






















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 Wilhelmi¹,
Macjonathan Okereke ¹, Mehrdad Asgari ⁵, Juliane Eberhardt ⁶,
Amir Mohammad Elahi ⁷, Christina Glaubitz , Maximilian Greiner¹,
Caroline T. Holick ¹, Tim Hoffmann ¹, Lea C. Klepsch ¹, Yannik Köster ¹,
Fabian Alexander Kreth ^{8,9}, Jakob Meyer¹, 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}, 

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

²QpiVolta Technologies Pvt Ltd

³Laboratory of Artificial Chemical Intelligence (LIAC), Institut des Sciences et Ingénierie Chimiques,
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

⁴National Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de
Lausanne (EPFL), Lausanne, Switzerland

⁵Department of Chemical Engineering & Biotechnology, University of Cambridge, Philippa Fawcett
Drive, Cambridge CB3 0AS, United Kingdom

⁶Macromolecular Chemistry, University of Bayreuth, 95447 Bayreuth, Germany

⁷Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Ecole
Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland

⁸Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena,
Philosophenweg 7a, 07743 Jena, Germany

⁹Institute for Technical Chemistry and Environmental Chemistry (ITUC), Friedrich Schiller University
Jena, Philosophenweg 7a, 07743 Jena, Germany

¹⁰Intel Labs


¹¹Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena,
Germany

¹²Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14,
07743 Jena, Germany

¹³Theoretical Chemistry, Technische Universität Dresden, Dresden 01062, Germany

¹⁴OpenBioML.org

¹⁵Stability.AI

mail@kjablonka.com

*These authors contributed equally.

March 29, 2024

Abstract

Large language models (LLMs) have gained widespread interest due to their ability to process human language and perform tasks they have not been explicitly trained on. 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 chemical sciences subfields, 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 systematically.

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 show the ability to solve a wide variety of tasks they have not been explicitly trained on.^{bommasani2021opportunities, andersson2023frontier} 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, zhang2024fine} 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, zhang2024fine}

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, mou2024fine} 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-et-al-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 7059 question-answer pairs manually (863) or semi-automatically (6196) compiled from diverse sources. Our corpus covers

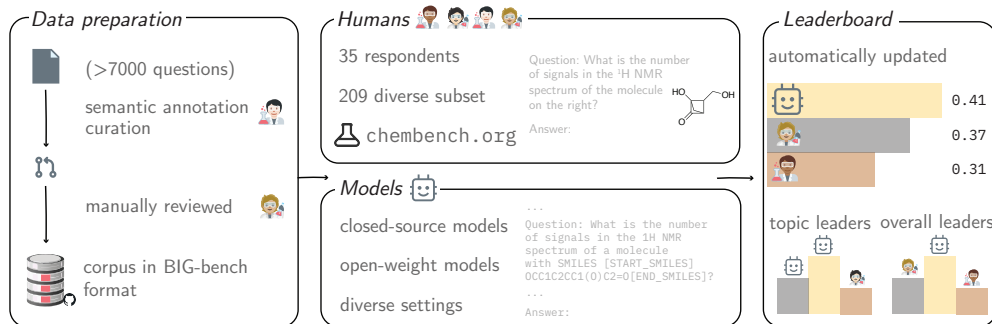


Figure 1: Overview of the ChemBench framework. The figure shows the different components of the ChemBench framework. The framework’s foundation is the benchmark corpus that consists of many questions and answers that we manually or semi-automatically compiled from various sources. We then used this corpus 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.

a large fraction of the topics taught in undergraduate and graduate chemistry curricula. It can be used to evaluate any system that can return text (i.e., including tool-augmented systems).

To contextualize the scores, we also surveyed more than 35 experts in chemistry on a subset of the benchmark corpus to be able to compare the performance of current frontier models with (human) chemists of different specializations. Our results indicate that current frontier models perform “superhuman” on some aspects of chemistry but, in many cases, including safety-related ones, might be very misleading. We find that there are still numerous limitations for current models to overcome, such that they can be directly applied in autonomous systems for chemists. Moreover, we conclude that carefully created broad benchmarks represent an important stepping stone for progress in this field.

2 Results and Discussion

2.1 Benchmark corpus

To compile our benchmark corpus, we utilized a broad list of sources (see Section 4.1), 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 Section 4.5 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 11 topics. It is clear that a focus of ChemBench (by design) lies on safety-related aspects, which in Figure 3 appear as a large distinct clusters across the embedding space.

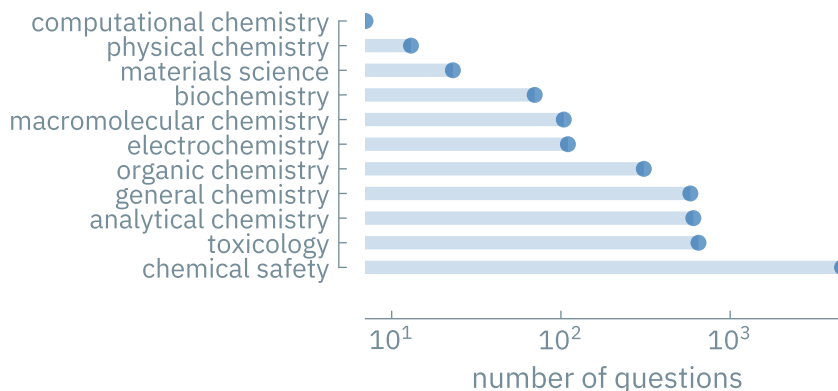


Figure 2: Number of questions for different topics. The topics have been assigned using a combination of a rule-based system (mostly based on the source the question has been sampled from) and a classifier operating on word embeddings of the questions. The figure shows that not all aspects of chemistry 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 (6202 MCQ questions and 857 open-ended questions).

“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**

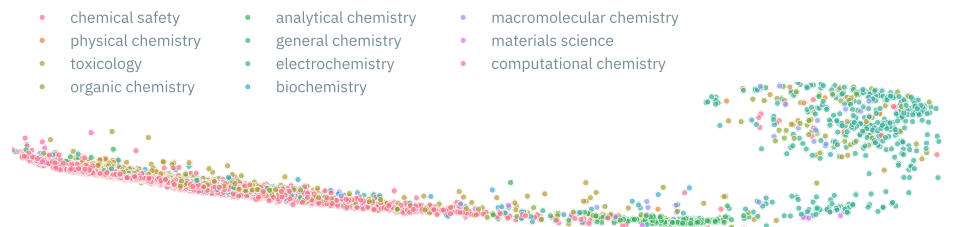


Figure 3: Principal component projection of embeddings of questions in the ChemBench corpus. To obtain this figure, we embedded questions and answers using the BART model^{bart} (using other embeddings, such as of OpenAI’s ada model, leads to qualitatively similar results). We then project the embeddings 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.

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 (209 questions) of the corpus that was curated to be a diverse and representative subset of the full corpus in which topics are more balanced than in the complete corpus (see Section 4.4 for details on the curation process). We also used this subset to seed the web application for the human baseline study.

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 benchmarking 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 aug-

ment 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.

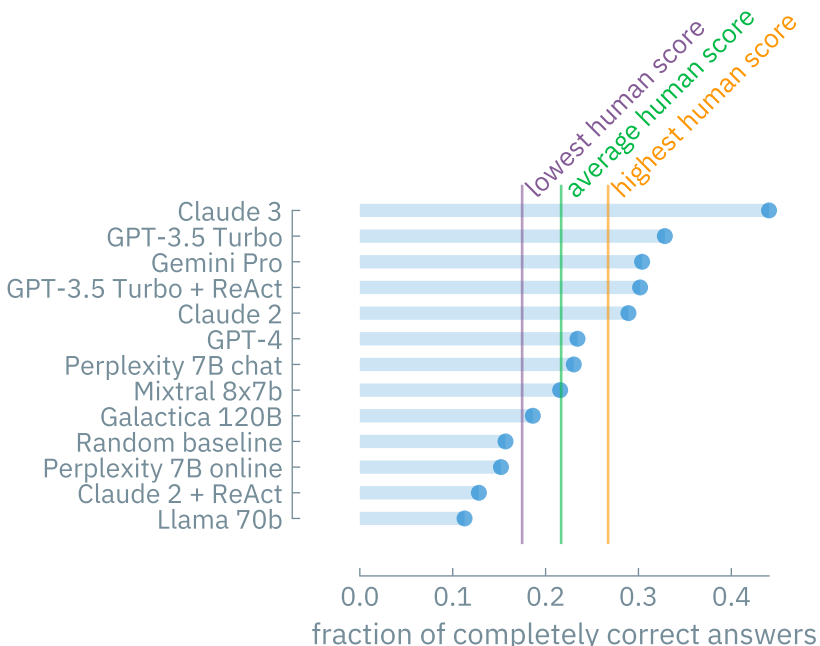


Figure 4: Performance of models and humans on the “tiny” subset of the ChemBench corpus. The figure shows the percentage of questions that the models answered completely correctly. We use horizontal bars to indicate the performance of various models and highlight statistics of the human performance. 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. Figure 10 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.

