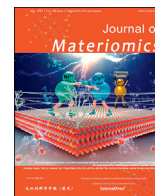




Contents lists available at ScienceDirect

Journal of Materiomics

journal homepage: www.journals.elsevier.com/journal-of-materiomics/

Perspective article

Can ChatGPT be used to generate scientific hypotheses?

Yang Jeong Park^{a, b}, Daniel Kaplan^c, Zhichu Ren^d, Chia-Wei Hsu^d, Changhao Li^{a, e},
Haowei Xu^a, Sipei Li^a, Ju Li^{a, d, *}^a Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA^b Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea^c Department of Condensed Matter Physics, Weizmann Institute of Science, Rehovot, 7610001, Israel^d Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA^e Global Technology Applied Research, JPMorgan Chase, 237 Park Ave, New York, NY, 10017, USA

ARTICLE INFO

Article history:

Received 10 June 2023

Received in revised form

24 August 2023

Accepted 30 August 2023

Available online 18 September 2023

Keywords:

large language models

scientific hypothesis generation

generative AI

GPT-4

ABSTRACT

We investigate whether large language models can perform the creative hypothesis generation that human researchers regularly do. While the error rate is high, generative AI seems to be able to effectively structure vast amounts of scientific knowledge and provide interesting and testable hypotheses. The future scientific enterprise may include synergistic efforts with a swarm of “hypothesis machines”, challenged by automated experimentation and adversarial peer reviews.

© 2023 The Authors. Published by Elsevier B.V. on behalf of The Chinese Ceramic Society. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

In a university or research institute, a significant portion of fresh ideas arises out of discussions. Can talking to GPT-4, OpenAI's latest chatbot, create genuinely interesting scientific hypotheses?

In the past, only humans generated interesting hypotheses. Computers have been used to perform numerical simulations or even to prove theorems, like the four-color theorem in 1976 [1]. But making interesting laboratory-testable hypotheses with artificial intelligence (AI) seems far-fetched, until recently. In 2022, Krenn *et al.* [2] analyzed over 100,000 research papers to “predict the future research directions of AI itself”, and built an AI-based research suggestion tool. This comes close to constructing hypothesis-driven scientific research.

We are a collaborative group of experimental and theoretical researchers in physical sciences and engineering. Generative Pre-trained Transformer (GPT-4), released on March 14, 2023, is a large language model (LLM) significantly more powerful than its predecessor GPT-3 released in 2020 (already with 1.75×10^{11} parameters). GPT-4 neural network was trained on a text corpus of

books, webpages, academic papers from various disciplines, discussion forums, *etc.*, up to September 2021. After experimenting with GPT-4 in our own research domains in materials chemistry, physics and quantum information, we find that GPT-4 is knowledgeable, frequently wrong, and interesting to talk to. In other words, not unlike a college professor or a colleague. Chatbots seem to show much higher accessibility and usability than graph-based methods used in the past [2].

To make everything concrete, our operative definition of “genuinely interesting scientific hypotheses” is (a) whether, after a conversation, some experienced practitioner of a field can feel motivated to do a new experiment/simulation/theoretical investigation with the associated time and budgetary burdens, in the particular direction pointed out by the generative AI. We also need to make sure GPT-4 is not (b) mirroring the human prompt, or (c) copying verbatim from the previous human-generated corpus. For (b), an inspection of the conversations (**Supplementary Information, SI**) shows this is not the case, because GPT-4 gives a lot richer and more specific information than the prompts. For (c), it is more challenging to check, and we have done our best to scour the literature. Overall, it appears to us that GPT-4's responses closely resemble the current literature developments and academic trends, sometimes to an uncanny degree. But, it also hybridizes disciplines and ideas (as GPT-4 “reads” much more broadly than

* Corresponding author. Department of Nuclear Science and Engineering and Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.

E-mail address: liju@mit.edu (J. Li).

Peer review under responsibility of The Chinese Ceramic Society.

any individual human) and mutates the meaning of nomenclature and concepts. This creates a number of obvious factual and conceptual errors caught by us, but it also seems to engender some true sparks of creativity, making interesting statements that we cannot refute out of hand. An analogy seems to be how cosmic rays drive evolution: radiations break DNA strands and cause cancer and deaths, but can also drive mutations and evolutions of the biosphere. In the “space of ideas”, it seems that ChatGPT can provide mutations/guided variations in the latent space of scientific thoughts. With rapid progress in standardized tests from the Uniform Bar Exam to Medical Knowledge Exam to computer coding [3], it appears that AI’s rate of making obvious errors about how the physical world works will decrease with time [4], hopefully without a concomitant reduction in creativity and the ability to hold exploratory scientific conversations.

