we visualize the questions in a two-dimensional space using a Principal Component Analysis (PCA) on the embeddings of the questions. In this representation, semantically similar questions are close to each other, and we color the points based on classification into output/num$_t$*opics.txttopics.ItisclearthatafocusofChemBench (bydesign)liesonsafety*−*relatedaspects,whichinFigure 3appearasalargedistinctclustersacrosstheembeddingspace.*

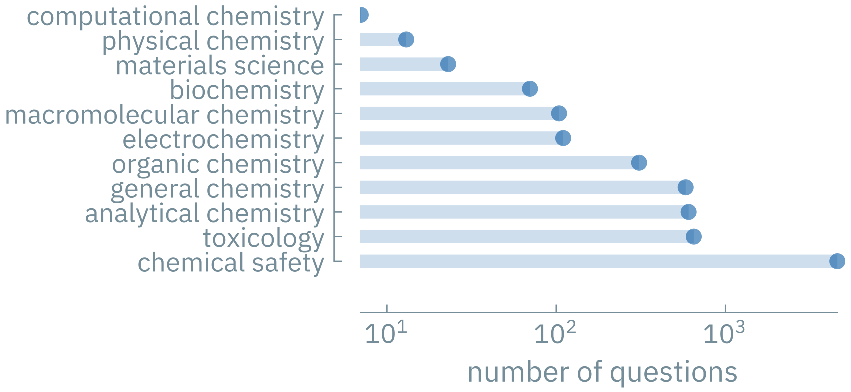

Figure 2: Number of questions for different topics. The topics have been assigned using a combination of a rule-based system (mostly based on the source the question has been sampled from) and a classifier operating on word embeddings of the questions. The figure shows that not all aspects of chemistry are equally represented in our corpus. The ChemBench corpus, by design, currently focuses on safety-related aspects, which is also evident in Figure 3. This figure represents the combined count of MCQ and open-ended questions.

While many existing benchmarks are designed around multiple-choice question (MCQ), this does not reflect the reality of chemistry education and research. For this reason, ChemBench samples both MCQ and open-ended questions (output/mcq$_q$*uestions.txtMCQquestionsan*
endedquestions).

"Tiny" subset It is important to note that a smaller subset of the corpus might be more practical for routine evaluations.[polo2024tinybenchmarks] For instance, **liang2023holistic** report costs of more than $10,000 for application programming interface (API) calls for a single evaluation on the widely used Holistic Evaluation of Language Models (HELM) benchmark. To address this, we also provide a subset (output/num$_t$*iny*$_q$*uestions.txtquestions*)*ofthecor*

2.2 Model evaluation

Benchmark suite design Because the text used in scientific settings differs from typical natural language, many models have been developed that deal with such text in a particular way. For instance, the Galactica model[taylor2022galactica] uses special tokenization or encoding procedures for molecules and equations. Current benchmark-

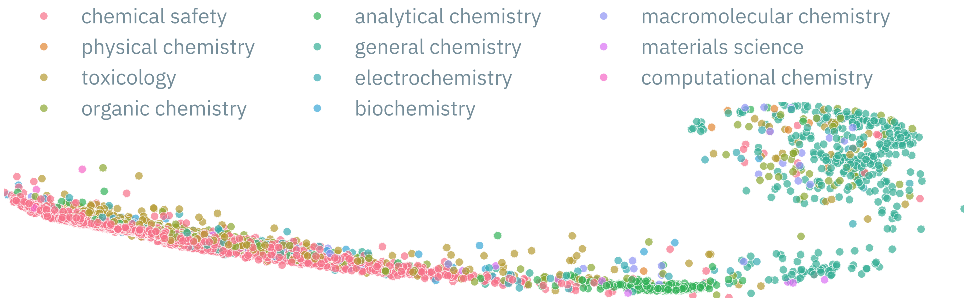

Figure 3: Principal component projection of embeddings of questions in the ChemBench corpus. To obtain this figure, we embedded questions and answers using the BART model[bart] (using other embeddings, such as of OpenAI's ada model, leads to qualitatively similar results). We then project the embeddings into a two-dimensional space using PCA. We color the points based on a classification into topics. Safety-related aspects cover a large part of the figure that is not covered by questions from other topics.

ing suites, however, do not account for such special treatment of scientific information. To address this, ChemBench encodes the semantic meaning of various parts of the question or answer. For instance, molecules represented in Simplified Molecular Input Line-Entry System (SMILES) are enclosed in `[START_SMILES] [\END_SMILES]` tags. This allows the model to treat the SMILES string differently from other text. ChemBench can seamlessly handle such special treatment in an easily extensible way because the questions are stored in an annotated format.

