nature chemistry

Article

https://doi.org/10.1038/s41557-025-01815-x

A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists

Received: 1 April 2024

Accepted: 26 March 2025

Published online: 20 May 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Large language models (LLMs) have gained widespread interest owing to their ability to process human language and perform tasks on which they have not been explicitly trained. However, we possess only a limited systematic understanding of the chemical capabilities of LLMs, which would be required to improve models and mitigate potential harm. Here we introduce ChemBench, an automated framework for evaluating the chemical knowledge and reasoning abilities of state-of-the-art LLMs against the expertise of chemists. We curated more than 2,700 question–answer pairs, evaluated leading open- and closed-source LLMs and found that the best models, on average, outperformed the best human chemists in our study. However, the models struggle with some basic tasks and provide overconfident predictions. These findings reveal LLMs' impressive chemical capabilities while emphasizing the need for further research to improve their safety and usefulness. They also suggest adapting chemistry education and show the value of benchmarking frameworks for evaluating LLMs in specific domains.

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[1,2], with the leading models now being able to pass the US Medical Licensing Examination[3] or other professional licensing exams. They also have been shown to design and autonomously perform chemical reactions when augmented with external tools such as web search and synthesis planners[4–7]. While some see 'sparks of artificial general intelligence (AGI)' in them[8], others see them as 'stochastic parrots'—that is, systems that only regurgitate what they have been trained on[9] and that show inherent limitations owing to the way they are trained[10]. 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[11–13].

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'[14]. This statement is motivated by a growing number of reports that use LLMs to predict properties of molecules or materials[2,15–19], optimize reactions[20,21], generate materials[22–25], extract information[26–33] or even to prototype systems that can autonomously perform experiments in the physical world based on commands provided in natural language[5–7].

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[34]. For instance, most insights in chemical research do not directly originate from data stored in databases but rather from the scientists interpreting the 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

e-mail: mail@kjablonka.com

https://doi.org/10.1038/s41557-025-01815-x

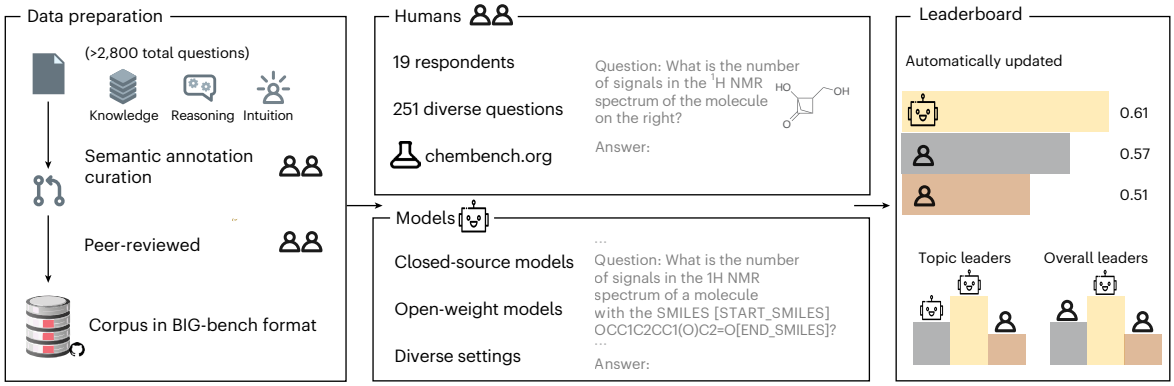


Fig. 1 | Overview of the ChemBench framework. The different components of the ChemBench framework. The framework's foundation is the benchmark corpus comprising thousands of questions and answers that we manually or semi-automatically compiled from various sources in a format based in the one introduced in the BIG-bench benchmark (Extended Data Fig. 1). Questions are classified on the basis of topics, required skills (reasoning, calculation, knowledge and intuition) and difficulty levels. 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 (Supplementary Note 15), which we propose as a foundation for evaluating future models.

copilot systems for chemists that can provide answers to questions or even suggest new experiments on the basis of vastly more information than a human could ever read.

However, the rapid increase in capabilities of chemical ML models led (even before the recent interest in LLMs) to concerns about the potential for the dual use of these technologies, for example, for the design of chemical weapons[35–40]. 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 (even though the synthesis would still require access to controlled physical resources and facilities). Still, it is essential to realize that the user base of LLMs is broader than that of chemistry and materials science experts who can critically reflect on every output these models produce. For example, many students frequently consult these tools—perhaps even to prepare chemical experiments[41]. 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 knowledge and reasoning capabilities are essential as they will determine the capabilities and limitations of their models in their work, for example, in copilot systems for chemists. Unfortunately, apart from exploratory reports, such as by prompting leading models with various scientific questions[13], there is little systematic evidence on how LLMs perform compared with 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[42] or the LM Eval Harness[43]. Among 204 tasks (such as linguistic puzzles), the former contains only 2 tasks classified as 'chemistry related', whereas the latter contains no specific chemistry tasks. Owing to the lack of widely accepted standard benchmarks, the developers of chemical language models[16,44–47] frequently utilize language-interfaced[48] tabular datasets such as the ones reported in MoleculeNet[49,50], Therapeutic Data Commons[51], safety databases[52] or MatBench[53]. In these cases, the models are evaluated on predicting very specific properties of molecules (for example, 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[54,55] or automatic text mining[56–58] have been proposed, none of them have been widely accepted. This is probably because they cannot automatically be used with black box (or tool-augmented) systems, do not cover a wide range of topics and skills 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[49].

In this work, we report a benchmarking framework (Fig. 1), which we call ChemBench, and use it to reveal the limitations of current frontier models for use in the chemical sciences. Our benchmark consists of 2,788 question–answer pairs compiled from diverse sources (1,039 manually generated and 1,749 semi-automatically generated). Our corpus measures reasoning, knowledge and intuition across 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 (that is, including tool-augmented systems).

To contextualize the scores, we also surveyed 19 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. In parts of the survey, the volunteers were also allowed to use tools, such as web search, to create a realistic setting.

