How to Detect and Defeat Molecular Mirage: A Metric-Driven Benchmark for Hallucination in LLM-based Molecular Comprehension

Anonymous ACL submission

Abstract

Large language models are increasingly used in scientific domains, especially for molecular understanding and analysis. However, existing models are affected by hallucination issues, resulting in errors in drug design and utilization. In this paper, we first analyze the sources of hallucination in LLMs for molecular comprehension tasks, specifically the knowledge shortcut phenomenon observed in the PubChem dataset. To evaluate hallucination in molecular comprehension tasks with computational efficiency, we introduce **Mol-Hallu**, a novel free-form evaluation metric that quantifies the degree of hallucination based on the scientific entailment relationship between generated text and actual molecular properties. Utilizing the Mol-Hallu metric, we reassess and analyze the extent of hallucination in various LLMs performing molecular comprehension tasks. Furthermore, the Hallucination Reduction Post-processing stage (HRPP) is proposed to alleviate molecular hallucinations, Experiments show the effectiveness of HRPP on decoder-only and encoder-decoder molecular LLMs. Our findings provide critical insights into mitigating hallucination and improving the reliability of LLMs in scientific applications.

1 Introduction

Large language models (LLMs) are regarded as foundation models in scientific fields due to their outstanding cross-domain generalization capability (Zhang et al., 2024a,b). In chemistry, LLMs are used for molecular property prediction (Lv et al., 2024; Qian et al., 2023) and molecular design (Flam-Shepherd et al., 2022; Grisoni, 2023). These models bridge the gap between molecular structural and property features and the natural language descriptions, facilitating multiple chemical applications including virtual screening, drug design, retrosynthesis planning, etc.

Although LLMs have shown powering generation capability in biochemistry domains, they suffer from hallucinations (Bang et al., 2023) which leads to the fabrication of non-existent facts or inappropriate molecular properties (Yao et al., 2023). Hallucinations often arise when new biochemical knowledge introduced during the supervised fine-tuning (SFT) stage conflicts with the model's pre-trained knowledge (Gekhman et al., 2024). The risky SFT strategy is frequently employed in various molecular LLMs (Pei et al., 2023; Fang et al., 2023; Yu et al., 2024), demonstrating the ubiquity of hallucinations.

Several studies on molecular LLMs analyze the hallucination phenomenon in molecule comprehension tasks. MoleculeQA (Lu et al., 2024b) and MoleculeTextQA (Laghuvarapu et al., 2024) construct multi-choice QA datasets to assess the hal-

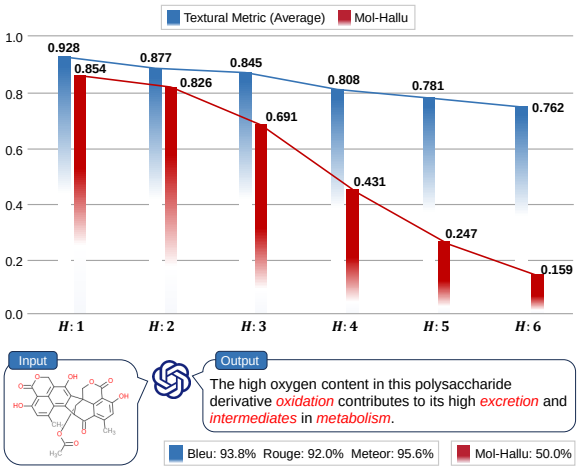


Figure 1: (1) The top figure shows the scoring curves of Mol-Hallu v.s. traditional metrics (BLEU, ROUGE, METEOR) across varying degrees of hallucination. $H : n$ indicates that samples contain n counterfactual errors, Mol-Hallu imposes an exponential penalty on hallucination errors in text., whereas traditional metrics fail to evaluate biochemical hallucination in texts reasonably. (2) The bottom figure proposes a biochemical sample that suffers severe hallucination (red are counterfactual entities) as an example. Mol-Hallu precisely reflects the hallucination degree in scientific texts compared to traditional metrics.

lucination issues in molecular LLMs. However, these approaches require additional datasets for fine-tuning in the context of fixed-form evaluation (Li et al., 2024b) and their multiple-choice question format is ill-suited for assessing the open-ended generation capabilities of large language models (Wang et al., 2023). To address this limitation, there is an urgent need for a free-form evaluation metric to quantify the degree of hallucination in molecular LLMs. Moreover, existing research has not yet analyzed the sources of hallucination in molecular LLMs or explored how to effectively mitigate these hallucinations.