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 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^{yao2023react}), 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 Section 4.5). 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., 10 percent correct answers for GPT-4) while this question appeared easier (25 percent correct) for trained humans. Importantly, the human experts are given a drawing of

the compounds, whereas models are only shown the SMILES string of a compound and have to use this to reason about the symmetry of the compound (i.e., to identify the number of diastereotopically distinct protons, which requires *reasoning* about the topology and structure of a molecule). These findings also shine an interesting light on the value of textbook-inspired questions. A subset of the questions in the ChemBench are based on textbooks targeted at undergraduate students. On those questions, the models tend to perform better than on some of our semi-automatically constructed tasks (see Figure 14). For instance, while the overall performance in the chemical safety topic is low, the models would pass the certification exam according to the German Chemical Prohibition Ordinance based on a subset of questions we sampled from the corresponding question bank (e.g., 71% correct answers for GPT-4, 67% for Claude 3, and 9% for the human experts). While those findings are impacted by the subset of questions we sampled, the results still highlight that good performance on such question bank or textbook questions does not necessarily translate to good performance on other questions that require more reasoning.

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 XX). 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](#), [golovneva2023](#)

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.

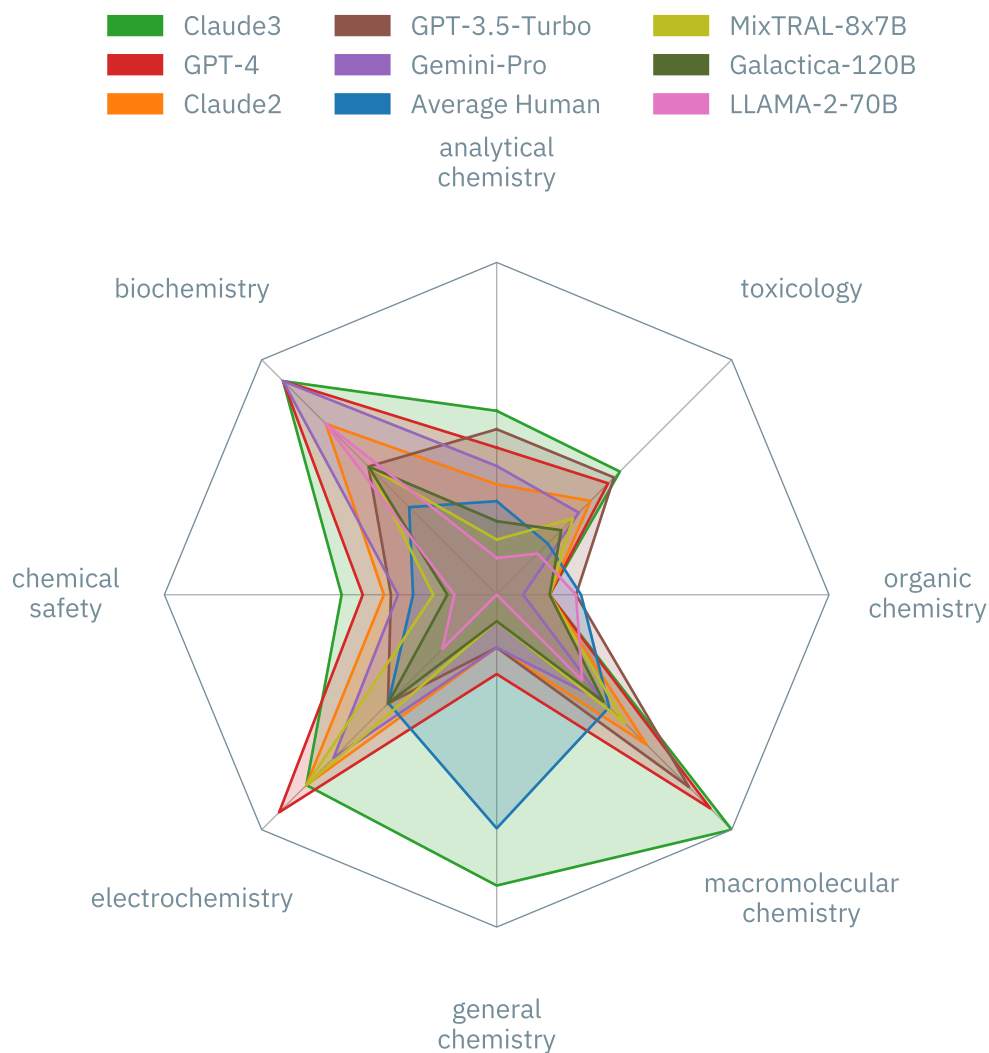


Figure 5: Performance of the models and humans on the different topics of the “tiny” subset of the ChemBench corpus. The radar plot shows the performance of the models and humans on the different topics of “tiny” subset of the ChemBench corpus. The performance is measured as the fraction of questions 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 Figure 14.

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.

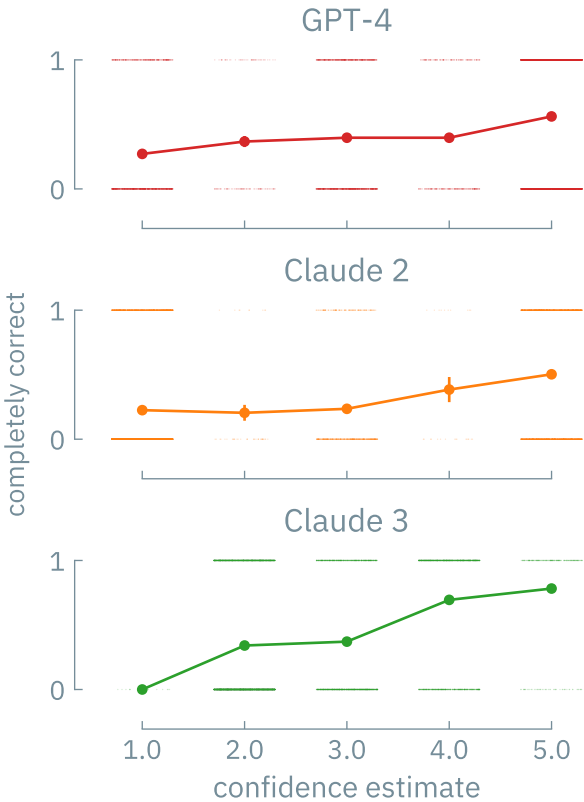


Figure 6: Relationship between confidence of the model in the answer and correctness. For this analysis, we used verbalized confidence estimates from the model. We prompted the models to return a confidence score on an ordinal scale to obtain those estimates. We then plot the correctness of the answers (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 3.97 (on a scale of 1–5) for the 120 questions it answered correctly and 3.57 for the 667 questions it answered incorrectly. While, on average, the verbalized confidence estimates from Claude 3 seem better calibrated (Figure 6), they are still misleading for the questions about Globally Harmonized System of Classification and Labelling of Chemicals (GHS) pictograms with an average score of 2.39 for correct answers and 2.34 for incorrect answers.

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 the development of better human-model interaction frameworks, given that models are not able to 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.

4 Methods

4.1 Curation workflow

For our dataset, we curated questions from existing exams or exercise sheets but also programmatically created new questions. Questions were added via Pull Requests on our GitHub repository and only merged into the corpus after passing manual review (Figure 7) as well as automated checks (e.g., for compliance with a standardized schema).

