What indeed can GPT models do in chemistry? A comprehensive benchmark on eight tasks

Taicheng Guo[*], **Kehan Guo**[*], **Bozhao Nan**, **Zhenwen Liang**, **Zhichun Guo**, **Nitesh V. Chawla**, **Olaf Wiest**, **Xiangliang Zhang**[†]
University of Notre Dame
{tguo2, kguo2, bnan, zliang6, zguo5, nchawla, owiest, xzhang33}@nd.edu

Abstract

Large Language Models (LLMs) with strong abilities in natural language processing tasks have emerged and have been rapidly applied in various kinds of areas such as science, finance and software engineering. However, the capability of LLMs to advance the field of chemistry remains unclear. In this paper, we establish a comprehensive benchmark containing 8 practical chemistry tasks, including 1) name prediction, 2) property prediction, 3) yield prediction, 4) reaction prediction, 5) retrosynthesis (prediction of reactants from products), 6) text-based molecule design, 7) molecule captioning, and 8) reagent selection. Our analysis draws on widely recognized datasets including BBBP, Tox21, PubChem, USPTO, and ChEBI, facilitating a broad exploration of the capacities of LLMs within the context of practical chemistry. Three GPT models (GPT-4, GPT-3.5, and Davinci-003) are evaluated for each chemistry task in zero-shot and few-shot in-context learning settings with carefully selected demonstration examples and specially crafted prompts. The key results of our investigation are 1) GPT-4 outperforms the other two models among the three evaluated; 2) GPT models exhibit less competitive performance in tasks demanding precise understanding of molecular SMILES representation, such as reaction prediction and retrosynthesis; 3) GPT models demonstrate strong capabilities in text-related explanation tasks such as molecule captioning; and 4) GPT models exhibit comparable or better performance to classical machine learning models when applied to chemical problems that can be transformed into classification or ranking tasks, such as property prediction, and yield prediction. In addition to the key findings from the comprehensive benchmark analysis, our work provides insights into the limitation of current LLMs and how different settings related to LLMs affect performance across various chemistry tasks. The code and datasets used in this study are available at https://github.com/ChemFoundationModels/ChemLLMBench. Our preliminary investigation provide the basis for further explore the potential of LLMs in chemistry and advancing applications of ML in chemistry.

1 Introduction

Large language models (LLMs) have recently demonstrated impressive reasoning abilities across a wide array of tasks. These tasks are not limited to natural language processing, but also extend to various language-related applications within scientific domains [45, 24, 19, 20]. Much of the research on the capacity of LLMs in science has been focused on tasks like answering of medical

[*]Both authors contribute equally to the work, under the support of NSF Center for Computer Assisted Synthesis (C-CAS). https://ccas.nd.edu.

[†]Corresponding author.

Preprint. Under review.

[24] and scientific questions [19, 20]. However, the exploration of their application to practical tasks in the field of chemistry remains underinvestigated. Although some studies [6, 22, 52, 37] have been conducted, they tend to focus on a specific case studies rather than a comprehensive or systematic evaluation. The exploration of LLMs' capabilities within the field of chemistry has the potential to revolutionize this domain and expedite research and development activities. Thus, the question, ***"What can LLMs like GPT models do in chemistry?"*** is a compelling topic of inquiry for both AI researchers and chemists. In order to address this knowledge gap, we have developed a comprehensive benchmark to provide a preliminary investigation into the abilities of GPT models across a diverse range of practical chemistry tasks. Our aim is to gain insights that will be beneficial for both AI researchers and chemists in order to advance the application of GPT models in chemistry. For AI researchers, we provide insights into the strengths, weaknesses, and limitations of LLMs in chemistry-related tasks, which can inform further development and refinement of different AI techniques for more effective applications within the field. For chemists, our study provides a better understanding of the extent to which they can rely on current GPT models. Utilizing our more extensive experimental setup, a broader range of chemistry tasks can be explored to further evaluate the capabilities of LLMs.

Our investigation focus on 8 practical chemistry tasks, which cover a wide spectrum of the chemistry domain. These include: 1) name prediction, 2) property prediction, 3) yield prediction, 4) reaction prediction, 5) retrosynthesis (prediction of reactants from products), 6) text-based molecule design, 7) molecule captioning, and 8) reagents selection. Our analysis draws on widely available datasets including BBBP, Tox21 [53], PubChem [26], USPTO [23, 42], ChEBI [15, 14], and so on. Three GPT models (GPT-4, GPT-3.5, and Davinci-003) [33] are evaluated for each chemistry task in zero-shot and few-shot in-context learning settings with carefully selected demonstration examples and specific prompts.

The **main findings** from the investigations are:

- GPT-4 outperforms the other two models among the three evaluated. The ranking of the three models on 8 tasks can be found in Table 1;
- GPT models exhibit less competitive performance in tasks demanding precise understanding of molecular SMILES representation, such as name prediction, reaction prediction and retrosynthesis;
- GPT models demonstrate strong capabilities in text-related explanation tasks such as molecule captioning;
- For chemical problems that can be converted to classification tasks or ranking tasks, such as property prediction, and yield prediction, GPT models can achieve competitive performance compared to baselines that use classical machine learning (ML) models as classifiers, or even better, as summarized in Table 1.

These conclusions are derived from conducting five repeated evaluations on each task, using the best evaluation setting that was discovered through a grid search on the validation set of each task.

Based on our analysis, specifically from formulating prompts and preparing demonstration examples for in-context learning, we have derived the following **recommendations** for AI researchers and chemists who are considering using LLMs for future chemistry tasks:

- In-context learning (ICL) achieves a significantly better performance than zero-shot prompting. It is thus crucial to develop methods for identifying in-context learning examples that have the potential to greatly enhance performance;
- ICL examples identified by similarity-based search are more useful than randomly sampled ones;
- Having a greater number of ICL examples is preferable over a limited number.

These suggestions are derived from analyzing the results of 12 different variants of each option and evaluating their performance on the validation set.

We highlight the **contributions** of this paper as follows:

- We are the first to establish a comprehensive benchmark to evaluate the abilities of LLMs on a wide range of chemistry tasks. These eight selected tasks, in consultation with chemists, not only encompass a broad spectrum of the chemistry domain but also demand various abilities such as comprehension, reasoning, and explanation using domain-specific chemistry

knowledge. The selected tasks and datasets thus provide a comprehensive and representative overview for benchmarking purposes.

- We provide a comprehensive experimental framework for testing of LLMs in chemistry tasks. To factor in the impact of prompts and demonstration examples in zero-shot and in-context learning, we have assessed multiple input options for the GPT models, focusing on the description of chemistry tasks. Five representative configurations were chosen based on their performance on a validation set, then these selected options were applied on the testing set. The conclusion is made from five repeated evaluations on each task since GPTs often yield different outputs at different API calls even though the input is the same. We thus believe that our benchmarking process is both reliable and systematic.
- Our investigations yield broader insights into the performance of GPT models on chemistry tasks. While confirming some anticipated outcomes, they also offer unexpected findings. Our work also led to practical recommendations that can guide AI researchers and chemists in leveraging GPT models more effectively in the future. We believe that this preliminary investigation offers crucial perspectives for AI researchers and chemists alike, deepening our comprehension of LLMs' capacities and fostering additional research at the intersection of LLMs and chemistry.

Table 1: The rank of three GPT models on eight chemistry tasks and performance highlight (NC: not competitive, C: competitive, SC: selectively competitive, acc: accuracy).

Task	GPT-4	GPT-3.5	Davinci-003	Performance highlight (comparing to baselines if any)
Name Prediction	1	2	3	**NC**: max. acc. 8% (Table 3)
Property Prediction	1	2	3	**SC**: outperform MoIR [47] (Table 5)
Yield Prediction	1	3	2	**C**: but 16-20% lower acc. than UAGNN [27] (Table 7)
Reaction Prediction	1	3	2	**NC**: 70% lower acc. than Chemformer [21] (Table 8)
Reagents Selection	2	1	3	**C**: 40-50% acc. (Table 9)
Retrosynthesis	2	3	1	**NC**: 40% lower acc. than Chemformer [21] (Table 10)
Text-Based Molecule Design	1	3	2	**SC**: better than MolT5-Large [15] (Table 11)
Molecule Captioning	1	2	1	**SC**: better than MolT5-Large [15] (Table 12)
Average rank	1.25	2.375	2.125	overall: 3 **SC**, 2 **C**, 3 **NC**

The paper is organized as follows: In section 2, we elaborate on the evaluation process, including an overview of the chemistry tasks, the utilized GPT models and prompts, and the validation and testing settings. Then we provide details of the evaluation on each chemistry task, in the following order: name prediction in section 3, property prediction in section 4, yield prediction in section 5, reaction prediction in section 6, reagents selection in section 7, retrosynthesis (prediction of reactants from products) in section 8, text-based molecule design in section 9, and molecule captioning in section 10. Finally, to answer the question *What indeed can GPT models do in chemistry?* we share our findings and discuss the constraints inherent to LLMs and how different settings related to LLMs affect performance across various chemistry tasks in section 11. Related works are presented in section 12, and the conclusions are summarized in section 13.

2 The Evaluation Process and Setting

The workflow of the evaluation process is shown in Fig. 1. In collaboration with chemists, we identify eight chemistry tasks. Following this, we generate, assess, and choose suitable prompts to forward to the GPT models. The acquired answers are then evaluated by selected metrics, both qualitatively and quantitatively.

