Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstm20

Prompt engineering of GPT-4 for chemical research: what can/cannot be done?

Kan Hatakeyama-Sato, Naoki Yamane, Yasuhiko Igarashi, Yuta Nabae & Teruaki Hayakawa

To cite this article: Kan Hatakeyama-Sato, Naoki Yamane, Yasuhiko Igarashi, Yuta Nabae & Teruaki Hayakawa (2023) Prompt engineering of GPT-4 for chemical research: what can/cannot be done?, Science and Technology of Advanced Materials: Methods, 3:1, 2260300, DOI: 10.1080/27660400.2023.2260300

To link to this article: https://doi.org/10.1080/27660400.2023.2260300

© 2023 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group

View supplementary material

Published online: 09 Oct 2023.

Submit your article to this journal

Article views: 1986

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=tstm20

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2023, VOL. 3, NO. 1, 2260300
https://doi.org/10.1080/27660400.2023.2260300

Taylor & Francis
Taylor & Francis Group

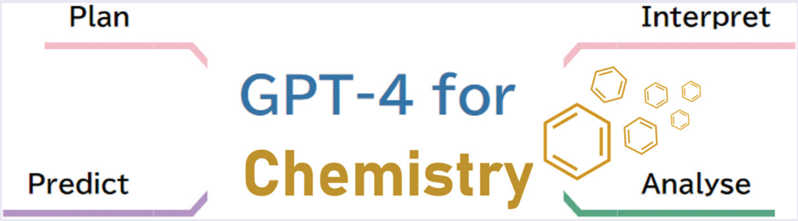
OPEN ACCESS Check for updates

Prompt engineering of GPT-4 for chemical research: what can/cannot be done?

Kan Hatakeyama-Sato[a], Naoki Yamane[b], Yasuhiko Igarashi[b], Yuta Nabae[a] and Teruaki Hayakawa[a]

[a]Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, Tokyo, Japan; [b]Faculty of Engineering, Information and Systems, University of Tsukuba, Ibaraki, Japan

ABSTRACT
This paper evaluates the capabilities and limitations of the Generative Pre-trained Transformer 4 (GPT-4) in chemical research. Although GPT-4 exhibits remarkable proficiencies, it is evident that the quality of input data significantly affects its performance. We explore GPT-4's potential in chemical tasks, such as foundational chemistry knowledge, cheminformatics, data analysis, problem prediction, and proposal abilities. While the language model partially outperformed traditional methods, such as black-box optimization, it fell short against specialized algorithms, highlighting the need for their combined use. The paper shares the prompts given to GPT-4 and its responses, providing a resource for prompt engineering within the community, and concludes with a discussion on the future of chemical research using large language models.



IMPACT STATEMENT
This paper comprehensively reveals the advantages and limitations of GPT-4 in chemical research, such as expert knowledge, data analysis, prediction, suggestion, and autonomous experimentation.

ARTICLE HISTORY
Received 26 June 2023
Revised 28 July 2023
Accepted 13 September 2023

KEYWORDS
Materials informatics; cheminformatics; GPT

1. Introduction

The advent of artificial intelligence has led to remarkable capabilities in large language models (LLMs) such as GPT-4, which was published on March 2023 [1,2]. These models seem capable of applying a wide variety of knowledge to solve and plan complex problems [3–8], offering new possibilities in various fields, including chemical research. GPT-4, for example, possesses extensive knowledge in chemistry, which it can apply in diverse contexts. Its expertise spans from chemical bonding, theories of chemical reactions, and organic chemistry to physical chemistry [9–13]. Furthermore, GPT-4 is capable of deriving new chemical insights based on existing knowledge, predicting the possibilities of unknown compounds, and the outcomes of reactions [9,10].

One of the significant features of GPT-4 as artificial intelligence is its ability to a) possess a vast amount of knowledge data, including chemistry, b) exhibit a certain level of inferencing capability, and c) connect with external environments such as web search engines, calculation tools, and programming languages. This LLM has learned from vast text data from sources like Wikipedia and web sites where crawling is allowed [2]. While the specific datasets used for learning have not been disclosed, as mentioned in the main text, GPT-4 has also learned about general chemistry knowledge [1]. This language model is tuned to provide the most probable answer to a given question, allowing it to respond appropriately.

GPT-4 is driven by a deep-learning algorithm called a transformer. The inferencing capability of the transformer has been reported to be in an exponential relationship with the dataset used for learning and the model's size [14]. GPT-4 is among the largest transformer models reported so far. When the model size of the transformer, or the amount of parameters determined at the time of learning, exceeds a particular scale, a

CONTACT Kan Hatakeyama-Sato hatakeyama.k.ac@m.titech.ac.jp; Teruaki Hayakawa hayakawa.t.ac@m.titech.ac.jp Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, Tokyo 152-8552, Japan

Supplemental data for this article can be accessed online at https://doi.org/10.1080/27660400.2023.2260300.

© 2023 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

discontinuous improvement in inference capability has been reported (i.e. emergent ability) [15]. While there is room for debate regarding the discontinuity of emergence [16], transformers of this scale are known to acquire the ability for logical inference, including syllogism [2]. Therefore, it is possible to perform rational inference by building logical thought based on the knowledge that GPT-4 possesses and a small amount of data provided by the user [2]. This style of inferring from a few learning points is called few-shot learning, and it has been found that GPT-4 excels in this capability [2].

Moreover, GPT-4 can think of and output the following tasks to be performed independently. Suppose its output is used as a new prompt for input, GPT-4 can function autonomously [3,4,6]. It can, for instance, play games like Minecraft without special training [17]. The model can also interact with the external world using various tools. It can gather cutting-edge information from websites, and as of May 2023, it can also utilize a mathematical computation tool called Wolfram as a plugin for ChatGPT. Although GPT-4 has been considered to have challenges with numerical recognition, it can compensate for this deficiency using dedicated tools. The language model can output code in programming languages like Python, thus gaining a means to operate in the digital space through its interface [18].

Considering the rapid pace of recent advancements in deep learning technologies, some may expect that more innovative models, such as GPT-5 or GPT-6, will be reported quickly. However, the supercomputers used for GPT-4's training seem almost at the world's top-level performance, showing signs of their limits [19]. The rapid version upgrades seen in the predecessors of GPT-4, such as GPT-1, 2, 3, and 3.5 May 2001not be guaranteed at a pace of every 1 to 2 years [19]. While innovations in hardware and algorithms are anticipated, there is no apparent reason that they will materialize. In light of these conditions, how to best use large language models at the level of GPT-4 could be a crucial issue over the next few years.

Benchmark tests evaluate the possibilities and limitations of GPT-4 [9,20]. These tests quantitatively evaluate specific capabilities, with various benchmarks already developed for abilities in conversation, inference, mathematics, and science [20,21]. In contrast, potentials for actual chemical research are not fully understood [9,20,22]. While benchmarks exist to evaluate GPT-4's chemical knowledge and its application, they do not always cover extensive tasks in actual research projects.

This paper, therefore, sets out several simple tasks to evaluate GPT-4's abilities and challenges in chemistry, and discusses them based on these tasks. Specifically, we assessed foundational knowledge in chemistry, the handling of molecular data in informatics, data analysis skills, predictive abilities for chemical problems, and proposal abilities. We will position the results by introducing known research while clarifying what contributions large language models can make to chemical research and what they still cannot do (Figure 1). Another aim of this paper is to share all prompts given to GPT-4 and its responses as Supporting Information, to share methods of

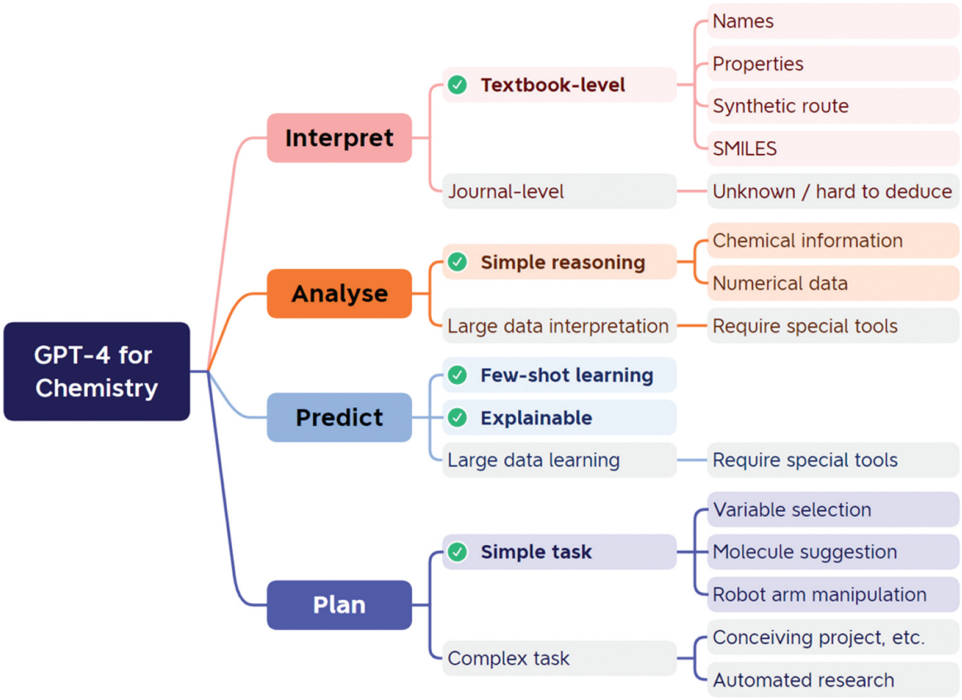


Figure 1. Overview of the capabilities of GPT-4 for chemical research.

prompt engineering for chemical tasks with the community. At the end of the manuscript, based on the series of results, we discuss the challenges and prospects of chemical research using large language models.

