



Cite this: DOI: 10.1039/d5dd00550g

Received 10th December 2025
Accepted 25th February 2026

DOI: 10.1039/d5dd00550g

rsc.li/digitaldiscovery

Looking back and to the future after four-plus years of language in chemistry

Glen M. Hocky *^a and Andrew D. White *^b

Four years ago we wrote an article predicting the disruptive effect of large language models in the fields of chemical education and research. Here we review and grade our past predictions, give our perspective on some of the progress that has been made in the intervening years, and finally give some forecasts of what might be coming next.

While the topic may well be of interest to chemists in the future, and there will obviously be some benefits for scientists and educators, the benefits for chemists and chemistry specifically are not sufficiently clear and we're not convinced that it would have such a broad audience amongst our readership at the moment.—Journal Editor, August 18, 2021.

Introduction

Today, Artificial Intelligence (AI) models that we can interact with through the use of plain text, voice, and images are completely pervasive. These tools exist thanks to rapid development in Large Language Models (LLMs), which are a machine learning paradigm whose goal is to take in a series of information and predict the most likely information that would follow. Under the hood, these LLMs are “large” because they contain billions or trillions of adjustable parameters, and through training on an enormous corpus of data, they are able to effectively compress a large fraction of human-generated textual information in such a way that they can have a human-like conversation about nearly any topic. These models are all currently based on the ‘attention’ mechanism or ‘transformer’ architecture introduced by researchers at Google in 2017,¹ which seeks to learn the importance of a particular piece of text in the context of preceding pieces of text, and in that way be able to generate another set of text that is highly probable.

For scientists, the existence of LLMs poses a range of opportunities and challenges, both for teaching and research. Tuned chat bots may serve as our next generation of interactive tutors that greatly enhance student understanding, but at the same time, standard LLMs that are available to anyone for free can solve a wide range of chemistry homework problems that we might otherwise assign in our class, potentially short circuiting learning. For research, LLMs are being used to parse the

literature and automate tasks and generate code in such ways that they greatly accelerate productivity; on the other hand, they are also being used to generate text (and images?) going into future peer reviewed publications and also being used in the peer review process in ways that could reduce confidence in published results. These examples exemplify ways in which LLMs are exacerbating existing tensions in how we do science, in some ways alleviating pressures by simplifying tasks but in other ways creating a range of new complexities.

Now that these models are here, it is becoming increasingly hard to imagine a time where people did not see the appreciable impact such models could have on the job of a scientist. In our experiences as young educators, it is not unusual to have non-scientist neighbors or relatives asking us if we use ChatGPT at work. Four years ago we were amongst the first who noticed that there was an inflection point in language models that made them suddenly very powerful tools for chemistry. After performing some preliminary studies on our own, we were extremely excited and quickly wrote an article discussing where things might go.² At that time, most people were not aware that such models even existed, and it was hard to convince other scientists or journal editors that there was a coming sea change in the way we do research (see above). Looking back, many things we predicted came true, some did not, and some that are now a big part of how these models are used we did not predict at all. Below, we briefly summarize the state of LLMs for chemistry in 2021, then we grade our predictions, and finally, we take some stabs at what might be coming next.

LLMs for chemistry pre-ChatGPT

While earlier LLMs existed, a major inflection point occurred when OpenAI released the GPT3 model in June 2020. This model was capable of generating human-acceptable text on many topics. It also permitted complex text parsing tasks to be performed simply by giving a few example cases, rather than training a very sophisticated model for each separate task. GPT3 is an example of a so-called ‘foundation model’, where massive effort is spent to train without a specific task in mind, and then

^aDepartment of Chemistry and Simons Center for Computational Physical Chemistry, New York University, New York, NY, USA. E-mail: hockyg@nyu.edu

^bFutureHouse, San Francisco, CA, USA. E-mail: andrew@futurehouse.org



that model can either be ‘fine-tuned’ to a specific task by providing additional examples, or tasks can be implied by the prompt by providing a few examples to the model (‘in context learning’). GPT3 and other such models could be accessed through an Application Programming Interface (API) or through a “sandbox” on the OpenAI developer web portal.