Since many widely utilized systems only provide access to text completions (and not the raw model outputs), ChemBench is designed to operate on text completions. This is also important given the growing number of tool-augmented systems that are deemed essential for building chemical copilot systems. Such systems can augment the capabilities of LLMs through the use of external tools such as search APIs or code executors.[schick2024toolformer, karpas2022mrkl, yao2022react] In those cases, the LLM that returns the probabilities for various tokens (that are often used for model evaluations[Fourrier_Habib_Launay_Wolf]) is only a part of the whole system, and it is not clear how to interpret the probabilities in the context of the whole system. The text completions, however, are the system's final outputs, which would also be used in a real-world application. Hence, we use them for our evaluations.

System performance To understand the current capabilities of LLMs in the chemical sciences, we evaluated a wide range of leading models[Huggingface] on the ChemBench corpus, including systems augmented with external tools. An overview of the results

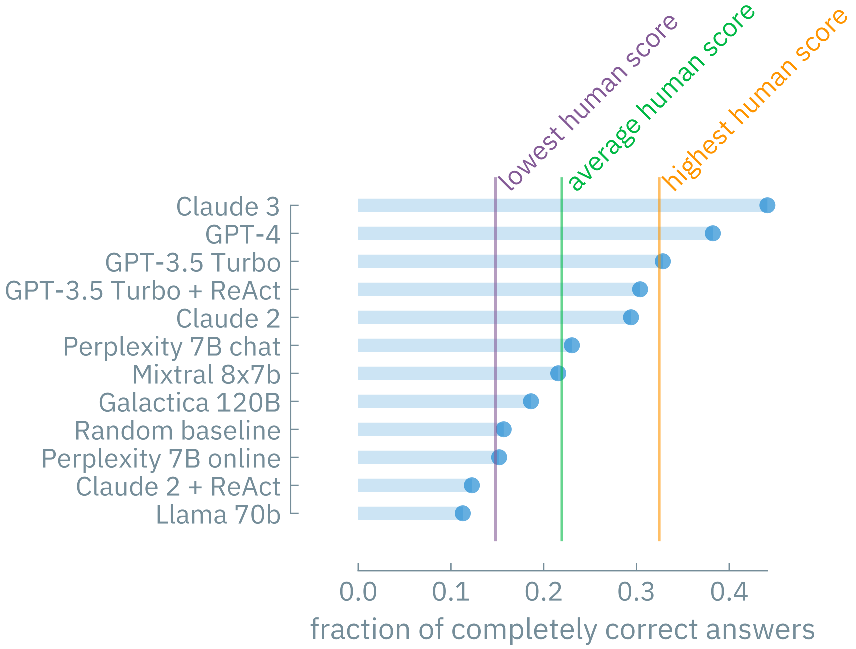

Figure 4: Performance of models and humans on the "tiny" subset of the ChemBench corpus. The figure shows the percentage of questions that the models answered completely correctly. We use horizontal bars to indicate the performance of various models and highlight statistics of the human performance. Since the humans did not answer all the questions, this plot is based on the subset of questions that most humans answered. The evaluation we use here is very strict as it only considers a question answered completely correctly or completely incorrectly, partially correct answers are also considered incorrect. **??** provides an overview of the performance of various models on the entire corpus. Systems with "ReAct" in the name are tool augmented, i.e., they can call external tools such as web search or Python code executors to better answer the questions. However, we limit those systems to a maximum of ten calls to the LLM. This constraint led the systems to often not find the correct answer within the specified number of calls. In this case, we consider the answer as incorrect.

of this evaluation is shown in Figure 4. In this figure, we show the percentage of questions that the models answered correctly. Moreover, we show the worst, best, and average performance of the human experts in our study, which we obtained via a custom web application (`chembench.org`) that we used to survey the experts. Remarkably, the figure shows that the leading LLM, Claude 3, outperforms the best human in our study in this overall metric and vastly, by more than a factor of two, exceeds the average performance of the experts in our study. Many other models also outperform the average human performance. Interestingly, the Galactica model, which was trained specifically for scientific applications, underperformed compared to many advanced commercial and some open-source models and barely exceeds the random baseline.