To ensure that the questions do not enter a training dataset, we use the same canary string as the BigBench project. This requires that LLM developers filter their training dataset for this canary string.^{openai2024gpt4, srivastava2022beyond}

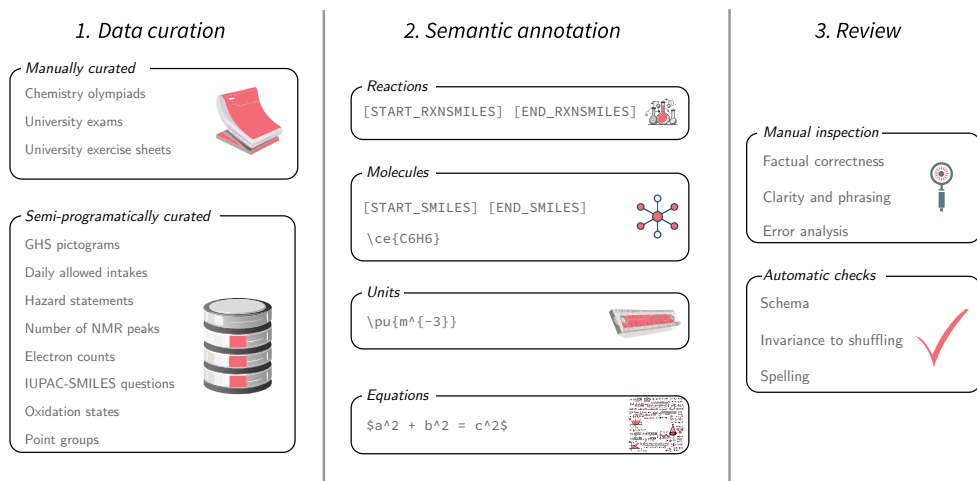


Figure 7: Overview of the workflow for the assembly of the ChemBench corpus. To assemble the ChemBench corpus, we first collected questions from various sources. Some tasks were manually curated, others semi-programmatically. We added semantic annotations for all questions to make them compatible with systems that use special processing for modalities that are not conventional natural text. We reviewed the questions using manual and automatic methods before adding them to the corpus.

Manually curated questions Manually curated questions were sourced from various sources, including university exams, exercises, and question banks. Table 1 provides an overview of the sources of the manually curated questions.

Table 1: Sources of manually curated questions. The table shows the sources and a brief description as well as the number of the manually curated questions.

source	description	question count
NMR spectroscopy	Questions are based on images shared via the Twitter account NMRspectroscopy	5
Analytical chemistry	Questions are based on exercises in master student examination at the FSU Jena, Germany	22
General chemistry	Questions originate from examinations for grade 10 at the Marist Comprehensive Academy Uturu, Nigeria	6
	Questions originate from the Bachelor of Science (for chemistry) courses at the TUM, Germany	2 + 5
Functional materials and nanomaterials	Questions are based on exercises in a seminar conducted at the FSU Jena, Germany	8
Combustion engineering	Based on a Master of Science examination paper at the University of Magdeburg, Germany	1
Materials' synthesis	Questions are based on seminars on material synthesis from FSU Jena, Germany	11
Chemistry olympiad	National and International Olympiads held in US, UK and Moldova	193
Organic reactivity	Based on exams in the Bachelor of Science in Chemistry at the TUM, Germany	34
Periodic table	Manually created	15
Polymer chemistry	Based on examinations at the University of Hamburg, Germany	17
Textbook questions	Various textbooks covering biomolecular science, drug synthesis, molecular structure, organic chemistry, and X-ray crystallography	202

Safety	Based on the question bank published by the Federal/State Working Group on Chemical Safety (BLAC), which is used for the expertise examination (“Sachkundeprüfung”)	90
	Based on toxicology exams at the LMU Munich, Germany	11
	Based on toxicology exams at the University of Vienna, Austria	19
	Based on toxicology exams at WWU Munster, Germany	17
	Based on exercises for lectures in high-energy density materials at LMU Munich, Germany	15
	Based on pharmacology exams at the University of Vienna, Austria	18
	Lab safety quizzes based on various sources	3 + 5 + 55 + 14 + 10

Semi-programmatically generated questions In addition to the manually curated questions, we also generated questions programmatically. An overview of the sources of the semi-programmatically generated questions is provided in Table 2.

Table 2: Sources of semi-programmatically generated questions. The table shows the sources and a brief description as well as the number of the semi-programmatically generated questions.

source	description	question count
Number of isomers	MAYGEN ^{Yirik_2021} was used to compute the number of isomers for a set of SMILES extracted from the ZINC dataset ^{Irwin_2012}	24
Total electron count of molecules	Electron counts based on the data from https://www.cheminfo.org/	25
Oxidation states	Oxidation states questions based on the data from https://www.cheminfo.org/	156

Chemical reactivity	Questions are framed based on the information from the Cameo Chemicals website	652
Number of NMR signals	Molecules are sampled from the ZINC database ^{Irwin_2012} , OpenChemLib ^{openchemlib} is used to compute the number of diastereotopically distinct hydrogen atoms	520
Point group of molecules	Our ChemCaption tool is used to assign the point group using spglib ^{spglib} and then each case was manually checked to select well-defined cases	16
IUPAC-SMILES pairs	Sampled from the PubChem ^{pubchem} database	175 + 175
PubChem ^{pubchem} safety data	Daily allowable intakes according to the World Health Organization	101
	Definitions of hazard statements	88
	GHS classification of chemicals mined through the API	787
Safety	Materials' compatibility	3196
	Chemical compatibility	297

4.2 Model evaluation workflow

Prompting We employ distinct prompt templates tailored for completion and instruction-tuned models to maintain consistency with the training. We impose constraints on the models within these templates to receive responses in a specific format so that robust, fair, and consistent parsing can be performed, as explained below. Certain models are trained with special annotations and \LaTeX syntax for scientific notations, chemical reactions, or symbols embedded within the text. For example, all the SMILES representations are encapsulated within `[START_SMILES]` `[\END_SMILES]` in Galactica^{taylor2022galactica}. Our prompting strategy consistently adheres to these details in a model-specific manner by post-processing \LaTeX syntax, chemical symbols, chemical equations, and physical units (by either adding or removing wrappers). This step can be easily customized in our codebase.

Parsing Our parsing workflow is 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 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 using regular expressions. If these hard-coded parsing steps fail, we use a LLM, e.g., Claude 2, to parse the completion.

Manual validation of parsing performance As stated, we tried to account for the variation in outputs with custom regular expressions. We selected a large, diverse subset of questions (10 per topic for all model reports) and manually investigated where the parsed output does not match the actual answer intended by the model. We found that for MCQ questions, the parsing was accurate in 99.76% of the cases, while for floating point questions, the parsing was accurate in 99.17% of the cases. The models generating errors are pplx-7b-chat and Mixtral-8x7b, which in the specified fraction of cases fail to follow the prompt.

Models

Completion models We used Galactica (120b)^{taylor2022galactica} with the default settings.

Instruction-tuned models In addition, we used Claude 2, Claude3 (Opus),^{anthropicClaudeModelFamily2024} GPT-4,^{openai2024gpt4} GPT-3.5-turbo,^{brown2020language} Gemini Pro,^{gemini} Mixtral-8x7b,^{jiang2024mixtral} Llama2 (70b),^{touvron2023llama} as well as the 7B chat model from Perplexity.AI.

Tool augmented models In addition to directly prompting LLMs, we also investigated the performance of tool-augmented systems. For this, we investigated the 7B parameter “online” model of Perplexity.AI. In addition, we utilized gpt-3.5-turbo as well as Claude 2, with ReAct-style tool augmentation.^{yao2023react} The latter two models had access to WolframAlpha, the ArXiv API, a Python interpreter, and web search (using DuckDuckGo). We implemented the systems using Langchain with the default prompts and constrained the system to a maximum of ten LLM calls.

4.3 Confidence estimate

To estimate the models’ confidence, we prompted them with the question (and answer options for MCQ) and the task to rate their confidence to produce the correct answer on a scale from 1 to 5. We decided to use verbalized confidence estimates^{xiong2023llms} since we found those closer to current practical use cases than other prompting strategies, which might be more suitable when implemented in systems.

4.4 Human baseline

Question selection Since we anticipated that we would not be able to collect enough responses for every question to allow for a meaningful statistical analysis, we decided to show a relevant subset of all questions to the human scorers. For selecting the subset, we decided to address 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 up to 13 questions per source that all LLMs (model names) from an initial scoring round did not answer correctly. In addition, we picked 100 diverse ones using greedy MaxMin sampling on the embeddings on the questions computed using BART. We added five random questions about the number of NMR signals and the point group of molecules.

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 not to use external tools other than a

calculator and 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.

Another aspect of requiring unsupervised question answering is that in real life, humans have tools and can use them to answer any question of interest. In our current study, we prompted users not to use those tools.

Participants Users were open to reporting about their experience in chemistry. Overall, 33 did so. Out of those, 6 reported to have been awarded a Ph.D. 4 are beyond a first postdoc, 18 have a master’s degree, and 5 have a bachelor’s degree.

