Quantifying Uncertainty in Large Language Models: Applications in Molecular Chemistry Tasks

Zizhang Chen[*]
Department of Computer Science
Brandeis University
Waltham, MA 02453
zizhang2@brandeis.edu

Pengyu Hong
Department of Computer Science
Brandeis University
Waltham, MA 02453
hongpeng@brandeis.edu

Sandeep Madireddy[†]
Mathematics and Computer Science Division
Argonne National Laboratory
smadireddy@anl.gov

Abstract

Large language models (LLMs) have exhibited impressive reasoning capabilities and proficiency in answering complex questions. However, they are prone to generating inaccurate or fabricated responses, a phenomenon commonly referred to as hallucination. This issue is particularly critical in high-stakes fields such as molecular chemistry, where errors can have significant consequences. It is essential to implement robust uncertainty quantification methods that enable us to evaluate the reliability of outputs generated by large language models. In this work, we present a novel Question Rephrasing technique to assess the input uncertainty of LLMs, which refers to the uncertainty arising from equivalent variations of the inputs provided to LLMs. This technique is integrated with sampling methods that measure the output uncertainty of LLMs, thereby offering a more comprehensive uncertainty assessment. We validated our approach to property prediction and reaction prediction for molecular chemistry tasks.

1 Introduction

In recent years, Large Language Models (LLMs), such as GPT (Achiam et al., 2023), Claude Anthropic (2024), and Llama Touvron et al. (2023), have demonstrated remarkable success in various tasks. Pre-trained on vast amounts of data and boosted with billions of parameters, these LLMs demonstrated impressive capabilities across a range of scientific domains, including chemistry Guo et al. (2023a), biology Agathokleous et al. (2023), and physics Nguyen et al. (2023). Despite their successes, a critical aspect that remains under-explored is the uncertainty inherent in the predictions produced by these LLMs. Understanding and quantifying uncertainty in LLM outputs is crucial for several reasons. It aids in informed decision-making, enhances user trust, and ensures the safety and reliability of AI systems (Sun et al., 2024). Moreover, transparency about model uncertainty fosters responsible AI deployment.

Inspired by the practice in psychological assessments, where clinicians ask the same question in different ways to test a patient's understanding and consistency of responses, we propose a technique, termed *Question Rephrasing*, to quantify the uncertainty of the answer produced by an LLM in response to a question. Essentially, given an initial question, the *Question Rephrasing* technique

[*] Corresponding Author
[†] Corresponding Author

Statistical Foundations of LLMs and Foundation Models Workshop at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

involves rephrasing the question while maximally preserving its original meaning and then submitting the rephrased question to the LLM. The consistency between the LLM's answers before and after rephrasing is evaluated to quantify the uncertainty of the LLM with respect to the input variations. In addition, a sampling approach is adopted that repeatedly queries the LLM with the same input to assess the output uncertainty of the LLM.

In our experiments, we applied our method to quantify the uncertainty of GPT-3.5/4 (Achiam et al., 2023) on two tasks in the Chemistry domain: property prediction and forward reaction prediction analogous to classification and text generation tasks, respectively. We found that GPT-4 was sensitive to *Question Rephrasing*, and the output uncertainty could serve as a valuable indicator for the accuracy and reliability of the LLM's response.

2 Background and Related Work

2.1 Textual representation of molecules

The textual representation of molecular structures is fundamental for applying language models to chemistry-related tasks. Prominent among these representations are the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988; O'Boyle, 2012) and the International Union of Pure and Applied Chemistry (IUPAC) (Panico et al., 1993; Leigh, 2011) nomenclature. Currently, no standardized rules are in place for assigning common names to chemical compounds. IUPAC provides a universally recognized method for naming chemical entities, whereas SMILES offers a more compact, machine-readable format that has recently facilitated significant advancements in applying language models to chemistry (Xu et al., 2017; Ross et al., 2022; Wu et al., 2023; Fang et al., 2024). Given its ease of use and compatibility with various machine learning workflows, we used the SMILES notation as the primary method for representing molecular structures.