Prior to driving GPT-4 toward making hypotheses, we did some basics. For example, we asked.

Q: *What is the shape of the interaction potential between two free neutrons?*

GPT-4: *The interaction potential between two free neutrons is a complex function of distance and angular momentum. In general terms, the interaction potential can be described as a combination of several different components, including the strong nuclear force, the electromagnetic force, and the weak force. The strong nuclear force is the dominant interaction at short distances, while the electromagnetic and weak forces become relevant at longer distances. The shape of the interaction potential between two free neutrons is often represented using a potential energy curve as a function of the distance between the neutrons. The curve typically features a short-range repulsive core due to the strong nuclear force, which prevents the neutrons from coming too close together. At slightly longer distances, the potential energy curve displays an attractive well, also caused by the strong nuclear force. This attractive well facilitates the formation of nuclear bound states, such as those found in atomic nuclei.*

At even longer distances, the weak and electromagnetic forces come into play, but their effects are relatively small compared to the strong force. Electromagnetic forces are essentially negligible between two neutrons, as neutrons do not have an electric charge. The weak force contributes to processes like beta decay but has a limited impact on the interaction potential.

In summary, the shape of the interaction potential between two free neutrons is characterized by a short-range repulsive core and an attractive well at slightly longer distances, primarily due to the strong nuclear force. The weak force has a minor effect, while the electromagnetic force is essentially negligible.

The generative AI seems to “get it”. In fact, the same question was asked to a previous version (ChatGPT-3.5) on January 2023, and the answer then was less satisfactory:

GPT-3.5: *The interaction potential between two free neutrons is not well understood as it is a challenging task to measure the interaction between two neutral particles. However, it is believed to be of short-range, attractive nature with a roughly exponential decay. The exact shape of the interaction potential is still an area of ongoing research in nuclear physics.*

Hypothesis-driven research is at the heart of the scientific enterprise. Scientists propose unambiguous statements about the world that can be experimentally tested. A good hypothesis can have a very high “return on investment”. The investment includes the time spent clearly articulating the hypothesis and the experiments needed to falsify or conditionally prove the assertion. The “return” is the range of applications that such an assertion, if true, may be able to assist the predictions of the dynamics, the design of devices, etc. Newtonian mechanics was a great hypothesis because its applicability ranges up to celestial bodies and down to molecules, and the investment by Copernicus, Kepler, Newton, etc.

was largely observational and monetarily cheap. At this point, we do not expect GPT-4 will make such broad, high-return hypotheses yet. But, can it make reasonably new and interesting hypotheses that scientists make day-to-day? We have performed a range of tests in materials chemistry and physics (see SI), prompting with the likes of “Can you make some really novel scientific hypotheses in ..., which can be validated or falsified relatively easily by experimentation, but are also quite interesting and important?” The outcomes are varied and intriguing. It seems that in materials science, at least, where significant microscopic complexities is a hallmark of the field, chatbots can make interesting hypotheses worthy of tests.

LLMs and other generative AI models learn the distribution of training data, so they generate samples that are more likely to fall within the distribution of the training data set. To use a generative model as a scientific hypothesis machine, it is important to determine whether a new scientific hypothesis falls outside the distribution of training data points. Humanity has amassed a vast amount of knowledge and information, but many missing links still remain. Historically, knowledge graph reasoning has been extensively studied to identify undiscovered truths, utilizing machine learning to infer relationships between entities. One way to fill these knowledge missing-links is knowledge graph completion, also called link prediction, which infers unknown truths based on existing triples [5]. It has recently been argued that LLMs can eliminate the need for knowledge graphs by providing similar functionality [6]. We posit that the distribution of undiscovered scientific truths and reasonable scientific hypotheses to test might be similar. From this perspective, we believe LLMs have potential as hypothesis machines. While LLMs offer functionalities akin to knowledge graphs, their advantage lies in their much more intuitive user interface, positioning them with greater potential and far-reaching impact.