Chemistry tasks. In order to explore the abilities of GPT models in the field of chemistry, we concentrate on three fundamental capabilities: comprehension, reasoning, and elucidation. We examine these competencies through eight diverse and broadly acknowledged practical chemistry tasks. These tasks are summarized in Table 2, in terms of the *task type* from the perspective of machine learning, the *dataset* used for evaluation, as well as the *evaluation metrics*. The *#ICL candidates* refers to the number of candidate examples, from which we select k demonstration examples, either

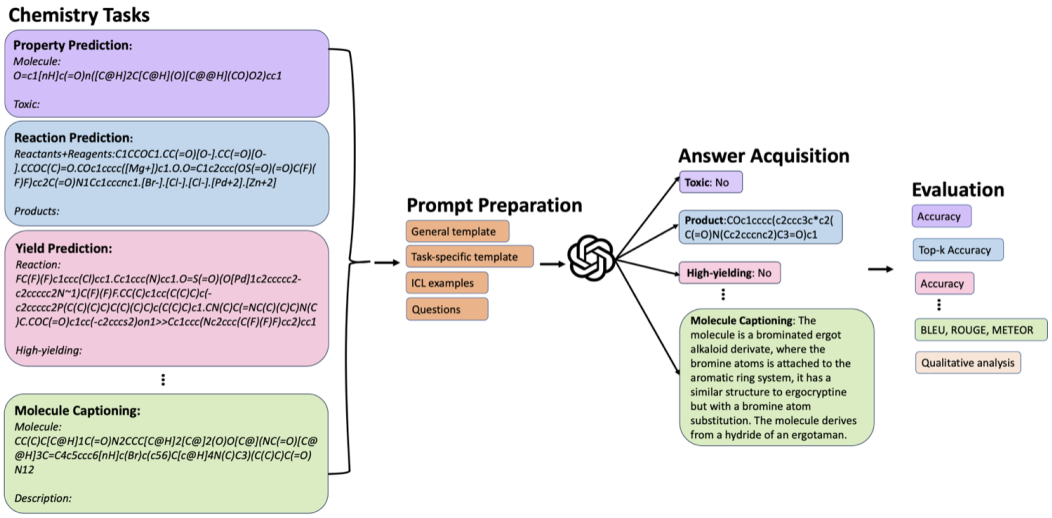


Figure 1: Overview of the evaluation process

randomly or based on similarity searches. These candidate sets are the training set used in classical machine learning models, e.g., in training classifiers, or generative models. We set the test set of 100 instances, randomly sampled from the original testing dataset (non-overlapping with the training set). To reduce the influence of the LLMs randomness on the results, each evaluation experiment is repeated five times and the mean and variance are reported.

Table 2: The statistics of all tasks, datasets, the number of ICL/test samples, and evaluation metrics

Ability	Task	Task Type	Dataset	#ICL candidates	#test	Evaluation Metrics
Understanding	Name Prediction	Generation	PubChem	500	100	Accuracy
	Property Prediction	Classification	BBBP, HIV, BACE, Tox21, ClinTox	2053, 41127, 1514, 8014, 1484	100	Accuracy, F1 score
Reasoning	Yield Prediction	Classification	Butchward-Hartwig, Suzuki-Miyaura	3957, 5650	100	Accuracy
	Reaction Prediction	Generation	USPTO-Mixed	409035	100	Accuracy, Validity
	Reagents Selection	Ranking	Suzuki-Miyaura	5760	100	Accuracy
	Retrosynthesis	Generation	USPTO-50k	40029	100	Accuracy, Validity
	Text-Based Molecule Design	Generation	ChEBI-20	26407	100	BLEU, Exact Match, etc
Explaining	Molecule Captioning	Generation	ChEBI-20	26407	100	BLEU, ROUGE, etc

GPT models. For all tasks, we evaluate the performance of three popular GPT models: GPT-4, GPT-3.5, and Davinci-003.

Zero-shot prompt. For each task, we apply a standardized zero-shot prompt template. As shown in Figure 2, we instruct the GPT models to act in the capacity of a chemist. The content inside the brackets is tailored to each task, adapting to its specific inputs and outputs. The responses from GPT models are confined to only returning the desired output without any explanations.

You are an expert chemist. Given the [reactants SMILES / molecular description / ...]: [Input], predict the [reaction product SMILES / molecule SMILES / ...] using your experienced chemical [reaction prediction / chemical molecule design / ...] knowledge. No explanations and other information. Only return the [product SMILES / designed molecular SMILES].

Figure 2: The standardized zero-shot prompt template for all tasks.

Task-specific ICL prompt. ICL is a new paradigm for GPT models where predictions are based solely on contexts enriched with a few demonstration examples [13]. This paper specifically defines ICL as a few-shot in-context learning approach, excluding the zero-shot paradigm. In order to thoroughly examine the capacities of GPT models within each chemistry-specific task, we design a

task-specific ICL prompt template. As shown in Figure 3. The format of the template is similar to that used in [37]. We also partition our template into four parts: {General Template}{Task-Specific Template}{ICL}{Question}. The {General Template} is almost the same as the zero-shot prompt, instructing the GPT models to play the role of a chemist and specify the chemistry task with its corresponding input and output. Considering that the responses for chemistry-related tasks must be accurate and chemically reasonable, it's crucial to prevent GPT models from generating hallucinated information. To this end, we introduce the {Task-Specific Template} which consists of three main components: [Input explanation], [Output Explanation], and [Output Restrictions], specifically designed to reduce hallucinations. These components are tailored to each task. The {ICL} part is a straightforward concatenation of the demonstration examples and it follows the structure "[Input]: [Input_content] [Output]: [Output_content]". The [Input] and [Output] denote the specific names of each task's input and output, respectively. For example, in the reaction prediction task, the [Input] would be "Reactants+Reagents" and the [Input_content] would be the actual SMILES of reactants and reagents. The [Output] would be "Products" and the [Output_content] would be the SMILES of products. Detailed ICL prompts for each task will be presented in their respective sections that follow. The last {Question} part presents the testing case for GPT models to respond to.

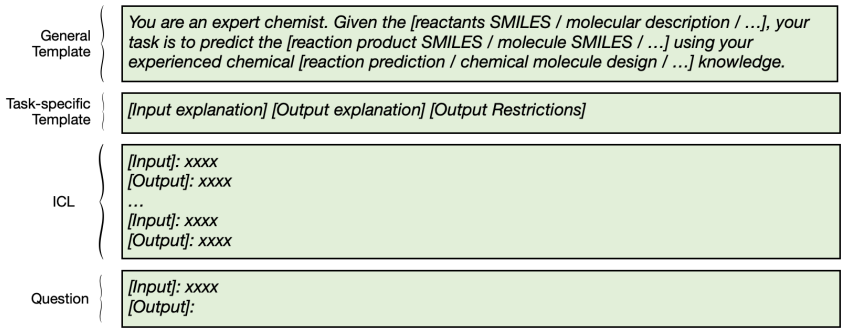


Figure 3: An ICL prompt template for all tasks.

ICL strategies. To investigate the impact of the quality and quantity of ICL examples on the performance of each task, we explore two ICL strategies. The quality is determined by the retrieval methods employed for finding similar examples to the sample in question. We conduct a grid search across two strategies: {Random, Scaffold}. In the Random strategy, we randomly select k examples from the ICL candidate pool. In the Scaffold strategy, if the [Input_content] is a molecule SMILES, we use Tanimoto Similarity [44] from Morgan Fingerprint [32] with radius=2 to calculate the molecular scaffold similarity to find the top-k similar molecule SMILES. If the [Input_content] is a description such as IUPAC name or others, we use Python's built-in difflib.SequenceMatcher tool [38] to find the top-k similar strings. To explore the influence of the quantity of ICL examples on the performance, we also perform a grid search for k, the number of ICL examples, in each task.

Experiment setup strategy. In property prediction and yield prediction tasks, we perform the grid search of k in {4, 8}. In name prediction, reaction prediction, and retrosynthesis tasks, we perform the grid search of k in {5, 20}. In text-based molecule design and molecule captioning tasks, we perform the grid search of k in {5, 10} because of the maximum token limitation of GPT models. To reduce the time consumption of API requests caused by testing on the large test set, we first construct a validation set of size 30 which is randomly sampled from the original training set. Then we search k and retrieval strategies ({Random, Scaffold}) on the validation set. Based on the validation set results, we choose 5 representative choices and test them on the test set of size 100, which is randomly sampled from the original test set. For each evaluation task, we run evaluation 5 times and report the mean and variance.

3 Name Prediction

For one molecule, there are different kinds of chemical names such as SMILES, IUPAC name, and molecular formula. To investigate whether the GPT models have the basic chemical name understanding ability, we construct 4 chemical name prediction tasks including SMILES to IUPAC

name translation (smiles2iupac), IUPAC name to SMILES translation (iupac2smiles), SMILES to molecule formula translation (smiles2formula), and molecule formula to SMILES translation (formula2smiles). We collect 630 molecules and their corresponding names including SMILES, IUPAC name, and molecule formula from PubChem [26]. We randomly sample 500 molecules as the ICL candidates, and other 30 molecules as the validation set, and other 100 molecules as the test set. For all name translation tasks, we use the exact match accuracy as the metric to evaluate the performance.

ICL Prompt. One example of the smiles2iupac prediction is shown in Figure 4. For other name translation tasks, we only change the underlined parts which represent different tasks and their corresponding input names and output names.

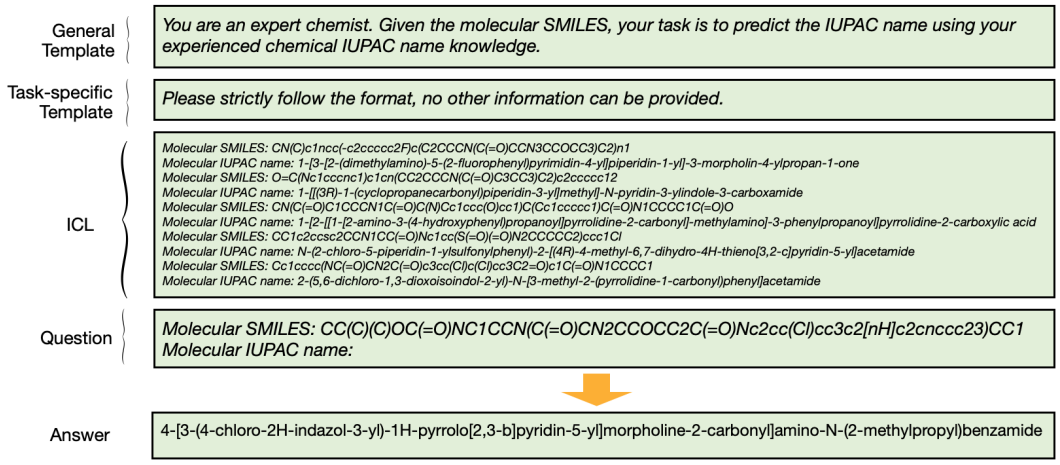


Figure 4: An ICL prompt example for smiles2iupac prediction

Results. The results are reported in Table 3 (we only report representative methods along with their optimal prompt settings via grid search on validation set). In all four name prediction tasks, the accuracy of the best method is extremely low (0.014 in the iupac2smiles task and 0.086 in the smiles2formula task) or even 0 (in the smiles2iupac and formula2smiles tasks). This indicates the GPT models lack basic chemical name understanding ability. The accuracy of Davinci-003 is considerably inferior to other models.

Table 3: The accuracy (↑) of GPT models in 4 different name prediction tasks. The best GPT model is in bold font. Here k is the number of examples used in few-shot ICL.