Comparison with models To compare the performance of humans (who might have answered only some questions) with the performance of models (which answered all questions), we focussed on questions that at least four humans answered and limited the pool of human scorers to those who answered at least 100 questions (i.e., 17 humans). The latter threshold was chosen to limit it to humans who seriously attempted to answer a part of the questions systematically. This analysis might lead to potential biases, most likely in favor of humans, as they were allowed to skip questions. 4 humans answered more than 200 questions. For the analysis, we treated each human as a model. We computed the topic aggregated averages per human for analyses grouped by topic and then averaged over all humans.

4.5 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 the output of sequence classification models. We use this information to make the assignment for questions that can easily be assigned to a topic based on the source (e.g., number of NMR signals, chemical compatibility, toxicology exam questions). For the remaining ones, e.g., from chemistry olympiad questions, we use zero-shot sequence classification^{zeroshotsequence} using the BART model^{bart, FacebookBART}, 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

The code and data for ChemBench is available at <https://github.com/lamalab-org/chem-bench>. The code for the app for our human baseline study is available at

<https://github.com/lamalab-org/chem-bench-app>. To ensure reproducibility, this manuscript was generated using the `showyourwork` framework.^{Luger2021} The code to rebuild the paper (including code for all figures and numbers next to which there is a GitHub icon) can be found at . To facilitate reproduction, some intermediate analysis results are cached at <http://dx.doi.org/10.5072/zenodo.34706>.

Acknowledgements

This work was supported by the Carl Zeiss Foundation, a “Talent Fund” of the “Life” profile line of the Friedrich Schiller University Jena and the FAIRmat consortium. We also thank Stability.AI for the access to its HPC cluster and donations.

M.A. expresses gratitude to the European Research Council (ERC) for evaluating the project with the reference number 101106377 titled “CLARIFIER” and accepting it for funding under the HORIZON TMA MSCA Postdoctoral Fellowships - European Fellowships. Furthermore, M.A. acknowledges the funding provided by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (Grant Reference: EP/Y023447/1; Organization Reference: 101106377).

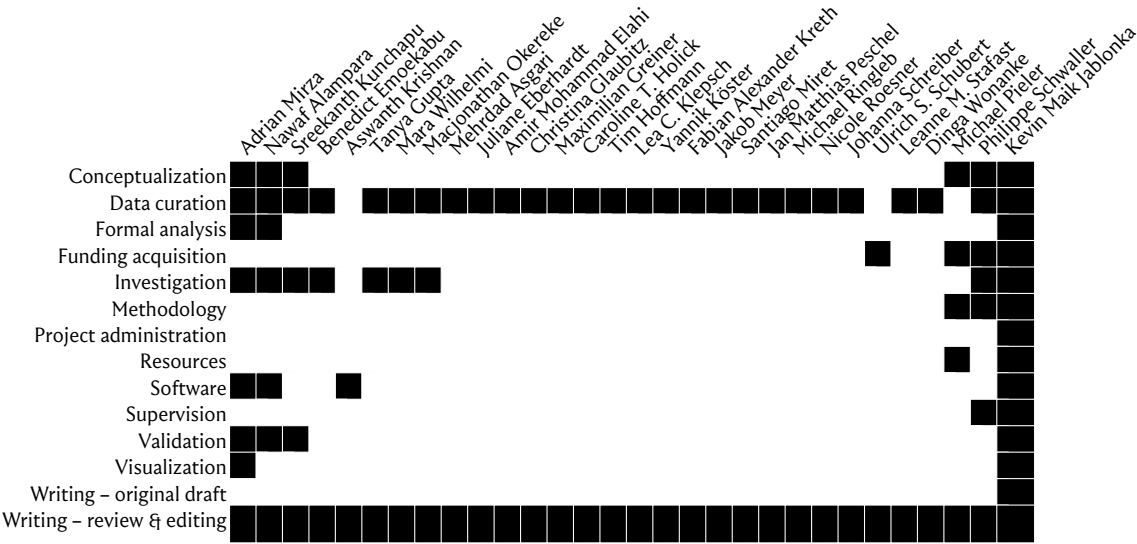
M.R. and U.S.S. thanks the “Deutsche Forschungsgemeinschaft” for funding under the regime of the priority programme SPP 2363 “Utilization and Development of Machine Learning for Molecular Applications – Molecular Machine Learning” (SCHU 1229/63-1; project number 497115849).

In addition, we thank the OpenBioML.org community and their ChemNLP project team for valuable discussions. Moreover, we thank Pepe Márquez for discussions and support and Julian Kimmig for feedback on the web app. In addition, we acknowledge support from Sandeep Kumar with an initial prototype of the web app. We thank Bastian Rieck for developing the \LaTeX -credit package (<https://github.com/Pseudomanifold/latex-credits>).

Conflicts of interest

K.M.J. is a paid consultant for OpenAI (as part of the red teaming network). M.P. is an employee of Stability.AI, and A.M. and N.A. are paid contractors of Stability.AI.

Author contributions



A Appendix

A.1 Desired properties of a chemistry benchmark

- *End-to-end automation.* For model development, the evaluations must be run many times (e.g., on regular intervals of a training run). Approaches that rely on humans scoring the answers of a system^{Schulze_Balhorn_2024, ai4science2023impact, castro2023large} can thus not be used.
- *Careful validation by experts.* Manual curation is needed to minimize the number of incorrect or unanswerable questions.^{northcutt2021pervasive} This is motivated by the observation that many widely used benchmarks are plagued by noisiness.^{Frye_2023, Awg}
- *Usable with models that support special treatment of molecules.* Some models, such as Galactica^{taylor2022galactica}, use special tokenization or encoding procedures for molecules or equations. The benchmark system must encode the semantic meaning of various parts of the question or answer to support this.
- *Usable with black box systems.* Many relevant systems do not provide access to model weights or 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 APIs or code executors.^{schick2024toolformer, karpas2022mrkl, yao2022react} 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 MCQs.* In real-world chemistry, as well as higher-level university education, multiple-choice questions are seldom utilized. Yet, most benchmarking frameworks focus on the MCQ setting because of the ease of evaluation. Realistic evaluations must measure capabilities beyond answering MCQ.
- *Cover a diverse set of topics.* Chemistry, as the “central science”, bridges multiple disciplines.^{Aspuru_Guzik_2018} To even just approximate “chemistry capabilities” the topics covered by a chemistry benchmark must be very diverse.

A.2 Related work

Existing benchmarks such as those from **guo2023large**, **sun2023scieval**, **Schulze_Balhorn_2024**, **Cai_2024** fail to comply with most of the requirements stipulated above. While these benchmarks could provide valuable insights in the short term, they cannot follow the rapid additions to the LLM space. ChemBench aims to correct this through a set of developments: compatibility with BigBench, end-to-end automation, a particular focus on chemical safety, employment of diverse prompting strategies, and specialized

notation for molecules and mathematical symbols. Moreover, our robust framework, including the platform `chembench.org`, will engage the community in open-source contributions.

A.3 Benchmark corpus

To ensure maximal interoperability with existing benchmarks or tools, we curated the data in an extended form of the widely used BigBench format.^{srivastava2022beyond} This also implies that future baselines can be built on top of our infrastructure if saved in the same format.

Figure 8 shows the distribution of the Flesch-Kincaid reading ease scores of the questions. We see that the questions are generally complex to read.

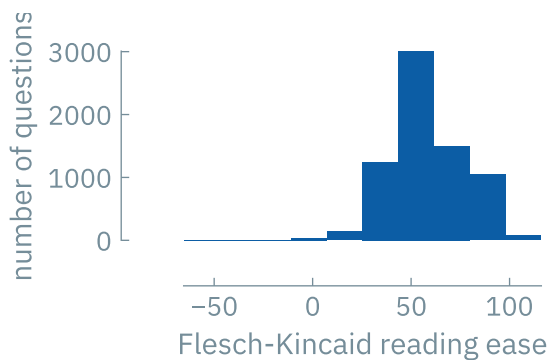


Figure 8: Distribution of Flesch-Kincaid reading ease scores of the questions. The Flesch-Kincaid reading ease score^{flesch1948new} measures how easy a text is to read. It is calculated based on the average number of syllables per word and words per sentence. The higher the score, the easier the text is to read. The distribution of the questions’ scores is shown in the histogram.

Figure 9 shows that most questions in our corpus are MCQ. A substantial fraction, in contrast to other benchmarks, is open-ended.