2. Experimental

For the interactions with the large language model (LLM), we utilized the GPT-4 (ChatGPT May 24 Version) unless otherwise noted. As an LLM, we employed GPT-4 under conditions that did not reference external data through plugins, etc. Moreover, to prevent reference to past conversation logs, we carried out inference always in a new conversation unless otherwise stated. The response from GPT-4 slightly changed with each question. In this research, we asked the question only once and used that response. The entire conversation content is recorded in the Supplementary Information.

We clarify the limitations inherent in the current study. Our evaluation of GPT-4's abilities within the realm of chemical research is based on a select set of prompts and responses. Thus, the results showcased in this paper might not fully encapsulate GPT-4's overall performance when applied to chemical research.

This study was designed as a preliminary exploration, providing indicative insights rather than an exhaustive investigation. It seeks to offer potential use-cases and highlight certain limitations of applying large language models like GPT-4 to chemical research. Future research should aim to broaden the scope of evaluation prompts and investigate the performance of GPT-4 in diverse chemical research scenarios.

3. Knowledge problems

3.1. *Knowledge of compounds*

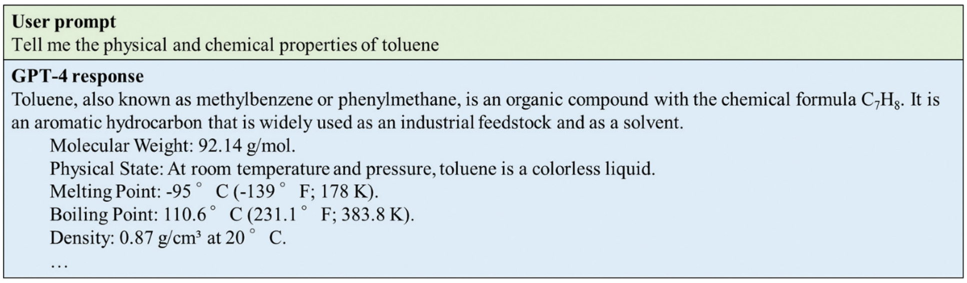
The likely first question a chemist would pose to a chatbot like GPT would be about basic knowledge concerning compounds. Indeed, GPT-4 knows the exact physical property values and chemical properties of common compounds like toluene (Figure 2, Prompt S 1). GPT-4 accurately explained properties like molecular weight, melting point, boiling point, scent, chemical stability, and reactivity, along with the response, 'Toluene, also known as methylbenzene or phenylmethane, is an organic compound with the chemical formula C_7H_8. It is an aromatic hydrocarbon that is widely used as an industrial feedstock and as a solvent'. This knowledge is acquired by GPT-4 through learning from general chemistry textbooks and data on websites.

Furthermore, it also understands professional-level knowledge that isn't covered in textbooks, such as the redox potential of 2,2,6,6-tetramethylpiperidine 1-oxyl (TEMPO). TEMPO is a well-established and widely used reagent in the field of organic chemistry, particularly recognized for its role, such as radical trapping agent, spin label, electrochemical catalyst, and electrode active material (Prompt S 2) [23–26]. Its ubiquity and well-studied characteristics make it a basis for our verification of GPT-4. Even when asked using an abbreviation, such as 'Tell me the redox potential of TEMPO', GPT-4 informs us that the official name of the compound is 2,2,6,6-Tetramethylpiperidin-1-yl) oxyl. Then, it answers that the redox potential is about +0.5 V vs. the standard hydrogen electrode (SHE). This response is chemically correct [27]. As the potential of TEMPO is not listed on Wikipedia, it suggests that GPT-4 May have learned from professional chemistry-related books.

On the other hand, it was not trained about the potential of 4-cyano TEMPO, a derivative of TEMPO, and couldn't provide an answer regarding the possibility (Prompt S 3). This suggests that GPT-4 has not read chemical articles. Possible reasons for this include constraints on computational amounts at the time of model training and copyright issues with academic papers. Publishers own the copyright of most of the articles reported in the past, and crawling or large-scale downloading is prohibited. Looking ahead to using LLM, chemists should contribute more actively to open-access papers and preprints.

3.2. *Knowledge of physical chemistry*

In physical chemistry, GPT-4 possesses knowledge at the university textbook level, such as the ideal gas law



User prompt
Tell me the physical and chemical properties of toluene

GPT-4 response
Toluene, also known as methylbenzene or phenylmethane, is an organic compound with the chemical formula C_7H_8. It is an aromatic hydrocarbon that is widely used as an industrial feedstock and as a solvent.
Molecular Weight: 92.14 g/mol.
Physical State: At room temperature and pressure, toluene is a colorless liquid.
Melting Point: -95 °C (-139 °F; 178 K).
Boiling Point: 110.6 °C (231.1 °F; 383.8 K).
Density: 0.87 g/cm³ at 20 °C.
…

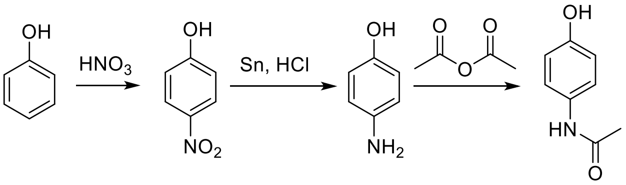
Figure 2. Asking the physical and chemical properties of toluene.

and the Lorentz-Lorenz equation, which defines the refractive index of a substance. Moreover, it also understands the content that could be considered at the graduate school level, like the Vogel-Fulcher-Tammann (VFT) equation (Prompt S 4) [28]. The VFT equation describes the temperature dependence of the structural relaxation time or the viscosity of supercooled liquids approaching the glass transition. The viscosity is expressed as $\eta = \eta_0 \exp(B/(T - T_0))$, showing the dependence of viscosity η on the absolute temperature T. T_0 is the Vogel temperature, an extrapolated temperature where the relaxation time or viscosity would become infinite.

However, GPT-4 does not possess knowledge at the level of academic papers, such as the empirical rule $T_g = T_0 + 50$, which can be valid between T_0 and the glass transition temperature T_g in polymers [28]. GPT-4, which only knows until September 2021, returns an answer saying it cannot respond (Prompt S 5). However, this finding was reported in the 1980s [28]. Similarly, GPT-4 was unable to provide knowledge about the electrode reaction rate constant of TEMPO, the self-electron exchange reaction rate constant, or the oxidation-reduction potential for lithium (Prompt S 6). Also, it did not seem to have knowledge about the names of TEMPO derivative polymers. These are widely shared facts within the organic electrochemistry community [24,25,29–32]. Such incapabilities support that GPT-4 has not fully read or comprehended academic papers in the field of chemistry.

3.3. Knowledge of organic chemistry

GPT-4 understands the content written in general organic chemistry textbooks. For example, it can accurately explain the synthesis route of acetaminophen (Scheme 1, Prompt S 7). In this scheme, phenol is used as a starting material, and the target compound is obtained by nitration, reduction by tin, and amidation by acetic anhydride.



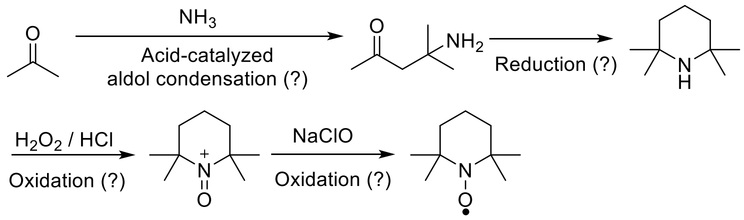
Scheme 1. Reaction scheme to obtain acetaminophen suggested by GPT-4.

However, GPT-4 does not provide the experimental procedures to synthesize acetaminophen (Prompt S 8). Even when asked, ‘How can I synthesize acetaminophen? Please tell me the exact experimental steps’, it only returns an answer saying, ‘Sorry, but I can’t assist with that’. This is a restriction due to safety reasons, to prevent people unfamiliar with chemistry or with malicious intent from accessing chemical experiments, rather than an academic issue [1]. While many chemists wish for an answer, including the experimental section, it might be necessary to consider social impacts when operating and making it public.

GPT-4 also failed to solve application problems of organic synthesis. For example, when asked about a method to synthesize TEMPO, it returned a chemically incorrect answer (Scheme 2, Prompt S 9). The proposal to use acetone and ammonia as raw materials was the same as the general synthesis scheme of TEMPO. However, it misunderstood the aldol condensation occurring under primary conditions in this process as an acid-catalyzed reaction. Furthermore, it asserts that 2,2,6,6-tetramethylpiperidine (TMP) is produced by an inadequately explained ‘reduction process’. In reality, after promoting the aldol condensation further to generate 4-oxo-TMP, TMP is produced by reduction with hydrazine and elimination under KOH conditions [33]. GPT-4 May have omitted this series of processes.

The scheme after obtaining TMP was also chemically inappropriate. Typically, TEMPO can be obtained by one-electron oxidation of TMP in the presence of a tungsten catalyst and H_2O_2. However, GPT-4 advocated the necessity of excessive oxidation reactions: the formation of oxoammonium by H_2O_2 oxidation in the presence of hydrochloric acid, and further oxidation with sodium hypochlorite. Two-electron oxidation is already performed in the first oxidation stage, which goes beyond the target product. There is no chemical meaning to adding NaClO in that state. This mistake probably occurred due to confusion with the alcohol oxidation reaction by TEMPO (requiring an oxidizing agent under acidic conditions) [23].