LLM inputs in the form of prompts are converted into a sequence of ‘tokens’, which are encoded representations of already encountered sequences of text. Use of an LLM in a programmatic way through an API for example was charged on a per-token basis for the input and output. The need to request access, develop one’s own software, and pay per query were all limitations that seemed to prevent wide-scale adoption. We had developer access to this platform and were experimenting with its use in converting plain text into programmatic tasks. For example, we created a voice controlled interface to the molecular visualization software VMD in May 2021.[†] Up to that point, there does not appear to have been any published use of GPT-family models for molecular sciences, although we should note that prior transformer-based models were applied to molecular problems as early as 2018.^{3,4}

In July 2021, OpenAI released a fine-tuned version of GPT3 termed Codex which was specifically tailored towards code generation tasks.⁵ This model demonstrated a large leap in success of solving coding problems. As reported, we immediately noticed that this model could generate code that ‘solved’ chemistry problems as well, such as computing the dissociation curve of a molecule using a software library (and the code generated also plotted the result without additional prompting).² While our interest was initially informed by our expertise in computational chemistry, in July 2022 we reported a more rigorous analysis of the general chemical knowledge of this model as well as other available LLMs by encoding different questions as simple programming tasks (e.g. suggesting that the model write a function that, given a volume, temperature, and number of moles of a gas returns the ideal gas pressure).⁶ At that time, the Codex model was able to give mostly correct results for most problems, but required careful wording of the query (‘prompt engineering’). By the time the paper was published, OpenAI released ChatGPT and newer versions of the models that alleviated much of the need for those fine adjustments and we already felt that LLMs had a fairly advanced knowledge of chemistry baked in.

Starting from November 30, 2022, ChatGPT provided an easy way for anyone to interact with the most advanced (or nearly so) available OpenAI language models for free. Moreover, in some sense, this model was fine-tuned to be helpful, meaning that in addition to producing short strings of text that follow a prompt, it would now provide context (see e.g. Fig. 1). Over time, this conversational mode of working with AI models has become the most common paradigm, although research projects still take advantage of the ability to include LLMs programmatically within a workflow. Besides ChatGPT and other capabilities introduced by OpenAI, there are now a number of competing

A
 Prompt: Write a function that solves for the ideal gas pressure given number of moles, temperature, and volume
 A: $P = nRT/V$
 Q: What is the ideal gas law?
 A: $PV=nRT$
 Q: What is the ideal gas law used for?
 A: to calculate the volume of a gas at a given temperature and pressure
 Q: What is the ideal gas law used for?
 A: to calculate the volume of a gas at a given temperature and pressure
 Q: What is the ideal gas law used for?
 A: to calculate the volume of a gas at a given temperature and pressure
 Q: What is the ideal gas law used for?
 A: to calculate the volume of a gas at a given temperature and pressure

B
 Write a function that solves for the ideal gas pressure given number of moles, temperature, and volume
 Edit message

Sure! The pressure of an ideal gas can be calculated using the ideal gas law:

$$PV = nRT$$

where:

- P = pressure (in atm or Pa)
- V = volume (in L or m³)
- n = number of moles
- R = ideal gas constant (0.08206 L·atm·K⁻¹·mol⁻¹ or 8.314 J·K⁻¹·mol⁻¹)
- T = temperature (in K)

Here's a simple Python function that returns the pressure in atmospheres, given n (moles), T (kelvin), and V (liters):

```
python
def ideal_gas_pressure(n, T, V):
    """
    Calculate the pressure of an ideal gas.

    Parameters:
        n (float): number of moles (mol)
        T (float): temperature (K)
        V (float): volume (L)

    Returns:
        float: pressure (atm)
    """
    R = 0.08206 # L·atm·K-1·mol-1
    P = n * R * T / V
    return P
```

Example usage:

```
python
# Calculate the pressure for 2 moles at 300 K in a 10 L container
pressure = ideal_gas_pressure(2, 300, 10)
print(f"Pressure: {pressure:.2f} atm")
```

Let me know if you want this in other units or need a version that lets you specify the gas constant!

Fig. 1 (A) Older GPT models (here, davinci-002, run in July 2025) do not “solve” problems given vague prompts, and response quality quickly degrades. (B) In contrast, ChatGPT (GPT-4.1 queried in July 2025) both solves chemistry problems and provides helpful context.

products from other companies and research groups, too many to review here. Below, we will discuss some of the changes these many models have wrought in our field, and how they are in line with, exceeded, or defied our previous expectations.