Method	smiles2iupac	iupac2smiles	smiles2formula	formula2smiles
GPT-4 (zero-shot)	0	0.008±0.008	0.048± 0.022	0
GPT-4 (Scaffold, k=5)	0	**0.014±0.009**	0.058±0.015	0
GPT-4 (Scaffold, k=20)	0	0.012±0.004	**0.086±0.036**	0
GPT-4 (Random, k=20)	0	0.010±0.007	0.070±0.032	0
GPT-3.5 (Scaffold, k=20)	0	0.010±0.000	0.052±0.004	0
Davinci-003 (Scaffold, k=20)	0	0	0.006±0.005	0

Case studies. Example results generated by GPT-4 (Scaffold, k=20) method for each task is shown in Table 4. In all tasks, GPT-4 model gives the wrong answers. In the smiles2formula task, we can observe that GPT models cannot even recognize the number of Carbon and infer the correct number of Hydrogen, demonstrating the bad chemical understanding ability of GPT models.

4 Molecule Property Prediction

Molecule property prediction [16, 47] is a fundamental task in computational chemistry that has been gaining significant attention is recent years due to its potential for drug discovery, material science

Table 4: Example results generated by GPT-4 (Scaffold, k=20) method for different tasks

Task	Input	Ground Truth	Output of GPT-4 (Scaffold, k=20)
smiles2iupac	CCOC(=O)C(C(C)=O)=C(C)N	ethyl 2-acetyl-3-aminobut-2-enoate	ethyl 2-methyl-5-oxo-2-azahept-4-en-3-oate
iupac2smiles	ethyl 2-acetyl-3-aminobut-2-enoate	CCOC(=O)C(C(C)=O)=C(C)N	CCOC(=O)C=C(C)C(=N)C
smiles2formula	Cc1noc(CCn2cc[nH]c2=O)n1	C8H10N4O2	C9H10N4O2
formula2smiles	C8H10N4O2	Cc1noc(CCn2cc[nH]c2=O)n1	CCn1cnc(=O)c(=O)n1Cc1cccnc1

and other areas in the chemistry. The task involves using machine learning techniques [17] to predict the chemical and physical properties of a given molecule, based on its molecular structure. We aim to further explore the potential of LLMs in molecular property prediction and assess their performance on a set of benchmark datasets, such as BBBP, HIV, BACE, Tox21, and ClinTox, which were originally introduced by [53]. The datasets are made up of extensive collections of SMILES, paired with binary labels that highlight the particular property being evaluated, such as BBBP: Blood-Brain Barrier Penetration, HIV: inhibit HIV replication, BACE: bindings results for a set of inhibitors of human beta-secretas, Tox21: toxicity of compounds and ClinTox: drugs failed clinical trials for toxicity reasons. A comprehensive explanation of these datasets can be referenced in the original research conducted by [53]. For ICL, we either select k samples randomly, or search the top-k most analogous molecules using RDKit [28] to determine the TanimotoSimilarity. However, it's crucial to mention that using the latter method does not assure an even distribution among classes. In our study, we employ a strategic sampling method for two categories of datasets: balanced and highly imbalanced. For balanced datasets, such as BBBP and BACE, we randomly select 30 samples for the validation process and 100 samples for testing from the original dataset. Contrastingly, for datasets exhibiting substantial label imbalance (39684:1443 $\approx$ 28:1, take HIV datasets as a example), we select samples from the majority and minority classes to achieve a ratio of 4:1. This strategic approach enables us to maintain a representative sample for the evaluation process, despite the original high imbalance in the dataset. To evaluate the results, we use the classification *accuracy*, as well as *F1* score as the evaluation metric, due to the class imbalance. We take MoIR [47] as a baseline, which is a Graph Neural Network based model, taking molecule graphs as input, and predict the property as a binary classification problem..

ICL Prompt. Figure 5 illustrates a sample of our ICL prompt for property prediction. Within the task-specific template, we include a detailed explanation of the task forecasting the penetration of the brain-blood barrier to assist GPT models in comprehending the input SMILES from the BBBP dataset. Additionally, we establish certain constraints for the output to conform to the specific characteristics of the property prediction task.

Table 5: F1 (↑) score of GPT models and baseline in molecular property prediction tasks. k is the number of examples used in few-shot ICL. The best GPT model is in bold font, and the baseline is underlined.

	BBBP	BACE	HIV	Tox21	ClinTox
MoIR [47]	0.859	0.693	0.315	0.647	0.750
GPT-4 (zero-shot)	0.560 ±0.034	0.322±0.018	0.977±0.013	0.489±0.018	0.555±0.043
GPT-4 (Scaffold, k= 4)	0.498 ± 0.028	0.516 ± 0.024	0.818 ± 0.015	0.444 ± 0.004	0.731 ± 0.035
GPT-4 (Scaffold, k= 8)	**0.587±0.018**	**0.666±0.023**	0.797 ± 0.021	**0.563±0.008**	0.736±0.033
GPT-4 (random, k= 8)	0.469 ± 0.025	0.504 ± 0.020	**0.994 ± 0.006**	0.528±0.003	**0.924±0.000**
GPT-3.5 (Scaffold, k= 8)	0.463 ± 0.008	0.406 ± 0.011	0.807 ± 0.021	0.529 ± 0.021	0.369 ± 0.029
Davinci-003 (Scaffold, k= 8)	0.378 ± 0.024	0.649 ± 0.021	0.832 ± 0.020	0.518±0.009	0.850 ± 0.020

Results. The results are reported as F1 in Table 5, accuracy in Table 6. We observed that GPT models outperform the baseline model in terms of F1 on four out of five datasets. In the range of GPT models examined, GPT-4 surpasses both Davinci-003 and GPT-3.5 in predicting molecular properties. In our investigation, we have found evidence to support that the expansion of in-context learning (ICL) instances leads to a measurable enhancement in model performance. This underlines a direct relationship between the extent of ICL data and the predictive precision of our models. Concurrently, our research presents empirical evidence that scaffold sampling exceeds the performance of random sampling on three distinct datasets (BBBP, BACE, Tox21). A plausible explanation for this could be

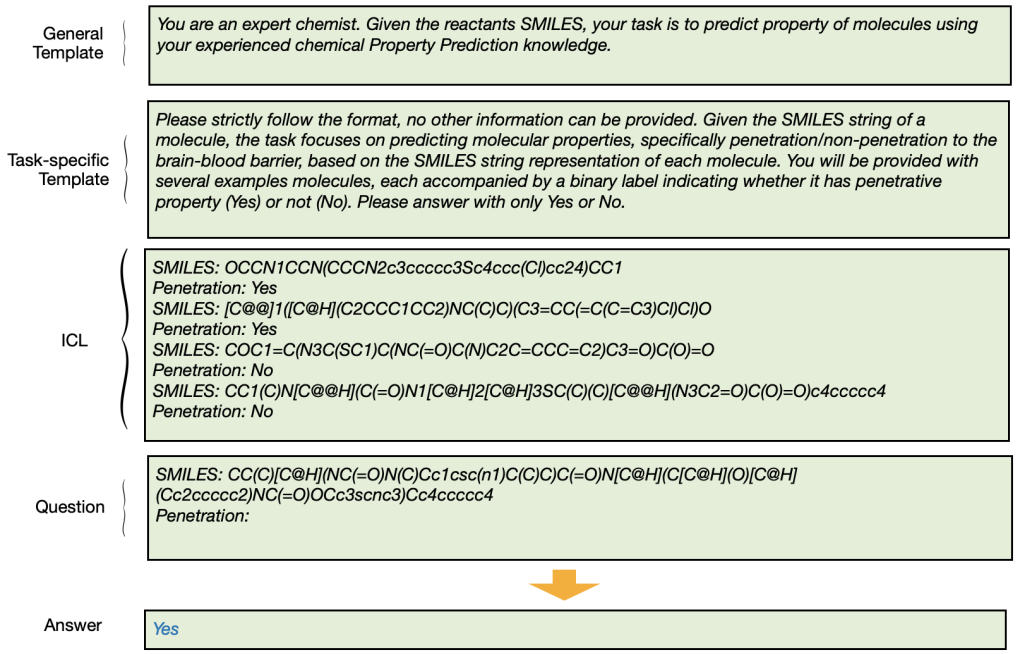


Figure 5: An ICL prompt example for property prediction

Table 6: Accuracy (↑) of GPT models and baseline in molecular property prediction tasks. k is the number of examples used in few-shot ICL. The best GPT model is in bold font, and the baseline is underlined.

	BBBP	BACE	HIV	Tox21	ClinTox
MoIR [47]	0.830	0.732	0.345	0.822	0.832
GPT-4 (zero-shot)	0.476 ± 0.036	0.499 ± 0.005	0.986 ± 0.007	0.518 ± 0.018	0.736 ± 0.027
GPT-4 (Scaffold, k= 4)	0.516 ± 0.022	0.514 ± 0.205	0.834 ± 0.014	0.457 ± 0.004	0.856 ± 0.014
GPT-4 (Scaffold, k= 8)	**0.614±0.016**	**0.679±0.205**	0.836 ± 0.020	0.737 ± 0.004	0.856 ± 0.014
GPT-4 (random, k= 8)	0.610 ± 0.021	0.588 ± 0.023	**0.996±0.004**	**0.874±0.003**	**0.930±0.010**
GPT-3.5 (Scaffold, k= 8)	0.463 ± 0.007	0.496 ± 0.016	0.864 ± 0.018	0.572 ± 0.026	0.578 ± 0.029
Davinci-003 (Scaffold, k= 8)	0.396 ± 0.023	0.650 ± 0.021	0.781 ± 0.004	0.682 ± 0.006	0.845 ± 0.010

the structural resemblances between the scaffold-sampled molecules and the query molecule, which potentially biases the GPT models towards more accurate decision.

5 Yield Prediction

Yield prediction [40] is a critical task in chemistry, specifically in the domain of synthetic chemistry, which involves the design and synthesis of new compounds for various applications, such as pharmaceuticals, materials, and catalysts. The yield prediction task aims to estimate the efficiency and effectiveness of a chemical reaction, primarily by quantifying the percentage of the desired product formed from the reactants. We use two High-Throughput experimentation(HTE) datasets: Butchward-Hartwig [1] and Suzuki-Miyaura dataset [39] for evaluation. These datasets comprise of reactions and their corresponding yields, which have been meticulously acquired through standardized and consistent experimental setups. This uniformity ensures that the data within each dataset is coherent, reducing the likelihood of discrepancies arising from variations in experimental procedures or conditions. We formulate the task of yield prediction as a binary classification problem, by determining whether a reaction is a high-yielding reaction or not. We used only random sampling for our ICL examples as reactions in those datasets belong to the same type. For every dataset, we randomly select 30 samples for the validation process and 100 samples for testing from the original dataset. To evaluate the results, we use the classification accuracy as the evaluation metric, with UAGNN [27] serving as baseline. UAGNN reports state-of-the-art preformance on yield prediction.