2.2 Chemistry tasks and LLMs

Recent literature highlights the expanding role of LLMs in molecular chemistry, particularly in enhancing predictive and generative tasks. Guo et al. (2023b) established benchmarks for evaluating LLMs in property and reaction outcome predictions, demonstrating their broad applicability. Zhong et al. (2024a) showed that while LLMs lag behind specialized machine learning models in processing geometric molecular data, they significantly enhance performance when integrated with these models. Zhong et al. (2024b) shows that LLMs as post-hoc correctors improves the accuracy of molecular property predictions after initial model training. Qian et al. (2023) and Jablonka et al. (2024) underscore the utility of LLMs in generating explanatory content for molecular structures and resolving complex chemical queries, enhancing both educational and practical applications. Luong & Singh (2024) found that transformer-based models like GPT and BERT exhibit high accuracy in reaction prediction and molecule generation.

2.3 Uncertainty quantification for black-box LLMs

The recent shift towards black-box LLMs, particularly in commercially deployed models such as GPT4 (Achiam et al., 2023), Claude 3 (Anthropic, 2023) and Gemini (Team et al., 2023), presents unique challenges for Uncertainty Quantification (UQ). Traditionally, UQ techniques have relied heavily on accessing the internal model parameters and predictions at a granular level, such as token probabilities and logits (Gal & Ghahramani, 2016; Malinin & Gales, 2018; Hu et al., 2023). However, the encapsulation of modern LLMs, often provided as API services, restricts such access. Recent studies Kuhn et al. (2023); Lin et al. (2023); Xiong et al. (2024) have started to address these limitations by innovating methods and pipelines that infer uncertainty directly from the text outputs generated by LLMs without requiring their internal workings. Kuhn et al.(2023) introduce semantic entropy, a novel metric to quantify uncertainty in LLMs that focuses on semantic equivalence, the concept that different phrases can express the same meaning. Later works (Lin et al., 2023; Xiong et al., 2024) introduce complex frameworks to refine black-box UQ methods comprising prompting strategies, sampling methods, and aggregation techniques. This work aims to quantify the black-box LLMs uncertainty on chemistry-related tasks.

3 Uncertainty Quantification in Molecular Chemistry Tasks

This section introduces and discusses UQ methods for chemistry-related tasks using black-box LLMs. We categorized our UQ metrics into two parts: **input uncertainty** and **output uncertainty**. Input uncertainty uses the *Question Rephrasing* strategy to assess LLM's sensitivity to variations in molecular representations. We systematically use the alternative SMILES representations of each input molecule in the prompt and investigate how these perturbations impact the LLM's output predictions. Since the alternative SMILES of the same molecule are used, we were able to guarantee that the semantics of the modified prompt remain the same. In addition, this method can test whether an LLM truly understands molecular representations in chemistry or is only able to perform string comparisons. Output uncertainty assesses the consistency of the output produced by an LLM, which is influenced purely by the model's inherent properties. We repeatedly query the model with identical input to create a distribution of the answers. We structured our pipelines based on existing UQ-related works (Prabhakaran et al., 2019; Lin et al., 2023; Kuhn et al., 2023). Below, we outline our UQ methods:

1. For a chemistry-related task t, given a SMILES representation x_i of the i-th molecule, generate a prompt P_{t,x_i} based on a task-specific template (see Section 3.1).
2. Generate a list of up to n SMILES variants of the molecule x_i: $L = \{x_i^1, x_i^2, ..., x_i^n\}$. We ask GPT-4 to rank the SMILES variants by its confidence to interpret their structures and choose the one, say $\hat{x}_i$, with the highest confidence to construct a prompt $P_{t,\hat{x}_i}$ by replacing x_i in P_{t,x_i} with $\hat{x}_i$ (see Section 3.2).
3. Ask the LLM to generate m responses for the prompt $P_{t,\hat{x}_i}$ and obtain $R_{t,\hat{x}_i} = \{r_{t,\hat{x}_i,1}, r_{t,\hat{x}_i,2}, ..., r_{t,\hat{x}_i,m}\}$.
4. Calculate the entropy-based uncertainty metrics U_{t,x_i} and $U_{t,\hat{x}_i}$ for R_{t,x_i} and $R_{t,\hat{x}_i}$, respectively.
5. Measure the input uncertainty by comparing U_{t,x_i} and $U_{t,\hat{x}_i}$ for all chosen x_i. Measure the output uncertainty by examining U_{t,x_i} and $U_{t,\hat{x}_i}$ separately.