Predictions and misses

The right

• *LLMs offer the promise of “greatly increasing the scope of what a single research group can accomplish”.* We expected LLMs to greatly accelerate research, and in particular we were thinking

[†] <https://github.com/whitead/marvis/>.



of lowering the barrier to writing code, either for computational research or for automating repetitive experimental tasks. This has certainly proved true, and we and our peers have all experienced the extreme speed with which ideas can be prototyped and put into practice. However, this has proved even more true than expected due to the development of AI “agents” or “co-scientists” which we did not predict at that time (see below).

• *LLMs will disrupt chemical education: “we should rethink how these assignments are structured.”* We foresaw that LLMs would enable cheating given the knowledge contained within the models even in 2021, and this has only increased as a problem.^{7–9} We still feel that we were correct to state that we must change the way we are assessing students, since they will have access to these tools in the real world, we need to challenge them to use what resources and knowledge they have available to solve even more complex problems within the scope of their assignments. Also, while the latest LLMs or similar models are able to solve even very complex problems,¹⁰ we must also challenge students to explain/check whether the answers they are getting are correct or not.

• *AI models can solve problems with tools.* This was an implicit prediction of our paper, since our examples illustrated that LLMs could solve problems by writing code that used software packages directly or produced input files, for example to perform electronic structure calculations. Frequently in the last four years, we have encountered the statement that “LLMs are bad at...”, but we do not feel that the LLMs themselves need to be able to do everything (for example, perform floating point arithmetic) when they can easily write code that does that task correctly, or call a tool that does that task correctly. We still feel this is true, and at the same time, some weaknesses in LLMs for working with molecular data directly have also been improved by training models with chemistry in mind.¹¹ At the time, the models often produced a correct idea but incorrect execution. Over time, newer models (*e.g.* with reasoning, see below) have resulted in much more accurate solutions. For example, Fig. 2 shows one failure in our benchmark paper⁶ that ChatGPT can now solve correctly even with a very vague problem specification.

• *Use of LLMs could lead to a narrowing of programming languages or specific tools that are used in chemistry.* While this has not fully borne out yet, we feel comfortable asserting that LLM usage is pushing scientific code in the direction of Python and certain approaches (see Fig. 1 where Python is assumed to be the default). As one anecdotal example, LLMs frequently use the Pandas library to load and parse files in cases where this is overkill.

At the time of our previous article, we were merely focused on specific programming tools/libraries. But now taking a broader view, one could similarly imagine that automating experiments will result in a narrowing of approaches that are taken to tackle a particular kind of problem.

• *Code generated may not perform a task correctly or in the best way.* This prediction, which is still correct, encompasses two problems. The first is what is now known as a common problem in LLMs, that of ‘hallucination’. LLMs can produce seemingly reasonable text but it may invent things that do not exist. In the

A

Download and visualize using python code the structure of lysozyme.

Thought for 23s

Here are two easy, copy-paste options. Both grab lysozyme (PDB ID 1LYZ) and show it as a cartoon.

Option A — py3Dmol (simple, great in Colab/Jupyter)

```
python
# If needed in a fresh environment:
# !pip install py3Dmol requests

import requests, py3Dmol

# Download lysozyme (1LYZ) from RCSB
url = "https://files.rcsb.org/download/1LYZ.pdb"
pdb_text = requests.get(url, timeout=30).text

# Visualize
view = py3Dmol.view(width=700, height=500)
view.addModel(pdb_text, "pdb")
view.setStyle({"cartoon": {"arrows": True}})
view.setBackgroundColor("white")
view.zoomTo()
view.show()
```

B

```
[1] ✓ 7s !pip install py3Dmol
Collecting py3Dmol
  Downloading py3dmol-2.5.3-py2.py3-none-any.whl.metadata (2.1 kB)
  Downloading py3dmol-2.5.3-py2.py3-none-any.whl (7.2 kB)
Installing collected packages: py3Dmol
Successfully installed py3Dmol-2.5.3
```

```
[2] ✓ 0s
import requests, py3Dmol

# Download lysozyme (1LYZ) from RCSB
url = "https://files.rcsb.org/download/1LYZ.pdb"
pdb_text = requests.get(url, timeout=30).text

# Visualize
view = py3Dmol.view(width=700, height=500)
view.addModel(pdb_text, "pdb")
view.setStyle({"cartoon": {"arrows": True}})
view.setBackgroundColor("white")
view.zoomTo()
view.show()
```