A.4 Model performance

We also evaluated the model performance on the entire ChemBench corpus. Figure 10 shows the fraction of questions that were answered completely correctly by the models. Note that this ranking differs from the one on the “tiny” subset.

Figure 11 shows the performance of the models on the different topics of the ChemBench corpus. The general pattern of performance varies significantly between the different topics and is also observed when the models are evaluated on the entire corpus. However, since some subjects are composed of questions from different

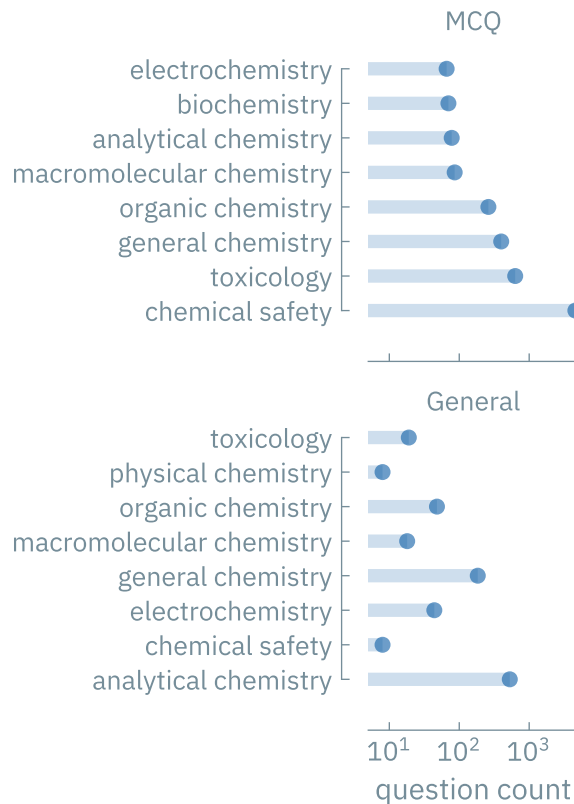


Figure 9: Number of multiple choice questions vs. open-ended questions per topic. The bar plot shows the number of MCQ and general questions per topic.

sources, the ranking of the models is, in some instances, different from the one on the “tiny” subset.

Figure 12 shows this data as a parallel coordinates plot. This visualization highlights the critical observation that the ranking of the models highly depends on the questions they are evaluated on. Only very broad benchmarks have the potential to provide a comprehensive view of a model’s capabilities. However, even in those cases, the weighting of the different topics is crucial. Hence, we believe that fine-grained analysis of model performance is vital for the development of future benchmarks.

To further investigate the performance of the models, we also compared the performance on different data sources. Compared to topics, this is a more fine-grained analysis, as topics can be composed of questions from different sources. In Figure 14, we see that the performance of the models varies significantly between the different data sources. Interestingly, the performance of the models on questions sourced

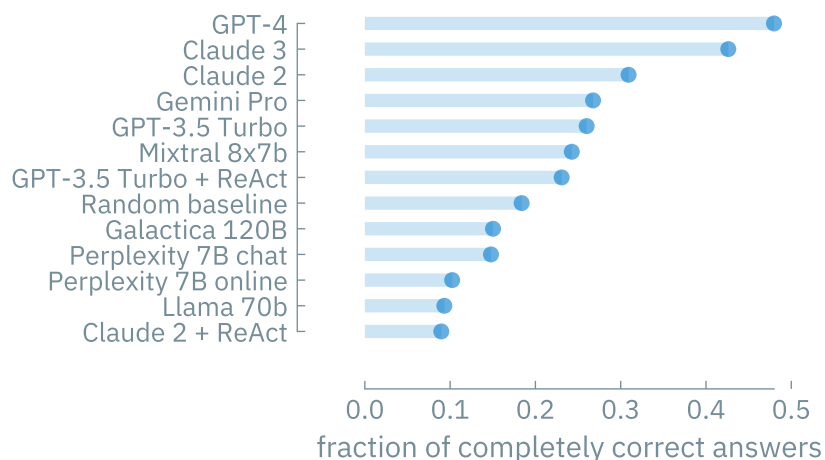


Figure 10: Overall performance of the models on the ChemBench corpus. The bar plot shows the fraction of questions that were answered completely correctly by the models. Scores computed on the entire ChemBench corpus.

based on textbooks seems to be better for our models than some of the semi-programmatically created tasks, such as questions about the number of signals in an NMR spectrum.

Figure 15 shows the same analysis on the “tiny” subset.

One might wonder if questions that are more difficult to parse lead to worse performance of the models. Figure 16 shows no clear correlation between the reading ease of the questions and the performance of the models.

A.5 Performance as a function of molecular features

To better understand if the performance of the models is correlated with specific features of the molecules, we analyzed the performance of the models as a function of the number of atoms and the complexity of the molecules. Figure 17 shows that the performance of the models is not correlated with the complexity of the molecules but rather with the number of atoms (Figure 18). The corresponding Spearman correlation coefficients are listed in ??.

A.6 Influence of scale

To obtain first insights in how the performance LLMs depends on scale, we tested the LLMs of the LLaMA series. Note that such analyses are difficult as models are typically not directly comparable in terms of dataset and training protocol.^{biderman2023pythia}

Figure 20 shows the performance of the LLaMA (chat) models with different parameter counts on the ChemBench corpus.

A.7 Human baseline

App To facilitate the collection of responses, we developed a responsive web application in Typescript using the Next.js^{nextjs} app router framework. This application handles serving the user interface and exposes various Representational State Transfer (REST) APIs for relevant operations. We utilize a MySQL^{mysql} database and Prisma object relational mapping (ORM)^{prisma} for efficient database management. The web application is styled with Tailwind CSS^{tailwindcss} using the shadcn/ui component library and uses NextAuth^{nextauth} for easy and secure user authentication and postMark for sending Emails. The application is hosted on the Vercel web hosting platform.

Statistics Appendix A.7 shows the distribution of scores our human scorers achieved.

We also recorded the time humans took to answer the questions. This time is the time from the question being displayed to the human to the human submitting the answer. Interestingly, we found no significant correlation between the experience of the human scorers and the performance on the questions (Appendix A.7, Spearman’s $\rho \approx 0.14$, and $p \approx 0.44$).

Additionally, we prompted users to provide additional information about their experience in chemistry. While we recorded fine-grained information, e.g., their specialization, we focused on the number of years since the first university-level chemistry course. Appendix A.7 shows that the experience of the human scorers was not significantly correlated with the correctness of their answers (Appendix A.7, Spearman’s $\rho \approx 0.14$, and $p \approx 0.44$).

A.8 Confidence estimates

Since it is important to understand if models can provide an indication of whether their answer might likely be incorrect, we prompted some of our top performing LLMs to return the confidence in providing a correct answer on an ordinal scale. This is similar to the verbalized confidence scores reported by [xiong2023llms](#). Figure 24 plots the distribution of those scores. We find that the models show different distributions of confidence scores, which, for some, are skewed to the extremes.

A.9 Leaderboard

Our leaderboard is based on the toolchain developed for Matbench.^{Dunn_2020} Briefly, the ChemBench pipeline produces standardized files in json format that contributors

can add via pull requests to the ChemBench repository. The Markdown tables and interactive plots are automatically generated and updated on the ChemBench website.