Given the considerable interest in tool-augmented systems, the mediocre performance of these systems (GPT-3.5 and Claude 2 augmented with tools) in our benchmark is striking. Their lack of performance, however, is partially because we limited the system to a maximum of ten LLM calls. With the default tool augmented setup (using the so-called ReAct method**yao2022react**), however, this often did not allow the system to identify the correct answer (e.g., because it repeatedly tried to search for the answer on the web and did not find a solution within ten calls to the LLM). This observation highlights the importance of communicating and measuring not only predictive performance but also computational cost (e.g., in terms of API calls) for tool-augmented systems.

Performance per topic To obtain a more detailed understanding of the performance of the models, we also analyzed the performance of the models in different subfields of the chemical sciences. For this analysis, we defined a set of topics (see **??**). We classified all questions in the ChemBench corpus into these topics based on hand-crafted rules and classifier models operating on the question texts. We then computed the percentage of questions the models or humans answered correctly for each topic. In this spider chart, the worst score for every dimension is zero (no question answered correctly), and the best score is one (all questions answered correctly). Thus, a larger colored area indicates a better performance. One can observe that this performance varies widely across models and topics. While macromolecular chemistry and biochemistry receive relatively high scores for many models, this is not the case for topics such as chemical safety or analytical chemistry. In the subfield of analytical chemistry the prediction of the number of signals observable in a Nuclear Magnetic Resonance
(NMR) spectrum proved difficult for the models (e.g., output/subset$_{s}cores/is_{n}umber_{n}mr_{p}eaks_{g}pt4.txtpercentc$
$4)whilethisquestionappearedeasier\ (output/human_{s}ubset_{s}cores/is_{n}umber_{n}mr_{p}eaks.txtpercentcorrect\ )fo$
$inspiredquestions.AsubsetofthequestionsintheChemBencharebasedontextbookstargetedatundergraduate$
$automaticallyconstructedtasks\ (see$**??**$).Forinstance,whiletheoverallperformanceinthechemicalsafetytopi$
$4,output/subset_{s}cores/is_{g}fk_{c}laude3.txt\%forClaude3,andoutput/human_{s}ubset_{s}cores/is_{g}fk.txt\%forthe$

We also gain insight into the models' struggles with chemical reasoning tasks when we examine their performance as a function of molecular descriptors. If the model would answer questions based on reasoning about the structures, one would expect the performance to depend on the complexity of the molecules. However, we find that the models' performance does not correlate with complexity indicators but rather trivially with the size of the compounds (see **??**). This indicates that the models may not be able to reason about the structures of the molecules (in the way one might expect) but instead rely on retrieval of fragments of the training data. Those observations mirror recent findings that LLMs struggle to “reverse” facts they have seen in training (i.e., generalize from “A has feature B” to “B is a feature of A”).[berglund2023reversal, zhu2023physics, allen2023physics, golovneva202]

It is important to note that the model performance for some topics, however, is underestimated in the current evaluation. This is because models provided via APIs typically have safety mechanisms that prevent them from providing answers that the provider deems unsafe. For instance, models might refuse to provide answers about cyanides. To overcome this, direct access to the model weights would be required, and we strive to collaborate with the developers of frontier models to overcome this limitation in the future. This is facilitated with the tooling ChemBench provides, thanks to which contributors can automatically add new models in an open science fashion.