Results and discussion

Benchmark corpus

To compile our benchmark corpus, we utilized a broad list of sources (Methods), ranging from completely novel, manually crafted questions over 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 two scientists in addition to the original curator and automated checks. Importantly, our large pool of questions encompasses a wide range of topics and question types (Fig. 2). The topics range from general chemistry to more specialized fields such as inorganic, analytical or technical chemistry. We also classify the questions on the basis of what skills are required to answer them. Here, we distinguish between questions that require knowledge, reasoning, calculation, intuition or a combination of these. Moreover, the annotator also classifies the questions by difficulty to allow for a more nuanced evaluation of the models' capabilities.

While many existing benchmarks are designed around multiple-choice questions (MCQ), this does not reflect the reality of chemistry education and research. For this reason, ChemBench samples both

https://doi.org/10.1038/s41557-025-01815-x

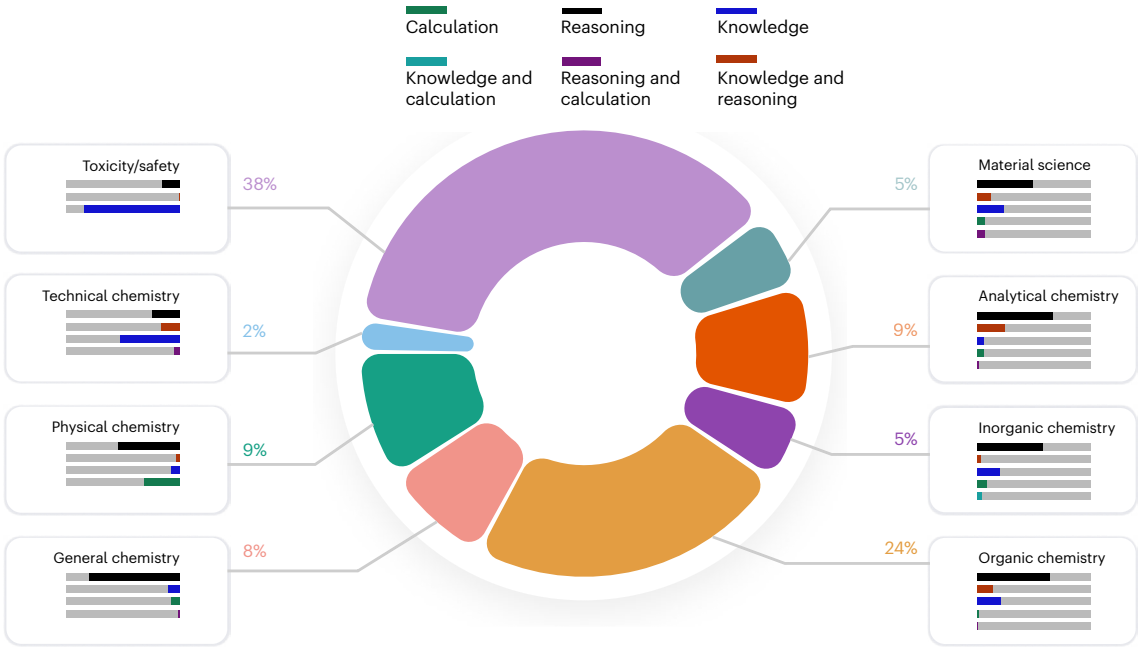


Fig. 2 | Distribution of topics and required skills. The distribution of questions across various chemistry topics, along with the primary skills required to address them. The topics were manually classified, showing a varied representation across different aspects of chemistry. Each topic is associated with a combination of three key skills: calculation, reasoning and knowledge, as indicated by the coloured bars. ChemBench samples encompass diverse topics and diverse skills, setting a high bar for LLMs to demonstrate human-competitive performance across a wide range of chemistry tasks.

MCQ and open-ended questions (2,544 MCQ and 244 open-ended questions). In addition, ChemBench samples different skills on various difficulty levels: from basic knowledge questions (as knowledge underpins reasoning processes[59,60]) to complex reasoning tasks (such as finding out which ions are in a sample given a description of observations). We also include questions about chemical intuition, as demonstrating human-aligned preferences is relevant for applications, such as hypothesis generation or optimization tasks[61].

ChemBench-Mini. It is important to note that a smaller subset of the corpus might be more practical for routine evaluations[62]. For instance, Liang et al.[63] report costs of more than US$10,000 for application programming interface (API) calls for a single evaluation on the widely used Holistic Evaluation of Language Models benchmark. To address this, we also provide a subset (ChemBench-Mini, 236 questions) of the corpus that was curated to be a diverse and representative subset of the full corpus. While it is impossible to comprehensively represent the full corpus in a subset, we aimed to include a maximally diverse set of questions and a more balanced distribution of topics and skills (see Methods for details on the curation process). Our human volunteers answered all the questions in this subset.

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[64] uses special 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 (for example, chemicals, units or equations) 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 LLM 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[65–67]. In those cases, the LLM which returns the probabilities for various tokens (that is, text fragments) represents only one component and it is not clear how to interpret those probabilities in the context of the entire 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[68].

Overall system performance. To understand the current capabilities of LLMs in the chemical sciences, we evaluated a wide range of leading models[69] on the ChemBench corpus, including systems augmented with external tools. An overview of the results of this evaluation is presented in Fig. 3 (all results can be found in Supplementary Fig. 4 and Supplementary Table 5). In Fig. 3, we show the percentage of questions that the models answered correctly. Moreover, we show the worst, best and average performance of the 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, o1-preview, outperforms the best human in our study in this overall metric by almost a factor of two. Many other models also outperform the average human performance. Interestingly, Llama-3.1-405B-Instruct shows performance that is close to the leading proprietary models, indicating that new open-source models can also be competitive with the best proprietary models in chemical settings.