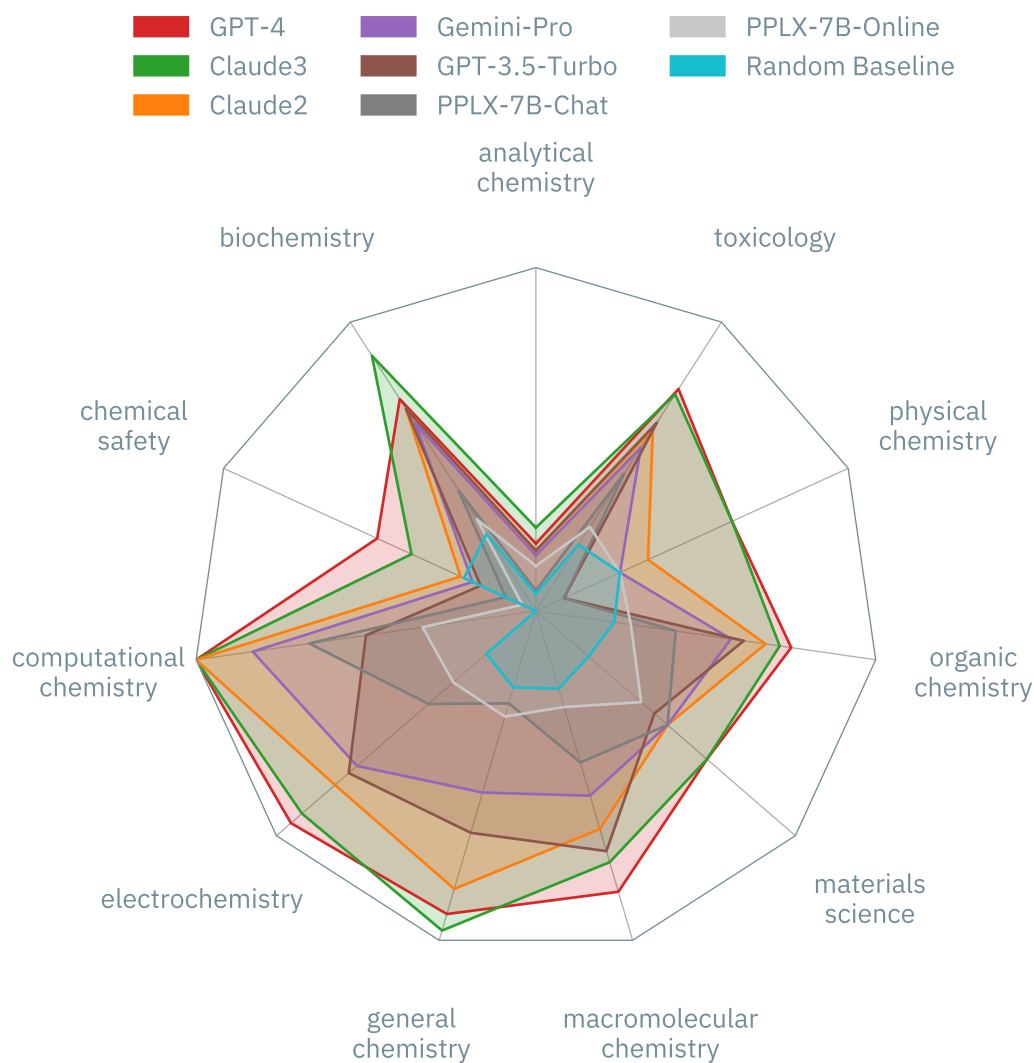


Figure 11: Performance of the models on the different topics of the ChemBench corpus. The radar plot shows the performance of the models on the different topics of the ChemBench corpus. The performance is measured as the fraction of questions answered completely correctly by the models. A score of 1 indicates that all questions were answered completely correctly, while a score of 0 indicates that none were answered completely correctly.

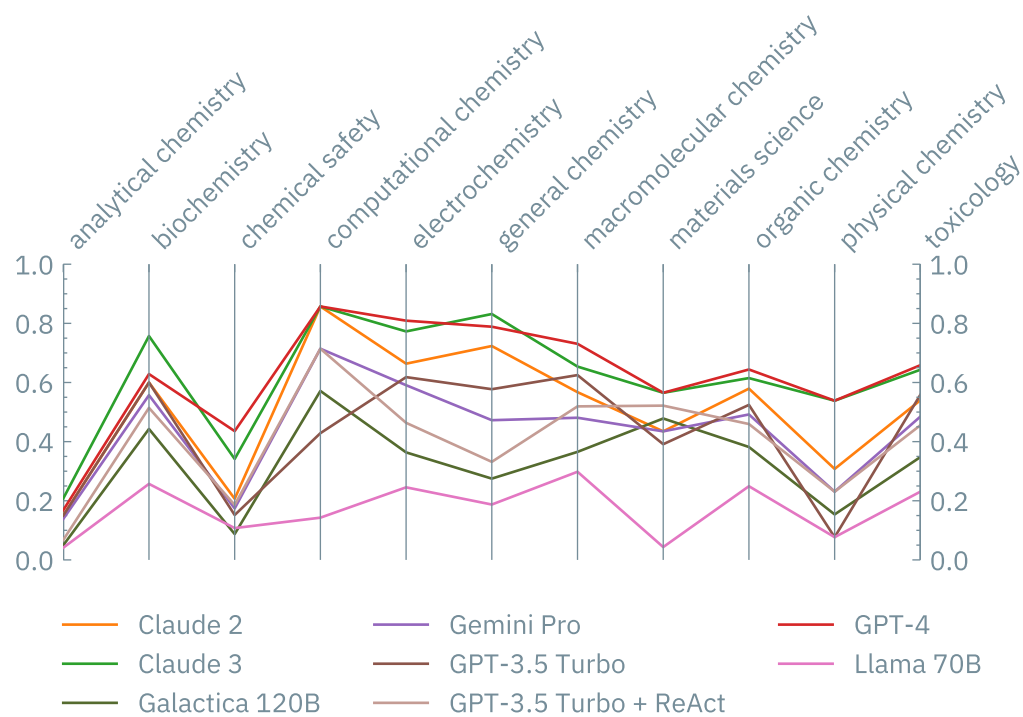


Figure 12: Performance of the models on the different topics of the ChemBench corpus. The parallel coordinates plot shows the performance of the models on the different topics of the ChemBench corpus. The performance is measured as the fraction of questions answered completely correctly by the models.

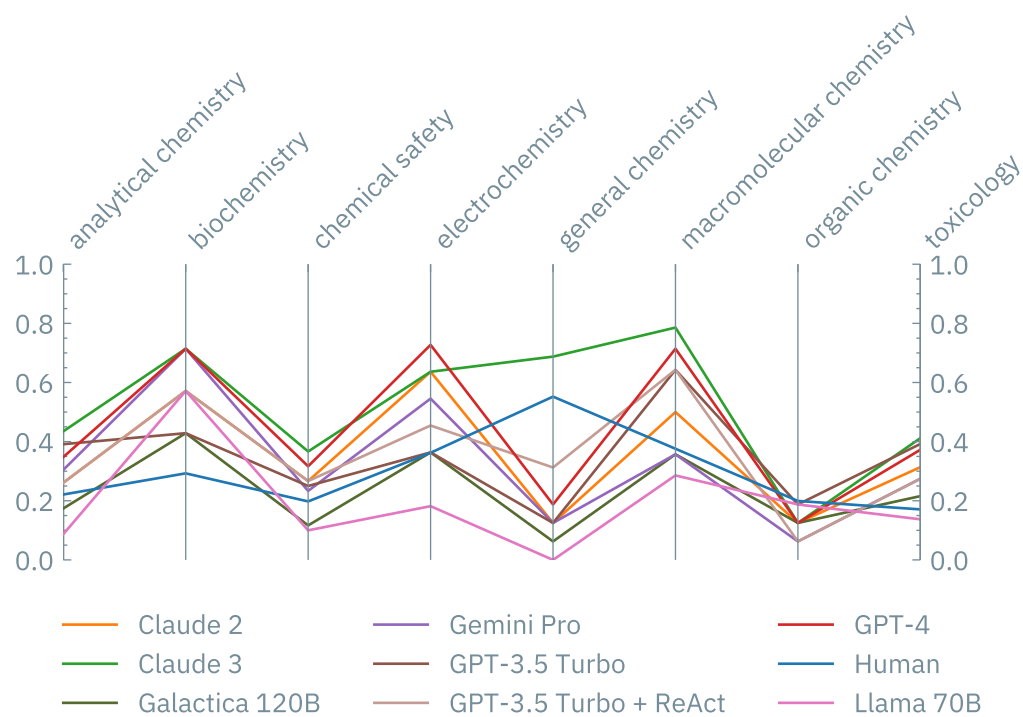


Figure 13: Performance of the models on the different topics of the “tiny” subset. The parallel coordinates plot shows the performance of the models on the different topics of the “tiny” subset. The performance is measured as the fraction of questions answered completely correctly by the models.