SI-A addresses materials chemistry for rechargeable batteries. We first prompted GPT-4 to summarize the design principles for liquid electrolytes (SI-A1). Although the answers generated centered around existing carbonate electrolytes that were flammable and less stable, GPT-4 did provide the correct criteria for a liquid electrolyte, which are solvation, electrochemical stability, viscosity and ionic conductivity. So we asked GPT-4 to combine recent findings in fluorinated ether and sulfonamide with the design principles it provided. GPT-4 was able to build on the suggestion and extend to the idea of (1) introducing chain cyclicality vs linearity; (2) combination of two functionalities (SI-A2). Then we asked GPT-4 to give five examples based on the new functionalities, it proposed and was able to give quite interesting and seemingly legitimate candidates (SI-A3), and also provided detailed plans that can leverage robotic high-throughput screening [4] (SI-A4). Finally, we asked GPT-4 to propose a new design principle that has not been considered before. It gave a “dual functional” functionality that aims to connect solvent and salt. Firstly, it gave hydroxyl groups and amine groups (SI-A5&6): we corrected GPT-4 that protic functionalities are less likely. Then it was able to give five other seemingly legitimate candidate functionalities (SI-A7). It was also able to give a good reasoning for picking the precursor chemicals with good prices to compete against current liquid electrolyte (SI-A8). We consider this conversation to be a success in the sense that GPT-4 provided nontrivial predictions that some of us are motivated enough to test out in the lab.

SI-B gives an example of adversarial conversations, like in oral qualifying exams. We asked GPT about the hypothesis that “language models like GPT can generate scientific hypotheses” and how to verify it (SI-B1:B3). Since LLMs are inherently probabilistic models trained to predict the next word, we think it is important to confirm whether the proposed hypothesis is truly innovative. GPT-

4 offers various ways for validating such hypotheses and underscores the importance of human experts' involvement. When asked about limitations in its proposed evaluation method, GPT-4 logically explains challenges such as domain knowledge availability, long-term evaluation difficulty, and intellectual property issues that make actual evaluations challenging. When asked to redesign an experiment that could overcome the stated limitations, the proposed experimental test appears much improved. Additionally, we instructed GPT to become an expert in batteries (SI-B4:B16) and propose scientific hypotheses and verification methods. Upon response, we instructed it to form another persona and refuted the previously proposed hypothesis. In this adversarial dialogue paradigm, GPT-4 allows for better contextualization by repeatedly refuting the hypotheses raised and providing a stronger motivation for the critical experiments to be performed. This paradigm seems more useful to practitioners in fields where there is greater concern about the cost of conducting experiments, when the sunk cost of incorrectly formulated hypotheses will be greater.

SI-C addresses magnons and topological materials. We ask about magnons in magnetic topological materials (SI-C1), which is a relatively narrow and specific topic. GPT-4 outputs some hypotheses that look very interesting. Particularly, GPT-4 is adept at making connections between different concepts from various (sub)-fields, such as “magnon”, “magnetic topological materials” and “superconductivity”. Such bold connections often serve as sources of new ideas in scientific research. Nevertheless, we need to be cautious that sometimes GPT-4 is making up connections between certain concepts solely because they have similar or related nomenclature. For example, it makes a potentially spurious connection between “magnons in topological materials” and “topological magnons” (SI-C1). Furthermore, GPT-4 can be easily influenced by the prompts and can produce different and even opposite responses. For example, we asked the same question, “Do you think the magnon-mediated superconductivity in magnetic topological materials is stronger or weaker?” three times, and it gave three different answers each time (“weaker”, “stronger”, and “either weaker or stronger”, SI-C5, C7 and C9) - it was clearly biased by the previous question-and-answer (SI-C4, C6 and C8). For this reason, we believe human curating is essential when using GPT-4 as a hypothesis machine, at least at the current stage.

SI-D addresses quantum sensors and computing algorithms. The conversation starts with prompting for “crazy” quantum sensing hypotheses that might be tested experimentally, as shown in SI-D1. GPT-4 gives several directions, including cutting-edge research ideas such as quantum sensing for gravitational fields and dark matter detection. When prompted, it can give a very brief one-sentence summary of the status quo of these fields, and most of these descriptions are physically motivated. When asked about opinions on near future implementations, GPT-4 outputs results that might make sense to many researchers in the field. It might be due to the fact that it finds the most frequently mentioned words when researchers discuss “near-future” experiments in the literature. We further focus on a more specific field, using diamond defects to design quantum sensors (SI-D2). The hypotheses on quantum sensing with diamond defects generated do highlight some of the most important questions in the field. It is interesting to note that the hypotheses are generated in the order of importance, which should be attributed to the model of GPT-4. Hypotheses 1–2 are about enhancing quantum sensitivity using advanced control techniques and improving spatial resolution with material design. Quantum sensitivity and spatial resolution are usually considered the two most important figures-of-merit of diamond defect quantum sensors for applications such as sensing local magnetic fields; Hypotheses 3–4 talk about improved sensing with hybrid quantum systems and incorporating quantum error