Notably, we find that models are still limited in their ability to answer knowledge-intensive questions (Supplementary Table 5); that is, they did not memorize the relevant facts. Our results indicate that this is not a limitation that could be overcome by simple application of retrieval augmented generation systems such as PaperQA2. This is probably because the required knowledge cannot easily be accessed via papers (which is the only type of external knowledge PaperQA2 has access to) but rather by lookup in specialized databases (for

https://doi.org/10.1038/s41557-025-01815-x

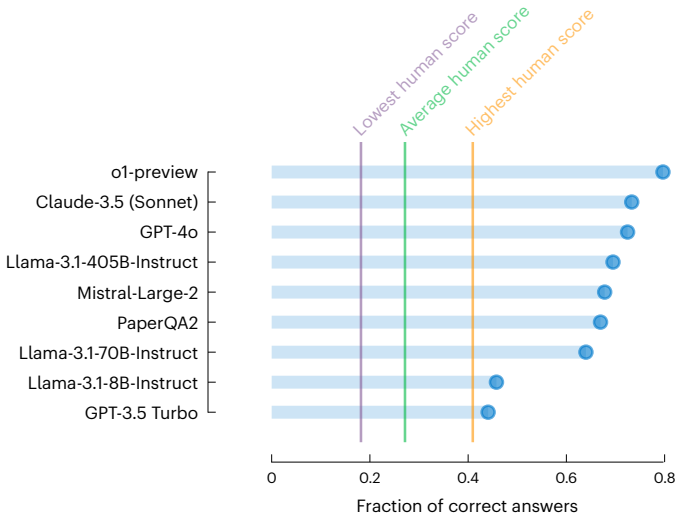


Fig. 3 | Performance of models and humans on ChemBench-Mini. The percentage of questions that the models answered correctly. Horizontal bars indicate the performance of various models and highlight statistics of human performance. The evaluation we use here is very strict as it only considers a question answered correctly or incorrectly, partially correct answers are also considered incorrect. Supplementary Fig. 3 provides an overview of the performance of various models on the entire corpus. PaperQA2 (ref. 33) is an agentic system that can also search the literature to obtain an answer. We find that the best models outperform all humans in our study when averaged over all questions (even though humans had access to tools, such as web search and ChemDraw, for a subset of the questions).

example, PubChem and Gestis), which the humans in our study also used to answer such questions (Supplementary Fig. 17). This indicates that there is still room for improving chemical LLMs by training them on more specialized data sources or integrating them with specialized databases.

In addition, our analysis shows that the performance of models is correlated with their size (Supplementary Fig. 11). This is in line with observations in other domains, but also indicates that chemical LLMs could, to some extent, be further improved by scaling them up.

Performance per topic. To obtain a more detailed understanding of the performance of the models, we also analysed the performance of the models in different subfields of the chemical sciences. For this analysis, we defined a set of topics (Methods) and classified all questions in the ChemBench corpus into these topics. We then computed the percentage of questions that the models or experts answered correctly for each topic and present them in Fig. 4. 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 coloured area indicates a better performance.

One can observe that this performance varies across models and topics. While general and technical chemistry receive relatively high scores for many models, this is not the case for topics such as toxicity and safety or analytical chemistry.

In the subfield of analytical chemistry, the prediction of the number of signals observable in a nuclear magnetic resonance spectrum proved difficult even for the best models (for example, 22% correct answers for o1-preview). Importantly, while the human experts are given a drawing of the compounds, the models are only shown the SMILES string of a compound and have to use this to reason about the symmetry of the compound (that is, to identify the number of diasterotopically 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 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 (Supplementary Fig. 5). 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 on the basis of a subset of questions we sampled from the corresponding question bank (for example, 71% correct answers for GPT-4, 61% for Claude-3.5 (Sonnet) and 3% 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 or are further away from the training corpus[10]. The findings also underline that such exams might have been a good surrogate for the general performance of skills for humans, but their applicability in the face of systems that can consume vast amounts of data is up for debate.

We also gain insight into the models' struggles with chemical reasoning tasks by examining their performance as a function of molecular descriptors. If the model would answer questions after 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 (Supplementary Note 5). 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 the proximity of the molecules to the training data[10].

It is important to note that the model performance for some topics, however, is slightly 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. Statistics on the frequency of such refusals are presented in Supplementary Table 8. 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 by the tooling ChemBench provides, thanks to which contributors can automatically add new models in an open science fashion.

Judging chemical preference. One interesting finding of recent research is that foundation models can judge interestingness or human preferences in some domains[61,70]. If models could do so for chemical compounds, this would open opportunities for novel optimization

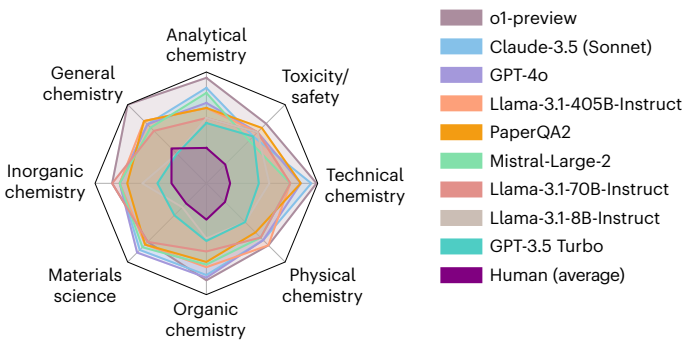


Fig. 4 | Performance of the models and humans on the different topics on ChemBench-Mini. The radar plot shows the performance of the models and humans on the different topics of ChemBench-Mini. Performance is measured as the fraction of questions that were answered correctly by the models. The best score for every dimension is 1 (all questions answered correctly) and the worst is 0 (no question answered correctly). A larger coloured area indicates a better performance. This figure shows the performance on ChemBench-Mini. The performance of models on the entire corpus is presented in Supplementary Fig. 3.