It takes the graphs of reactants and products as input, and learns representation of these molecules through a graph neural network, and then predicts the scaled yield .

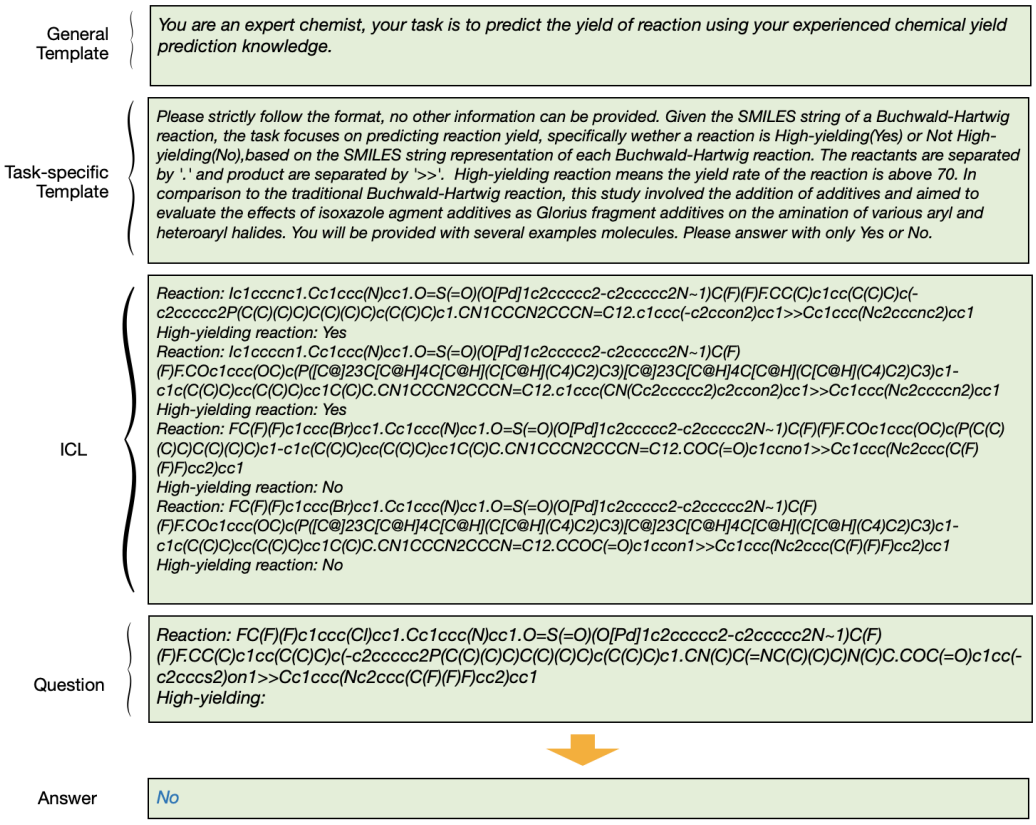


Figure 6: An ICL prompt example for yield prediction

ICL prompt. We show our ICL prompt for yield prediction with an example from Butchward-Hartwig dataset. As described in Figure 6, we incorporate an input explanation (wherein the reactants are separated by '.' and the products are split by '>>') to assist large language models. Additionally, output restrictions are enforced to ensure the generation of valid results.

Results. The results are presented in Table 7. Our analysis reveals that in the task of yield prediction, GPT models perform below the established baseline model, UAGNN. However, it's worth noting that the UAGNN model was trained on the full training dataset including thousands of examples. Considering the spectrum of GPT models under scrutiny, GPT-4 emerges as the superior model, overshadowing both Davinci-003 and GPT-3.5 in predicting reaction yields. In the process of our investigation, we unearthed supporting evidence that signifies the role of ICL instances in the enhancement of model performance. This suggests an inherent correlation between the quantity of ICL data and the predictive accuracy of the models under consideration. This phenomenon is particularly in the case of GPT-4, we observed a significant improvement in performance when the number of ICL examples was increased from 4 to 8, both in the Butchward-Hartwig and Suzuki-coupling reactions. This indicates that even within the same model architecture, the amount of contextual data can significantly influence the predictive capabilities.

6 Reaction Prediction

Reaction prediction is a central task in the field of chemistry, with significant implications for drug discovery, materials science, and the development of novel synthetic routes. Given a set of reactants, the goal of this task is to predict the most likely products formed during a chemical reaction [43, 10].

Table 7: Accuracy (↑) of yield prediction task. k is the number of examples used in few-shot ICL. The best GPT model is in bold font, and the baseline is underlined.

	Butchward-Hartwig	Suzuki-coupling
UAGNN [27]	<u>0.965</u>	<u>0.957</u>
GPT-4 (zero-shot)	0.322 ± 0.034	0.214 ± 0.019
GPT-4 (random, k= 8)	**0.800±0.008**	**0.764±0.013**
GPT-4 (random, k= 4)	0.574 ± 0.045	0.324 ± 0.018
GPT-3.5 (random, k= 8)	0.585 ± 0.045	0.542 ± 0.011
Davinci-003 (random, k= 8)	0.467 ± 0.013	0.341 ± 0.017

In this task, we use the widely-adopted USPTO-MIT dataset [23] for evaluating the performance of GPT models. This dataset contains approximately 470,000 chemical reactions extracted from US patents. In the experiment, we use the USPTO Mixed dataset where the reactants and reagents strings are splitted by '.'. We randomly sample 30 samples from the original validation set for validation and 100 samples from the original test set for testing. We use the Top-1 Accuracy as the evaluation metric and Chemformer [21] as the baseline due to its superior performance among the machine learning solutions for reaction prediction. Chemformer is a seq2seq model trained to predict the output product when given reactants and reagents as inputs. We also report the percentage of invalid SMILES generated by each method.

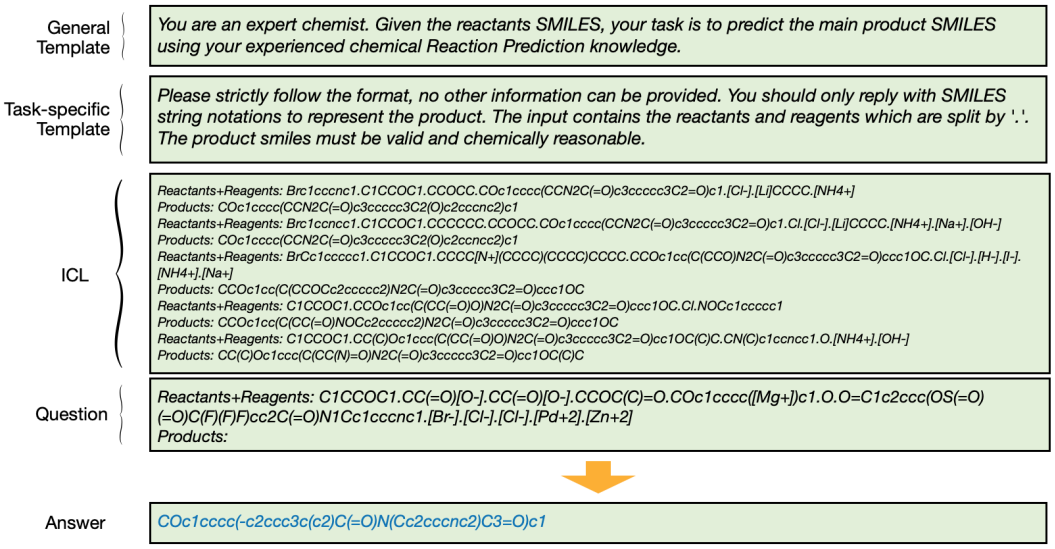


Figure 7: An ICL prompt example for reaction prediction

ICL Prompt. One example of our ICL prompt for reaction prediction is shown in Figure 7. Given the nature of the reaction prediction task and the characteristics of the USPTO-MIT dataset, we enhance the task-specific Template with an input explanation (stating that the input includes reactants and reagents, which are separated by '.') to assist the GPT models in understanding the input SMILES. Moreover, we incorporate output restrictions to guide GPT models in generating chemically valid and reasonable products.

Results. The results are reported in Table 8. We can observe that compared to the baseline, the performance of GPT models is considerably inferior, especially for the Zero-shot prompting (Top-1 Accuracy is only 0.004 and it generates 17.4% invalid SMILES). The less competitive results of GPT models can be attributed to the lack of an in-depth understanding of the SMILES strings that represent reactants and products, as well as the reaction process that transforms reactants into products. It's also worth mentioning that the high accuracy achieved by Chemformer is due to its training on the complete dataset. More conclusions and detailed analysis are summarized in the section 11 Discussion.

Table 8: The performance of GPT models and baseline in the reaction prediction task. k is the number of examples used in few-shot ICL. The best GPT model is in bold font, and the baseline is underlined.

Method	Top-1 Accuracy (↑)	Invalid SMILES (↓)
Chemformer [21]	0.938	0%
GPT-4 (zero-shot)	0.004 ± 0.005	17.4% ± 3.9%
GPT-4 (Scaffold, k=20)	**0.230 ± 0.022**	7.0% ± 1.6%
GPT-4 (Random, k=20)	0.012 ± 0.008	8.4% ± 1.5%
GPT-4 (Scaffold, k=5)	0.182 ± 0.015	**6.6% ± 1.5%**
GPT-3.5 (Scaffold, k=20)	0.184 ± 0.005	15.6% ± 2.3%
Davinci-003 (Scaffold, k=20)	0.218 ± 0.008	11.4% ± 2.7%

7 Reagents Selection

Reagents selection, also known as reagent recommendation, involves the identification and proposal of the most fitting reagents for a specific chemical reaction or process. Compared to other prediction and generation tasks, these selection tasks might be more fitting for LLMs and carry extensive implications. Reagent recommendation can markedly enhance reaction design by pinpointing optimal reagents and conditions for a given reaction, thereby augmenting efficiency and effectiveness in both academic and industrial settings. Drawing from a vast corpus of chemical knowledge, GPT models may be able to generate suggestions, leading to chemical reactions with a greater likelihood of yielding superior results.

In this study, we formulate four reaction component selection task from the Suzuki High-Throughput Experimentation (HTE) dataset. The dataset, created by Perera et al.[34], evaluates the Suzuki coupling of 5 electrophiles and 7 nucleophiles across a matrix of 11 ligands (with one blank), 7 bases (with one blank), and 4 solvents, resulting in a reaction screening dataset comprising 5,760 data points. The task of reagents selection can be divided into three categories: Reactant selection, Ligand Selection and Solvent selection. For validation, 30 examples were randomly sampled, while 100 examples were used for testing, all taken from the original datasets. Top-1 Accuracy serves as the assessment metric for both reactant and solvent selection, while Top-50% is utilized for ligand selection, as the upper half of the ligands in the list typically provide satisfactory yields in chemical reactions. This task is newly emergent in the field of chemistry, and as such, there are no established baselines yet.