In the subsequent subsections, we provide detailed explanations of our UQ methods.

3.1 Prompt design for molecular chemistry tasks

It was shown that LLMs exhibited a certain degree of zero-shot learning capabilities (Brown et al., 2020). Here, we adopted and modified the structured approach delineated in the recent Chemistry LLM benchmark study Guo et al. (2023b) to design chemistry task-specific prompt completion pairs using In-Context Learning (ICL) samples. Motivated by the OpenAI prompt guide (Shieh, 2023) and the benchmark paper Guo et al. (2023b), we designed our prompts to consist of three parts: 1. Chemistry role-playing prompts with task-specific instructions. 2. Few shot ICL samples were constructed using k-scaffold sampling. 3. Questions to be answered for the target SMILES. Table 1 showcases the prompt design for the toxicity prediction task.

Table 1: An example of prompts for chemistry-related tasks.

Role: *You are an expert Chemist specializing in Chemical Property Prediction.*
Task: *Given the SMILES representation of a molecule, use your expertise to predict the molecular properties based on its structure...*
ICL samples: *For the following SMILES, determine if each molecule contains a toxicity compound, answering only with "Yes" or "No". A few examples are provided:* *SMILES: Few-shot example 1* *Contain toxicity compound: Yes* ... *SMILES: Few-shot example p* *Contain toxicity compound: No*
Question: *SMILES: target smiles* *Contain toxicity compound: [Provide an answer based on analysis]* *Please strictly answer with "Yes" or "No".*

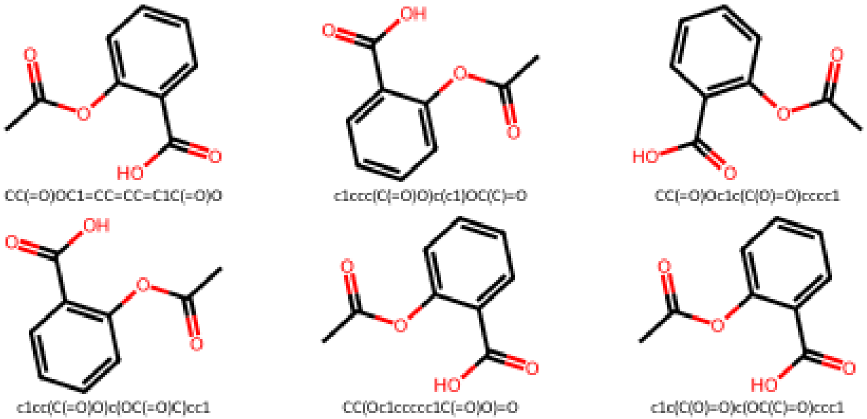
CC(=O)OC1=CC=CC=C1C(=O)O
c1ccc(C(=O)O)c(c1)OC(C)=O
CC(=O)Oc1c(C(O)=O)cccc1
c1cc(C(=O)O)c(OC(=O)C)cc1
CC(Oc1ccccc1C(=O)O)=O
c1c(C(O)=O)c(OC(C)=O)ccc1

Figure 1: SMILES representation variants of Aspirin. While all structures depict the same molecule, their SMILES representations are different, which introduces input variations. **Top left**: Canonical SMILES representation of Aspirin. **Rest**: Five SMILES variations of Aspirin.