https://doi.org/10.1038/s41557-025-01815-x

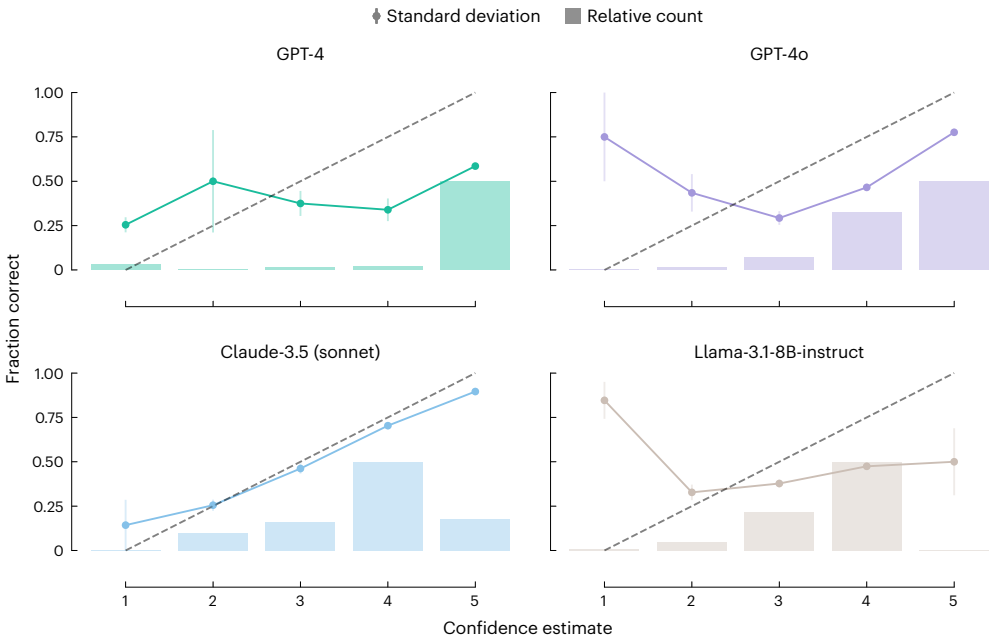


Fig. 5 | Reliability and distribution of confidence estimates. For this analysis, we used verbalized confidence estimates from the model. The models were prompted to return a confidence score on an ordinal scale to obtain those estimates. The line plot shows the average fraction of correctly answered questions for each confidence level. The bar plot shows the distribution of confidence estimates. The error bars indicate the standard deviation for each confidence level (for which the number of samples is given by the height of the bar). A confidence estimate would be well calibrated if the average fraction of correctly answered questions increases with the confidence level. The dashed black line indicates this ideal behaviour, which would be monotonically increasing correctness with higher levels of confidence. We use colours to distinguish the different models, as indicated in the titles of the subplots. We find that most models are not well calibrated and provide misleading confidence estimates.

approaches. Such open-ended tasks, however, depend on an external observer defining what interestingness is[71]. Here, we posed models the same question that Choung et al.[72] asked chemists at a drug company: 'which of the two compounds do you prefer?' (in the context of an early virtual screening campaign setting; see Supplementary Table 2 for an example). Despite chemists demonstrating a reasonable level of inter-rater agreement, our models largely fail to align with expert chemists' preferences. Their performance is often indistinguishable from random guessing, even though these same models excel in other tasks in ChemBench (Supplementary Table 5). This indicates that using preference tuning for chemical settings could be a promising approach to explore in future research.

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.

To investigate this, we prompted[68] some of the top-performing models to estimate, on an ordinal scale, their confidence in their ability to answer the question correctly (see Methods for details on the methodology and comparison to logit-based approaches).

In Fig. 5, we show that for some models, there is no meaningful 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 with trustworthy uncertainty estimates, this is a concerning observation highlighting the need for critical reasoning in the interpretation of the model's outputs[34,73]. For example, for the questions about the safety profile of compounds, GPT-4 reported a confidence of 1.0 (on a scale of 1–5) for the one question it answered correctly and 4.0 for the six questions it answered incorrectly. While, on average, the verbalized confidence estimates from Claude-3.5 (Sonnet) seem better calibrated (Fig. 5), they are still misleading in some cases. For example, for the questions about the labelling of chemicals (GHS) pictograms Claude-3.5 (Sonnet) returns an average score of 2.0 for correct answers and 1.83 for incorrect answers.

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 that 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 questions 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 requiring more reasoning about chemical structures (for example, number of isomers or nuclear magnetic resonance peaks). 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 (when trained on the right training corpus).

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.

Although our findings indicate many areas for further improvement of LLM-based systems, such as agents (more discussion in Supplementary Note 11), 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 might be far from reasoning like a chemist, our ChemBench framework will be a stepping stone for developing systems that come closer to this goal.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41557-025-01815-x.

References

1. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
2. Zhong, Z., Zhou, K. & Mottin, D. Benchmarking large language models for molecule prediction tasks. Preprint at https://doi.org/10.48550/arXiv.2403.05075 (2024).
3. Kung, T. H. et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
4. OpenAI et al. *Gpt-4 technical report*. (2024); https://doi.org/10.48550/arXiv.2303.08774
5. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
6. M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
7. Darvish, K. et al. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. *Matter* **8**, 101897 (2025).
8. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with gpt-4. Preprint at https://doi.org/10.48550/arXiv.2303.12712 (2023).
9. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM conference on fairness, accountability, and transparency*, 610–623 (Association for Computing Machinery, 2021).
10. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D. & Griffiths, T. L. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proc. Natl Acad. Sci. USA* **121**, e2322420121 (2024).
11. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2021).
12. Anderljung, M. et al. Frontier ai regulation: managing emerging risks to public safety. Preprint at https://doi.org/10.48550/arXiv.2307.03718 (2023).
13. Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. Preprint at https://doi.org/10.48550/arXiv.2311.07361 (2023).
14. White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **7**, 457–458 (2023).
15. Jablonka, K. M. et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250 (2023).
16. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
17. Xie, Z. et al. Fine-tuning gpt-3 for machine learning electronic and functional properties of organic molecules. *Chem. Sci.* **15**, 500–510 (2024).
18. Liao, C., Yu, Y., Mei, Y. & Wei, Y. From words to molecules: a survey of large language models in chemistry. Preprint at https://doi.org/10.48550/arXiv.2402.01439 (2024).
19. Zhang, D. et al. Chemllm: a chemical large language model. Preprint at https://doi.org/10.48550/arXiv.2402.06852(2024).
20. Ramos, M. C., Michtavy, S. S., Porosoff, M. D. & White, A. D. Bayesian optimization of catalysts with in-context learning. Preprint at https://doi.org/10.48550/arXiv.2304.05341 (2023).
21. Kristiadi, A. et al. A sober look at LLMs for material discovery: are they actually good for bayesian optimization over molecules? In *Proc. 41st International Conference on Machine Learning* 1025 (JMLR.org, 2024).
22. Rubungo, A. N., Arnold, C., Rand, B. P. & Dieng, A. B. Llm-prop: predicting physical and electronic properties of crystalline solids from their text descriptions. Preprint at https://doi.org/10.48550/arXiv.2310.14029 (2023).
23. Flam-Shepherd, D. & Aspuru-Guzik, A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. Preprint at https://doi.org/10.48550/arXiv.2305.05708 (2023).
24. Gruver, N. et al. Fine-tuned language models generate stable inorganic materials as text. In *Twelfth International Conference on Learning Representations* (2024); https://openreview.net/forum?id=vN9fpfqoP1
25. Alampara, N., Miret, S. & Jablonka, K. M. Mattext: do language models need more than text & scale for materials modeling? Preprint at https://doi.org/10.48550/arXiv.2406.17295 (2024).
26. Patiny, L. & Godin, G. Automatic extraction of fair data from publications using llm. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2023-05v1b-v2 (2023).
27. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
28. Zheng, Z. et al. Image and data mining in reticular chemistry powered by gpt-4v. *Digit. Discov.* **3**, 491–501 (2024).
29. Lála, J. et al. Paperqa: retrieval-augmented generative agent for scientific research. Preprint at https://doi.org/10.48550/arXiv.2312.07559 (2023).
30. Caufield, J. H. et al. Structured prompt interrogation and recursive extraction of semantics (spires): a method for populating knowledge bases using zero-shot learning. *Bioinformatics* **40**, btae104 (2024).
31. Gupta, T. et al. DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long papers) 13465–13483 (Association for Computational Linguistics, 2023).
32. Schilling-Wilhelmi, M. et al. From text to insight: large language models for chemical data extraction. *Chem. Soc. Rev.* **54**, 1125–1150 (2025).
33. Skarlinski, M. D. et al. Language agents achieve superhuman synthesis of scientific knowledge. Preprint at https://doi.org/10.48550/arXiv.2409.13740 (2024).
34. Miret, S. & Krishnan, N. Are llms ready for real-world materials discovery? Preprint at https://doi.org/10.48550/arXiv.2402.05200 (2024).
35. Gopal, A. et al. Will releasing the weights of future large language models grant widespread access to pandemic agents? Preprint at https://doi.org/10.48550/arXiv.2310.18233 (2023).