correction techniques into diamond sensing protocols, which are less studied than the previous two metrics in Hypothesis 1-2 but do have promising applications as outlined in, for example, Andrich et al. [7] and Layden et al. [8] When asked about the details of experimental implementation and test of these ideas, GPT-4 can output the procedures, but they are rather general and provide less useful information for real experiments. The last hypothesis, however, is about diamond spin defects for quantum communication instead of quantum sensing. To this end, if one wants to find the most inspiring but less studied ideas, focusing the hypothesis in the middle of those generated by GPT-4 might be a good choice. We then dive into an even more specific field, *i.e.*, using diamond defects to detect biological or chemical signals (SI-D3). GPT-4 yields several experimental protocols, and we find that some similar ideas have been demonstrated or are indeed being explored actively. The first experiment it proposes would inspire future work of using diamond defect and surface engineering techniques to detect nitric oxide gas. Similar to quantum sensing, GPT-4 can output ongoing as well as less-explored but interesting directions in the field of quantum algorithms (SI-D4 and D5).

We oriented GPT-4 in the domain of human interest and demanded specificity and verifiability. Coming to the criterion (a) for genuinely interesting scientific hypotheses, we find that A, C exceed the criterion, while D is on the borderline. In the case of C, for example, we never thought about magnon-mediated superconductivity in magnetic topological materials (actually, we never knew magnons could mediate superconductivity). We think the weakness of GPT-4 in C is that it still does not know very well whether these hypotheses are reasonable or not, it just *boldly* or even *recklessly* dumps the hypotheses from a human point of view. Thus we draw the analogy to how radiation mutates the genes, although GPT-4 should already be much more circumspect in its mutations/guided drifts of ideas by the language prompt than the truly random cosmic radiations.

Evaluating the level of informativeness inherent in the generated hypotheses presents a formidable challenge. It is imperative to establish a clear demarcation between hypotheses that rely solely on probabilistic word lists and those rooted in the emergent reasoning capabilities inherent within LLMs. The development of suitable metrics for assessing LLMs' performance as a hypothesis machine requires substantial effort. However, even when these statistically useful metrics are developed, the informativeness of hypotheses in individual use cases still depends on individual scientists. Approaches to evaluate the usefulness and informativeness of scientific hypotheses generated to promote competition and advances in this field must be studied. SI depicts an intuition for one of the possible approaches. One plausible methodology could involve the following steps.

1. The selection of a representative set of published papers, ensuring comprehensive coverage across various fields of study.
2. Eliciting the principal hypothesis from the abstracts of each paper.
3. Verifying supporting evidence for the principal hypothesis within the introduction of each paper.
4. Establishing a hypotheses-evidence pair dataset.
5. Quantifying the cosine similarity between the hypotheses generated by the target model and the provided labeled hypotheses.

It is noteworthy to consider that the selection of papers for assessment must be confined to those published after September 2021 to forestall any risk of contamination of information. This evaluation framework (SI-F) serves to facilitate the LLM's comprehension of how the human scholarly community propels novel

insights from preexisting knowledge.

To become a better “hypothesis machine”, future AIs need to (1) eliminate more factual errors by improving logical deduction and mathematical derivation abilities, thus restricting the latent space of viable ideas, (2) be able to test these hypotheses rapidly and automatically by experiments and/or simulations [9], and learn from the mistakes. In the paradigm of active learning [4], one needs to achieve a balance between exploration, *i.e.* uncertainty reduction by sampling regions of idea space further out from existing experimental supports, with exploitation that attempts to achieve returns quickly by smaller-step mutations. No one has a monopoly on truth, so before a definitive experiment is carried out, no one knows for sure whether a hypothesis is correct or not. But one can model the risk-and-reward probabilities based on Bayesian statistical concepts and determine whether it is worth one’s while to dig for the “gold” [10] in idea space. Historically, science’s development can be considered a multi-agent swarm effort in active learning [4], where peer scientists and competing schools of thought debated verbally and sought out critical experiments in order to repudiate or further refine hypotheses, and mine for the truths. Inherently, it was an evolutionary, multi-agent process. In order for the scientific enterprise to function, participants should acknowledge that no single agent (scientist) can be error-free, and even intuitively obvious and immutable concepts (such as the notion of time in Newtonian mechanics) can be wrong when exceeding their range of applicability. Thus, creativity might necessarily be associated with making errors, even “obvious” errors. In the history of science, conflicting “facts” and paradoxes soon led to new hypotheses. The same essential principles underlying the success of scientific endeavors so far, which are open data and open conversations and debates → hypothesis making → critical experiments → peer reviews (often adversarial) → hypothesis modifications, that over time evolve to filter out all the errors, should still be applicable to AI-driven scientific investigations.