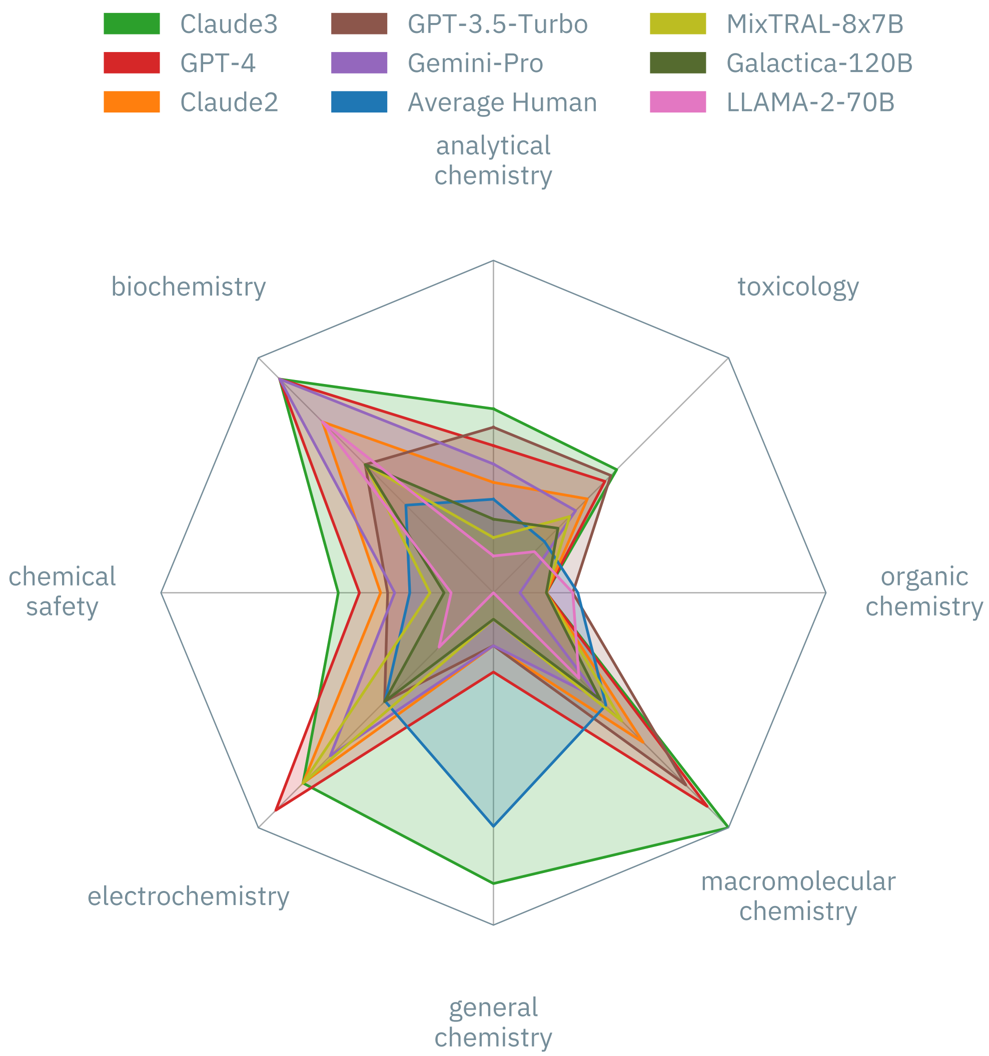

Figure 5: Performance of the models and humans on the different topics of the "tiny" subset of the ChemBench corpus. The radar plot shows the performance of the models and humans on the different topics of "tiny" subset of the ChemBench corpus. The performance is measured as the fraction of questions that were answered completely correctly by the models. The best score for every dimension is one (all questions answered correctly), and the worst is zero (no question answered completely correctly). A larger colored area indicates a better performance. This figure shows the performance on the subset of questions that were answered by humans. The performance of models on the entire corpus is shown in **??**.

Confidence estimates One might wonder whether the models can estimate if they can answer a question correctly. If they could do so, incorrect answers would be less problematic as one could detect when an answer is incorrect. To investigate this, we prompted**xiong2023llms** some of the top-performing models to estimate, on an ordinal scale, their confidence in their ability to answer the question correctly.