36. Ganguli, D. et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. Preprint at https://doi.org/10.48550/arXiv.2209.07858 (2022).
37. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191 (2022).
38. Campbell, Q. L., Herington, J. & White, A. D. Censoring chemical data to mitigate dual use risk. Preprint at https://doi.org/10.48550/arXiv.2304.10510 (2023).
39. Moulange, R., Langenkamp, M., Alexanian, T., Curtis, S. & Livingston, M. Towards responsible governance of biological design tools. Preprint at https://doi.org/10.48550/arXiv.2311.15936 (2023).
40. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. A teachable moment for dual-use. *Nat. Mach. Intell.* **4**, 607–607 (2022).
41. One-third of college students used chatgpt for schoolwork during the 2022-23 academic date. *Intelligent.com* https://www.intelligent.com/one-third-of-college-students-used-chatgpt-for-schoolwork-during-the-2022-23-academic-date/ (2023).
42. Srivastava, A. et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023); https://openreview.net/forum?id=uyTL5Bvosj
43. Gao, L. et al. A framework for few-shot language model evaluation version v0.4.0. *Zenodo* https://zenodo.org/records/10256836 (2023).
44. Guo, T. et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. Preprint at https://doi.org/10.48550/arXiv.2305.18365 (2023).
45. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: towards chemical foundation models. Preprint at https://doi.org/10.48550/arXiv.2209.01712 (2022).
46. Cai, X. et al. Comprehensive evaluation of molecule property prediction with chatgpt. *Methods* **222**, 133–141 (2024).
47. Frey, N. C. et al. Neural scaling of deep chemical models. *Nat. Mach. Intell.* **5**, 1297–1305 (2023).
48. Dinh, T. et al. Lift: language-interfaced fine-tuning for non-language machine learning tasks. *Adv. Neural Inf. Process. Syst.* **35**, 11763–11784 (2022).
49. Wu, Z. et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
50. Huang, Y. et al. Chemeval: a comprehensive multi-level chemical evaluation for large language models. Preprint at https://doi.org/10.48550/arXiv.2409.13989 (2024).
51. Huang, K. et al. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *NeurIPS Datasets and Benchmarks* (2021); https://www.semanticscholar.org/paper/Therapeutics-Data-Commons%3A-Machine-Learning-and-for-Huang-Fu/54ca116f1e9a45768a3a2c47a4608ff34adefa0c
52. Zhao, H. et al. Chemsafetybench: benchmarking llm safety on chemistry domain. Preprint at https://doi.org/10.48550/arXiv.2411.16736 (2024).
53. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).
54. Zaki, M. & Krishnan, N. A. Mascqa: investigating materials science knowledge of large language models. *Digit. Discov.* **3**, 313–327 (2024).
55. Arora, D., Singh, H., & Mausam. Have LLMs advanced enough? A challenging problem solving benchmark for large language models. in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H., Pino, J. & Bali, K.) 7527–7543 (Association for Computational Linguistics, 2023); https://aclanthology.org/2023.emnlp-main.468/
56. Song, Y., Miret, S., Zhang, H. & Liu, B. Honeybee: progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H. et al.) 5724–5739 (Association for Computational Linguistics, 2023).
57. Wei, Z. et al. Chemistryqa: a complex question answering dataset from chemistry. *OpenReview* https://openreview.net/forum?id=oeHTRAehiFF (2021).
58. Song, Y., Miret, S. & Liu, B. Matsci-nlp: evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) 3621–3639 (Association for Computational Linguistics, 2023).
59. Hu, X. et al. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations* (2024); https://openreview.net/forum?id=9OevMUdods
60. Bloom, B. *Taxonomy of Educational Objectives: the Classification of Educational Goals* (Longmans, 1956).
61. Zhang, J., Lehman, J., Stanley, K. & Clune, J. OMNI: Open-endedness via models of human notions of interestingness. In *Twelfth International Conference on Learning Representations* (2024); https://openreview.net/forum?id=AgM3MzT99c
62. Polo, F. M. et al. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proc. 41st International Conference on Machine Learning* (JMLR.org, 2024).
63. Liang, P. et al. Holistic evaluation of language models. *Transactions on Machine Learning Research* (2023); https://openreview.net/forum?id=iO4LZibEqW
64. Taylor, R. et al. Galactica: a large language model for science. Preprint at https://doi.org/10.48550/arXiv.2211.09085 (2022).
65. Schick, T. et al. Toolformer: language models can teach themselves to use tools. *Adv. Neural Inf. Proc. Syst.* **36**, 68539–68551 (2024).
66. Karpas, E. et al. Mrkl systems: a modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. Preprint at https://doi.org/10.48550/arXiv.2205.00445 (2022).
67. Yao, S. et al. ReAct: Synergizing reasoning and acting in language models. In *Eleventh International Conference on Learning Representations* (OpenReview.net, 2023).
68. Xiong, M. et al. Can llms express their uncertainty? An empirical evaluation of confidence elicitation in llms. In *Twelfth International Conference on Learning Representations* (OpenReview.net, 2024); https://openreview.net/forum?id=gjeQKFxFpZ
69. Beeching, E. et al. Open llm leaderboard. *Hugging Face* https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (2023).
70. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
71. Hughes, E. et al. Position: Open-endedness is essential for artificial superhuman intelligence. In *Proc. 41st International Conference on Machine Learning* Vol. 235 (eds Salakhutdinov, R. et al.) 20597–20616 (PMLR, 2024); https://proceedings.mlr.press/v235/hughes24a.html
72. Choung, O.-H., Vianello, R., Segler, M., Stiefl, N. & Jiménez-Luna, J. Extracting medicinal chemistry intuition via preference machine learning. *Nat. Commun.* **14**, 6651 (2023).
73. Li, B. et al. Trustworthy ai: from principles to practices. *ACM Comput. Surv.* **55**, 1–46 (2023).