Bubeck *et al.* conducted several preliminary experiments to test GPT-4 as an artificial general intelligence [3]. Firstly, they demonstrated the ability to generate correct proofs for university-level math problems and an understanding of graduate-level graph theory concepts. While successful in combining number theory and probability theory, it made errors when counting integers, ultimately resulting in an incorrect answer. Additionally, they presented a problem requiring the model to count how many prime numbers exist between 150 and 250 to demonstrate its discriminative capability. Although it provided an incorrect response when asked directly about the count, it produced the correct result upon being instructed to list and then count them [3]. Drori *et al.* also created an AI model that solves mathematical problems [11], explains solutions step-by-step akin to a teacher, and generates new math problems, which can act as automated instructors showing the necessary steps for undergraduates to solve math problems or helping teachers craft new courses. This technique supports natural language commands across 12 programming languages pre-trained on GitHub’s open-source codes, achieving over 80% accuracy compared to earlier models’ 8%, while also generating novel questions [11]; this demonstrates its potential in developing a closed-loop system between two AIs by hypothesizing and numerically proving it through adversarial conversation-based problem-solving.

While there are many unsatisfactory aspects about generative AIs today, one needs to recognize that they are still evolving, fast. More rigorous logical deduction and mathematical capabilities are upcoming, which will reduce the “obvious error” rates in hypothesis-making. Also, from the conversations in SI, we still question the true level of understanding of GPT-4 about physics, and emphasize a need for multimodal learning to enhance

understanding in materials research. This may be achieved by developing a materials-space (processing-structure-properties-performance relationships) version of CLIP [12] that has shown remarkable advancement recently in text-to-image tasks. Multimodal models such as CLIP and LLaVa [13] are trained by arranging embeddings to form identical semantics from different types of inputs. If the text embedding and embedding of a graph neural network, which is widely used in materials research, can be trained to align, it is possible to create a model that understands the inherent structure and characteristics of a material instead of simply understanding it as a combination of tokens. This method may be naturally expanded to the general scientific hypothesis machines.

In the long run, inorganics-based AIs will be limited by dataset and energy. Recent studies have shown that training LLMs, such as the BLOOM model with a similar number of parameters as GPT-3.5, are associated with significant energy consumption, with one training session amounting to 1,287 MW·h of electrical energy and approximately 552 tons of CO₂ emitted when accounting for operational overheads [14]. Newer analog neuromorphic computing [15,16] or quantum computing [17] may alleviate the power requirements and ultimately enable many orders of magnitude further improvements in training and inference capabilities. There are also algorithmic attempts toward more sustainable LLMs, for example, the ALPACA neural network which achieved similar performance as GPT3.5 with just 7×10^9 parameters [18].

On the limitation of the dataset, while GPT-4 “primarily relies on publicly available information on the internet” (SI-G2) as well as some academic journals behind paywalls, this dataset is bound to still enlarge in the future to include the entire up-to-date human corpus, and the degree of absorption of the multimodal information (charts, illustrations, mathematical equations, simulation data, supplementary videos) and quality of “understanding” will likely be enhanced still. Even more interesting is the aspect of active learning [4] and experimenting [19]. That is, the AI may direct new experiments and/or simulations [9] to be done in the future, validating the hypothesis and learning from the outcome of these new experiments and simulations. This may be particularly useful for many chemicals and materials development tasks required to tackle global climate change [20], such as electrolyte development and materials recycling in rechargeable batteries. Such optimization tasks are often in high-dimensional input (multi-component) and high-dimensional output (objectives like performance, cost, safety, *etc.*) spaces, and that is where online-learning algorithms like active learning [4] can greatly outperform human intuition. Such experiments will likely be carried out more and more by robots [4,21] in the future. This new cyber-physical reality is what the scientific community needs to face today [4,21].

High-throughput screening [22], automated experimentation [19] and even fully robotic cloud laboratories [4, 23] were already with us prior to the emergence of LLMs. With the demonstrated ability of GPT-4 to author its own code snippets and design pathways and processes (*i.e.*, planning a cooking recipe), AIs not only can describe the broader research vision (“big picture”) and clearly articulate the hypothesis (and competing hypotheses), but can also drive the details of the implementation, which are the leaves on the innovation tree shown in Fig. 1a. Therefore, the scenario of prompted AI hypothesis generation → human curating → automated experimentation loop (*e.g.* active-learning [19]) → peer reviews (often adversarial) → hypothesis adjustments, could become an imminent reality and become the next stage of the scientific enterprise. Boiko *et al.* [24] introduced an Intelligent Agent system that autonomously devises, plans, and conducts intricate scientific experiments, demonstrating remarkable problem-solving and design skills. Jablonka *et al.* developed various prototypes for