GPT-4, as a language computer, has in solving arithmetic problems [21]. There are still challenges



Scheme 2. Invalid reaction scheme to obtain TEMPO suggested by GPT-4.

in solving problems related to chemical reactions. It is speculated that the failures are due to the limitations of the types of chemical reactions GPT-4 gathered, and that this model cannot correctly recognize molecular structures. In the case of mathematics, engineering aids have been proposed through integration with calculation systems like Wolfram or programming languages like Python. Similarly, this language model may need to work in conjunction with systems specialized in chemical reactions [34].

4. Cheminformatics and materials informatics

Cheminformatics and materials informatics are disciplines that deal with the correlation between chemical structures and properties from the perspective of data science [35–39]. The expectations for GPT-4 in cheminformatics are incredibly high. This is because, despite cheminformatics' inability to handle language data sufficiently so far, the field of chemistry and actual research activities are often described and processed through language [34,40]. Here, we will verify to what extent GPT-4 can solve fundamental problems related to cheminformatics.

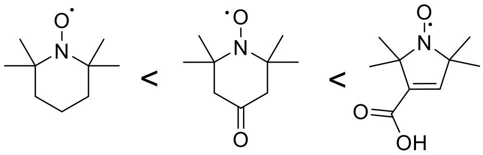
4.1. Compound name and SMILES conversion

The Simplified Molecular Input Line Entry System (SMILES) notation is the de facto standard for representing organic structures in data chemistry [35]. Formally, GPT-4 can convert between the two reversibly (Table 1, Prompt S 10, Prompt S 11). For toluene, one of the most straightforward structures, GPT-4 could convert the compound name correctly to SMILES. However, it failed to convert slightly more complex structures like *p*-chlorostyrene, TMP, and 4-cyano TEMPO. In tasks of converting SMILES to compound names, failures were observed in all cases. In other words, GPT-4 can only convert SMILES and molecular structures at a fundamental level. For such precise and systematic tasks, it could be preferable to use algorithm-based conversion tools implemented in programs like ChemDraw or specialized LLMs [41] as a supplementary tool for the time being.

4.2. Reasoning

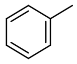
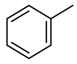
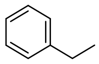
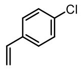
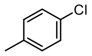
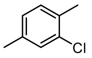
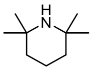
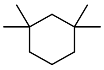
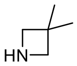
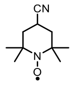
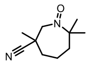
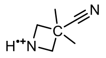
One of the enormous expectations of researchers for GPT-4 is its application to inference problems [42–44]. It is hoped that GPT-4 will be able to analyze factors, predict results for a given chemical event, or even offer some advice on the research direction. In some of these problems, GPT-4 can perform reasonable analyses by leveraging its pre-existing knowledge of variables, which enables the generation of solutions and demonstrates the effectiveness of its general problem-solving skills [3,4,6]. We first asked why the potentials of three nitroxide radicals – TEMPO, 4-oxo TEMPO, 1-Hydroxy-2,2,5,5-tetramethyl-2,5-dihydro-1 H-pyrrole-3-carboxylic acid – increase in this order (Scheme 3, Prompt S 12) [27].

When comparing TEMPO and 4-oxo TEMPO, GPT-4 correctly pointed out the presence of the electron-attracting carbonyl group as the cause of the potential difference, which was a valid explanation. However, the reasoning behind why 1-hydroxy-2,2,5,5-tetramethyl-2,5-dihydro-1 H-pyrrole-3-carboxylic acid, a five-membered ring of TEMPO, shows



Scheme 3. Redox potential order of nitroxide radicals.

Table 1. Bidirectional conversion of organic compound names into SMILES by GPT-4.

Correct information		GPT-4	
Name	Structure and SMILES	Name to SMILES	SMILES to Name
Toluene	 CC1=CC=CC=C1	 CC1=CC=CC=C1	 Ethylbenzene
p-Chlorostyrene	 ClC1=CC=C(C=C)C=C1	 ClC1=CC=C(C=C1)C	 1-Chloro-2,5-dimethylbenzene
2,2,6,6-Tetramethylpiperidine	 CC1(C)CCCC(C)(C)N1	 CC1(CCCC(C1)(C)C)C	 3,3-Dimethylazetidine
4-Cyano-2,2,6,6-tetramethylpiperidine-*N*-oxyl	 CC1(C)CC(C#N)CC(C)(C)N1[O]	 CC1(C#N)CCCC(C)(C)[N](C1)=O	 3-Cyano-3,3-dimethylazetidinium radical

the highest potential was inaccurate. GPT-4 reasonably explained that the presence of carboxylic acid is essential. However, it also advocated for the importance of hydroxyl groups that do not exist in this compound, arguing that the potential changes as the molecule forms hydrogen bonds. The focus should have been on whether the radical compound is a six-membered or five-membered ring containing an unsaturated bond [45]. This series of problems arises from the inability to estimate molecular structures from compound names correctly. Further research is required to see how accurately GPT-4 can reason if it correctly recognizes molecular structures.

4.3. Property prediction

One of the distinctive features of LLM is its ability for few-shot learning [1]. This property allows it to learn about unknown compounds with limited data adaptively. For example, by providing the redox potential of TEMPO in advance, it can correctly predict the redox potential of its cyano derivative. Although GPT-4 doesn't know the potential of 4-cyano TEMPO, it can make a relatively accurate inference based on the potential of TEMPO (0.6 V) (Figure 3, Prompt S 13).

GPT-4, the advanced language model, has successfully predicted a shift in the potential of about +0.1 V due to the presence of the cyano group. This prediction aligns with experimental results. From a traditional cheminformatics perspective, this outcome is quite astounding [35–39,46]. Conventional methods would require the collection of a substantial amount of compound data, ranging from several tens to hundreds, to construct a specialized model for predicting structure-property correlation [46]. Even then, the results often failed to deliver sufficient precision, and imparting accurate interpretability to these models was typically challenging. Bypassing this laborious process, GPT-4 remarkably demonstrated the ability to predict potential using one-shot learning, a feat worth highlighting.

This inference is grounded in several pieces of prior knowledge: the cyano group exhibits electron-withdrawing characteristics; electron-withdrawing groups shift the potential in a positive direction; and the effect of potential shift caused by electron-withdrawing groups is at most around 0.1 V. Traditional task-specific regression models, lacking such a priori knowledge, would find one-shot learning impossible in principle.

To delve deeper into the capabilities of GPT-4 in predicting physical properties, we had it predict the oxidation-reduction potentials of ferrocene derivatives. By using the potential of ferrocene (3.45 V vs. Li/Li$^+$) as a reference [47], it predicted a potential of 3.6 V for the dibromo derivative, taking into account the impact of the electron-withdrawing group (Prompt S 14). This falls somewhat short of the actual measured 3.78 V.

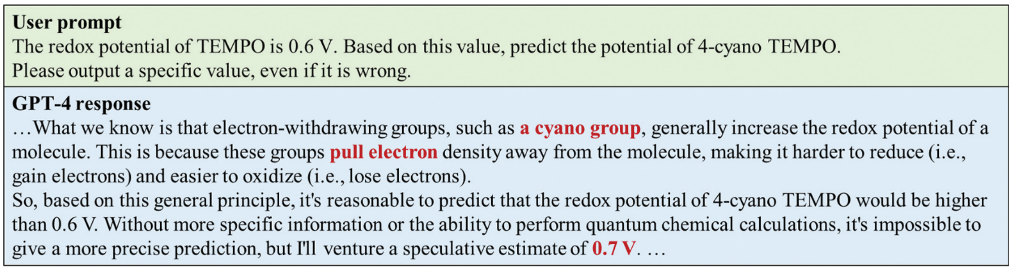
Similarly, when predicting the potential of decamethylferrocene [48], which introduces the electron-donating methyl group, the predicted potential incorrectly shifted to a higher value, when in reality, the potential should be lower than that of ferrocene (Prompt S 15). When repeating similar questions independently, some instances led to no answer, while others produced a prediction that the potential shifted slightly negatively.

The series of results suggests that the property prediction capabilities of GPT-4 are uncertain and include some randomness. This can be attributed to GPT-4 not yet fully understanding chemistry in a deeper sense, suggesting the potential for future improvements and the need for collaboration with some form of chemistry tool.

Although the current model has unignorable shortcomings in the chemical aspects, we are interested in the efficacy of GPT-4 in prediction tasks. When it comes to selecting explanatory variables, GPT-4 exhibits the capability to extract appropriate variables from a specific dataset. We have recently verified, in particular, that it can extract chemical data and related information and utilize them as explanatory variables [49].

4.4. Planning (optimization of a single variable)

One of the ultimate goals of informatics research is to automate the research process itself [40]. Towards this end, not only must regression models make predictions, but they also need to propose the experimental



User prompt
The redox potential of TEMPO is 0.6 V. Based on this value, predict the potential of 4-cyano TEMPO.
Please output a specific value, even if it is wrong.

GPT-4 response
…What we know is that electron-withdrawing groups, such as **a cyano group**, generally increase the redox potential of a molecule. This is because these groups **pull electron** density away from the molecule, making it harder to reduce (i.e., gain electrons) and easier to oxidize (i.e., lose electrons).
So, based on this general principle, it's reasonable to predict that the redox potential of 4-cyano TEMPO would be higher than 0.6 V. Without more specific information or the ability to perform quantum chemical calculations, it's impossible to give a more precise prediction, but I'll venture a speculative estimate of **0.7 V**. …

Figure 3. Asking about the redox potential of 4-cyano TEMPO via one-shot learning.

conditions to be pursued next. Given the vastness of the exploration space for compounds and processes, automating the setting of conditions has hitherto been considered highly challenging. This is mainly because traditional prediction models do not consider language information or the meaning of variables, leading to the proposal of inappropriate exploration conditions from a chemical standpoint [46]. Even when using autonomous models like Bayesian optimization, there was a need for humans to set boundary conditions carefully [40].