Adrian Mirza[1,2,18], Nawaf Alampara[1,18], Sreekanth Kunchapu[1,18], Martiño Ríos-García[1,3,18], Benedict Emoekabu[1], Aswanth Krishnan[4], Tanya Gupta[5,6], Mara Schilling-Wilhelmi[1], Macjonathan Okereke[1], Anagha Aneesh[1], Mehrdad Asgari[7], Juliane Eberhardt[8], Amir Mohammad Elahi[9], Hani M. Elbeheiry[10], María Victoria Gil[3], Christina Glaubitz[19], Maximilian Greiner[1], Caroline T. Holick[1,11], Tim Hoffmann[1,11], Abdelrahman Ibrahim[1], Lea C. Klepsch[1,11], Yannik Köster[1,11], Fabian Alexander Kreth[12,13], Jakob Meyer[1], Santiago Miret[14], Jan Matthias Peschel[1], Michael Ringleb[1,11], Nicole C. Roesner[1,11], Johanna Schreiber[1,11], Ulrich S. Schubert[1,2,11,13], Leanne M. Stafast[1,11], A. D. Dinga Wonanke[15], Michael Pieler[16,17], Philippe Schwaller[5,6] & Kevin Maik Jablonka[1,2,11,13] ✉

[1]Laboratory of Organic and Macromolecular Chemistry, Friedrich Schiller University Jena, Jena, Germany. [2]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Jena, Germany. [3]Institute of Carbon Science and Technology, CSIC, Oviedo, Spain. [4]QpiVolta Technologies Pvt Ltd, Bengaluru, India. [5]Laboratory of Artificial Chemical Intelligence, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [6]National Centre of Competence in Research Catalysis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [7]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. [8]Macromolecular Chemistry, University of Bayreuth, Bayreuth, Germany. [9]Laboratory of Molecular Simulation, Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne, Sion, Switzerland. [10]Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Jena, Germany. [11]Jena Center for Soft Matter, Friedrich Schiller University Jena, Jena, Germany. [12]Institute for Technical Chemistry and Environmental Chemistry, Friedrich Schiller University Jena, Jena, Germany. [13]Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Jena, Germany. [14]Intel Labs, Hillsboro, OR, USA. [15]Theoretical Chemistry, Technische Universität Dresden, Dresden, Germany. [16]OpenBioML.org, London, UK. [17]Stability.AI, London, UK. [18]These authors contributed equally: Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García. [19]Unaffiliated author: Christina Glaubitz. ✉e-mail: mail@kjablonka.com

Methods

Curation workflow

For our dataset, we curated questions from existing exams or exercise sheets but also programmatically created new questions (see Supplementary Table 3 for more details). Questions were added via Pull Requests on our GitHub repository and only merged into the corpus after passing manual review (Extended Data Fig. 1) as well as automated checks (for example, 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[4,42].

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

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 Supplementary Table 3.

Chemical preference data. These questions assess the ability to establish a 'preference', such as favouring a specific molecule. Chemical preference is of major importance in drug discovery projects, where the optimization process to reach the desired molecular properties is a process that takes several years within a chemist's career. Our data corpus is adapted from the published dataset by Choung et al.[72], which consists of more than 5,000 question–answer pairs about chemical intuition. To build the dataset, they presented 35 medicinal chemists with two different molecules, asking them what molecule they would like to continue with when imaging an early virtual screening campaign setting. The question was designed so the scientists do not spend much time answering it, relying only on their feelings or 'chemical preference'.

To understand whether the capabilities of the leading models align with the preferences of professional chemists, we randomly selected 1,000 data points from the original dataset to create a meaningful evaluation set, where molecules are represented as SMILES. To ablate the effect of different molecular representations, we only considered questions for which we could obtain International Union of Pure and Applied Chemistry names for both the molecules present.

Model evaluation workflow

A graphical overview of the pipeline is presented in Supplementary Fig. 12.

Prompting. We employ distinct prompt templates tailored for completion and instruction-tuned models to maintain consistency with the training. As explained later, 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. 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[64]. 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, and we provide presets for the models we evaluated.

Parsing. Our parsing workflow is multistep and primarily based on regular expressions. In the case of instruction-tuned models, we first identify the [ANSWER][\ANSWER] environment that 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 MCQ) 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, for example, '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, for example, Claude-3.5 (Sonnet), to parse the completion (Supplementary Note 8 provides more details on this step).

Models. For all models, we performed inference using greedy decoding (that is, temperature 0). We used the API endpoints provided by the model developers and those provided by Groq. PaperQA2 was used (in August 2024) via an API provided by FutureHouse.

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[68] since we found those to be closer to current practical use cases than other prompting strategies, which might be more suitable when implemented in systems. In addition, this approach captures semantic uncertainty, which is not the same as the probability of a token being given a sequence of tokens (that is, the uncertainty one obtains from logit-based approaches). On top of that, many proprietary models do not provide access to the logits, making this approach more general. In Supplementary Note 12, we provide more details and comparisons with a logit-based approach.