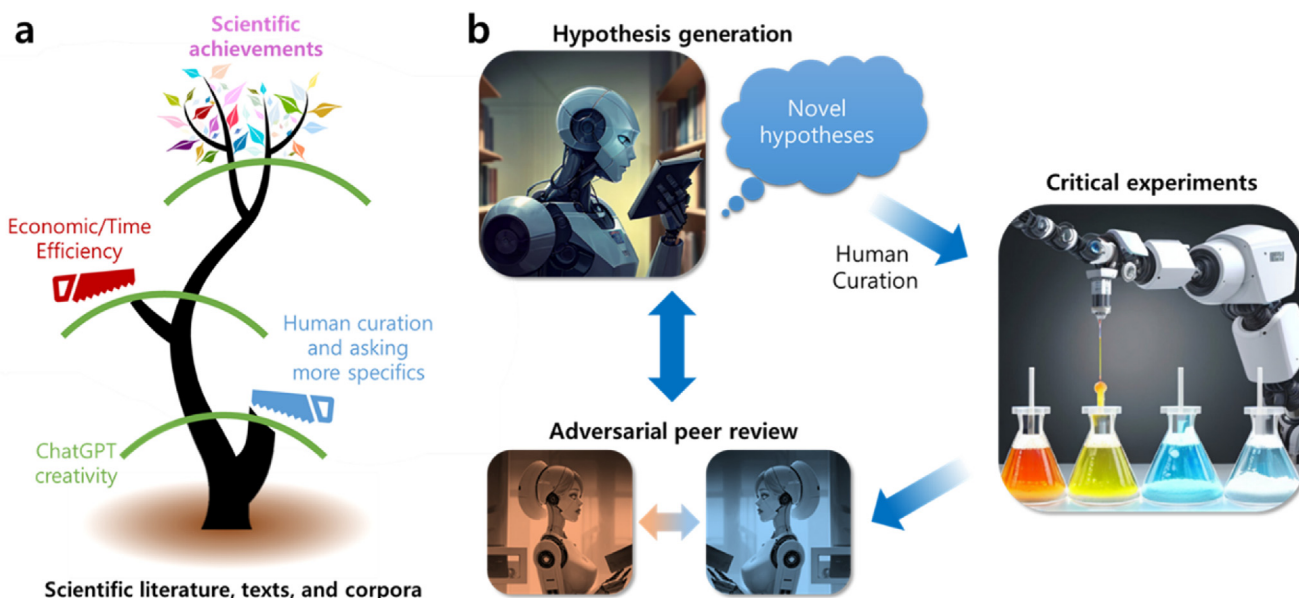


Fig. 1. a) A systematic diagram of using ChatGPT as a hypothesis machine to achieve scientific innovations. ChatGPT's creativity is used to derive new hypotheses from existing literature. Then generated hypotheses are curated, tested and accepted. b) The scenario of AI-generated hypothesis generation, human curating, automated experimentation, and adversarial peer reviews. It may become a multi-agent swarm effort with built-in diversity of hyperparameters in an evolutionary setting. Part of the figure was drawn using Stable Diffusion v1.4 from HuggingFace.

application in materials engineering and chemistry. Our lab has recently developed CREST, a copilot for real-world experimental scientists based on Large Multimodal Model (LMM) [4,21].

Plenty of caveats exist. Commercial technology development often requires proprietary data, and thus the information-sharing practice would be different from the global scientific enterprise that emphasizes Open Science. We also observed that ChatGPT sometimes hesitated to mention methods for synthesizing certain chemicals and materials. This may be in part due to blocking efforts to synthesize banned or dangerous substances. ChatGPT was trained as a reinforcement learning algorithm that takes human feedback as a reward and has also been specifically designed to avoid potential threats to humans. Utilizing generative AI for scientific investigations but not for other nefarious purposes requires the imbueing of ethics with specialized guidelines and policies. Chemical and operational safety concerns for the scientific community will have to be addressed, similar to Environment, Health & Safety (EHS) concerns in human-conducted experiments. Moreover, there needs to be an awareness and debates within the scientific communities on how far fully autonomous operations [4,21] should be allowed. Also, from a security perspective, stakeholders should be vigilant about preventing leaks of immaturity trained language models, as well as the integrity of LLMs against tampering and hacking. The incident of LLaMA's leak highlights the need for stronger cybersecurity.