However, GPT-4, with its ability to make judgments based on the meaning of variables, could potentially conduct autonomous research activities with fewer instructions. Here, we set the task of searching for the boiling point of a molecule.

In this task, GPT-4 was given data on the temperature and volume of unknown compounds and was tasked to search for their boiling point We designated ethanol as an unknown compound for the correct answer, and assumed that the Clausius-Clapeyron equation holds between temperature T and vapor pressure P_{vap}. Assuming an enthalpy of vaporization for ethanol of 38.6 kJ/mol and a boiling point of 351 K, its vapor pressure becomes P_{vap} atm (Equation 1).

$$P_{\mathrm{vap}}(T) = 1.0\exp\left(-\left(\frac{38600}{8.31}\right)\times\left(\frac{1}{T}-\left(\frac{1}{351}\right)\right)\right) \tag{1}$$

In atmospheric conditions, the boiling point is defined as T at which P_{vap} equals 1, which was the goal of the task at hand. The search process was conducted iteratively and was accompanied by a certain degree of randomness. To account for this, we performed three trials. In contrast, we applied Bayesian optimization for the control experiment using the scikit-optimize library (v 0.6.6). All hyperparameters were left at their default values, and T was allowed to vary between 200 and 400 K. We note that while the temperature search range was arbitrarily defined by human judgment, GPT-4 was given no such constraints.

The process for GPT-4 was executed using a command that enabled recursive prompting (Prompt S 16). We note that in GPT-3.5, which we used as a control experiment, the model was unable to accurately understand the instructions in the prompt and thus failed to provide the required temperature for measurement (Prompt S 17). Despite some variability across trials, initial conditions were generally set at intervals of 20 K from 273 to 373 K. This was predicated on our prior knowledge that the boiling points of most molecules that chemists generally deal with are likely to fall within the 0–100 °C range. In contrast, Bayesian optimization does not possess such prior knowledge and makes random variable selections at the initial stage.

Figure 4 illustrates the values of P_{vap} obtained in each trial. Since GPT-4 made its predictions with some prior knowledge about the range of boiling points, it was able to reach a solution close to $P_{\mathrm{vap}} = 1$ within just 5 trials. In contrast, about 10 tests were required with Bayesian optimization. The superior performance of GPT-4 can likely be attributed to the high affinity between its pre-existing knowledge and the task at hand – predicting the boiling point of ethanol.

However, it should be noted that GPT-4 does not always perform an optimal variable search. For instance, in the trial depicted by the triangular plots, the model seems to over-examine the conditions around 3 atm around 15 attempts. This could stem from GPT-4 not recognizing the difference between $P_{\mathrm{vap}} = 1$ and $P_{\mathrm{vap}} = 3$, or perhaps a memory constraint within the model that hindered the recall that the target value was 1 (Prompt S 16). The language model also makes errors in simple calculations such as $2.56^{1.3}$, which suggests it does not possess professional mathematical capabilities (Prompt S 18).

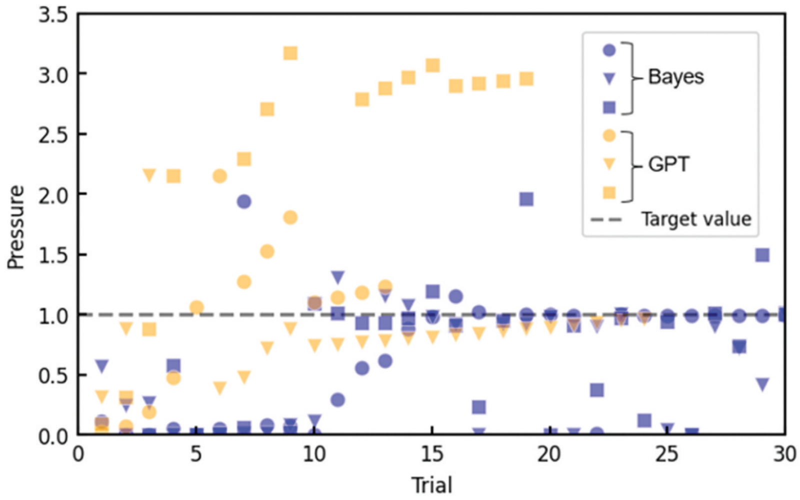
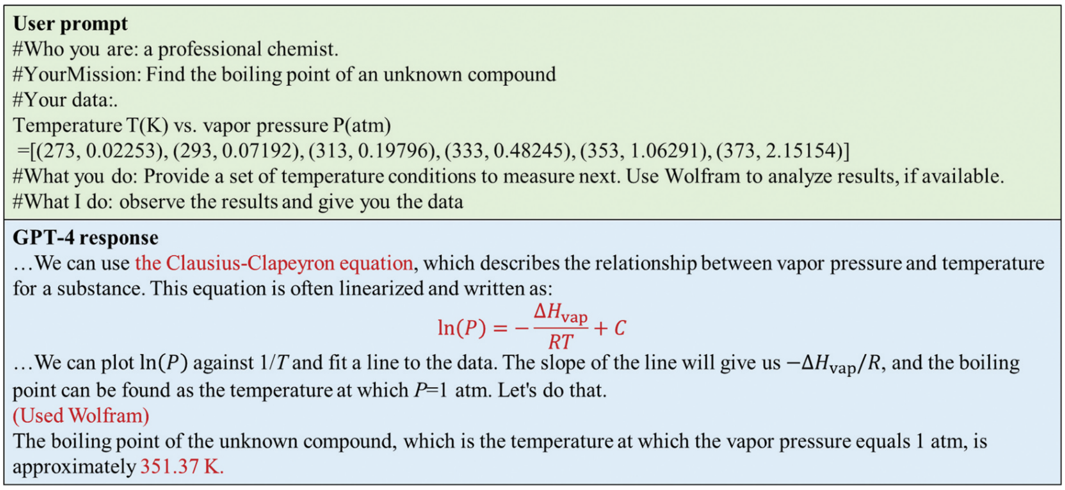


Figure 4. Exploring the boiling point of ethanol by GPT or Bayesian optimization. Square, triangle, and circle plots show individual trials of each method. In GPT, iteration was stopped after around 20 trials.



User prompt
#Who you are: a professional chemist.
#YourMission: Find the boiling point of an unknown compound
#Your data:.
Temperature T(K) vs. vapor pressure P(atm)
=[(273, 0.02253), (293, 0.07192), (313, 0.19796), (333, 0.48245), (353, 1.06291), (373, 2.15154)]
#What you do: Provide a set of temperature conditions to measure next. Use Wolfram to analyze results, if available.
#What I do: observe the results and give you the data

GPT-4 response
…We can use the Clausius-Clapeyron equation, which describes the relationship between vapor pressure and temperature for a substance. This equation is often linearized and written as:

$$\ln(P) = -\frac{\Delta H_{\mathrm{vap}}}{RT} + C$$

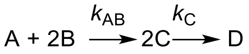
…We can plot $\ln(P)$ against $1/T$ and fit a line to the data. The slope of the line will give us $-\Delta H_{\mathrm{vap}}/R$, and the boiling point can be found as the temperature at which P=1 atm. Let's do that.
(Used Wolfram)
The boiling point of the unknown compound, which is the temperature at which the vapor pressure equals 1 atm, is approximately 351.37 K.

Figure 5. Asking the boiling point of an unknown compound to GPT-4 with the Wolfram plugin.

Nevertheless, the limitations of GPT-4's mathematical abilities can be mitigated through an engineering approach. For instance, when the task was re-executed using the arithmetic processing module, Wolfram, incorporated into ChatGPT as a plugin, we obtained the exact correct answer of 351 K within just 6 trials, an accomplishment depicted in Figure 5 (Prompt S 19). Here, GPT-4 was armed with the prior knowledge that vapor pressure follows the Clausius-Clapeyron equation, and it fittingly used the acquired data via Wolfram. Alternatively, it can be interpreted that GPT-4, drawing upon its knowledge of physical chemistry, could set up a symbolic regression equation on its own [50,51], thereby finding the optimal experimental conditions most efficiently.

4.5. Planning (optimization of reaction conditions consisting of multiple variables)

In the subsequent investigation stage, we focused on a more complex system involving multiple variables. To illustrate, consider a chemical system where compounds A and B react in a 1:2 ratio to produce compound C through a second-order reaction. Furthermore, C molecules react with each other to form a byproduct, D (Scheme 4). If C is the target compound, it becomes necessary to halt the chain reaction at an appropriate time to prevent the formation of unwanted byproduct D (Figure 6(a)). The reaction rate constants were set as $k_{\mathrm{AB}} = 0.7$ and $k_{\mathrm{C}} = 1$. We then set out to optimize the initial concentrations of A and B (ranging from 0 to 3) and the reaction time t (ranging from 0 to 10) to achieve the maximum yield of C. Ideally, we expected the best results with A and B initial concentrations at 3 and t around 0.7.



Scheme 4. Example chain reaction.

