Are large language models superhuman chemists?

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

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

March 30, 2024

Abstract

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

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

We curated more than 7,000 question-answer pairs for a wide array of 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,[1,2] with the leading models now being able to pass in some evaluations the United States Medical Licensing Examination.[3]

References

1. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165.* arXiv: `2005.14165 [cs.CL]`.
2. Zhong, Z., Zhou, K. & Mottin, D. Benchmarking Large Language Models for Molecule Prediction Tasks. *arXiv preprint arXiv:2403.05075.* arXiv: `2403.05075 [cs.LG]`.
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.