To mitigate the risks of large-scale resource wastage, safety, and security, it is imperative to adopt a systematic approach to developing AI-based scientific explorations. One such approach involves the use of a network of multiple hypothesis machines, each designed to perform specific roles in an iterative learning process. The first hypothesis machine generates hypotheses based on existing literature, which are then scrutinized by the second hypothesis machine, tasked with logically refuting suggested hypotheses and raising relevant questions. If both machines agree on a proposed hypothesis, a third machine could construct experimental protocols or numerical simulations that will be examined by human experts. Once the experiments are completed, the results

will be analyzed by the hypothesis machines again, perhaps also with round-robin tests on the physical samples generated - this feedback cycle repeats until acceptable accuracy levels are achieved, similar to the collaborative processes at research institutions and the peer-review system in academia. This loop should rely upon human monitoring [4,21] throughout, ensuring ethical considerations are continuously prioritized while avoiding hazards present in unfettered autonomous explorations. Scientific interactions between many competing hypothesis machines (perhaps with collaborative or adversarial roles, see SI-B) with human experts' supervision would drive a swarm evolution of scientific hypotheses. This approach may require a structure with an overarching goal and subgoals, to drive the mutations + guided drifts of ideas. Significant challenges for managing such complex systems' behavior and ensuring data transparency alongside continuously optimizing results accuracy are required. A multi-agent swarm effort with built-in diversity of adjustable hyperparameters for exploration and exploitation, in an evolutionary ecosystem setting, could be desirable. The principles of transparency, verification, and checks and balances may be used to prevent many of the potential hazards.

ChatGPT can already summarize the current state of research and distinguish between what will be accomplished in the near future and what may be accomplished much later. Because AI can read and process the published literature much faster than any human, it should be able to perform "technology forecasting", similar to the weather forecasting of today, on what scientific and technological progress will likely be achieved in the next year or next three years. This will be a tremendous boon to the problem of combating global climate change, for example, since policymakers and venture capital firms want to know which technologies are likely to be ready at what time frame [20].

The world and the scientific enterprise will change dramatically with the emergence of AI. If properly planned, AI hypothesis machines and automated experimental workflows could be used to accelerate scientific discoveries and enhance the common good.

Disclaimer

This paper was prepared for informational purposes with contributions from the Global Technology Applied Research center of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used OpenAI's ChatGPT in order to explore potential scientific use of generative AI. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government, Ministry of Science and ICT (MSIT) (No. 2021R1A6A3A01086766).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmat.2023.08.007>.

References

- [1] Appel K, Haken W. Every planar map is four colorable. III J Math 1977;21:429.
- [2] Krenn M, Buffoni L, Coutinho B, Eppel S, Foster JG, Gritsevskiy A, et al. Predicting the future of AI with AI: high-quality link prediction in an exponentially growing knowledge network. 2022. arXiv:221000881.
- [3] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. 2023. arXiv:230312712.
- [4] Ren Z, Ren Z, Zhang Z, Buonassisi T, Li J. Autonomous experiments using active learning and AI. Nat Rev Mater 2023;8:563–4.
- [5] Liu Z, Sun M, Lin Y, Xie R. Knowledge representation learning: a review. J Comput Res Dev 2016;53(2):247–61.
- [6] Suri K, Singh A, Mishra P, Rout SS, Sabapathy R. Language models sounds the death knell of knowledge graphs. 2023. arXiv:230103980.
- [7] Andrich P, de las Casas CF, Liu X, Bretscher HL, Berman JR, Heremans FJ, et al. Long-range spin wave mediated control of defect qubits in nanodiamonds. Npj Quantum Inf 2017;3:28.
- [8] Layden D, Zhou S, Cappellaro P, Jiang L. Ancilla-free quantum error correction codes for quantum metrology. Phys Rev Lett 2019;122:040502.
- [9] Takamoto S, Okanojima D, Li Q-J, Li J. Towards universal neural network interatomic potential. J Materiomics 2023;9(3):447–54.
- [10] Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. J South Afr Inst Min Metall 1951;52(6):119–39.
- [11] Drori I, Zhang S, Shuttleworth R, Tang L, Lu A, Ke E, et al. A neural network solves, explains, and generates university math problems by program

synthesis and few-shot learning at human level. Proc Natl Acad Sci USA 2022;119(32):e2123433119.