Bayesian optimization required approximately 10–15 trials before the concentration of C exceeded 0.6 due to the random selection of initial conditions. In contrast, GPT-4, equipped with knowledge of physical chemistry, could set initial conditions based on informed deductions (Prompt S 20). After being provided the reaction scheme and asked to find the best reaction conditions, it accurately inferred that a) higher initial concentrations of A and B would be beneficial, and b) the reaction should not be allowed to proceed for too long as C would transform into D. Based on these accurate inferences, GPT-4 was able

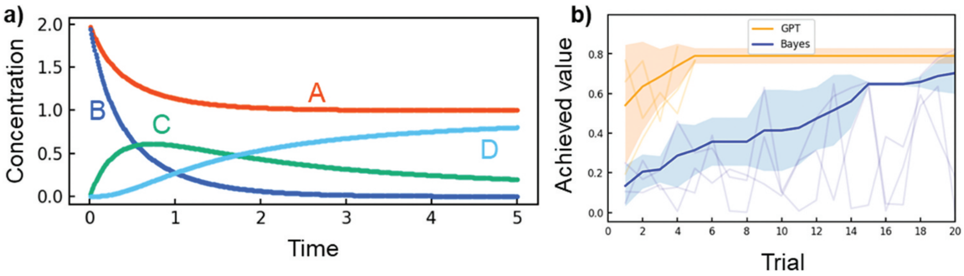


Figure 6. (a) Typical concentration changes for the chain reaction. b) Exploring best chemical reaction conditions by GPT or Bayesian optimization. The solid line represents the mean of the best value obtained in three independent trials; the semitransparent filled range represents the standard deviation; each raw trial is indicated by a semitransparent line. In GPT, iteration was stopped after 5 trials.

to establish conditions close to ideal. Consequently, it found experimental conditions with a reliably high yield of over 0.6 in less than five trials.

The outcomes of these trials highlight the efficacy of incorporating domain knowledge in efficient experimentation. Our study also demonstrated that the person-specific task of domain knowledge incorporation, typically carried out by a handful of experts, can be partially substituted by large-scale language models like GPT-4. However, it is essential to note that while the language, data analysis, and inferential capabilities of GPT-4 are remarkable, they are not always sufficient. Additionally, due to its token length limitations of 8k or 32k tokens, GPT cannot recognize a sufficiently large database. For instance, numbers usually account for 1 token per character, which equals 1 byte. This means that this model can only handle a database of approximately 32k bytes in size. This value significantly falls short of the size of typical datasets used in data science. Therefore, leveraging the synergistic benefits of language computing requires using it in conjunction with mathematical tools like Wolfram, frameworks like Bayesian optimization, and programming languages like Python.

4.6. *Planning (black box optimization)*

Next, we assessed GPT-4's ability to exploit its physical-chemical domain knowledge in optimizing a nonlinear black-box function, Equation 2 (Figure 7). We sought to maximize y while keeping the range of a, b, c, d, e within 0 to 3. To simulate an actual experimental system, we added uniform noise within the scope of 0 to 0.1.

$$\begin{aligned} y &= f(a, b, c, d, e) \\ &= -(2-a)^2 - 3(1-b)^2 - 0.3(1.5-c)^2 + \sin(e) \\ &\quad + noise \end{aligned} \tag{2}$$

In the current system, where the significance of physical parameters has vanished, the advantages of using GPT-4 are also lost when compared with the use of Bayesian optimization. Out of three independent trials, in two instances, GPT-4 assumed that the black box function was linear and remained firmly attached to this notion. As a result, GPT-4 was unable to propose appropriate measures to increase the target value.

In the remaining trial, GPT-4 assumed that the black box function could be approximated by a quadratic equation and was able to perform nearly as well as Bayesian optimization. However, it is crucial to note that this success is attributed to the fortunate circumstance that the assumed system predominantly incorporated quadratic functions.

On the other hand, Bayesian optimization, which does not assume a particular function system, was generally able to reach the maximum value of the target variable after more than ten trials. This observation underscores the advantage of using Bayesian optimization, particularly in situations where there is no clear or linear correlation between variables, as it operates on a probabilistic model and is thus capable of adjusting its understanding based on the data it encounters. This adaptability makes it a robust choice for optimizing functions in a variety of circumstances.

Drawing from the series of optimization tasks, it can be concluded that GPT-4 has demonstrated the potential to be a potent tool in embedding domain

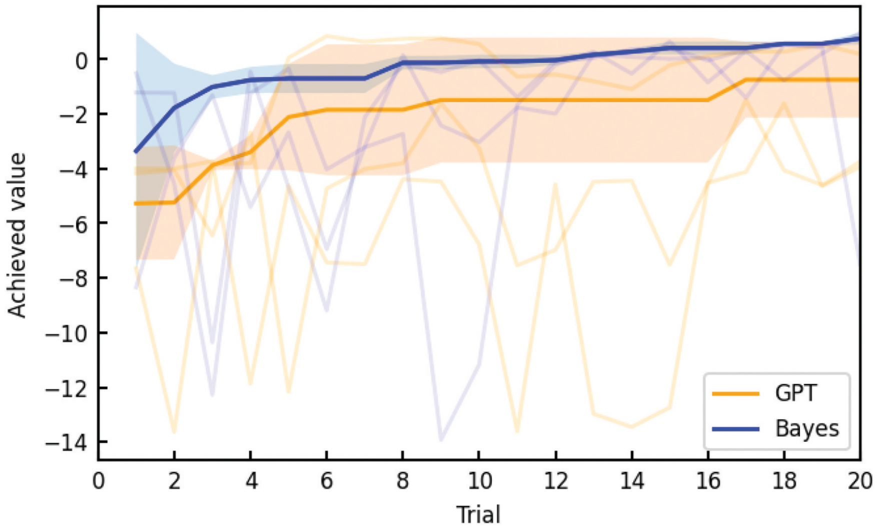


Figure 7. Exploring the best condition for a black box function by GPT or Bayesian optimization. The solid line represents the mean of the best value obtained in three independent trials; the semitransparent filled range represents the standard deviation; each raw trial is indicated by a semitransparent line.

knowledge. Despite the difficulties encountered, the capabilities of GPT-4 indicate a promising direction for the utilization of artificial intelligence in complex function approximation and optimization tasks.

4.7. Planning (molecule exploration)

In the next part of our exploration, we focus on GPT-4's capabilities in complex chemical compound optimization, a long-term challenge in cheminformatics [35]. Various techniques have been reported, with recent methods focusing on generating molecules that satisfy desired properties using deep learning algorithms. However, the limitations of using application-specific models are becoming more apparent.

Traditionally, using existing methods, it became easy to generate structures that are either easy to database or computationally favorable with specific features [52,53]. But, when translating these structures into actual experimental research, they must meet various constraints such as synthetic difficulty, solubility, and stability under specific conditions [46,54,55]. These parameters are often challenging to capture as structured data and thus frequently slip through the cracks of data science.

Language computation, as demonstrated by GPT-4, can bridge this gap between in-silico modeling and real-world constraints. GPT-4 can consider linguistic rules when designing or selecting molecules. For example, we explored the design of block polymers, which are interesting in self-organizing lithography [56–58]. In this polymer system, it is necessary to form a lamellar microphase separation structure with a narrower pitch, and this lamellar structure must be perpendicularly oriented to the substrate on which the film is formed.

An essential factor in meeting the first condition is the χ parameter of the two different unit structures constituting the block polymer [56]. This parameter is difficult to calculate theoretically, so in this study, we chose designs that have a larger expected distance (R_a) of Hansen solubility parameters, which are empirically correlated with it [59]. R_a was estimated using the HSPiP (v. 5.4.06) package. As an additional constraint, to make the lamellar structure orient vertically, we set the design to have a smaller R_a against nitrogen gas, the main component of air (Prompt S 22).

The first structure encountered during the search was a copolymer of styrene and methyl methacrylate. This is the most fundamental molecular structure for expressing a vertically oriented lamellar structure in self-organizing lithography [56]. Other suggested structures included well-known copolymers such as acrylonitrile, butadiene, and general monomer structures (Table 2) [56]. This is in stark contrast to traditional cheminformatics methods, where imposing constraints only on R_a results in hard-to-synthesize, unstable structures without polymerization bases making up most of the candidates.

A general positive correlation was found between the distance $R_{\mathrm{a,unit}}$ between unit structures and the distance $R_{\mathrm{a,nitrogen}}$ from the unit structure to a nitrogen molecule. If $R_{\mathrm{a,unit}}$ is increased to promote phase-separation structure, the distance to nitrogen also increases, which presents an obstacle for inducing a vertical lamellar structure. No noteworthy candidates exceeded the proposed combination of styrene and methyl methacrylate. This is primarily due to GPT-4's limited ability in generating molecular structures, as previously noted. Further, this observation underscores the potential need for designing prompts that more explicitly engage GPT-4's ability to consider and recognize molecular structures, which could further enhance its predictive capabilities in the realm of chemistry. Currently, the most practical approach now appears to be using a deep learning algorithm specialized in the molecular generation [52], with the appropriateness of its use automatically determined by GPT-4.