ICL prompt. One example of our ICL prompt for reagents selection is shown in Figure 8. Considering the structure of the dataset and the characteristics of the reagents, we provide detailed task description and an answer template to guide GPT models towards the desired output.

Results. Our results are presented in Table 9. From the table, it's evident that GPT-4 and GPT-3.5 perform comparatively well in reagent selection tasks. This suggests a promising potential for GPT models in the realm of reagent selection.

Table 9: Accuracy (↑) of GPT models in the reagent selection tasks. For Reactant Selection and Solvent selection task, we report the mean (and standard deviation) of the Top-1 Accuracy score and we report the Top-50% accuracy score for the Ligand Selection task. The best GPT model is in bold font, and the baseline is underlined.

	Reactant Selection	Solvent Selection	Ligand Selection
GPT-4 (zero-shot)	0.299 ±0.029	**0.526**±0.012	**0.534**±0.059
GPT-3.5 (zero-shot)	**0.400**±0.038	0.368±0.034	0.436 ± 0.020
Davinci-003 (zero-shot)	0.178 ± 0.034	0.463 ± 0.014	0.432 ± 0.020

8 Retrosynthesis

Retrosynthesis planning is a crucial task in synthetic organic chemistry that involves identifying efficient synthetic pathways for a target molecule by recursively transforming it into simpler precursor molecules. In contrast to reaction prediction, retrosynthesis planning involves a reverse extrapolation

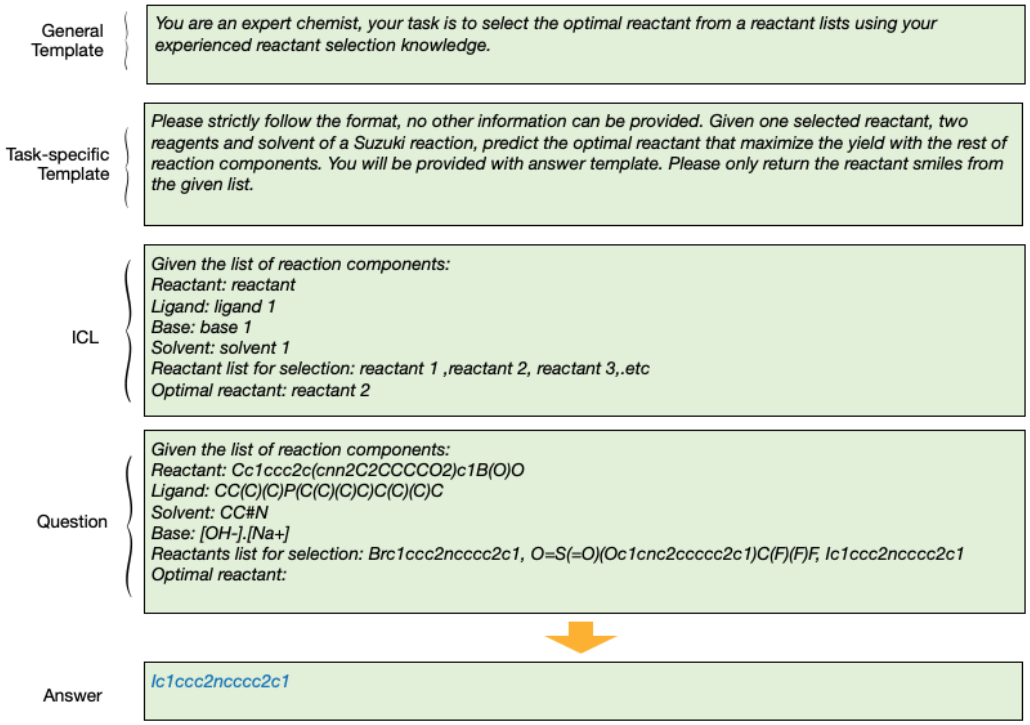


Figure 8: An ICL prompt example for reagents selection

from the target molecule to identify the readily available reactants for its synthesis. In this study, we use the USPTO-50k dataset [42], which contains 50,037 chemical reactions. In our experiment, we use the data splitting as [15] and we the training set which contains 40,029 reactions as the ICL candidates. The metric and baseline are the same as the reaction prediction.

ICL Prompt. One example of our ICL prompt for reaction prediction is shown in Figure 9. As discussed in the reaction prediction task, we also add the task-specific template to help GPT models understand the input and restrict the output.

Table 10: The performance of GPT models and baseline in Retrosynthesis task. The best GPT model is in bold font, and the baseline is underlined.

Method	Top-1 Accuracy (↑)	Invalid SMILES (↓)
Chemformer [21]	0.536	0%
GPT-4 (zero-shot)	0.006 ± 0.005	$20.6\% \pm 4.7\%$
GPT-4 (Scaffold, k=20)	0.096 ± 0.013	$10.4\% \pm 3.4\%$
GPT-4 (Scaffold, k=5)	0.114 ± 0.013	$11.0\% \pm 1.2\%$
GPT-4 (Random, k=20)	0.012 ± 0.011	$18.2\% \pm 4.2\%$
GPT-3.5 (Scaffold, k=20)	0.022 ± 0.004	$6.4\% \pm 1.3\%$
Davinci-003 (Scaffold, k=20)	$\mathbf{0.122 \pm 0.013}$	$\mathbf{6.0\% \pm 1.2\%}$

Results. The results are reported in Table 10. The performance of GPT models is also inferior than the baseline due to the lack of an in-depth understanding of the SMILES strings that represent reactants and products. Detailed analysis are summarized in the later section 11 Discussion.

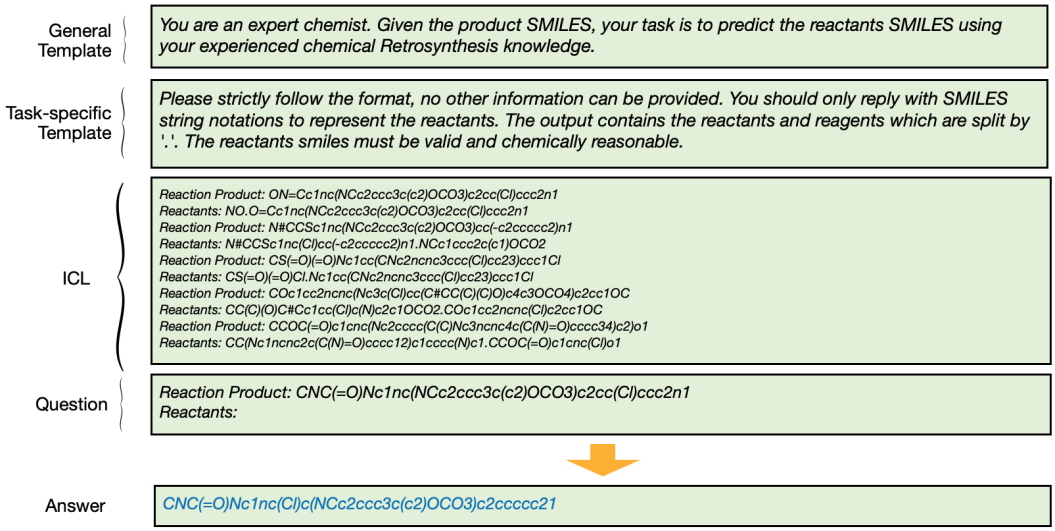


Figure 9: An ICL prompt example for Retrosynthesis

9 Text-Based Molecule Design

Text-Based Molecule Design is a novel task in computational chemistry and drug discovery. It involves generating new molecules with desired molecule descriptions. In our experiment, we employ the ChEBI-20 dataset which consists of 33,010 molecule-description pairs. The dataset is split into 80/10/10% as the training/validation/test set [15]. We use the training set which contains 26407 molecule-description pairs as the ICL candidates. For comparison, we use the MolT5-Large [15] as the baseline. MolT5-Large is the initial effort to investigate the translation between molecules and text, including tasks such as text-based molecule design and molecule captioning. It builds upon T5 [36], an encoder-decoder Transformer model, and benefits from pretraining on a large amount of dataset. To comprehensively evaluate the performance, we employ three different types of metrics. The first type of metric is the chemical similarity between the ground-truth molecules and generated molecules, measured by FTS (fingerprint Tanimoto Similarity) [44] in terms of MACCS [38], RDK [28], and Morgan [11]. Secondly, we also use FCD (Fréchet ChemNet Distance) [35] which allows comparing molecules based on the latent information used to predict the activity of molecules [15]. Since the generated molecules are in SMILES string format, we also employ natural language processing metrics including BLEU, Exact Match [15], and Levenshtein distance [31] between the ground-truth molecules and generated molecules SMILES. Finally, to evaluate whether generated molecules are valid, we use RDKIT [28] to check the validity of generated molecules and report the percent of the valid molecules.

ICL Prompt. One ICL prompt example for text-based molecule design is shown in Figure 10.

Table 11: The performance of GPT models and baseline in the Text-Based Molecule Design task. The best GPT model is in bold font and the baseline is underlined.