3.2 Input Uncertainty: Sensitivity Analysis

We investigated input uncertainty by analyzing the sensitivity of a black-box LLM to changes in inputs. For each ICL prompt P_{t,x_i} of a chemistry task t, we rephrased it by replacing the SMILES representation x_i with its equivalent SMILES to generate a new prompt. Specifically, we first obtained the structure of the molecule s_i of x_i using RDKit (Landrum et al., 2013, 2020). Then, we obtained a list of up to n distinct SMILES representations $L = \{x_i^1, x_i^2, ..., x_i^n\}$ for the structure s_i. For better illustration, we use Aspirin as an example to showcase this step (see Figure 1). We then prompted GPT-4 to rank the obtained SMILES variants by its confidence in interpreting the structures from those SMILES variants (see Table 2). The SMILES variant $\hat{x}$ with the highest confidence score was chosen to construct a new prompt $P_{t,\hat{x}_i}$ by replacing x_i in P_{t,x_i} with $\hat{x}_i$. The LLM was then asked to generate responses for the prompts $P_{t,x}$ and $P_{t,\hat{x}}$ separately. We then evaluated the responses produced by LLM for $P_{t,x}$ and $P_{t,\hat{x}}$. *Accuracy* was the metric used in the molecule classification tasks, and *exact match accuracy* was the metric used in the tasks that generate SMILES.

Table 2: Prompt template for generating SMILE confidence score

Role: *As an expert in chemistry with a thorough understanding of SMILES notation.*
Questions: *Can you rank your confidence score in the following smiles for interpreting its structures? [please output the exact smile string]:* *variation SMILES 1* *variation SMILES 2* ... *variation SMILES n*

3.3 Output uncertainty: Uncertainty Quantification from Structure Similarly

In this section, we explain the entropy-based metrics for measuring the output uncertainty of black-box LLMs, focusing on classification and generation tasks in the chemistry domain.
For classification tasks, the LLM's responses $R_{t,x_i} = \{r_{t,x_i,1}, r_{t,x_i,2}, ..., r_{t,x_i,m}\}$ of the molecule x_i can be interpreted as a set of classification results, where each response $r_{t,x_i,j}$ is a class label predicted by LLMs from a set of possible classes $C = \{c_1, c_2, \ldots, c_k\}$. Here, k is the number of classes that appear in the prediction outputs. The probability of each class $c_j \in C$ can be calculated as the percentage of c_j appearing in R_{t,x_i}:

$$P(c_j) = \frac{|\{r_{t,x_i} = c_j : r_{t,x_i} \in R_{t,x_i}\}|}{|R_{t,x_i}|} \tag{1}$$

where $|\{r_{t,x_i} = c_j : r_{t,x_i} \in R_{t,x_i}\}|$ counts the number of times that class c_j appears in R_{t,x_i}. The uncertainty score U_{t,x_i} is formulated as:

$$U_{t,x_i} = -\sum_{j=1}^{k} P(c_j) \log P(c_j) \tag{2}$$

For all generation tasks that produce the SMILES representation, we measured the similarity between the generated SMILES using the Tanimoto Similarity (Butina, 1999; Chung et al., 2019) based on their molecular fingerprints, which can be obtained with RDKit (Landrum et al., 2013). Sometimes an LLM may generate invalid SMILES representations. We set the similarity between an invalid SMILES and any other SMILES to be an infinitely small number ϵ. Once we obtain the pairwise similarity between all SMILES generated for a specific molecule x_i, we applied hierarchical clustering to group the generated SMILES into g clusters $S = \{s_1, s_2, \ldots, s_g\}$. The probability of a cluster $s_j \in S$ is calculated as its percentage in R_{t,x_i}:

$$P(s_j) = \frac{|\{r_{t,x_i} \in R_{t,x_i} : r_{t,x_i} = s_j\}|}{m} \tag{3}$$

Without loss of generality, the uncertainty score U_{t,x_i} can be formulated as follows:

$$U(R_{t,x_i} \mid S) = -\sum_{j=1}^{g} P(s_j) \log P(s_j) \tag{4}$$

4 Experiments

Following Kuhn et al. (2023); Lin et al. (2023), we evaluate our output uncertainty metric by utilizing it to predict whether LLM can correctly generate an answer. We plot the Receiver operating characteristic curve (ROC) and calculate the Area under the ROC Curve (AUC) score. An AUC score of 0.5 indicates that the uncertainty metrics are no better than a random classifier, whereas a high AUC score indicates that the metrics can help us determine whether to trust the model's response. We evaluated the input uncertainty by comparing the model performances across different inputs. A significant increase or decrease in model performance may indicate that the model is sensitive to its input and, thus, less likely to be trusted.