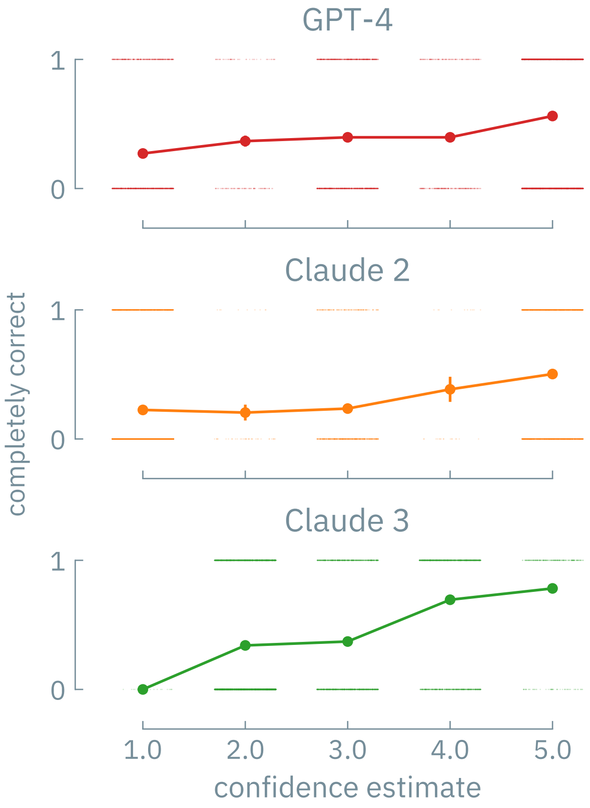

Figure 6: Relationship between confidence of the model in the answer and correctness. For this analysis, we used verbalized confidence estimates from the model. We prompted the models to return a confidence score on an ordinal scale to obtain those estimates. We then plot the correctness of the answers (which is calculated as the mean value of all answers being either completely correct (1) or completely incorrect (0)) against the confidence score. The stripplots show the individual data points, a strong density indicating a higher number of points. The lines show the mean, and the error bars show the standard error of the mean. The figure shows that the reliability of the confidence estimates varies widely across models. While the ones for GPT-4 and Claude 2 are not particularly reliable, the ones for Claude 3 follow the expected trend.

In Figure 6, we show that for some models, there is no significant correlation between the estimated difficulty and whether the models answered the question correctly or not. For applications in which humans might rely on the models to provide answers, this is a concerning observation highlighting the need for critical reasoning in the interpretation of the model's outputs.[Li_2023, miret2024llms] For example, for the questions about the safety profile of compounds, GPT-4 reported an average confidence of output/model$_{c}onfidence_{p}erformance/gpt4_{i}s_{p}ictograms_{a}verage_{c}onfidence_{c}orrect_{o}verall.txt(ond$
$-5)fortheoutput/model_{c}onfidence_{p}erformance/gpt4_{i}s_{p}ictograms_{n}um_{c}orrect_{o}verall.txtquestionsitansu$

3 Conclusions

On the one hand, our findings underline the impressive capabilities of LLMs in the chemical sciences: Leading models outperform domain experts in specific chemistry questions on many topics. On the other hand, there are still striking limitations. For very relevant topics the answers models provide are wrong. On top of that, many models are not able to reliably estimate their own limitations. Yet, the success of the models in our evaluations perhaps also reveals more about the limitations of the exams we use to evaluate models—and chemists—than about the models themselves. For instance, while models perform well on many textbook questions, they struggle with questions that require some more reasoning. Given that the models outperformed the average human in our study, we need to rethink how we teach and examine chemistry. Critical reasoning is increasingly essential, and rote solving of problems or memorization of facts is a domain in which LLMs will continue to outperform humans.

Our findings also highlight the nuanced trade-off between breadth and depth of evaluation frameworks. The analysis of model performance on different topics shows that models' performance varies widely across the subfields they are tested on. However, even within a topic, the performance of models can vary widely depending on the type of question and the reasoning required to answer it.

The current evaluation frameworks for chemical LLMs are primarily designed to measure the performance of the models on specific property prediction tasks. They cannot be used to evaluate reasoning or systems built for scientific applications. Thus, we had little understanding of the capabilities of LLMs in the chemical sciences. Our work shows that carefully curated benchmarks can provide a more nuanced understanding of the capabilities of LLMs in the chemical sciences. Importantly, our findings also illustrate that more focus is required in developing better human-model interaction frameworks, given that models cannot estimate their limitations.

While our findings indicate many areas for further improvement of LLM-based systems, it is also important to realize that clearly defined metrics have been the key to the progress of many fields of ML, such as computer vision. Although current systems are far from reasoning like a chemist, our ChemBench framework will be a stepping stone for developing systems that might come closer to this goal.