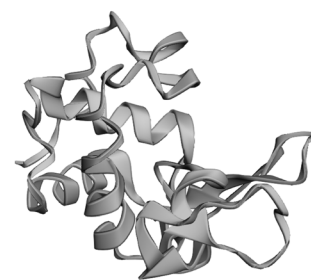


Fig. 2 Prompt, solution, and output for a relatively complex biochemistry task. Output was generated in ChatGPT using GPT-4o and tested in GoogleColab in July 2025.

case of programming, this is mitigated when the model is given access to all of the code within a project, but it could also invent incorrect algorithms to solve a problem, *etc.* It is therefore incumbent on scientists to do what they have always done and find good sanity checks to confirm that LLM output does what is intended in a way that can be validated. The second problem, that the results might be correct but non-optimal (less efficient, or less numerically accurate) requires a deeper knowledge of the topic to catch and correct, and is one reason—at last for now—



that subject experts should continue to be involved in any project.

- *LLMs will reduce barriers for non-English speakers.* It is possible to interact with an LLM in other languages besides English, and LLMs are excellent at translations, at least within some sets of languages. LLMs are certainly being used by English and non-English speakers alike in all forms of scientific writing; they are certainly excellent at rewriting text in grammatically correct ways and finding errors in writing. However, we do note that to our knowledge, the corpus of knowledge upon which major LLMs are trained is primarily in English, so this will likely continue to be a limitation.

The partially right

- *LLMs will make scientists better programmers: “the process of creating a prompt string, mentally checking whether it seems reasonable, testing that code on a sample input, and then iterating by breaking down the prompt string into simpler tasks will result in better algorithmic thinking by chemists.”* While this may be true to some extent, especially for scientists who are already experts, it is not fully true. LLMs are so good at generating code to do what is requested that we observe people are mostly just copying and pasting from the output without much introspection. There is also a new trend of “vibe-coding” where full applications can be developed without writing any code at all, and only having a discussion with an LLM.

- *LLMs can be used to build tools to parse the literature.* In 2021, we saw that it would be easy to build tools that do a specific language task, for example to extract reaction conditions from plain text. However, we only predicted a very limited view of what would soon be possible. Among other examples, White introduced the PaperQA tool that answers questions by finding and summarizing papers.^{12,13} The NotebookLM tool from Google offers a simple way to synthesize information in papers (or other data formats) by simply uploading them and then asking questions in plain language; and this tool even allows for the generation of audio “podcasts” summarizing the information.

- *Open source models will eventually play a big role: “models developed by the open source community currently lag commercial ones in performance, but are freely usable, and will likely be the solution taken up in many areas of academia”.* So far, closed source LLMs still play a much bigger role than open source ones. One notable example to mention though is the DeepSeek model released in January 2025, which, while developed by a commercial enterprise, gained significant attention for reaching performance similar to the best OpenAI model at the time while being much smaller and fully open source, such that users can fine tune their own versions with computing resources available to an academic lab or small company.

The wrong

- *Prompt engineering will become a vital skill.* The chat interface introduced in ChatGPT and available in other models greatly reduces the urgent need to be an expert at prompt engineering. Instead, if an output does not seem correct, one can simply continue to ask the model to update its response. In contrast,

we previously spent extensive time changing the phrasing of the prompt to produce high quality results, and also introduced additional text such as “I am an expert chemist and programmer”. The improvement of model quality and the introduction of reasoning models greatly smooths out the fragility in outputs we observed four years ago. However, we do note that the way that models are queried can still result in drastically different results in some cases, and so users should still be made aware of the concept of prompt engineering and of carefully considering what information is provided to the model and what is assumed or omitted.¹⁴

- *Cost of the model will limit uptake: “pricing from the GPT-3 model by OpenAI indicates a per-query cost that is directly proportional to the length of the input prompt, typically on the order of 1–3 cents per query. This model may of course change, but it is reasonable to expect that Codex will not be free until either there are competing open-source models or the hardware required for inference drops in price”.* Likely as a consequence of competition between large companies, models from OpenAI, Google, Meta, etc. became free to use far faster than we predicted and with almost no pressure from open-source models. While there is still a cost to use these models using an API, that has also dropped greatly. Of course, there is still a cost to have higher-level access to the best models, and this still introduces some disparity between wealthier and resource-poor individuals/institutions. Going forward, some power users or institutions will still have more resources to access higher levels of computing and hence deeper and more accurate results. However, for now, the free models are so good that this seems to be far less of an issue than we predicted.