4.1 Property Prediction

We used five datasets (BBBP, HIV, BACE, Tox21, and ClinTox (Wu et al., 2018)) and the associated tasks to investigate the capabilities of our method to quantify the uncertainty of Black-box LLMs (specifically GPT-4) on predicting molecular properties. These datasets, sourced from the corresponding established databases and scientific literature, are primarily used in training machine learning models to predict binary molecular properties from their SMILES representations. For each dataset, adapted from the experimental settings of (Guo et al., 2023b), we randomly sampled the 100 molecules as a test set and constructed the prompts using ICL samples querying from the rest of the dataset. For each prompt, we repeatedly generated 5 responses and calculated the uncertainty score from Equation (2), here, denoted as Class Entropy, and used to predict whether GPT-4 can generate the correct answers. In addition, we reformulate the input SMILES and re-run the experiments following the methods mentioned in Section 3.2.
The prediction and uncertainty quantification results are presented in Table 3 and Figure 2. We noticed a slight decrease in model performance (except BP) when using reformed SMILES over the original SMILES input in Table 3. This indicates GPT's relatively high confidence among the input invariants. In addition, according to Figure 2, the AUC score for the original SMILES spans between 0.546 and 0.774, indicating a moderate trustworthiness in using the output uncertainty score to predict the GPT's response correctness.

4.2 Forward Reaction Prediction

We utilize the USPTO-MIT dataset (Schneider et al., 2016; Jin et al., 2017) to evaluate our uncertainty quantification metrics. The test set is constructed by randomly sampling 100 reaction-product pairs, while the remaining data are used to query the in-context learning (ICL) samples. For evaluations, we employ GPT-4 and GPT-3.5 Turbo to generate responses. We repeatedly generate 3, 10, 15, and 20 responses for each prompt. We first calculate the accuracy score by performing an exact

Table 3: Property prediction results of GPT-4 using original input SMILES (Orig. SMILES) and reformulated SMILES (Reform. SMILES) on five datasets. The evaluation metrics include Accuracy and F1 score. The average Class Entropy (C. E) is also reported.

Model	$GPT-4$ (Orig. SMILES)			$GPT-4$ (Reform. SMILES)		
Eval. metric	Acc.	F1	U.Q	Acc.	F1	U.Q
BACE	0.750	0.766	0.150	0.660 ↓	0.638↓	0.398
BBBP	0.690	0.756	0.290	0.700 ↑	0.795 ↑	0.415
ClinTox	0.820	0.357	0.319	0.833 ↓	0.285 ↓	0.427
HIV	0.910	0.471	0.060	0.763 ↓	0.350 ↓	0.292
Tox21	0.707	0.522	0.105	0.533 ↓	0.416 ↓	0.290

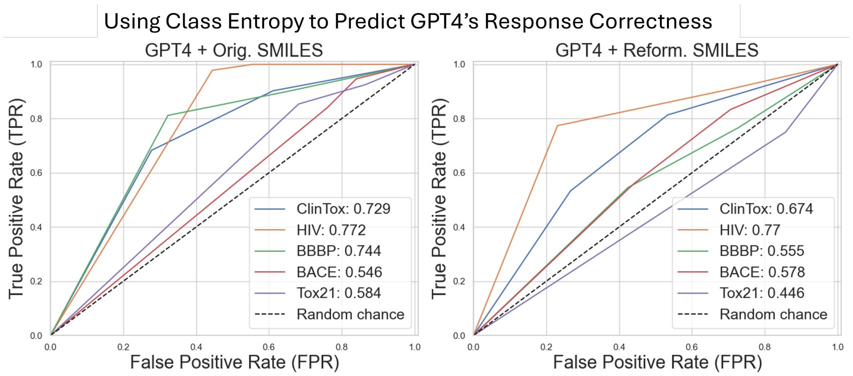

Figure 2: ROC curve for evaluating the in predicting the correctness of the GPT using our uncertainty score.

match comparison between the generated SMILES and the ground-truth SMILES. We then calculate the output uncertainty metric and use it to predict whether the response from black-box LLMs is correct. We then derived the AUC score for each set of responses. In addition, we perform the input uncertainty analysis by reformulating the input SMILES as we mentioned in Section 3.2 and repeat the above steps.