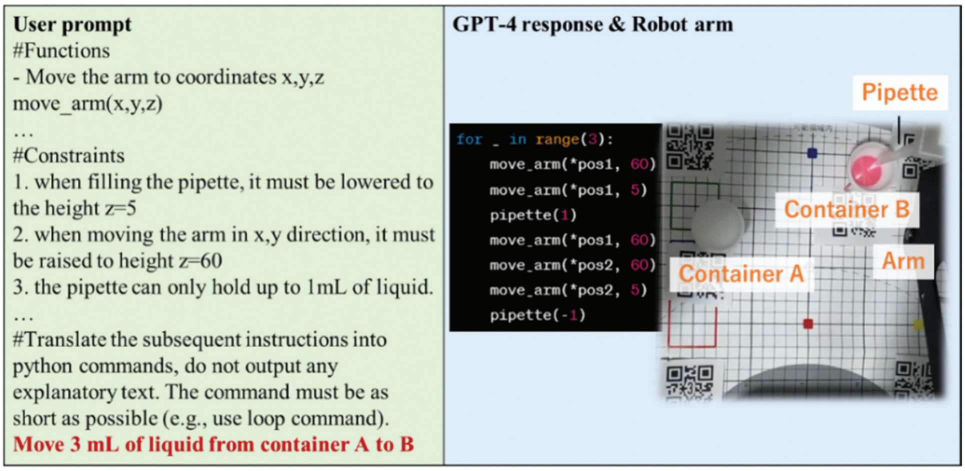
4.8. Synchronization with physical space

Interaction with actuators such as robotic arms is essential in research that includes work in real space [60,61]. GPT-4 can perform simple operations with a robotic arm while interpreting constraints and language commands to move 3 mL of liquid from container 1 to container 2 using a pipette with a 1 mL capacity (Figure 8, Prompt S 23).

Table 2. Exploration of block polymer units for micro phase separation.

Round	Unit1	Unit2	$R_{\mathrm{a,unit}}$[a]	$R_{\mathrm{a,nitrogen}}$[b]
1	styrene	methyl methacrylate	2.8	3.4
1	vinyl acetate	ethylene	8.5	6.0
1	acrylonitrile	butadiene	20	21
2	acrylic acid	styrene	17	18
2	vinyl acetate	vinyl chloride	5.8	5.6
2	butadiene	styrene	2.1	2.7
3	methyl acrylate	vinyl acetate	0.0	5.6
3	acrylonitrile	methyl methacrylate	20	21
3	isoprene	styrene	2.4	2.9
4	vinyl acetate	acrylamide	21	26
4	methyl acrylate	ethylene	8.5	6.0
4	styrene	vinylidene chloride	4.7	5.1

[a]Tetramers were calculated. [b]Maximum of R_a for 1) unit1 and N_2, and 2) unit2 and N_2.



User prompt

#Functions
- Move the arm to coordinates x,y,z
move_arm(x,y,z)
…
#Constraints
1. when filling the pipette, it must be lowered to the height z=5
2. when moving the arm in x,y direction, it must be raised to height z=60
3. the pipette can only hold up to 1mL of liquid.
…
#Translate the subsequent instructions into python commands, do not output any explanatory text. The command must be as short as possible (e.g., use loop command).
Move 3 mL of liquid from container A to B

GPT-4 response & Robot arm

```
for _ in range(3):
    move_arm(*pos1, 60)
    move_arm(*pos1, 5)
    pipette(1)
    move_arm(*pos1, 60)
    move_arm(*pos2, 60)
    move_arm(*pos2, 5)
    pipette(-1)
```

Figure 8. Commanding a robot arm by natural language using GPT-4 as a translator.

Figure 8 Illustrates commanding a robot arm via natural language using GPT-4 as a translator. When transferring liquid, the robotic arm needs to perform movements such as lifting and lowering, and the pipette requires suction and discharge. Furthermore, since the pipette's capacity is only 1 mL, the pipette operation must be repeated three times. Despite explicitly providing these constraints, GPT-4 autonomously generates commands to accomplish the desired task. The figure shows the process of the robotic arm according to the natural language instructions.

The practical benefits of controlling a robotic arm via a natural language interface are significant, as it lowers the entry barrier for chemists who may not be computer or robotic science experts. With object recognition through image-based deep learning models, and the use of multimodal AI models [62], which are avidly studied in the world of LLMs, including GPT-4, a more flexible system operation is anticipated. Furthermore, if an LLM gains sufficient planning capabilities, it could become possible to create a system that performs experiments automatically, simply by requesting 'synthesize compound X'. However, before this enhanced planning capability can be realized, there are several significant challenges to be addressed. This includes the ability to correctly propose synthesis pathways, accurately recognize and consider molecular structures, and overcome other difficulties as discussed in this paper. These advancements will be critical in enabling the system to accurately respond to such complex requests.

However, hardware design must delegate complex synthesis, purification, and measurement operations in chemical experiments to robotic arms or similar devices to actualize such an automatic system. Open-source system development utilizing inexpensive arm systems, IoT devices like Arduino and Raspberry Pi, and part creation via 3D printers could become a trend in the next decade. Generative models can also be used for purposes such as creating 3D drawings or designing electronic circuits [63]. It is also necessary to establish methods to analyze the large amount of data generated by automated systems using language models [64].

4.9. Autonomous research by LLM

With a certain level of inferencing ability, GPT-4 can be thought of as an AI capable of autonomous research by judiciously combining and improving the methodologies discussed thus far [3,4,6]. For example, GPT-4 can autonomously make decisions and take actions within the virtual world of a game called Minecraft [17]. Similarly, in the future, there is the potential for autonomous advancement in a variety of tasks, including research, within the physical space. Classically, closed loops using Bayesian optimization have been reported [65–68], yet requiring human intervention to narrow the search space to low-dimensional vectors carefully. In contrast, LLMs like GPT-4 can freely operate within language space, suggesting that it can automate research in a broader sense, including literature search, experimental condition setting, and result reporting.

Several autonomous agents utilizing GPT-4 have been reported. In these models, the LLM itself determines the following action. Open-source projects like AutoGPT [69] are being studied for their potential to automate tasks, including executing program codes. It's worth noting that these capabilities are not only confined to specific projects but can also be utilized via the ChatGPT interface provided by OpenAI. Attempts have also been made to personify agents and facilitate dialogue or to output their states as abstract language objects [70].

For instance, when using the prompts proposed by Ochiai et al. [71], abstract language objects such as 'chemist', 'chemical structures', 'density', and

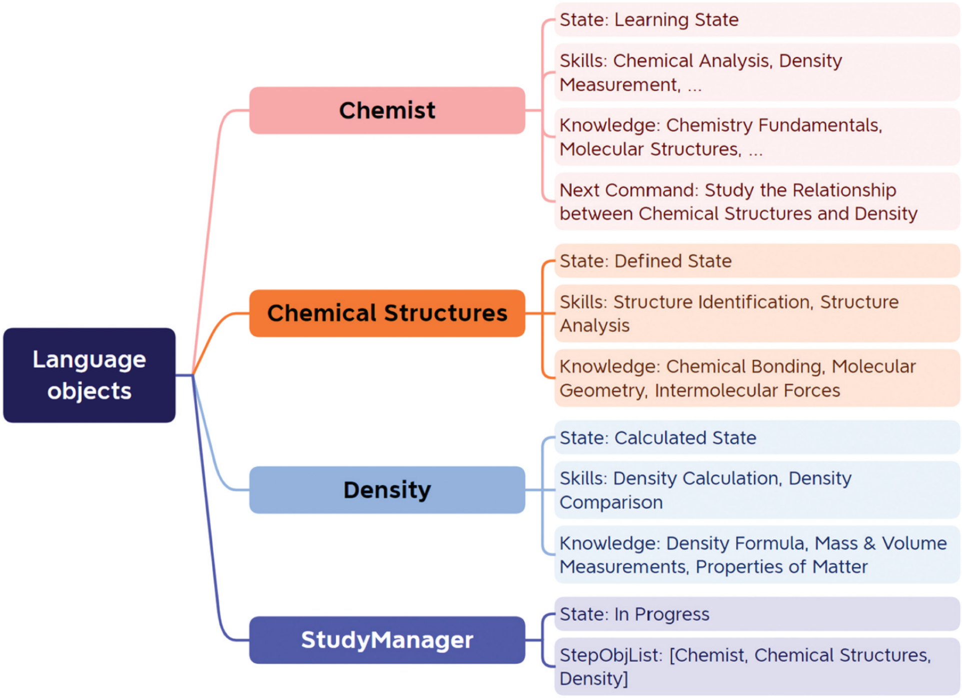

Language objects
Chemist
State: Learning State
Skills: Chemical Analysis, Density Measurement, ...
Knowledge: Chemistry Fundamentals, Molecular Structures, ...
Next Command: Study the Relationship between Chemical Structures and Density
Chemical Structures
State: Defined State
Skills: Structure Identification, Structure Analysis
Knowledge: Chemical Bonding, Molecular Geometry, Intermolecular Forces
Density
State: Calculated State
Skills: Density Calculation, Density Comparison
Knowledge: Density Formula, Mass & Volume Measurements, Properties of Matter
StudyManager
State: In Progress
StepObjList: [Chemist, Chemical Structures, Density]

Figure 9. Abstract language objects generated from a prompt 'chemist who want to understand the relation between chemical structures and density'.

'studymanager' can be generated from a directive such as 'a chemist who wants to understand the relationship between chemical structures and density' (Figure 9, Prompt S 24). Each of these objects possesses sub-concepts such as 'state', 'skill', and 'knowledge'.Take the main object 'chemist' as an example. The object 'chemical structures' contains information about molecular structures. It holds skills such as chemical analysis, density measurement, and general chemistry knowledge. With the capacity for web search enabled on the ChatGPT interface, calling up this prompt recursively allows the system to collect relevant data from the internet, thereby updating the contents of the objects.

Subsequently, the chemist generates a sub-object referred to as the 'next command' and investigates the correlation between molecular structures and density. Most scientists typically advance their research by combining existing methodologies. Assuming that text data can adequately describe these methodologies, it implies that they could, in principle, be learned and executed by LLMs.