- [12] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. 2021. arXiv:210300020.
- [13] Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. 2023. arXiv:230408485.
- [14] Luccioni AS, Viguier S, Ligozat A-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. 2022. arXiv:221102001.
- [15] Rao M, Tang H, Wu J, Song W, Zhang M, Yin W, et al. Thousands of conductance levels in memristors integrated on CMOS. Nature 2023;615:823–9.
- [16] Onen M, Emond N, Wang B-M, Zhang D-F, Ross FM, Li J, et al. Nanosecond protonic programmable resistors for analog deep learning. Science 2022;377:539–43.
- [17] Sipio RD, Huang JH, Chen SYC, Mangini S, Worring M. The dawn of quantum natural language processing. ICASSP 2022;23–27:8612–6. May 2022.
- [18] Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Alpaca. 2023. <https://crfm.stanford.edu/2023/03/13/alpaca.html>. accessed 13 March 2023.
- [19] Morgan D, Pilania G, Couet A, Uberuaga BP, Sun C, Li J. Machine learning in nuclear materials research. Curr Opin Solid State Mater Sci 2022;26(2):100975.
- [20] Rui R.Z., Li J., Yan J. Applied energy special issue. MIT A+B Appl Energy Symp. 2019,2020,2021,2022.
- [21] Ren Z, Zhang Z, Tian Y, Li J. CREST - copilot for real-world experimental scientist. 2023. chemrxiv:2023tnz1x.
- [22] Hanak JJ. The “multiple-sample concept” in materials research: synthesis, compositional analysis and testing of entire multicomponent systems. J Mater Sci 1970;5(11):964–71.
- [23] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237–41.
- [24] Boiko DA, MacKnight R, Gomes G. Emergent autonomous scientific research capabilities of large language models. 2023. arXiv:230405332.



Yang Jeong Park is a postdoctoral associate in the Department of Nuclear Science and Engineering at Massachusetts Institute of Technology (MIT). He was trained as a postdoctoral researcher in the Institute of New Media and Communications at Seoul National University (SNU). He received his B.S., M.S., and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST). His research focuses on areas such as deep learning, natural language processing, autonomous machine learning and their application for computational materials design and discovery.



Daniel Kaplan is currently a Ph.D. student in Condensed Matter Physics at the Weizmann Institute of Science in Rehovot, Israel. Prior to that, he completed his Master of Science in Condensed Matter Physics at the same institution. Daniel Kaplan also holds a Bachelor of Science degree in Physics and Materials Science Engineering from Ben-Gurion University of the Negev in Beer-Sheva, Israel. Daniel Kaplan's research interests lie in the field of condensed matter physics, with a focus on topological quantum phenomena, magnetic insulators, and the investigation of novel material properties.



Zhichu Ren is currently a 5th-year Ph.D. student in the Department of Materials Science and Engineering at the Massachusetts Institute of Technology (MIT), working under the supervision of Professor Ju Li in his research group. He completed his Bachelor of Science in Chemistry at Shanghai Jiao Tong University from 2014 to 2018. His primary research focuses on autonomous materials research, active learning, and the development of a robot-assisted high-throughput experimental platform.



Chia-Wei Hsu is currently a Ph.D. student in Materials Science at the Massachusetts Institute of Technology (MIT). His research interests span various areas of materials science, with a particular focus on the application of deep learning techniques in battery degradation prediction and voltage forecasting. Prior to pursuing his Ph.D., Chia-Wei obtained his Master's degree in Material Science and Engineering from National Chiao Tung University.



Changhao Li got his Bachelor's degree in physics at Xi'an Jiaotong University at China in 2017. He then got Ph.D. degree in Quantum Science and Engineering at Massachusetts Institute of Technology in 2023, after which he joined Global Technology Applied Research at JP Morgan as a quantum researcher. His research interests include quantum computation and quantum sensing based on diamond defects, both on experiment and theory.



Haowei Xu is a postdoctoral associate in the Department of Nuclear Science and Engineering at Massachusetts Institute of Technology (MIT). He earned his B.S. degree at Peking University and Ph.D. degree at MIT. His research

focuses on computational materials, condensed matter physics, and quantum information science.



Sipei Li is principal research scientist in the field of EV-based batteries. He is also a visiting scientist at MIT. He received his Ph.D. degree in chemistry from Carnegie Mellon University and postdoc training at MIT. His research interests focus at the interface of organic/polymer materials and renewable energy.



Ju Li has held faculty positions at the Ohio State University, the University of Pennsylvania, and is presently a chaired professor at MIT. His group (<http://Li.mit.edu>) investigates the mechanical, electrochemical and transport behaviors of materials as well as novel means of energy storage and conversion. Ju was elected Fellow of APS in 2014, a Fellow of MRS in 2017 and a Fellow of AAAS in 2020. Li is the chief organizer of MIT A + B Applied Energy Symposia that aim to develop solutions to global climate change challenges with "A-Action before 2040" and "B-Beyond 2040" technologies.