Human baseline

Question selection. Several design choices were made when selecting ChemBench-Mini. Firstly, from the full dataset, we kept all the questions labelled as advanced. In this way, we can obtain a deeper insight into the capabilities of LLMs on advanced tasks when compared with actual chemists. Secondly, we sample a maximum of three questions across all possible combinations of categories (that is, knowledge or reasoning) and topics (for example, organic chemistry and physical chemistry). Thirdly, we do not include any intuition questions in this subset because the intended use of ChemBench-Mini is to provide a fast and fair evaluation of LLMs independent of any human baseline. In total, 236 questions have been sampled for ChemBench-Mini. Then, this set is divided into two subsets on the basis of the aforementioned combinations. One of the question subsets allows tool use, and the other does not.

Study design. Human volunteers were asked the questions in a custom-built web interface (Supplementary Note 10), which rendered chemicals and equations. Questions were shown in random order, and volunteers were not allowed to skip questions. For a subset of the questions, the volunteers were allowed to use external tools (excluding other LLM or asking other people) to answer the questions. Before answering questions, volunteers were asked to provide information about their education and experience in chemistry. The study was conducted in English.

Human volunteers. Users were open to reporting about their experience in chemistry. Overall, 16 did so. Out of those, 2 are beyond a first postdoc, 13 have a master's degree (and are currently enroled in Ph.D. studies) and 1 has a bachelor's degree. For the analysis, we excluded volunteers with less than 2 years of experience in chemistry after their first university-level course in chemistry.

Comparison with models. 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.

The performance metrics reported for models in the main text are computed on the same questions that the humans answered. Metrics for the entire corpus are reported in Supplementary Note 4.

Data annotation

In the curation of our dataset, we manually assigned difficulty levels and required skills to each question. We used the following guidelines for these annotations: calculation is required if answering a question would require the use of a calculator, knowledge is required if answering a question requires non-trivial knowledge of facts (for example, the H/P statements of chemicals). Reasoning is required if answering a question requires multiple reasoning steps. Basic questions only require those skills up to the high school level. Advanced questions would require an expert multiple minutes or hours to answer.

Inclusion and ethics statement

The authors confirm that they have complied with all relevant ethical regulations, according to the Ethics Commission of the Friedrich Schiller University Jena (which decided that the study is ethically safe). Informed consent was obtained from all volunteers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data for ChemBench is available via GitHub at https://github.com/lamalab-org/chembench and via Zenodo at https://zenodo.org/records/14010212 (ref. 74).

Code availability

The code for ChemBench is available via GitHub at https://github.com/lamalab-org/chembench and via Zenodo at https://zenodo.org/records/14010212 (ref. 74) The code for the app for our human baseline study is available via GiHub at https://github.com/lamalab-org/chem-bench-app.

References

74. Mirza, A. et al. chem-bench version v0.2.0. *Zenodo* https://doi.org/10.5281/zenodo.14010212 (2024).

Acknowledgements

This work was supported by the Carl Zeiss Foundation, and a 'Talent Fund' of the 'Life' profile line of the Friedrich Schiller University Jena. In addition, M.S.-W.'s work was supported by Intel and Merck via the AWASES programme. Parts of A.M.'s work were supported as part of the 'SOL-AI' project funded by the Helmholtz Foundation model initiative. K.M.J. is part of the NFDI consortium FAIRmat funded by the Deutsche Forschungsgemeinschaft (the German Research Foundation) project no. 460197019. K.M.J. thanks FutureHouse (a non-profit research organization supported by the generosity of Eric and Wendy Schmidt) for supporting PaperQA2 runs via access to the API. We also thank Stability.AI for the access to its HPC cluster. M.R.-G. and M.V.G. acknowledge financial support from the Spanish Agencia Estatal de Investigación (AEI) through grants TED2021-131693B-I00 and CNS2022-135474, funded by Ministerio de Ciencia, Innovación y Universidades (MICIU)/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. M.V.G. acknowledges support from the Spanish National Research Council through the Programme for internationalization i-LINK 2023 (project no. ILINK23047). A.A. gratefully acknowledges financial support for this research by the Fulbright US Student Programme, which is sponsored by the US Department of State and German-American Fulbright Commission. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the Fulbright Programme, the Government of the USA or the German-American Fulbright Commission. M.A. expresses gratitude to the European Research Council for evaluating the project with the reference no. 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 under the UK government's Horizon Europe funding guarantee (grant reference EP/Y023447/1; organization reference 101106377). M.R. and U.S.S. thank 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 no. 497115849). A.D.D.W. acknowledges funding from the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101107360. P.S. acknowledges support from the National Centre of Competence in Research Catalysis (grant no. 225147), a National Centre of Competence in Research grant funded by the Swiss National Science Foundation. In addition, we thank the OpenBioML.org community and their ChemNLP project team for valuable discussions. Moreover, we thank P. Márquez for discussions and support and J. Kimmig for feedback on the web app. In addition, we acknowledge support from S. Kumar with an initial prototype of the web app. We thank B. Smit for feedback on an early version of the manuscript.

Author contributions

A.M., N.A., M.R.-G. and K.M.J. contributed to the software development of the benchmarking framework. K.M.J. wrote the article with help from A.M., N.A., S.K., and M.R.-G. A.K. wrote the software for chembench.org. A.M., N.A., S.K., M.R.-G., K.M.J., T.G., M.S.-W., M.V.G., B.E. and M.O. contributed to the generation of the question dataset. A.A., A.M.E., M.A., J.E., H.M.E., M.V.G., M.G., C.T.H., C.G., T.H., A.I., L.C.K., Y.K., F.A.K., J.M., S.M., J.M.P., M.R., N.C.R., J.S., L.M.S. and A.D.D.W. answered the question dataset for the human benchmark tests. U.S.S. and P.S. contributed to supervision and funding acquisition. K.M.J. directed the project and conceptualized it with P.S., M.P., A.M., N.A., M.R.-G. and S.K. All authors reviewed and edited the manuscript.

Funding

Open access funding provided by Friedrich-Schiller-Universität Jena.

Competing interests

K.M.J. has been a paid contractor for OpenAI (as part of the red teaming network). M.P. is an employee of Stability.AI, and A.M. and N.A. were paid contractors of Stability.AI. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41557-025-01815-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-025-01815-x.

Correspondence and requests for materials should be addressed to Kevin Maik Jablonka.