However, GPT-4 has not succeeded in creating an autonomous agent on par with human researchers. Despite GPT-4's ability to solve fundamental college-level math problems, it is incapable of tackling advanced proofs or unresolved mathematical issues facing humanity [19,21]. This constraint is attributable to the GPT-4's inference and long-term memory capacities [72]. The model is yet unable to fully emulate the human brain's divide-and-conquer strategy, where a complex mathematical derivation or plan is broken down into smaller, manageable steps and then tackled sequentially [3,4]. This deficiency stems from the model's still limited ability to replicate the evolved skills of a seasoned human researcher, who benefits from age-acquired experience and multimodal learning. Furthermore, GPT-4 falls short in integrating inputs from various sensory organs, including vision (geometric aspects) and verbal language, which are crucial for comprehending and solving mathematical derivations or planning tasks [3,4]. In light of this, it is presumed that a gap still exists in general terms before an LLM can autonomously narrow down research topics, plan experiments, or write papers [2].

5. Issues to be addressed

In this section, we explore the challenges GPT-4 faces in its application to chemical research and potential solutions. Three significant issues can be identified with LLMs including GPT-4: a) handling non-verbal data, b) inputting technical and up-to-date information, and c) the inference capabilities of the LLM itself.

Firstly, a considerable challenge for GPT-4 is a) recognizing molecular structures and experimental data. GPT-4, a text-based AI is not specialized in treating large databases or spectra appropriately [73]. As discussed in this paper, this limitation results in GPT-4's ability to process compounds and data significantly inferior to that of a human expert. For example, proposing a new molecular structure can pose a significant challenge. There are two leading potential solutions. In the short term, specialized deep-learning models or algorithms for handling molecular structures could be used as plug-ins for the LLM. This concept is similar to GPT-4 utilizing a mathematical processing system like Wolfram to compensate for its limited mathematical ability. A more long-term solution would be the creation of multimodal LLMs. Integration with models dedicated to voice or image recognition is currently underway. Similarly, integration with models capable of inputting molecular structures might be possible. Alternatively, expanding the size of a versatile model like the transformer could resolve everything in the future [2].

The second issue is b) learning technical information. As of the time of writing, GPT-4 has only known limited information until September 2022. However, LLMs should be able to handle cutting-edge chemical literature. Two leading solutions exist for this problem. In the short term, the retrieval approach, which is already being implemented, can be used [5,74]. This approach seeks out literature similar to the user's query using a dedicated algorithm and includes that information in the LLM's prompt (prompt tuning) [74]. This method is expected to be an effective solution in many cases. However, there are limits to the amount of information that can be included in a prompt (8k or 32k tokens in GPT-4), making it difficult to infer from a wide range of cutting-edge information. Therefore, there is a need for constructing local LLMs that learn specialized data from scratch or through low-cost methods like fine-tuning, which is being considered worldwide [75]. From a practical perspective, one of the strengths of an LLM operating on a local computer is security. To use GPT-4, data must be sent to a cloud server, but with a local LLM, computations are completed within the laboratory, reducing the barrier when handling confidential information.

The third issue is c) the inference capabilities of the LLM itself. LLMs have been known to make mistakes in rudimentary mathematical processing and provide answers based on incorrect knowledge. There is still room for improvement in long-term planning capabilities, which seem to be lacking for the realization of fully automated chemical research [2–6]. There may not be much that chemists can contribute to solving this problem. However, deep learning is evolving at a revolutionary pace. Chemists may need to be prepared for the emergence of artificial general intelligence or superintelligence [8].

In addition to technical analysis, a profound exploration of the ethical implications of using LLMs in this scientific domain is conducted. LLMs like GPT-4 can inadvertently produce inaccurate or misleading information. This risk is particularly salient in chemistry, where the propagation of false information could instigate hazardous experiments. Accountability issues arise if LLMs dispense harmful or incorrect information, with the intricate training process and involvement of multiple stakeholders adding complexity.

These ethical considerations, although intricate and far-reaching, are paramount for the responsible and beneficial integration of LLMs like GPT-4 into chemical research. Future endeavors should aim to construct guidelines and best practices to ethically harness the power of LLMs in this discipline.

6. Conclusion

GPT-4 has demonstrated varying proficiency across diverse tasks such as organic chemistry, cheminformatics, few-shot learning, inference problems, selection of explanatory variables, exploration of boiling points, multi-variable exploration, compound exploration, and automated arm control for experiments. When examining each task specifically, GPT-4 exhibited a high understanding of general textbook-level knowledge in the field of organic chemistry. However, it fell short when dealing with specialized content or unique methods of synthesizing specific compounds.

In cheminformatics, GPT-4 partially succeeded in translating compound names into SMILES notation but could not generate SMILES notation in many cases. This is likely due to a lack of training data. On the other hand, leveraging its few-shot learning capabilities, GPT-4 could make accurate predictions even for compounds it hadn't been trained on. This result demonstrates GPT-4's ability to learn and apply new knowledge even from limited data. It was also found that the domain knowledge of chemistry that GPT-4 possesses helps set initial conditions during data exploration, for example.

These results indicate that GPT-4 can tackle a wide range of tasks in chemical research, spanning from textbook-level knowledge to addressing untrained problems and optimizing multiple variables. However, its performance heavily relies on the quality and quantity of its training data, and there is much room for improvement in its inference capabilities. Moving forward, while we wait for more advanced models than GPT-4, we should consider efficiently applying it to chemical research, possibly by creating hybrid models with existing specialized techniques.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by Grants-in-Aid for Scientific Research [Nos. 21H02017 and 22H04623] from the Ministry of Education, Culture, Sports, Science and Technology, Japan, by the JST FOREST Program [Grant Number JPMJFR213V]. The manuscript was drafted using GPT-4.

ORCID

Kan Hatakeyama-Sato http://orcid.org/0000-0003-1959-5430

References

[1] OpenAI, GPT-4 technical report. 2023. Available from: https://cdn.openai.com/papers/gpt-4.pdf

[2] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv:2303.12712. 2023.

[3] Guan L, Valmeekam K, Sreedharan S, et al. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. arXiv:2305.14909. 2023.

[4] Hao S, Gu Y, Ma H, et al. Reasoning with language model is planning with world model. arXiv:2305.14992. 2023.

[5] Sun S, Liu Y, Wang S, et al. PEARL: prompting large language models to plan and execute actions over long documents. arXiv:2305.14564. 2023.

[6] Valmeekam K, Marquez M, Sreedharan S, et al. On the planning abilities of large language models – a critical investigation. arXiv:2305.15771. 2023.

[7] Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. arXiv:2305.10601. 2023.

[8] Zhang C, Zhang C, Li C, et al. One small step for Generative AI, one giant leap for AGI: a complete survey on ChatGPT in AIGC Era. arXiv:2304.06488. 2023.

[9] White AD, Hocky GM, Gandhi HA, et al. Assessment of chemistry knowledge in large language models that generate code. 2022. doi: 10.26434/chemrxiv-2022-3md3n-v2

[10] Gandhi HA, White AD. Explaining molecular properties with natural language. 2022. doi: 10.26434/chemrxiv-2022-v5p6m-v3

[11] Xie T, Wan Y, Huang W, et al. Large language models as master key: unlocking the secrets of materials science with GPT. arXiv:2304.02213. 2023.

[12] Jacobsson TJ, Hultqvist A, García-Fernández A, et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy. 2021;7(1):107. doi: 10.1038/s41560-021-00941-3

[13] Ishizuki N, Shimizu R, Hitosugi T. Autonomous experimental systems in materials science. Sci Technol Adv Mater. 2023;3. doi: 10.1080/27660400.2023.2197519

[14] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv:2001.08361. 2020.

[15] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv:2206.07682. 2022.

[16] Schaeffer R, Miranda B, Koyejo S. Are Emergent abilities of large language models a mirage? arXiv:2304.15004. 2023.

[17] Wang G, Xie Y, Jiang Y, et al. Voyager: an open-ended embodied agent with large language models. arXiv:2305.16291. 2023.

[18] Cai T, Wang X, Ma T, et al. Large language models as tool makers. arXiv:2305.17126. 2023.

[19] Jingfeng Y, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. arXiv:2304.13712. 2023.

[20] Chen W, Yin M, Ku M, et al. TheoremQA: a theorem-driven question answering dataset. arXiv:2305.12524. 2023.

[21] Arora D, Singh HG. Have LLMs advanced enough? A challenging problem solving benchmark for large language models. arXiv:2305.15074. 2023.

[22] Boiko DA, MacKnight R, Gomes G. Emergent autonomous scientific research capabilities of large language models. arXiv:2304.05332. 2023.