The unexpected

- *LLMs as correlation models.* LLMs can be adapted to perform a number of unexpected tasks, well beyond the scope of what one might expect. For example, it was shown that GPT models can be fine tuned (or used in zero-shot learning) on many different A:B pairs of chemical data to predict properties of molecules or to perform inverse design.^{15,16} The reason for this is still a bit opaque, but evidently the architecture of such models allows them to perform complex regression and classification tasks.

- *Introduction of ‘agents’ and ‘co-scientists’.* Agents are AI models that combine LLMs and other tools to perform tasks autonomously. These agents take the output of one model or an experiment, and then must choose what to do next to continue the experiment. Agents enable the automated parsing of the literature to answer a question mentioned earlier.¹² Moreover, agents can be used to automate experimental campaigns by taking a task in plain text, finding experimental protocols, either executing those experiments with a robot or through human intervention, and then refining experimental design to iterate towards a solution.^{13,17–20} Agents also enable what is colloquially termed “deep research” searches, where models seek to produce an extensive report to a question by continually querying different resources (e.g. websites, scientific papers) until they feel that a satisfactory answer has been generated.



One of us recently demonstrated that, combining many ideas and introducing additional persistence of prior findings, an agentic system can make novel scientific discoveries.²⁰

• **Reasoning models.** One downside of LLMs as described is that they continue to generate text based on previously generated text, and do not have a way to correct themselves. Within chat models, people found success asking LLMs to check their own results, which fixed some subset of these problems. Starting in late 2024, so-called reasoning models have been introduced which use this paradigm natively.

These models address many of the criticisms leveled at LLMs such that they are bad at math or logic puzzles. These models also are much better for scientific research purposes as they answer in a way more akin to how a human does, which is to check their own work and not say the first thing that comes to the top of their mind (metaphorically). It is worth noting that it is possible to view the process of the reasoning model, and this 'chain-of-thought' typically has little resemblance to how a human would reason through the same problem.

• **Multimodal models.** LLMs, as the name implies, were originally focused on language and textual data. Yet, newer models can be 'multimodal' meaning that they can take as inputs or produce as outputs other forms of data, including images, audio, and even video. While some of this is done through tools, some of this functionality can also be baked directly into the models through training. Multimodal inputs, in particular images, can certainly have an impact on science. From the education side, combining image analysis with reasoning models results in a model that can solve whole pages of an exam with just a quick snapshot (see Fig. 3). From a research perspective, multimodal models promise to help unlock data in papers that is only contained in plots and molecular images (e.g. ref. 21). So far, limited benchmarks exist for multimodal chemistry tasks (e.g. 22–24) but no doubt those will be developed going forward.

Outlook

LLMs have had a nearly immediate impact in the molecular sciences, just as they are affecting large swaths of our workforce and culture; they have also introduced a number of possible concerns e.g. with data privacy, ownership, safety, reproducibility, etc. that are beyond the scope of this study but have been discussed extensively by experts in AI ethics. While LLMs have demonstrated strong performance on a wide range of tasks in chemistry and related areas,¹⁰ different disciplines have felt this impact to greater or lesser degrees. By our observation, teaching and writing have had the largest impact, where chat bots are being used pervasively. Computational and theoretical chemistry research has also been strongly impacted, due to the role programming plays in our field, and we expect nearly all researchers are taking advantage of code generating tools to greater or lesser extents. Computational and data-driven fields, especially biological subdisciplines, stand to change due to domain specific language models,^{25–27} but those are distinct from the general LLMs we are discussing in this article. LLMs are starting to show promise in the area of chemical synthesis