We present our results in Table 4. We observe that GPT-3.5/4 performed poorly on reaction prediction tasks. In addition, our output uncertainty metrics are reliable indicators of the correctness of GPT's responses (AUC score ranges from 0.86 to 0.99). We also observed a substantial decline in model performance on reaction prediction tasks when presented with the variations in molecular representation, demonstrating the LLMs' weakness in understanding basic chemistry knowledge.

Table 4: Reaction prediction performances of GPTs and AUC scores of output uncertainty metrics

Method	Top-1 Acc.	AUC-3	AUC-10	AUC-15	AUC-20
GPT-4 + Orig.	0.250	0.864	0.919	0.915	0.927
GPT-4 + Reform	0.070 ↓	0.972	0.941	0.958	0.993
GPT-3.5 + Orig	0.186	0.904	0.899	0.924	0.943
GPT-3.5 + Reform	0.036 ↓	0.919	1.000	1.000	1.000

5 Conclusions

In this work, we introduce a novel *Question Rephrasing* technique for uncertainty quantification in LLMs, specifically applied to chemistry tasks. By integrating input and output uncertainty assessments, we enhanced the ability to comprehensively evaluate the reliability of LLMs. We applied our approach to quantify the trustworthiness of LLMs in molecular chemistry. Experiment results show that GPT-3.5/4 exhibits sensitivity to input variations, and entropy-based metrics can effectively capture the output uncertainty of GPT-3.5/4, enabling the prediction of the correctness of LLM responses. Our experimental results underscore the need to enhance LLMs' understanding of basic chemistry knowledge. We believe that our approach and the discovery in this study help pave the way for developing more reliable and transparent AI systems for scientific applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Evgenios Agathokleous, Costas J Saitanis, Chao Fang, and Zhen Yu. Use of chatgpt: What does it mean for biology and environmental science? *Science of The Total Environment*, 888:164154, 2023.

Anthropic. Introducing the claude-3 family. 2023. URL https://www.anthropic.com/news/claude-3-family.

AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

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.

Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

Neo Christopher Chung, BłaŻej Miasojedow, Michał Startek, and Anna Gambin. Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data. *BMC bioinformatics*, 20(Suppl 15):644, 2019.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tlsdsb6l9n.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Taicheng Guo, Kehan Guo, Zhengwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, Xiangliang Zhang, et al. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365*, 16, 2023a.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b. URL https://openreview.net/forum?id=1ngbR3SZHW.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*, 2023.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pp. 1–9, 2024.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

Greg Landrum et al. Rdkit/rdkit: 2020 release, 2020. URL https://doi.org/10.5281/zenodo.3732262.

Geoffrey J Leigh. *Principles of chemical nomenclature: a guide to IUPAC recommendations*. Royal Society of Chemistry, 2011.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2023.

Kha-Dinh Luong and Ambuj Singh. Application of transformers in cheminformatics. *Journal of Chemical Information and Modeling*, 2024.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, et al. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*, 2023.

Noel M O'Boyle. Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. *Journal of cheminformatics*, 4:1–14, 2012.

R Panico, WH Powell, and Jean-Claude Richer. *A guide to IUPAC Nomenclature of Organic Compounds*, volume 2. Blackwell Scientific Publications, Oxford, 1993.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, November 2019. URL https://aclanthology.org/D19-1578.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. Can large language models empower molecular property prediction? *arXiv preprint arXiv:2307.07443*, 2023.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

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.

Jessica Shieh. Best practices for prompt engineering with openai api. *OpenAI https://help. openai. com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api*, 2023.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5312–5320, 2023.

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.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 285–294, 2017.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Benchmarking large language models for molecule prediction tasks. *arXiv preprint arXiv:2403.05075*, 2024a.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Harnessing large language models as post-hoc correctors. *arXiv preprint arXiv:2402.13414*, 2024b.