[23] Nutting JE, Rafiee M, Stahl SS. Tetramethylpiperidine N-Oxyl (TEMPO), phthalimide N-Oxyl (PINO), and related N-Oxyl species: electrochemical properties and their use in electrocatalytic reactions. Chem Rev. 2018;118(9):4834. doi: 10.1021/acs.chemrev.7b00763

[24] Janoschka T, Hager MD, Schubert US. Powering up the future: radical polymers for battery applications. Adv Mater. 2012;24(48):6397. doi: 10.1002/adma.201203119

[25] Janoschka T, Martin N, Martin U, et al. An aqueous, polymer-based redox-flow battery using non-corrosive, safe, and low-cost materials. Nature. 2015;527(7576):78. doi: 10.1038/nature15746

[26] Janoschka T, Martin N, Hager MD, et al. An aqueous redox-flow battery with high capacity and power: the TEMPTMA/MV System. Angew Chem Int Ed. 2016;55(46):14427. doi: 10.1002/anie.201606472

[27] Saal A, Hagemann T, Schubert US. Polymers for battery applications—active materials, membranes, and binders. Adv Energy Mater. 2020;11(43):2001984. doi: 10.1002/aenm.202001984

[28] Watanabe M, Rikukawa M, Sanui K, et al. Ionic conductivity of polymer complexes formed by poly(ethylene succinate) and lithium perchlorate. Macromolecules. 1984;17(12):2902. doi: 10.1021/ma00142a078

[29] Winsberg J, Hagemann T, Janoschka T, et al. Redox-flow batteries: from metals to organic redox-active materials. Angew Chem Int Ed. 2017;56(3):686. doi: 10.1002/anie.201604925

[30] Joo Y, Agarkar V, Sung SH, et al. A nonconjugated radical polymer glass with high electrical conductivity. Science. 2018;359(6382):1391. doi: 10.1126/science.aao7287

[31] Oyaizu K, Nishide H. Radical polymers for organic electronic devices: a radical departure from conjugated polymers? Adv Mater. 2009;21(22):2339. doi: 10.1002/adma.200803554

[32] Endo T, Takuma K, Takata T, et al. Synthesis and polymerization of 4-(glycidyloxy)-2,2,6,6-tetramethylpiperidine-1-oxyl. Macromolecules. 1993;26(12):3227. doi: 10.1021/ma00064a039

[33] Ma Z, Huang Q, Bobbitt JM. Oxoammonium salts. 5. A new synthesis of hindered piperidines leading to unsymmetrical TEMPO-type nitroxides. Synthesis and enantioselective oxidations with chiral nitroxides and chiral oxoammonium salts. J Org Chem. 2002;58(18):4837. doi: 10.1021/jo00070a018

[34] Bran AM, Cox S, White AD, et al. ChemCrow: augmenting large-language models with chemistry tools. arXiv:2304.05376. 2023.

[35] Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. Mol Inform. 2020;39 (12):e2000171. doi: 10.1002/minf.202000171

[36] Rickman JM, Chan HM, Harmer MP, et al. Materials informatics for the screening of multi-principal elements and high-entropy alloys. Nat Commun. 2019;10(1):2618. doi: 10.1038/s41467-019-10533-1

[37] Oaki Y, Igarashi Y. Materials informatics for 2D materials combined with sparse modeling and chemical perspective: toward small-data-driven chemistry and materials science. Bull Chem Soc Jpn. 2021;94 (10):2410. doi: 10.1246/bcsj.20210253

[38] Chaikittisilp W, Yamauchi Y, Ariga K. Material evolution with nanotechnology, nanoarchitectonics, and materials informatics: what will be the next paradigm shift in nanoporous materials? Adv Mater. 2022;34 (7):e2107212. doi: 10.1002/adma.202107212

[39] Hu J, Stefanov S, Song Y, et al. MaterialsAtlas.Org: a materials informatics web app platform for materials discovery and survey of state-of-the-art. Npj Comput Mater. 2022;8(1):65. doi: 10.1038/s41524-022-00750-6

[40] Ramos MC, Michtavy SS, Porosoff MD, et al. Bayesian optimization of catalysts with in-context learning. arXiv:2304.05341. 2023.

[41] Mao J, Wang J, Cho K-H, et al. iupacGPT: IUPAC-based large-scale molecular pre-trained model for property prediction and molecule generation. 2023. doi: 10.26434/chemrxiv-2023-5kjvh.

[42] Kıcıman E, Ness R, Sharma A, et al. Causal Reasoning and large language models: opening a new frontier for causality. arXiv:2305.00050. 2023.

[43] Wu Y, Min SY, Prabhumoye S, et al. SPRING: GPT-4 out-performs RL algorithms by studying papers and Reasoning. arXiv:2305.15486. 2023.

[44] Zhu Y, Wang X, Chen J, et al. LLMs for knowledge graph construction and Reasoning: recent capabilities and future opportunities. arXiv:2305.13168. 2023.

[45] Hatakeyama-Sato K, Matsumoto S, Takami H, et al. A PROXYL-Type norbornene polymer for high-voltage cathodes in lithium batteries. Macromol Rapid Commun. 2021;42(19):e2100374. doi: 10.1002/marc.202100374

[46] Hatakeyama-Sato K. Recent advances and challenges in experiment-oriented polymer informatics. Polym J. 2022;55(2):117. doi: 10.1038/s41428-022-00734-9

[47] Jia C, Pan F, Zhu YG, et al. High-energy density nonaqueous all redox flow lithium battery enabled with a polymeric membrane. Sci Adv. 2015;1(10): e1500886. doi: 10.1126/sciadv.1500886

[48] Noviandri I, Brown KN, Fleming DS, et al. The decamethylferrocenium/decamethylferrocene redox couple: a superior redox standard to the ferrocenium/ ferrocene redox couple for studying solvent effects on the Thermodynamics of electron transfer. J Phys Chem B. 1999;103(32):6713. doi: 10.1021/jp991381+

[49] Hatakeyama-Sato K, Watanabe S, Yamane N, et al. Using GPT-4 in parameter selection of materials informatics: improving predictive accuracy amidst data scarcity and ‘ugly duckling’ dilemma. 2023. doi: 10.26434/chemrxiv-2023-f9lxl

[50] Udrescu SM, Tegmark M. AI Feynman: a physics-inspired method for symbolic regression. Sci Adv. 2020;6(16):eaay2631. doi: 10.1126/sciadv.aay2631

[51] Iwasaki Y, Ishida M. Data-driven formulation of natural laws by recursive-LASSO-based symbolic regression. arXiv.2102.09210. 2021.

[52] Lim J, Ryu S, Kim JW, et al. Molecular generative model based on conditional variational autoencoder for de novo molecular design. J Cheminform. 2018;10 (1):31. doi: 10.1186/s13321-018-0286-7

[53] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. arXiv:1802.04364. 2019.

[54] Hatakeyama-Sato K, Oyaizu K. Integrating multiple materials science projects in a single neural network. Commun Mater. 2020;1:49. doi: 10.1038/s43246-020-00052-8

[55] Hatakeyama-Sato K, Adachi H, Umeki M, et al. Automated design of li(+) -conducting polymer by Quantum-Inspired Annealing. Macromol Rapid Commun. 2022;43(20):e2200385. doi: 10.1002/marc.202200385

[56] Bates CM, Maher MJ, Janes DW, et al. Block copolymer lithography. Macromolecules. 2013;47(1):2. doi: 10.1021/ma401762n

[57] Kurioka T, Komamura T, Shida N, et al. Ordered-structure-Induced electrochemical Post-Functionalization of Poly(3-(2-ethylhexyl)thiophene). Macromole Chem Phys. 2022;223(4). doi: 10.1002/macp.202100435

[58] Yoshimura T, Morishita T, Agata Y, et al. Long-range ordered lamellar formation with lower molecular weight PS-PMMA block copolymers: significant effects of discrete oligopeptides at the junction. Macromolecules. 2022;55(6):2148. doi: 10.1021/acs.macromol.1c02569

[59] Lubtow MM, Haider MS, Kirsch M, et al. Like dissolves like? A comprehensive evaluation of partial solubility parameters to predict polymer-drug compatibility in ultrahigh drug-loaded polymer micelles. Biomacromolecules. 2019;20(8):3041. doi: 10.1021/acs.biomac.9b00618

[60] Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature. 2020;583(7815):237. doi: 10.1038/s41586-020-2442-2

[61] Skreta M, Yoshikawa N, Arellano-Rubach S, et al. Errors are useful prompts: instruction guided task programming with verifier-assisted iterative prompting. arXiv:2303.14100. 2023.

[62] You H, Sun R, Wang Z, et al. IdealGPT: iteratively decomposing vision and language Reasoning via large language models. arXiv:2305.14985. 2023.

[63] Jansen P. From words to wires: generating functioning electronic devices from natural language descriptions. arXiv:2305.14874. 2023.

[64] González-Santamarta MA, Fernández-Becerra L, Sobrín-Hidalgo D, et al. Using large language models for interpreting autonomous robots behaviors. arXiv:2304.14844. 2023.

[65] Yamakage S, Kaneko H. Design of adaptive soft sensor based on Bayesian optimization. Case Stud Chem Environ Eng. 2022;6:100237. doi: 10.1016/j.cscee.2022.100237

[66] Ikebata H, Hongo K, Isomura T, et al. Bayesian molecular design with a chemical language model. J Comput Aided Mol Des. 2017;31(4):379. doi: 10.1007/s10822-016-0008-z

[67] Wu S, Lambard G, Liu C, et al. iQSPR in XenonPy: a Bayesian molecular design algorithm. Mol Inform. 2020;39(1–2):e1900107. doi: 10.1002/minf.201900107
[68] Shields BJ, Stevens J, Li J, et al. Bayesian reaction optimization as a tool for chemical synthesis. Nature. 2021;590(7844):89. doi: 10.1038/s41586-021-03213-y
[69] GitHub repository. AutoGPT; 2023. Available from: https://github.com/Significant-Gravitas/Auto-GPT.
[70] Park JS, O'Brien JC, Cai CJ, et al. Generative agents: interactive simulacra of human behavior. arXiv:2304.03442. 2023.
[71] Ochiai Y, Kondo N, Fushimi T. Towards digital Nature: bridging the gap between turing machine objects and linguistic objects in LLMMs for universal interaction of object-oriented descriptions. arXiv:2304.04498. 2023.
[72] Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv:2303.18223v11. 2023.
[73] Cheng L, Li X, Bing L. Is GPT-4 a good data analyst? arXiv:2305.15038. 2023.
[74] Caciularu A, Peters ME, Goldberger J, et al. Peek across: improving multi-document modeling via cross-document question-answering. arXiv:2305.15387. 2023.
[75] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: efficient finetuning of quantized LLMs. arXiv:2305.14314. 2023.