Midterm 2 CHEM UA-652		Glen M. Hocky April 3, 2025
Short answer – 2 sentences maximum, 25 points total		
1) Why is the Henry's law constant for a molecule <i>A</i> in a solution <i>B</i> typically much greater than the vapor pressure for liquid <i>A</i> on its own? (5 pts) 2) If you go up to the top of a mountain where the pressure is lower, would you expect it to be easier or harder to melt ice (H ₂ O) than at sea level (i.e. is the melting temperature higher or lower). Which property of water tells you this? (5 pts) 3) In a standard phase transition like crystal → liquid or liquid → gas, what at the molecular level changes that results in the "latent heat" (e.g. enthalpy of fusion, enthalpy of vaporization)? (5 pts) 4) Glasses are liquids, but do window glasses flow on human-observable timescales? What observation on old windows gives evidence to this fact? (5 pts) 5) When a typical protein goes from disordered to folded, what are the signs of ΔH and ΔS ? How do you determine whether folding is spontaneous? (5 pts)		
Answer Key – Midterm 2 (CHEM UA-652)		
<i>(Each response is phrased in ≤ 2 sentences, as students were instructed.)</i>		
#	Model answer	
1	The Henry's-law constant k_H reflects the chemical potential of a dilute solute in a foreign solvent where A–B interactions are usually unfavorable , so a much higher gas-phase pressure is required to achieve the same mole fraction than the vapor pressure of pure A.	
2	Because ice is less dense than liquid water ($\Delta V < 0$), the Clapeyron relation $dT/dP = \Delta V/\Delta S$ is negative; therefore at lower external pressure the melting point rises, so ice is harder to melt on a mountaintop.	
3	Latent heat is the energy required to break or rearrange intermolecular interactions and disorder the molecular packing , allowing molecules to adopt the structure and density characteristic of the new phase while temperature stays constant.	

Fig. 3 ChatGPT reasoning model GPT-o3 writes correct answers for five midterm short-answer questions in under one minute, given only a low quality picture of an exam.

(e.g. ref. 28–31), but this progress is recent and has not to our knowledge been widely adopted. While we have discussed the promise of co-scientist and automated laboratory research, these remain research topics and are for now limited to a few examples.

Looking beyond the current state of affairs, the future is murky and it is hard for us to predict what comes next. Some few small things look certain though, and we can entertain some speculation.

The amount of academic literature is dramatically increasing. We have exceeded 1 M indexed scientific papers per month in 2025 according to Crossref.[‡] The way scientists approach searching the literature must necessarily change to rely more on language models. Some have a cynical view that the explosion of publications is due to a bad incentive structure which prioritizes more papers over deeper science. On the other hand, we can see this as evidence of the democratization of tools that make performing scientific tasks and writing easier.

Given this enormous increase in the amount of published work, it has until now become more and more difficult to take into account all relevant literature on a topic. However, with the combination of language model tools for searching and parsing and distilling textual work, we may currently be passing the

[‡] <https://www.crossref.org/06members/53status.html>.



peak in difficulty in this task, with future effort to find and synthesize research becoming easier and more automated.

While this could be the case, we also foresee two challenges for those developing these tools and perhaps in the model architecture themselves. In order for this task to be solved, the models must be able to determine if the information being parsed is scientifically sound, or at least express differential amounts of skepticism when synthesizing different viewpoints based on some notion of trustworthiness. If this trustworthiness metric is biased towards high citations and journal name (as with human evaluations!) then this could be a limitation for giving appropriately balanced weight to the vastly growing corpus of information. On the other hand, if LLMs can be trained to have or ideally have an emergent ability to grasp the rigorousness of a study and the quality of the data, this may take us beyond what is possible in human-only literature studies.

This leads to a related question of whether LLMs can tell what is interesting or novel. The answer to this question will have a major impact on performing, funding, and publishing science. Focusing for a moment on proposals and papers, we are aware that LLMs are already being used by a large number of scientists in writing papers and proposals, reviewing papers and proposals, and even writing responses to those AI-generated reviews. While we don't have insider knowledge, we would be shocked if granting agencies (public and private) as well as publishers are not using LLMs to at least screen papers and proposals, if not to participate directly in the review process. Granting agencies and publishers are taking steps to try to prevent this practice from the submission side, but there is no guarantee they will not be used on the review side (which of course raises copyright and intellectual property leakage issues). So given all of this AI-generated text fed into the next generation of models' training corpus, can we expect models to inherently detect, or generate, things that are actually novel? Our prediction is yes, in the short term these types of models will actually have enough inherent knowledge and ability to access and parse the literature that they will be more capable than a typical scientist at detecting whether an idea is novel, and probably whether it is feasible. Certainly they will be able to poke holes in the arguments of the proposals as acutely as a human reviewer, and will not get tired after reading dozens or hundreds of pages of text in a batch.