Method	BLEU (↑)	Exact (↑)	Levenshtein (↓)	Validity (↑)	MACCS FTS (↑)	RDK FTS (↑)	Morgan FTS (↑)	FCD (↓)
MolT5-Large [15]	0.601	0.290	41.600	0.940	0.879	0.797	0.752	5.394
GPT-4 (zero-shot)	0.490±0.017	0.046±0.009	47.418±1.668	0.758±0.015	0.733±0.020	0.514±0.021	0.432±0.014	11.913±0.972
GPT-4 (Scaffold, k=10)	**0.816±0.004**	**0.174±0.029**	**21.160±0.600**	0.888±0.023	**0.867±0.005**	0.738±0.010	**0.672±0.013**	6.224±0.449
GPT-4 (Scaffold, k=5)	0.815±0.011	0.164±0.018	21.862±1.768	0.874±0.030	0.865±0.015	**0.741±0.023**	0.670±0.028	**5.843±0.515**
GPT-4 (Random, k=10)	0.602±0.016	0.060±0.007	42.390±1.008	0.770±0.030	0.762±0.013	0.548±0.017	0.475±0.015	10.594±0.414
GPT-3.5 (Scaffold, k=10)	0.479±0.156	0.094±0.011	82.008±40.354	0.854±0.059	0.833±0.006	0.686±0.016	0.585±0.013	8.341±0.607
Davinci-003 (Scaffold, k=10)	0.741±0.011	0.100±0.010	25.648±2.186	**0.936±0.009**	0.783±0.014	0.648±0.004	0.560±0.010	8.335±0.310

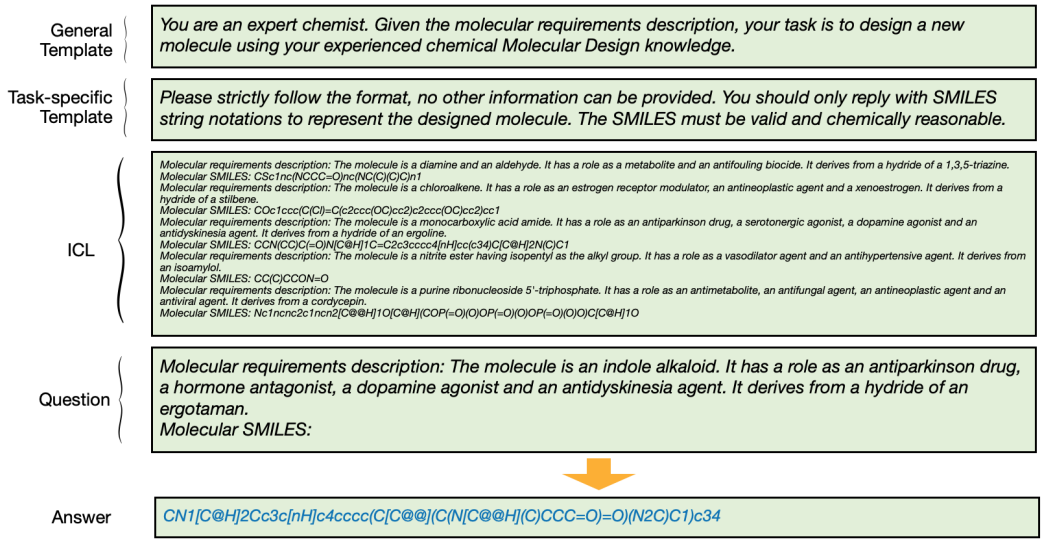


Figure 10: An ICL prompt example for Text-Based Molecule Design

Results. The results are reported in Table 11. We can observe that the best ICL prompting GPT models (GPT-4 and Davinci-003) can achieve competitive performance or even outperform the baseline in some metrics (BLEU, Levenshtein). Although the GPT models significantly underperform the baseline in terms of exact match and Morgan FTS metrics, it's important to note that we only utilize a maximum of 10 examples, which is substantially less than the training set (comprising 26,407 training examples) used for the baseline. These results demonstrate the strong few-shot text-based molecule design ability of GPT models. Last, not being exactly the same as the ground truth doesn't necessarily mean it's incorrect, especially in the context of molecular design. The molecules generated by GPT models may still be useful and can serve as alternatives to the ground truth, given they fulfill the requirements described in the input text and a majority (over 89%) are chemically valid.

Case studies. We select three different types of molecules (organic molecule without rings, organic molecule with ring, and metal atom) as examples, and show the generated molecules in Figure 11. We observe that the structure of molecules generated by the GPT-4 (Scaffold, k=10) method is more similar to the ground truth compared to Davinci-003, GPT-4 (zero-shot), and even the baseline. Additionally, for metal atoms design, GPT models outperform the baseline which wrongly generates the SMILES instead of the metal atom. These cases show promising results of the molecule design ability of GPT models. However, evaluating whether the generated molecules are helpful such as molecule novelty in real-world scenarios is still a difficult problem. Thus we conclude that GPT models have excellent potential in molecule design and there are prospects for investigating this ability.

10 Molecule Captioning

Molecule captioning is an important task in computational chemistry, offering valuable insights and applications in various areas such as drug discovery, materials science, and chemical synthesis. Given a molecule as input, the goal of this task is to generate a textual description that accurately describes the key features, properties, and functional groups of the molecule. We also use the ChEBI-20 dataset and the training set of it as the ICL candidates as discussed in the Text-Based Molecule Design Section. We use traditional captioning metrics including BLEU, ROUGE, and METEOR for evaluation.

ICL Prompt. One example of our ICL prompt for molecule captioning is shown in Figure 12.

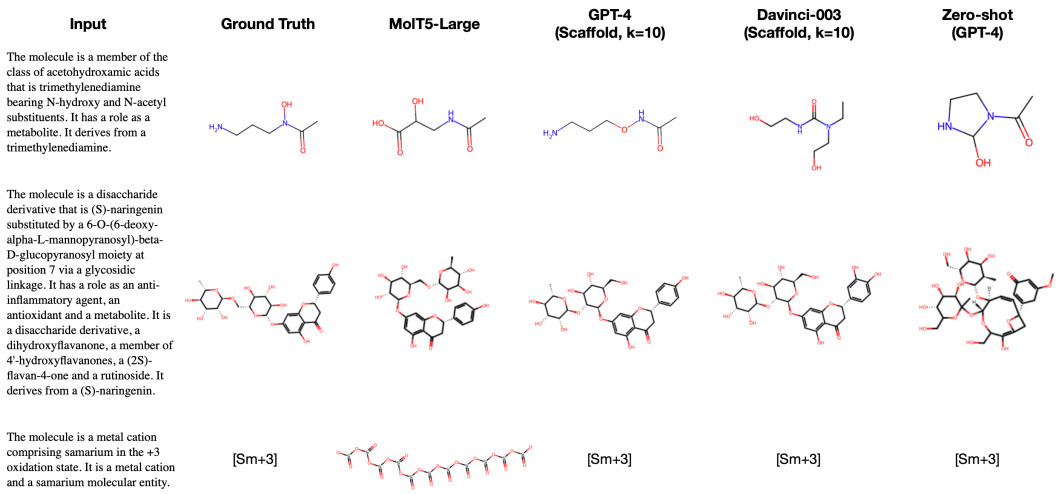


Figure 11: Examples of molecules generated by different models.

Results. The results are reported in Table 12. We can observe that the best ICL prompting GPT models (GPT-4 and Davinci-003) can achieve competitive performance or even outperform the baseline in some metrics (BLEU-2 and BLEU-4). This indicates the inspiring capability of the GPT models in the molecule captioning task.

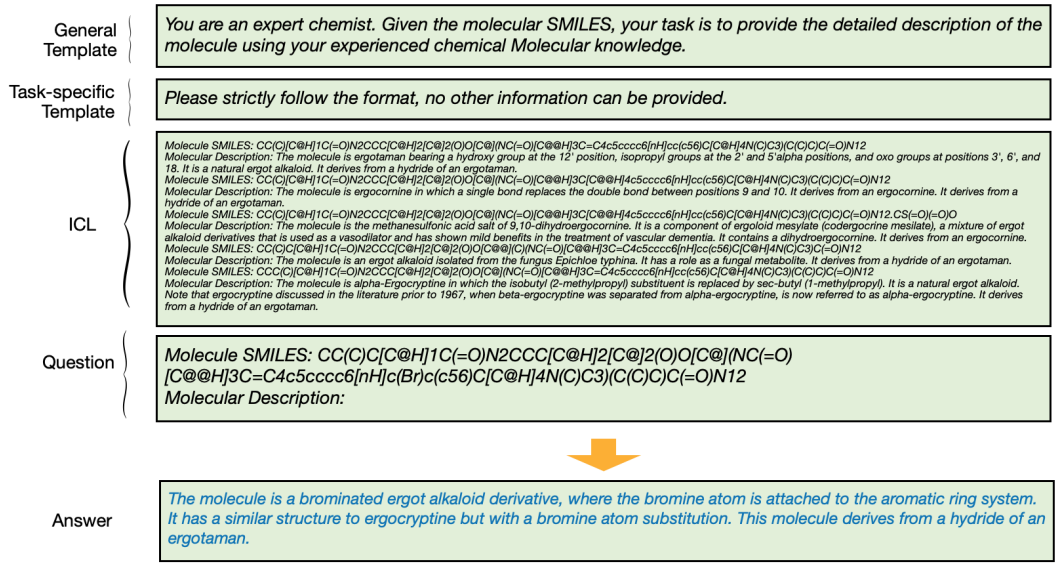


Figure 12: An ICL prompt example for molecule captioning

Case studies. Same as case studies in the Text-Based Molecule Design task, we also select three different types of molecules as examples, and the captions are shown in Figure 13. We observe that although the performance of the baseline is close to GPT models, the captions generated by the baseline contain more descriptions that violate the chemical facts. In contrast, the captions generated by GPT-4 models contain only a few inaccurate descriptions, highlighting the excellent explaining ability of GPT models. This highlights the limitations of applying traditional Natural Language Processing (NLP) evaluation metrics to this task. Therefore, it is necessary to create more suitable evaluation metrics for chemistry-related generation tasks.

Table 12: The performance of GPT models and baseline in the molecule captioning task. The best GPT model is in bold font and the baseline is underlined.

Method	BLEU-2 (↑)	BLEU-4 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	METEOR (↑)
MolT5-Large [15]	0.482	0.383	0.574	0.410	0.516	0.530
GPT-4 (zero-shot)	0.062±0.001	0.013±0.001	0.192±0.002	0.040±0.002	0.125±0.002	0.209±0.002
GPT-4 (Scaffold, k=10)	0.464±0.008	0.365±0.008	**0.545±0.003**	**0.362±0.003**	0.459±0.007	**0.519±0.005**
GPT-4 (Scaffold, k=5)	0.456±0.003	0.357±0.004	0.540±0.005	0.355±0.007	0.455±0.005	0.505±0.005
GPT-4 (Random, k=10)	0.260±0.007	0.140±0.007	0.393±0.004	0.180±0.006	0.309±0.004	0.320±0.007
GPT-3.5 (Scaffold, k=10)	0.468±0.010	0.368±0.010	0.534±0.005	0.355±0.007	0.457±0.006	0.497±0.005
Davinci-003 (Scaffold, k=10)	**0.488±0.011**	**0.391±0.012**	0.532±0.008	0.359±0.010	**0.465±0.008**	0.478±0.011