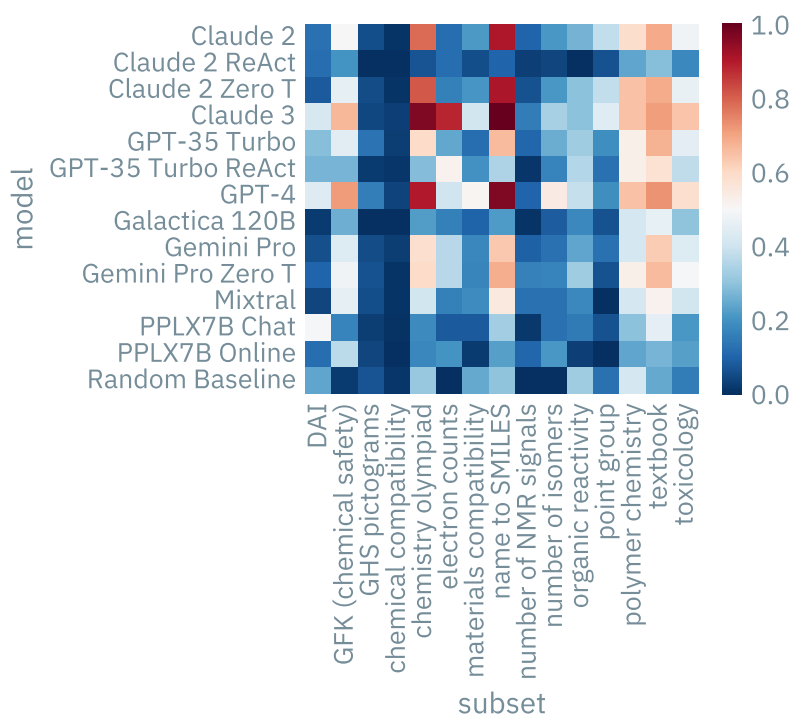


Figure 14: Fraction of completely correctly answered questions per data source. The heatmap shows, in color, the fraction of questions answered completely correctly by different systems for some of our data sources. The performance is measured as the fraction of questions answered completely correctly by the models. A score of one (red) indicates that all questions were answered completely correctly, while a score of zero (blue) indicates that none of the questions were answered completely correctly. We see that the performance of the models varies significantly between the different data sources. For instance, it is interesting to observe that questions sourced based on textbooks seem easier for our leading models than for humans. However, this performance does not correlate with performance on other sources, e.g., semi-programmatically created tasks such as questions about the number of signals in an NMR spectrum.

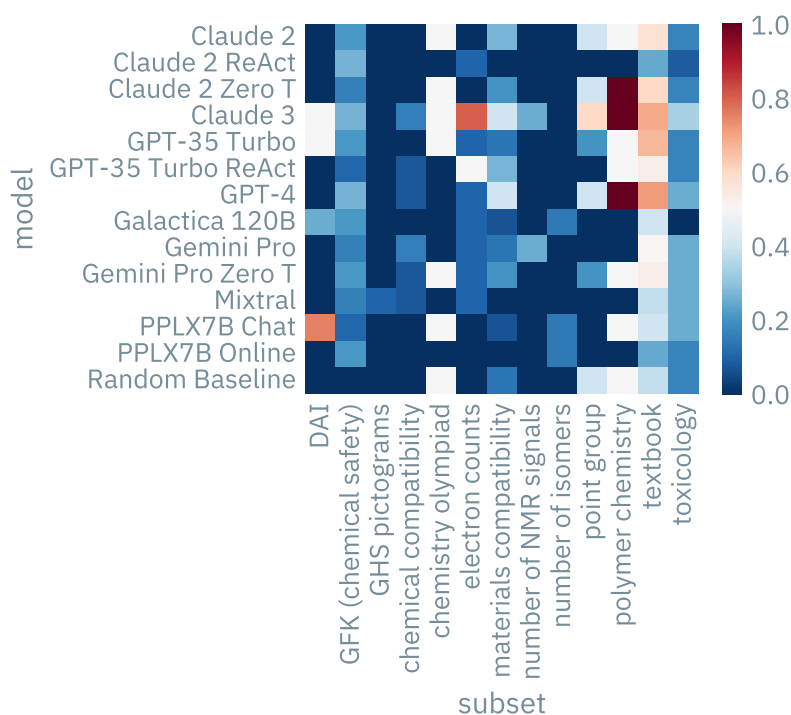


Figure 15: Fraction of completely correctly answered questions per data source on the “tiny” subset. The heatmap shows, in color, the fraction of questions answered completely correctly by different systems for some of our data sources. The performance is measured as the fraction of questions answered completely correctly by the models. A score of one (red) indicates that all questions were answered completely correctly, while a score of zero (blue) indicates that none were answered completely correctly. We see that the performance of the models varies significantly between the different data sources. For instance, it is interesting to observe that questions sourced based on textbooks seem easier for the leading models than for humans. However, this performance does not correlate with performance on other sources, e.g., semi-programmatically created tasks such as questions about the number of signals in an NMR spectrum.

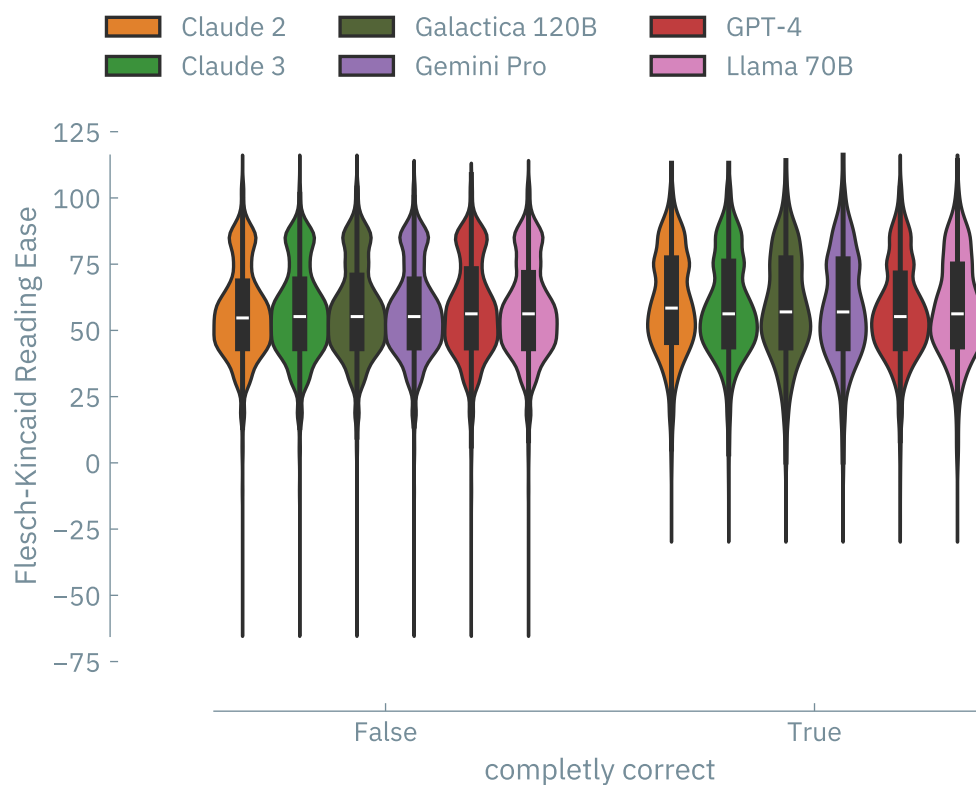


Figure 16: Model performance as a function of reading ease. The violin plots show the distribution of reading ease scores for questions answered completely correctly and those not. We do not observe a clear correlation between the reading ease of the questions and the performance of the models.

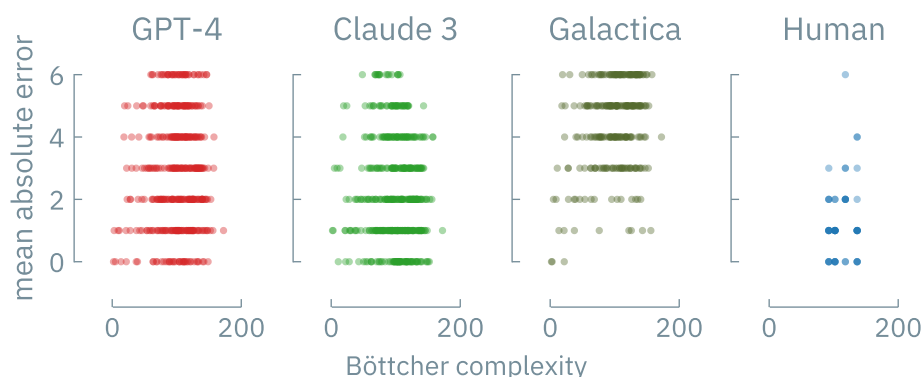


Figure 17: Dependence of the mean absolute error in predicting the number of NMR signals on the Böttcher complexity of the molecules. The complexity measure proposed by [Böttcher_2016](#) is an information-theoretic additive measure of compound complexity that follows chemical intuitions. The plot shows that for the LLMs, the predictive performance (measured as the mean absolute error in the prediction of the number of NMR signals) is not correlated with the complexity of the molecules. For inference based on reasoning, one would expect that the complexity of the molecule is a good predictor of the difficulty of the question.