Peer review information *Nature Chemistry* thanks Joshua Schrier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

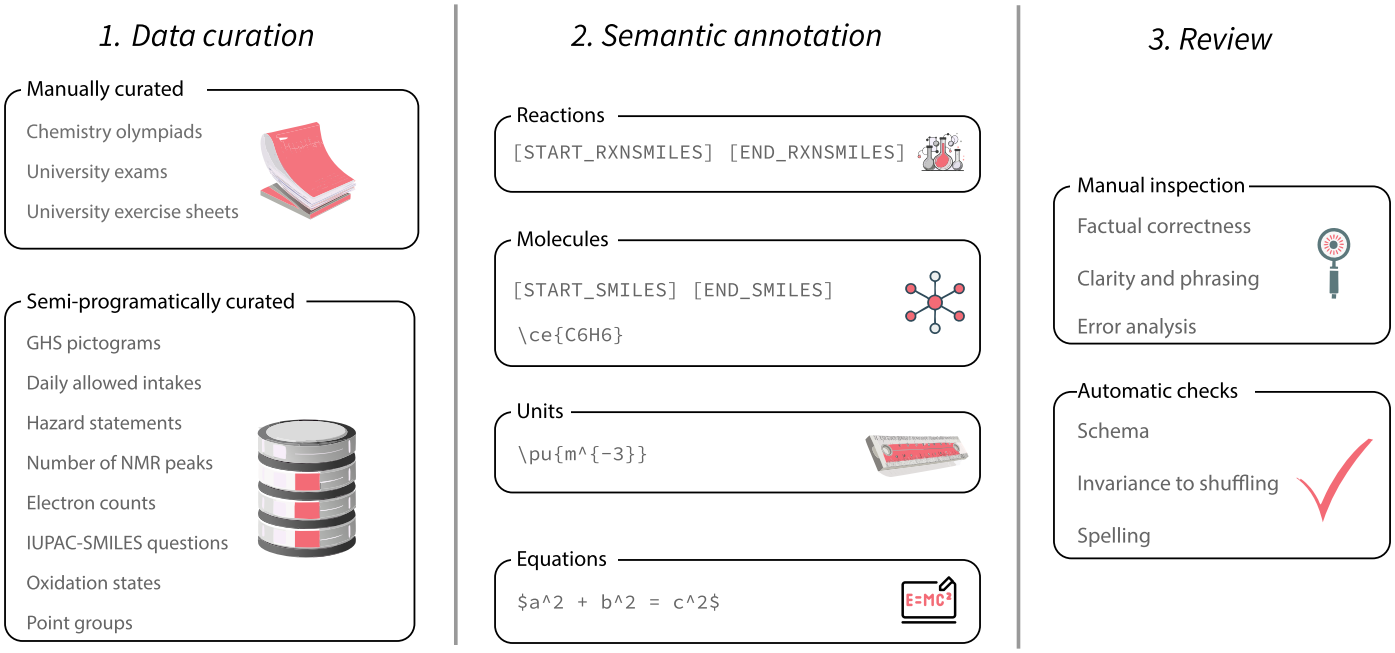
https://doi.org/10.1038/s41557-025-01815-x

Extended Data Table 1 | Overview of sources of the curated questions

Source	Count
Semi-automatically generated	1749
URL	375
Textbook	206
Exam	149
IChO	149
No source	139
Lectures	21

The table provides an overview of the types of sources the questions have been curated from. Detailed sources are available in the source data on GitHub. Questions without a source have been curated completely from scratch. Questions based on lecture notes or URLs have been curated based on content presented in those resources. All questions have been rephrased, annotated, and reviewed before being added to the corpus.

https://doi.org/10.1038/s41557-025-01815-x



Extended Data Fig. 1 | 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.

nature portfolio

Corresponding author(s): Jablonka

Last updated by author(s): Jan 27, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	For data analysis we used custom code that can be found at https://github.com/lamalab-org/chembench-paper and https://github.com/lamalab-org/chem-bench. We used version v0.2.0 to generate the results in the paper.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data for ChemBench is available at https://github.com/lamalab-org/chem-bench and archived on Zenodo under https://zenodo.org/records/14010212.74 A reproducible version of this manuscript, archived at was generated using the showyourwork framework.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex and gender have not been considered in the study design.
Reporting on race, ethnicity, or other socially relevant groupings	We did not collected fine-grained personal information to avoid dealing with personal information.
Population characteristics	Users were open to reporting about their experience in chemistry. Overall, 16 did so. Out of those, 2 are beyond a first postdoc, 13 have a master's degree (and are currently enrolled in Ph.D. studies), and 1 has a bachelor's degree. For the analysis, we excluded volunteers with less than two years of experience in chemistry after their first university-level course in chemistry.
Recruitment	The study was conducted as an only survey that was advertised via email to student and faculty bodies of the EPFL and the Friedrich-Schiller University Jena. The email made clear that participation is voluntarily and used for a benchmarking study. There is a potential self-selection bias for participants interested in LLMs and chemical questions.
Ethics oversight	The authors confirm to have complied with all relevant ethics regulations (no personal data was recorded). The institutional review board of the Friedrich-Schiller University of Jena was consulted.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Study type: Observational study evaluating human expert performance in chemistry-related questions Setting: Web-based survey using chembench.org platform Purpose: To benchmark human expert performance against LLM capabilities in chemistry
Research sample	Sample size: 19 expert chemists participated in the study Demographics (out of 16 who reported their experience): 2 were beyond first postdoc 13 had master's degrees and were enrolled in PhD studies 1 had a bachelor's degree
Sampling strategy	No sample size calculation was performed. We recruited as many participants as we could.
Data collection	Method: Custom web application (chembench.org) was used to survey the experts Format: Questions presented through web interface Molecules shown as rendered drawings and SMILES strings LaTeX equations and chemical equations rendered using MathJax Time taken to answer questions was recorded Tool usage was tracked
Timing	02.09.2024-13.09.2024
Data exclusions	Pre-established criteria: Volunteers with less than two years of experience in chemistry after their first university-level chemistry course were excluded from analysis Rationale: To ensure participants had sufficient expertise in chemistry
Non-participation	No participant dropped out
Randomization	Questions were presented to participants in random order

Randomization

Tool-use questions were separated from non-tool-use questions

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.