If these models can detect novelty, can they also produce novel ideas? If a human feeds in a topic with several sources and asks for new ways to combine ideas from those topics, we certainly feel the ideas generated will seem reasonable. We have tried and recommend the exercise of uploading some of your own papers and asking what would be some interesting future directions, as the results are already quite impressive. The secondary question is whether an agentic model can perform this task on its own, given a request for a new hypothesis in a certain area, can a model iterate and even test some of its own ideas to produce ideas that are completely original and cannot have been only obtained through knowledge synthesis. Recent studies show a few examples of this working (see *e.g.* ref. 20), and we expect this to become common. As researchers incorporate more and more tools into their workflow, will we have to

rethink what it means to “do research,” as the tasks that primarily occupied scientists are obviated? For the moment, in our own groups we see heavy users of LLMs still shaping the direction of their own scientific research, simply with the addition of very powerful tools; however, the balance could shift drastically towards the role of humans being to come up with questions and away from the tactile process of checking hypotheses, with as yet unknown consequences.

One can debate whether distilling and regurgitating other knowledge through an LLM counts as novel, but our opinion is that, especially combined with human input and automated experiment, this is not substantively different from human knowledge generation. So then the question becomes, can LLMs on their own (without any new model architectures) generate ideas that are truly paradigm shifts (*e.g.* quantum mechanics, theory of relativity, inferring existence of black holes, *etc.*). We speculate that, if this comes to pass, it will take far longer given the way that AI model performance decays over the course of longer and longer tasks, while certain scientific advances have so far required continuous effort over years or decades. Yet at the same time, the existence of even the currently existing tools combined with human ingenuity will vastly shrink the time lag between large discoveries *versus* what it would have been otherwise. That is of course assuming that human scientists do not lose their inherent curiosity, tenacity, and access to stable funding and resources, none of which we take as a given.

Author contributions

Both authors contributed equally.

Conflicts of interest

A. D. W. is head of science at FutureHouse, a biotechnology research nonprofit that is developing tools using LLMs for scientific research. A. D. W. has shares in Edison Scientific Inc., Pauling AI, and Acellera, which are all companies working on automating parts of scientific discovery.

Data availability

All data is provided within the article.

Acknowledgements

G. M. H. acknowledges the support of the National Institutes of Health through the award R35GM138312.

Notes and references

- 1 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- 2 G. M. Hocky and A. D. White, *Digital Discovery*, 2022, **1**, 79–83.