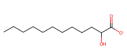
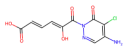
Input	Ground Truth	MolT5-Large	GPT-4 (Scaffold, k=10)	Davinci-003 (Scaffold, k=10)	Zero-shot (GPT-4)
	The molecule is a 2-hydroxy fatty acid anion resulting from the deprotonation of the carboxy group of 2-hydroxydodecanoic acid. The major microspecies at pH 7.3 It derives from a dodecanoate. It is a conjugate base of a 2-hydroxydodecanoic acid.	The molecule is an oxo fatty acid anion resulting from the deprotonation of the carboxy group of 2-oxododecanoic acid. The major species at pH 7.3. It is a conjugate base of a 2-oxododecanoic acid. It is a tautomer of a 2-hydroxydodec-2-enoate.	The molecule is a long-chain fatty acid also known as undecylic acid, with a carboxylate anion at one end, consisting of an eleven-carbon straight chain. It has a role as an antimicrobial agent and a fungicide. It is a saturated fatty acid and a long-chain fatty acid. It derives from a hydride of an undecane. It is a conjugate base of an undecanoic acid.	The molecule is a decanoic acid that is an organic carboxylic acid having a 10-carbon chain and a carboxyl group at the end. It is an aliphatic carboxylic acid and an acyclic fatty acid. It is the conjugate base of decanoic acid.	The given molecular SMILES, CCCCCCCCCCC(O)C(=O)[O-], represents an organic molecule with the following detailed description: 1. The molecule consists of a linear alkyl chain containing 11 carbon atoms, which makes it an undecyl group. The linear chain can be denoted as CH3(CH2)10-. This chain confers hydrophobic properties to the molecule. 2. There is a hydroxyl group (OH) attached to one of the terminal carbon atoms. This has a polar, hydrophilic character, thus adding an amphiphilic property to the molecule. 3. The other terminal carbon atom features a carboxylate group, which is written as COO-. This is an anionic, negatively charged functional group that imparts additional hydrophilicity to the molecule. Combining these attributes, the detailed molecule description would be an amphiphilic, undecyl hydroxyl carboxylate with an 11-carbon linear alkyl chain and terminal hydroxyl and carboxylate functional groups. This molecule is likely to exhibit surfactant-like properties due to its amphiphilic nature, which allows it to dissolve in both polar and nonpolar solvents.
	The molecule is an organochlorine compound and a pyridazinone. It derives from a muconic acid. It is a conjugate acid of a 5-amino-4-chloro-2-(2-oxidomuconoyl)pyridazin-3(2H)-one(2-).	The molecule is an oxo monocarboxylic acid that is 3,8,8-trichloroocta-2,4,7-trienoic acid substituted by a 4-chlorophenyl group at position 7, a hydroxy group at position 2 and an oxo group at position 6. It is a member of monochlorobenzenes, an oxo monocarboxylic acid and an enol. It is a conjugate acid of a 6-oxo-2-hydroxy-7-(4'-chlorophenyl)-3,8,8-trichloroocta-2E,4E,7E-trienoate.	The molecule is a pyridazinone that is pyridazin-3(2H)-one substituted by an amino group at position 5, a chloro group at position 4, and a 2-hydroxymuconoyl group at position 2. It is a metabolite of the herbicide chloridazone. It has a role as a marine xenobiotic metabolite. It is a pyridazinone, a primary arylamine, an organochlorine compound, and a carboxylic acid.	The molecule is a monocarboxylic acid anion arising from deprotonation of the carboxy and enol groups of 5-amino-4-chloro-2-(2-hydroxy-2-oxoprop-1-en-1-yl)pyridazin-3(2H)-one; major species at pH 7.3. It is a conjugate base of a 5-amino-4-chloro-2-(2-hydroxy-2-oxopropenyl)pyridazin-3(2H)-one.	The given molecular SMILES notation represents a chlorinated pyrimidine derivative with an extended conjugated system involving an ester, an -unsaturated ketone, and a vinyl group. Below is a detailed description of the molecule: 1. The core structure is a pyrimidine ring that consists of a 6-membered aromatic ring with nitrogen atoms at position 1 and 3. The nitrogen atom at position 1 (N1) is connected to a chlorine (Cl) atom at position 6, rendering it chlorinated. 2. There are two additional double bonds in the ring between C2 and N3, and between C4 and N1, making it fully conjugated and aromatic. 3. At position 5 (C5) of the pyrimidine ring, we have a carbonyl (C=O) group attached, making it a 2,4-dioxo-1,3-diazine derivative. 4. On the exocyclic carbonyl group, there is a single bond to an ester group (C(=O)O). The ester group is connected to a sp2 carbon which is part of a vinyl group (C=C). 5. This vinyl group (C=C)
[25Mg]	The molecule is the stable isotope of magnesium with relative atomic mass 24.985837, 10.0 atom percent natural abundance and nuclear spin 5/2.	The molecule is the stable isotope of magnesium with relative atomic mass 25.9723. The least abundant (3.09 atom percent) isotope of naturally occurring magnesium.	The molecule is the stable isotope of magnesium with relative atomic mass 24.985837, 10 atom percent natural abundance and nuclear spin 5/2. It is a magnesium isotope and a member of magnesium isotopes.	The molecule is the stable isotope of magnesium with relative atomic mass 24.3050, 78.99 atom percent natural abundance and nuclear spin 0.	The given SMILES notation [25Mg] represents an isotope of magnesium (Mg). Magnesium is an alkaline earth metal and has an atomic number of 12. The given notation shows magnesium with a mass number of 25, which means it has 13 neutrons (mass number = atomic number + neutrons, so neutrons = 25 - 12 = 13). Magnesium is a divalent cation; it has a positive 2 charge when it forms ions (Mg2+), and it typically forms ionic compounds with other elements. In its elemental form, magnesium is a lightweight, silvery-white metal, which is essential for living organisms and widely used in the industry. The description of the molecule using the SMILES notation [25Mg] focuses primarily on the isotope of magnesium with the mass number 25.

Figure 13: Examples captions generated by different models. Descriptions that violate chemical facts are marked in grey.

11 Discussion

The results described above allow us to approach the question **What can GPT models do in chemistry?** from several directions. We discuss the key findings from our comprehensive benchmark analysis, provide valuable insights by thoroughly analyzing the limitation of LLMs and how different settings related to LLMs affect performance across various chemistry tasks.

11.1 Limitation of LLMs on understanding molecular SMILES

A significant limitation of LLMs is their lack of understanding of molecular representations in SMILES strings, which in many cases leads to inaccurate or inconsistent results as shown in Section 3 for the translation of different ways to name molecules. SMILES (Simplified Molecular Input Line Entry System) [49, 50] is a widely used textual representation for chemical structures. For example, the SMILES string for ethanol, a simple alcohol, is "CCO". This string represents a molecule with two carbon atoms (C) connected by a single bond and an oxygen atom (O) connected to the second carbon atom. SMILES strings can serve as both input and output for LLMs, alongside other natural

language text. However, several issues make it challenging for LLMs to accurately understand and interpret SMILES strings: **1)** Hydrogen atoms are not explicitly represented in SMILES strings, as they can be inferred based on the standard bonding rules. LLMs frequently struggle to infer these implicit hydrogen atoms and may even fail at simple tasks like counting the number of atoms in a molecule [22, 6]. **2)** A given molecule can have multiple valid SMILES representations, which can lead to ambiguity if not properly processed or standardized. LLMs may thus fail to consistently recognize and compare molecular structures represented by different SMILES strings. **3)** LLMs do not have any inherent understanding of SMILES strings, and treat them as a sequence of characters or subwords. When processing long SMILES strings, LLMs rely on the byte-pair encoding tokenization technique, which can break the string into smaller pieces or subwords in ways that do not represent the molecular structure and properties of molecules represented by SMILES strings. Because many tasks in cheminformatics rely on the accurate representation of a molecule by SMILES strings, the non-competitive performance of GPT models in converting structures into SMILES strings (and vice versa) affects downstream tasks such as retrosynthesis, reaction and name prediction. Models that have an enhanced ability of handling molecular structures and their specific attributes or coupling to existing tools such as RDKit[28] will be needed.

11.2 Can GPT models outperform existing baselines in chemistry tasks?

Table 1 summarizes the performance of GPT models and allows a comparison of the top-performing GPT models to their respective baselines (if available and designate the performance of GPT models as NC (not competitive), C (competitive), or SC (selectively competitive). We discuss these performance categories below.

- **NC (not competitive)**. In tasks such as **reaction prediction**, and **retrosynthesis**, GPT models are worse than existing machine learning baselines trained on large amounts of training data, partially because of the limitation on understanding molecular SMILES strings. In **reaction prediction** and **retrosynthesis**, SMILES strings are present in both the input and output of the GPT models. Without an in-depth understanding of the SMILES strings that represent reactants and products, as well as the reaction process that transforms reactants into products, it will be difficult for GPT models to generate accurate responses. GPT models exhibit poor performance on the task of **name prediction** as well. This further validates the notion that GPT models struggle with understanding long strings in formats such as SMILES, IUPAC name, and molecule formula, and make correct translations between them.
- **C (Competitive)**. GPT models can achieve satisfactory results when the chemistry tasks are formulated into forms of classification (e.g., formatting **yield prediction** into a high-or-not classification, instead of regression) or ranking (as seen in **reagents selection**). This is understandable, because making choices is inherently simpler than generating products, reactants or names. GPT models can achieve an accuracy of 40% to 50% when asked to select the reactant or solvent or ligand from provided candidates. Although GPT-4's performance on **yield prediction** falls short compared to the baseline model UAGNN [27] (with 80% versus 96% on the Buchwald-Hartwig dataset, and 76% versus 96% on the Suzuki-coupling dataset), it demonstrates improved performance when given more demonstration examples within the few-shot in-context learning scenario. It is worth noting that the UAGNN model was trained on thousands of examples for these specific reactions.
- **SC (selectively competitive)**. GPT models are selectively competitive on two types of tasks.
  - In the **property prediction** task on some datasets (HIV, ClinTox), GPT models outperform the baseline significantly, achieving F1 scores and accuracy nearing 1. This might be due to the fact that the property labels to be predicted are included in the prompts, with GPT models being simply tasked with responding with *yes* or *no*. For example, the prompt includes *inhibit HIV replication* or *drugs failed clinical trials for toxicity reason*. In contrast, baselines employing machine learning models do not include the semantic meaning of these labels in their input. The input for these models only comprises molecular representations in graph form but no labels.
  - For tasks related to text, such as **text-based molecule design and molecule captioning**, GPT models exhibit strong performance due to their language generation capabilities. On the task of **text-based molecule design**, GPT models outperform the baseline when evaluated using natural language processing metrics such as BLEU and Levenshtein.

However, when it comes to exact match, the accuracy is less than 20%. This suggests that the molecules designed by GPT models may not be exactly the same as the ground truth. Particularly in the context of molecular design/generation, the exact match is a significant metric. Unlike in natural language generation where there is some allowance for deviation from the input, molecular design demands precise accuracy and chemical validity. However, not being precisely identical to the ground truth does not automatically invalidate a result. Molecules generated by GPT models may still prove to be beneficial and could potentially act as viable alternatives to the ground truth, provided they meet the requirements outlined in the input text and the majority (over 89%) are chemically valid. Nonetheless, assessing the true utility of these generated molecules, such as evaluating their novelty in real-world applications, can be a time-consuming undertaking.