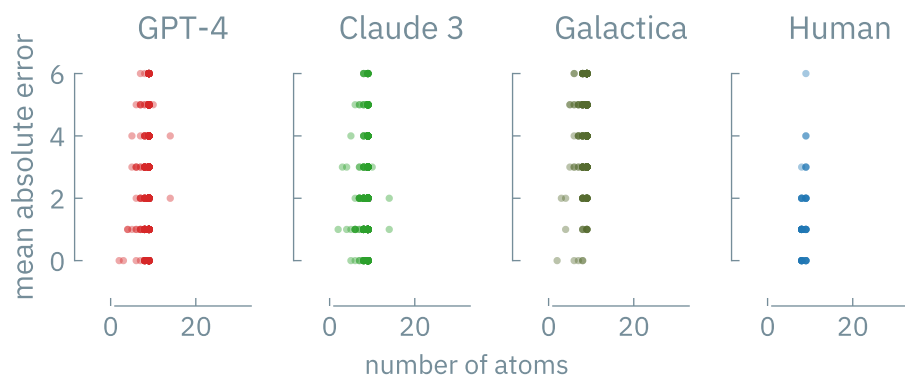


Figure 18: Dependence of the mean absolute error in predicting the number of NMR signals on the number of atoms. The plot shows that for the LLMs, the predictive performance (measured as the mean absolute error in the prediction of the number of NMR signals) is correlated with the number of atoms in the molecule. For reasoning-based inference, one would expect that the number of atoms in the molecules is not necessarily a good predictor, and certainly worse than complexity measures, of the difficulty of the question.

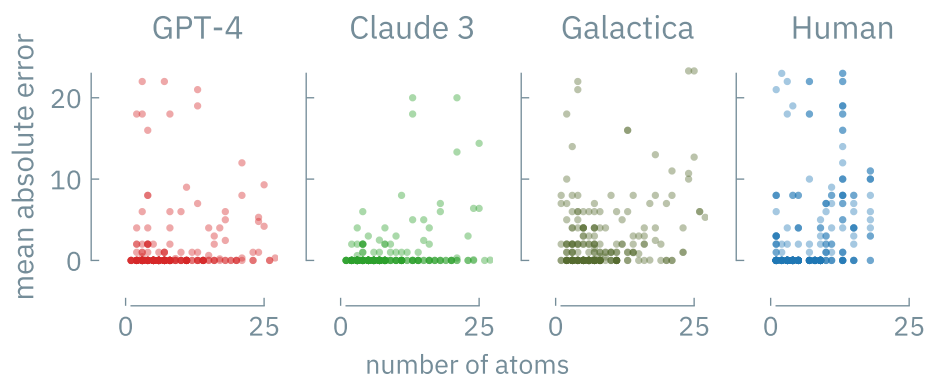


Figure 19: Dependence of the mean absolute error in predicting total electron counts on the number of atoms. The plot shows that for the LLMs, the predictive performance (measured as the mean absolute error in the prediction of the total electron counts) is correlated with the number of atoms in the molecule.

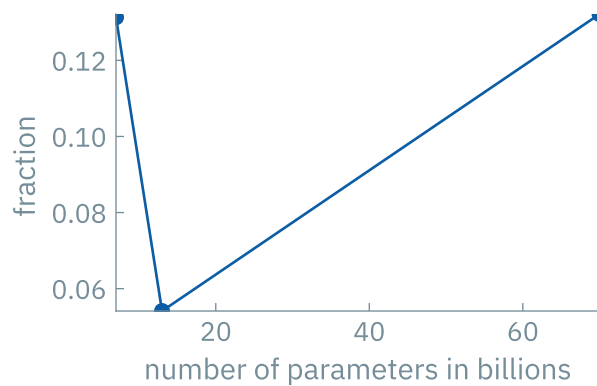


Figure 20: Performance of LLaMA models on the ChemBench corpus. The plot shows the fraction of correctly answered questions as a function of the model size.

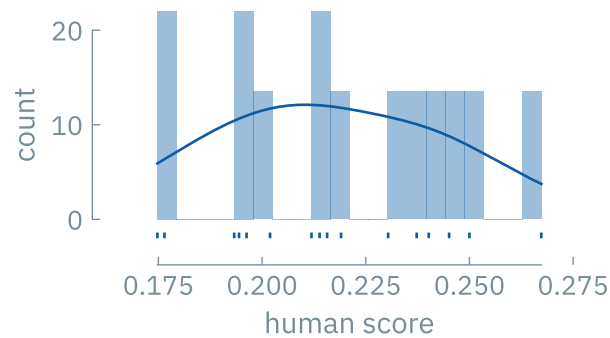


Figure 21: Distribution of human scores. The histogram and kernel density estimates show the fraction of questions answered completely correctly. Since the best possible score for each question is one and the worst possible score is zero, the values on this plot are between zero and one.

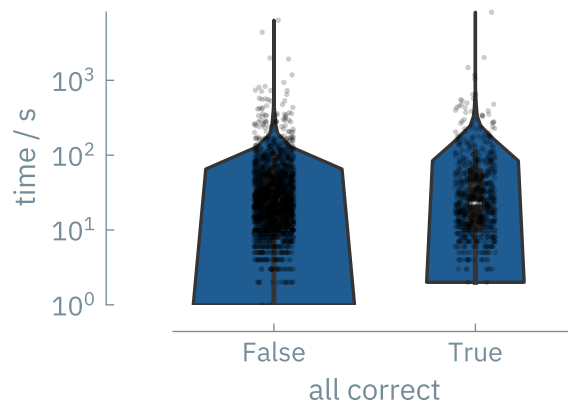


Figure 22: Time taken by human scorers to answer questions vs. correctness of their answers. From the plot, it is clear that there is no clear dependence of the correctness of the answers on the time taken by the human scorers to answer the questions.

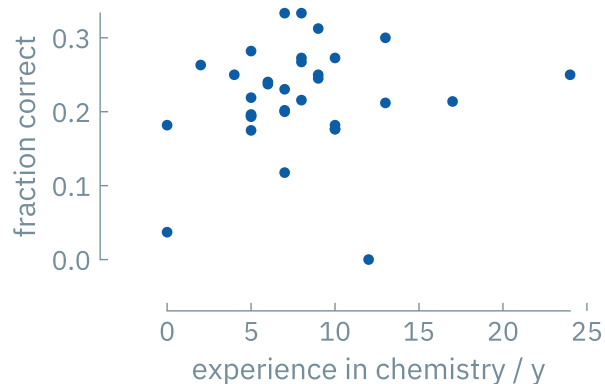


Figure 23: Experience of human scorers vs. correctness of their answers. The experience (in the number of years since the first university-level chemistry course) of the human scorers was not significantly correlated with the correctness of their answers.

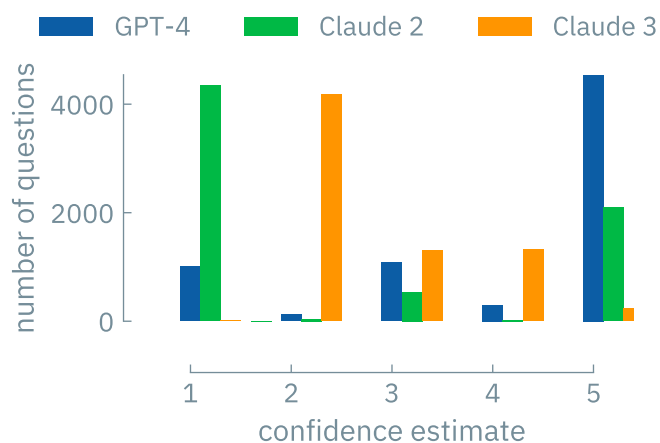


Figure 24: Distribution of confidence scores reported by LLMs. LLMs show different distributions of confidence scores. The confidence scores are reported on an ordinal scale from 1 to 5, with 1 indicating low confidence and 5 indicating high confidence. The bar plots show how many questions were answered with each confidence score.

Acronyms

AGI artificial general intelligence.

API application programming interface.

GHS Globally Harmonized System of Classification and Labelling of Chemicals.

HELM Holistic Evaluation of Language Models.

LLM large language model.

MCQ multiple-choice question.

ML machine learning.

NMR Nuclear Magnetic Resonance.

ORM object relational mapping.

PCA Principal Component Analysis.

REST Representational State Transfer.

SMILES Simplified Molecular Input Line-Entry System.