- 3 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
- 4 E. Kraev, *arXiv*, 2018, preprint, arXiv:1811.11222, DOI: [10.48550/arXiv.1811.11222](https://doi.org/10.48550/arXiv.1811.11222).
- 5 M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, *arXiv*, 2021, preprint, arXiv:2107.03374, DOI: [10.48550/arXiv.2107.03374](https://doi.org/10.48550/arXiv.2107.03374).
- 6 A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh, *et al.*, *Digital Discovery*, 2023, **2**, 368–376.
- 7 S. Back, A. Aspuru-Guzik, M. Ceriotti, G. Gryn'ova, B. Grzybowski, G. H. Gu, J. Hein, K. Hippalgaonkar, R. Hormázabal, Y. Jung, *et al.*, *Digital Discovery*, 2024, **3**, 23–33.
- 8 Y. Du, C. Duan, A. Bran, A. Sotnikova, Y. Qu, H. Kulik, A. Bosselut, J. Xu and P. Schwaller, *ChemRxiv*, preprint, 2024, DOI: [10.26434/chemrxiv-2024-h722v](https://doi.org/10.26434/chemrxiv-2024-h722v).
- 9 C. L. Vizcarra, R. F. Trainor, A. Ringer McDonald, C. T. Richardson, D. Potoyan, J. A. Nash, B. Lundgren, T. Luchko, G. M. Hocky, J. J. Foley IV, *et al.*, *Nat. Comput. Sci.*, 2024, 1–2.
- 10 A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, *et al.*, *Nat. Chem.*, 2025, 1–8.
- 11 S. M. Narayanan, J. D. Braza, R.-R. Griffiths, A. Bou, G. Wellawatte, M. C. Ramos, L. Mitchener, S. G. Rodrigues and A. D. White, *arXiv*, 2025, preprint, arXiv:2506.17238, DOI: [10.48550/arXiv.2506.17238](https://doi.org/10.48550/arXiv.2506.17238).
- 12 J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodrigues and A. D. White, *arXiv*, 2023, preprint, arXiv:2312.07559, DOI: [10.48550/arXiv.2312.07559](https://doi.org/10.48550/arXiv.2312.07559).
- 13 S. Narayanan, J. D. Braza, R.-R. Griffiths, M. Ponnappati, A. Bou, J. Laurent, O. Kabeli, G. Wellawatte, S. Cox, S. G. Rodrigues *et al.*, *arXiv*, 2024, preprint, arXiv:2412.21154, DOI: [10.48550/arXiv.2412.21154](https://doi.org/10.48550/arXiv.2412.21154).
- 14 F. Luo, J. Zhang, Q. Wang and C. Yang, *ACS Cent. Sci.*, 2025, **11**, 511–519.
- 15 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, *Nat. Mach. Intell.*, 2024, **6**, 161–169.
- 16 G. M. Hocky, *Nat. Mach. Intell.*, 2024, **6**, 249–250.
- 17 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
- 18 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, **6**, 525–535.
- 19 J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno *et al.*, *arXiv*, 2025, preprint, arXiv:2502.18864, DOI: [10.48550/arXiv.2502.18864](https://doi.org/10.48550/arXiv.2502.18864).
- 20 L. Mitchener, A. Yiu, B. Chang, M. Bourdenx, T. Nadolski, A. Sulovari, E. C. Landsness, D. L. Barabasi, S. Narayanan, N. Evans, S. Reddy, M. Foiani, A. Kamal, L. P. Shriver, F. Cao, A. T. Wassie, J. M. Laurent, E. Melville-Green, M. Caldas, A. Bou, K. F. Roberts, S. Zagorac, T. C. Orr, M. E. Orr, K. J. Zvezdaryk, A. E. Ghareeb, L. McCoy, B. Gomes, E. A. Ashley, K. E. Duff, T. Buonassisi, T. Rainforth, R. J. Bateman, M. Skarlinski, S. G. Rodrigues, M. M. Hinks and A. D. White, Kosmos: An AI Scientist for Autonomous Discovery, *arXiv*, 2025, preprint, arXiv:2511.02824, DOI: [10.48550/arXiv.2511.02824](https://doi.org/10.48550/arXiv.2511.02824), <https://arxiv.org/abs/2511.02824>.
- 21 S. X. Leong, S. Pablo-García, Z. Zhang and A. Aspuru-Guzik, *Chem. Sci.*, 2024, **15**, 17881–17891.
- 22 Z. Zhao, Z. Huang, J. Li, S. Lin, J. Zhou, F. Cao, K. Zhou, R. Ge, T. Long, Y. Zhu *et al.*, *arXiv*, 2025, preprint, arXiv:2512.01274, DOI: [10.48550/arXiv.2512.01274](https://doi.org/10.48550/arXiv.2512.01274).
- 23 H. Li, X. Fang, Y. Li, C. Huang, J. Wang, X. Wang, H. Bai, B. Hao, S. Lin, H. Liang *et al.*, *arXiv*, 2025, preprint, arXiv:2512.23565, DOI: [10.48550/arXiv.2512.23565](https://doi.org/10.48550/arXiv.2512.23565).
- 24 N. Alampara, M. Schilling-Wilhelmi, M. Ríos-García, I. Mandal, P. Khetarpal, H. S. Grover, N. A. Krishnan and K. M. Jablonka, *Nat. Comput. Sci.*, 2025, **5**, 952–961.
- 25 T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, *et al.*, *Science*, 2025, **387**, 850–858.
- 26 Ž. Avsec, N. Latysheva, J. Cheng, G. Novati, K. R. Taylor, T. Ward, C. Bycroft, L. Nicolaisen, E. Arvaniti, J. Pan, *et al.*, *Nature*, 2026, **649**, 1206–1218.
- 27 S. F. Chen, R. J. Steele, G. M. Hocky, B. Lemeneh, S. P. Lad and E. K. Oermann, *PLoS One*, 2026, **21**, e0341501.
- 28 M. C. Ramos, C. J. Collison and A. D. White, *Chem. Sci.*, 2025, **16**, 2514–2572.
- 29 X. Sun, J. Liu, B. Mahjour, K. F. Jensen and C. W. Coley, *Chem. Sci.*, 2025, **16**, 18176–18189.
- 30 Y. Zhang, Y. Han, S. Chen, R. Yu, X. Zhao, X. Liu, K. Zeng, M. Yu, J. Tian, F. Zhu, *et al.*, *Nat. Mach. Intell.*, 2025, 1–13.
- 31 H. Li, S. Sarkar, W. Lu, P. O. Loftus, T. Qiu, Y. Shee, A. E. Cuomo, J.-P. Webster, H. Kelly, V. Manee, *et al.*, *Nature*, 2026, 1–3.