11.3 The capability of different GPT models

As shown in Table 1, we can find that GPT-4 model shows better chemical understanding, reasoning, and explaining abilities than Davinci-003 and GPT-3.5. This further verifies the GPT-4 model outperforms the other two models in both basic and realistic scenarios [5].

11.4 The effects of the ICL

To investigate the effects of the ICL, we introduced ICL prompting and different ICL retrieval methods, and the different number of ICL examples in each task. Based on the experiments, we have the following three observations:

- In all tasks, the performance of ICL prompting is better than zero-shot prompting.
- In most tasks, using scaffold similarity to retrieve the most similar examples of the question as ICL examples achieves better performance than random sampling.
- In most tasks, using larger k (more ICL examples) usually achieves better performance than small k (fewer ICL examples).

These observations indicate that ICL prompting generally achieves better performance than Zero-shot prompting. The quality of quantity of ICL examples in ICL prompting plays an important role in the performance of ICL prompting [18, 29]. This may inspire that it is necessary to design more chemistry-specific ICL methods to build high-quality ICL examples to further improve the ICL prompting performance.

11.5 The limitations of current evaluation methods

Although in Text-Based Molecule Design and Molecule Captioning tasks, GPT models show competitive performance compared to the baseline in some metrics (BLEU, Levenshtein, ROUGE, FCD, etc), we observe that the exact match of GPT models is inferior to the baseline in the Text-Based Molecule Design task and the GPT models generate some descriptions which violate chemical facts. This divergence between metrics and real-world scenarios mainly arises because, unlike many natural language processing tasks that can be suitably evaluated by sentence-level matching evaluation metrics, chemistry-related tasks necessitate exact matching for SMILES and precise terminology in descriptions. These findings spotlight the limitations of current evaluation metrics and underscore the need for the development of chemistry-specific metrics.

11.6 Hallucination of GPT models in chemistry

Our evaluation experiments across various tasks reveal two primary types of hallucinations exhibited by GPT models in the domain of chemistry. The first type occurs when the input is given in SMILES format (e.g., name prediction); GPT models occasionally struggle with interpreting these SMILES correctly. For instance, they may fail to recognize the number of atoms or certain functional groups within molecules during name prediction tasks. The second type of hallucination arises when the expected output from GPT models should be in the form of SMILES (e.g., reaction prediction and retrosynthesis). Here, GPT models may produce molecules that are chemically unreasonable, suggesting a gap in understanding of what constitutes valid SMILES. Hallucination issues represent

a key challenge with GPT models, particularly in the field of chemistry which necessitates exact matching of SMILES and adherence to strict chemical facts [51]. Current GPT models need further investigation into this problem.

11.7 Prospects of GPT models for Chemistry

Overall, through an exhaustive set of experiments and analyses, we outline several promising avenues for the application of GPT models in the field of chemistry. While GPT models underperform relative to baselines across a majority of tasks, it's important to note that GPT models leverage only a few examples to solve chemistry problems, whereas baselines are trained on extensive, task-specific datasets and are limited to certain tasks. This observation provides valuable insights into the potential of GPT models' generalized intelligence in the domain of chemistry. The employment of advanced prompting techniques such as Chain-of-thought (CoT) [48], Decomposed Prompting [25] could potentially boost the capacity of GPT models to perform complex reasoning. On the other hand, GPT models display a considerable amount of hallucinations in chemistry tasks, indicating that current GPT models may not yet possess the necessary capabilities to solve practical chemistry problems effectively. However, with continuous development of GPT models and further research into methods to avoid hallucinations, we are optimistic that GPT models can significantly enhance their problem-solving abilities in the field of chemistry.

12 Related Work

12.1 Large Language Models

The rise of Large Language Models (LLMs) has marked a significant trend in recent natural language processing (NLP) research. This progress has been fuelled by milestones such as the introduction of GPT-3, a model with 175 billion parameters [4] and other notable advancements such as T0 [41], Flan-T5 [9], Galactica [45] and LLaMa [46]. These advancements have had a substantial impact on the NLP field. Recently, the release of GPT4 which is an evolution from GPT-3.5 series with Reinforcement Leaning with Human-Feedback(RLHF) [8] has drawn considerable attention for its improvements in language understanding, generation, planning and overall performance. Despite the vast potential of GPT models, existing research primarily centers on their performance within general NLP tasks. The scientific disciplines, notably chemistry, have received relatively less focus. The application of GPT models in these specialized domains presents an opportunity for significant advancements. Therefore, we conduct a comprehensive experimental analysis to evaluate the capability of GPT series model in chemistry-related tasks.

12.2 Large Language Model Evaluations

In recent years, the evaluation of LLMs like GPT has become a significant field of inquiry. [7] showed ChatGPT's proficiency in law exams, while technical aspects of GPT-4 were analyzed in [33]. Discourse and memorization aspects of GPT models have been studied [12], emphasizing the role of 'memorization' in quality output. GPT models are also applied in healthcare [11] and code generation tasks [30]. Specifically, in healthcare, the utility and safety of GPT models in clinical settings were explored. Meanwhile, for code generation, a rigorous evaluation of LLM-synthesized code was conducted. These studies underscored the complexity of achieving task-specific accuracy and functionality with GPT models. Lastly, AGIEval [54] assessed GPT models' general abilities but noted struggles in complex reasoning tasks.

Our work aligns with these evaluations but diverges in its focus on chemical tasks. To our knowledge, this is the first study to transform such tasks to suit LLM processing and to perform a comprehensive evaluation of these models' ability to tackle chemistry-related problems. We believe that this focus will significantly contribute to the field, expanding our understanding of GPT models' capabilities in specific scientific domains.

12.3 Large Language Model for Chemistry

Recent efforts integrating LLMs with the field of chemistry generally fall into two distinct categories. One category aims to create a chemistry agent with LLMs' by leveraging its planning ability to

utilize task-related tools. For example, Bran et al [3] developed ChemCrow, which augmented LLMs with chem-expert designed tools for downstream tasks such as organic synthesis, drug discovery, and materials design. Similarly, by leveraging the planning , design, execution ability of multiple LLMs, Boiko et al [2] developed an autonomous chemical agent to conduct chemical experiments. The other category involves directly use of LLMs for the downstream tasks in chemistry [22, 51, 6]. However, while some of work also studied the performance of LLMs in chemistry-related tasks, these studies do not offer a comprehensive evaluation of the capabilities of LLMs in chemistry-related tasks. As such, there is a notable gap for a benchmark that meticulously assesses the capability of LLMs within the field of chemistry. This benchmark is essential to not only identify the strengths and limitations of these models in a specialized scientific domain, but also to guide future improvements and applications.

13 Conclusion

In this paper, we summarize the required abilities of GPT models in chemistry and construct a comprehensive benchmark to evaluate the 3 most popular GPT models (GPT-4, GPT-3.5, Davinci-003) on 8 widely-used chemistry tasks. The experiment results show that GPT models perform less competitive in generative tasks which require in-depth understanding molecular SMILES strings, such as reaction prediction, name prediction, and retrosynthesis. GPT models show competitive performance in tasks that are in classification or ranking formats such as yield prediction and reagents selection. GPT models are selectively competitive on tasks involving text in prompts such as property prediction and text-based molecule design, or explainable tasks such as molecule captioning. These experiments indicate the potential of GPT models in chemistry tasks and the need for further improvement. Our study can provide valuable insights and we hope our work can inspire future research to explore the potential of GPT models in chemistry.

Acknowledgments and Disclosure of Funding

The research reported in this paper was supported by NSF Center for Computer Assisted Synthesis (C-CAS) No. CHE-2202693.

References

[1] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360 (6385):186–190, 2018.

[2] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

[3] Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[6] Cayque Monteiro Castro Nascimento and André Silva Pimentel. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655, 2023.

[7] Jonathan Choi, Kristin Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Journal of Legal Education*, 2023.

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[10] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5): 434–443, 2017.

[11] Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al. Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint arXiv:2304.13714*, 2023.

[12] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*, 2023.

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

[14] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47.

[15] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[16] Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. Few-shot graph learning for molecular property prediction. In *Proceedings of the Web Conference 2021*, pages 2559–2567, 2021.

[17] Zhichun Guo, Bozhao Nan, Yijun Tian, Olaf Wiest, Chuxu Zhang, and Nitesh V Chawla. Graph-based molecular representation learning. *arXiv preprint arXiv:2207.04869*, 2022.

[18] Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples, 2022.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[21] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

[22] Kevin Jablonka, Philippe Schwaller, Andrés Ortega-Guerrero, and Berend Smit. Is gpt-3 all you need for low-data discovery in chemistry. *10.26434/chemrxiv-2023-fw8n4*, 2023.

[23] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network, 2017.

[24] Rehan Ahmed Khan, Masood Jawaid, Aymen Rehan Khan, and Madiha Sajjad. Chatgpt-reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2):605, 2023.

[25] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

[26] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

[27] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, and Seokho Kang. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics*, 14:1–10, 2022.

[28] G. A. Landrum. Rdkit: Open-source cheminformatics software. http://www.rdkit.org, 2020.

[29] Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization, 2022.

[30] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023.

[31] Frederic P Miller, Agnes F Vandome, and John McBrewster. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance, 2009.

[32] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2): 107–113, 1965.

[33] OpenAI. Gpt-4 technical report, 2023.

[34] Damith Perera, Joseph W Tucker, Shalini Brahmbhatt, Christopher J Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W Sach. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434, 2018.

[35] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[37] Mayk Caldas Ramos, Shane S Michtavy, Marc D Porosoff, and Andrew D White. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341*, 2023.

[38] David Ratcliff, John W.; Metzener. Pattern matching: The gestalt approach, 1988.

[39] Brandon J Reizman, Yi-Ming Wang, Stephen L Buchwald, and Klavs F Jensen. Suzuki–miyaura cross-coupling optimization enabled by automated feedback. *Reaction chemistry & engineering*, 1(6):658–666, 2016.

[40] Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrzej M Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G Doyle, Nitesh V Chawla, et al. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 2023.

[41] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[42] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56 (12):2336–2346, 2016.

[43] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[44] Taffee T Tanimoto. Elementary mathematical theory of classification and prediction. *Journal of Biomedical Science and Engineering*, 1958.

[45] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[47] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888*, 2021.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[49] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[50] David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[51] A.D. White. The future of chemistry is language., 2023.

[52] Andrew D. White, Glen M. Hocky, Heta A. Gandhi, Mehrad Ansari, Sam Cox, Geemi P. Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and Willmor J. Peña Ccoa. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2:368–376, 2023. doi: 10.1039/D2DD00087C. URL http://dx.doi.org/10.1039/D2DD00087C.

[53] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[54] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.