https://doi.org/10.1038/s41570-023-00502-0

The future of chemistry is language

Andrew D. White

Large language models such as GPT-4 have been approaching human-level ability across many expert domains. GPT-4 can accomplish complex tasks in chemistry purely from English instructions, which may transform the future of chemistry.

Large language models (LLMs) predict an output sequence from an input sequence. For example, you could input "ethanol" and it will output CCO — the simplified molecular-input line-entry system (SMILES) representation of ethanol (SMILES are how a chemical structure is written as text). Like any machine learning model, LLMs are fit empirically with a large dataset — generally a large subset of the internet. While they were first pursued for tasks in natural language, such as translating English to French[1], they can now be used to identify objects in images[2], predict protein structure[3] and estimate reaction yields[4], and they are the technology behind the popular ChatGPT.

The recent release of GPT-4 has renewed interest in how LLMs, such as GPT-4, can be applied in chemistry[5]. I have been using the early versions of GPT-4 since 6 months before the release and believe they represent the future of the field — but not by replacing existing computational or experimental methods. Instead, LLMs will transform how we connect our data, computer programs and scientific literature, and how we plan experiments.

Like any emerging idea in chemistry, it will take time to see where LLMs will fit. They are already used in most modern reaction synthesis planner tools[4] and have started seeing applications in explaining molecular properties[6] — but where might LLMs go next? I believe LLMs are about to be stapled to every tool in chemistry. Akin to the creation of the internet, it is a foundational technology that will accelerate how fast a chemist can learn and use computational tools. It can be hard to see this today, just like it was hard to predict the effects of the internet, but LLMs are rapidly proving useful in many different areas, precisely because we have built so much software, data and science around natural language.

LLMs can answer questions, summarize text, change formats between files, learn to use programs, and plan and execute multi-step plans. LLMs can also be used for a "semantic" search — searching for information based on its meaning rather than exact matching of words. For example, LLMs can ingest one or multiple papers and answer questions grounded in that specific information. This means you can have a LLM read a 1,000-page PDF of regulations and ask it technical questions about the content. It even attributes its sources, as shown in Fig. 1, given the correct tools. Even if they have no direct impact on science, LLMs can write emails, summarize meetings, create action items from a transcript and query databases, all from natural language.

These capabilities are especially valuable in chemistry. LLMs can act as an interpreter, so that we can convert the predictions of black box models into natural language explanations[6]. We can also use LLMs to make IUPAC names, or even common names, as inputs to molecular prediction tasks[7]. You can ask "What is the solubility of 2-acetyloxybenzoic acid?" and get back a numerically correct answer. LLMs can also write computational chemistry code, lowering the barrier of entry for writing density functional theory (DFT) input files or analysing protein structure[8]. This removes the need to learn a new software library or application by reading documents; the user can just have a chat with a LLM to write code.

It is worth noting that LLMs are not going to replace DFT or your search for the best catalyst. What they will do is wrap the inputs and outputs to methods so that language is how information goes between people and tools. LLMs can even respond to their own faults, such as reading an error message or having a user tell it to fix a mistake. Imagine if all our tools, data and results are trivially interoperable, not because of some universal standard, but through natural language. Humans have refined natural language to reflect how we interact with

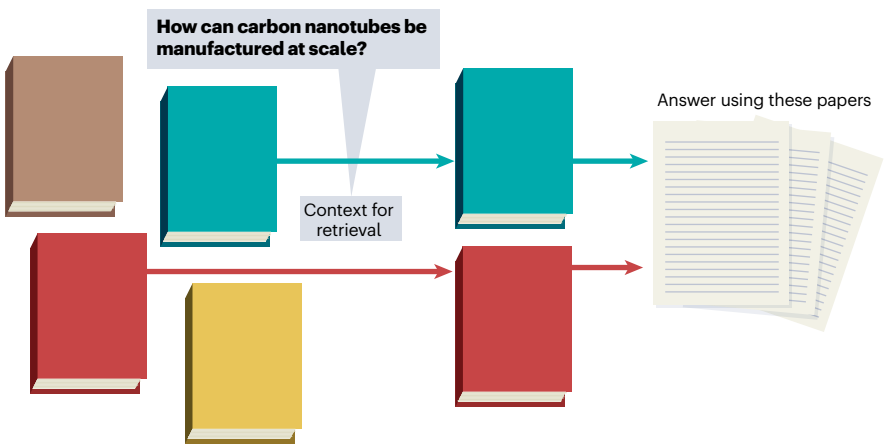

Fig. 1 | A question answered by a LLM augmented with documents to overcome hallucinations. By giving the large language model (LLM) access to real papers, it answers with real citations and context can be generated. Relevant papers are chosen based on the question, then the content (or summary) of those papers is given to the LLM as context to provide an answer.

our environment and tools. Similarly, we have shaped natural language around chemistry. Now, thanks to LLMs, a two-way conversation in natural language will improve both the inputs and outputs of chemical tools.

The hallucination problem is one of the key challenges with LLMs. This is where ChatGPT and similar models seem to make up facts, functions or citations. This reflects the short time frames to answer questions and limited to no access to outside sources. One solution is letting LLMs have access to the internet, Wikipedia, or curated document sets (such as your research group's Slack history), which helps LLMs ground their answers in evidence. Another tactic is to simply tell LLMs to slow down and show their workings — a so-called chain of thought. It is a strange feeling to be reading and writing research papers about how to better ask questions of an AI — a new area of research called prompt engineering.

Hyperconnected tools are the next step in LLMs. There are many frontiers that previously looked like multi-year projects, which are suddenly becoming more feasible. For example, unlocking historical chemical data — often locked in tables or images — may soon be accessible with LLMs for chemistry. We're also seeing some emergent behaviour of LLMs when we give them access to computer programs — as shown in the GPT-4 release information, where novel compounds can be designed.

How do we adapt and make use of LLMs? It is time to rethink our tools and experiments. We do not need better file formats, new interfaces to enter data, better schemas, or more training on niche, specialist tools. Instead, we can start using expressive natural language and let LLMs help bridge the gap from our intent to the tools of chemistry. It is also clear that LLMs are continuing to improve at a dramatic pace, and new abilities may emerge to further improve the ease of our interactions with technology in chemistry. Clear communication in natural language is about to be the most valuable technical skill as we enter this new phase of chemistry.

Andrew D. White
Department of Chemical Engineering, University of Rochester, Rochester, NY, USA.
e-mail: Andrew.white@rochester.edu

Published online: 19 May 2023

References

1. Vaswani, A. et al. Attention is all you need. *Proc. 31st Int. Conf. NIPS* 6000–6010 (2017).
2. Liu, Z. et al. Swin Transformer: Hierarchical vision transformer using shifted windows. *ICCV* 10012–10022 (2021).
3. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2022).
4. Schwaller, P. et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
5. OpenAI. GPT-4 Technical Report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
6. Gandhi, H. A. & White, A. D. Explaining molecular properties with natural language. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2022-v5p6m-v3 (2022).
7. Jablonka, K. M., Schwaller, P., Andres O. & Smit, B. Is GPT-3 all you need for chemistry? Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2023-fw8n4 (2023).
8. White, A. D. et al. Assessment of chemistry knowledge in large language models that generate code. *Dig. Discov.* **2**, 368–376 (2023).

Competing interests

A.D.W. was a paid consultant of OpenAI, the developers of GPT-4 mentioned in the article.