To alleviate these issues, we first analyze the source of hallucinations in molecular LLMs and propose **Mol-Hallu**, the first free-form evaluation metric specifically designed to assess hallucination. Our investigation focuses on the PubChemQA dataset (Li et al., 2024a), a widely recognized benchmark source from PubChem database (Wang et al., 2009) that aligns molecular structures with textual descriptions. We identify that knowledge shortcuts in this dataset hinder the alignment between molecular structures and biochemical entities, resulting in increased hallucinations. To quantify the extent of hallucinations, Mol-Hallu leverages the union of the answer and the molecular general description, rewarding correct biomedical entities. The union and intersection are computed using an entailment model to determine whether the molecular descriptions entail a given text n-gram. To enhance evaluation, we curated a chemical entity database by automatically annotating PubChem and ChEMBL (Mendez et al., 2019) datasets, to accurately retrieve biomedical entities from predicted texts. Fig.1 demonstrates the rationality of Mol-Hallu for hallucination evaluation compared to traditional metrics including BLEU (Papineni et al., 2002a), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

To mitigate the hallucination in current molecular LLMs, we propose the Hallucination Reduction Post-processing (HRPP) stage, which constructs a hallucination-sensitive preference dataset by leveraging our chemical entity database, thereby optimizing the accuracy of scientific entities in text generated by molecular LLMs. The HRPP approach has validated its effectiveness and generalizability under decoder-only and encoder-decoder language models, two basic paradigms of molecular LLMs. Our contributions are summarized as follows:

- We dive into the molecular hallucination issue and identify that bio-knowledge shortcuts in the dataset exacerbate LLM hallucination.
- To measure the hallucination in molecular comprehension with efficiency, we propose the first free-form evaluation metric, Mol-Hallu, which calculates the F1-score of scientific entities using entailment probability.
- We further propose the hallucination reduction post-processing stage to alleviate the molecular hallucination using the hallucination-sensitive preference dataset.

2 Related Works

2.1 LLMs for Molecular Comprehension

Large language models pretrained with biochemical scientific data have shown substantial success in molecular comprehension tasks (Feng et al., 2024). The molecular encoders capture 1D sequential features (Irwin et al., 2022; Edwards et al., 2022; Fang et al., 2023; Wang et al., 2019), 2D topological features (Rong et al., 2020; Ying et al., 2021; Wang et al., 2022), and 3D structural patterns (Liu et al., 2021; Zhou et al., 2023; Lu et al., 2024a) from the molecule. Related studies have adopted two primary strategies to bridge the heterogeneity gap between molecular and textual representations for enhanced comprehension. Firstly, the cross-modal contrastive learning strategy is applied to fine-tune molecular and textual encoders. MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2023a), and MolCA (Liu et al., 2023b) construct a joint representational space that aligns molecular features with their corresponding textual descriptions. As textual encoders grow in parameter size and inferential capability, some studies (Cao et al., 2025, 2024b; Hu et al., 2025) have turned to supervised fine-tuning using molecular-text datasets to establish a pooling layer that maps molecular representations into the textual space of LLMs. However, constrained by the feature bias of molecular encoders and the prior knowledge of LLMs, current molecular LLMs are plagued by significant hallucination issues.

2.2 Hallucination in Biochemical LLMs

Alongside the advancement in reasoning, LLM models often generate nonsensical or unfaithful content to the provided source, referred as *hallucination* (Bang et al., 2023; Maynez et al., 2020).

The source-reference divergence phenomenon (Ji et al., 2023) is the main cause of hallucination. The divergence comes from heuristic data collection (Parikh et al., 2020) and imperfect representation learning during the training procedure (Feng et al., 2020) or erroneous decoding when conducting inference (Dziri et al., 2021). In molecular comprehension tasks, molecular LLMs often generate counterfactual content, which can lead to adverse consequences such as misleading users, and ultimately undermine the reliability of LLMs in scientific applications (Lu et al., 2024b).

The evaluation of hallucinations in LLMs can be categorized into two main types: **(1) Fixed-form evaluation** and **(2) Free-form evaluation**. Fixed-form evaluation uses multi-choice QA datasets, such as MoleculeQA and MoleculeTextQA, to assess hallucinations. However, this method requires fine-tuning LLMs on hallucination datasets and uses a multi-choice format that differs from the open-ended nature of LLM tasks, making it less reflective of true hallucination extent. In contrast, free-form evaluation leverages automated functions for faster, more computationally efficient assessments. Hallucination detection methods also fall into two categories: **(1) Fact-checking-based methods**, which verify accuracy through external (Chern et al., 2023; Min et al., 2023) or internal knowledge (Kadavath et al., 2022; Dhuliawala et al., 2023), and **(2) Uncertainty estimation methods** (Varshney et al., 2023; Manakul et al., 2023), which detect hallucinations by quantifying model confidence without external references. Our work bridges these approaches by introducing a free-form evaluation metric for molecular comprehension tasks. This method leverages ground truth while avoiding the need for external retrieval or fine-tuning, providing an efficient and domain-specific solution for hallucination detection. Currently, there are no such metrics for hallucination assessment in biochemical LLMs (Rawte et al., 2023), which limits the effectiveness of large scientific models in drug discovery. To address this, we propose the first free-form evaluation metric focused on the entailment of scientific entities, enabling more reliable application in this domain.

3 Methodology

In this section, we propose the definition, the source, the Mol-Hallu evaluation metric, and the alleviation strategy for the molecular hallucination phenomenon.

3.1 Definition of Molecular Hallucination

Before delving into the source and evaluation of molecular hallucination, we first define the **Molecular Hallucination** as prediction texts that do not consist of the pharmacological or chemical properties of the molecule. Formally, given the molecule SMILES M and the question Q. The hallucination is that LLM $f_\theta(\cdot)$ outputs non-existent or counterfactual scientific entities E that do not satisfy the reality $\mathbb{T}$, where $\mathbb{T}$ is the ground-truth entity set without any non-existent facts.

3.2 Source of Molecular Hallucination

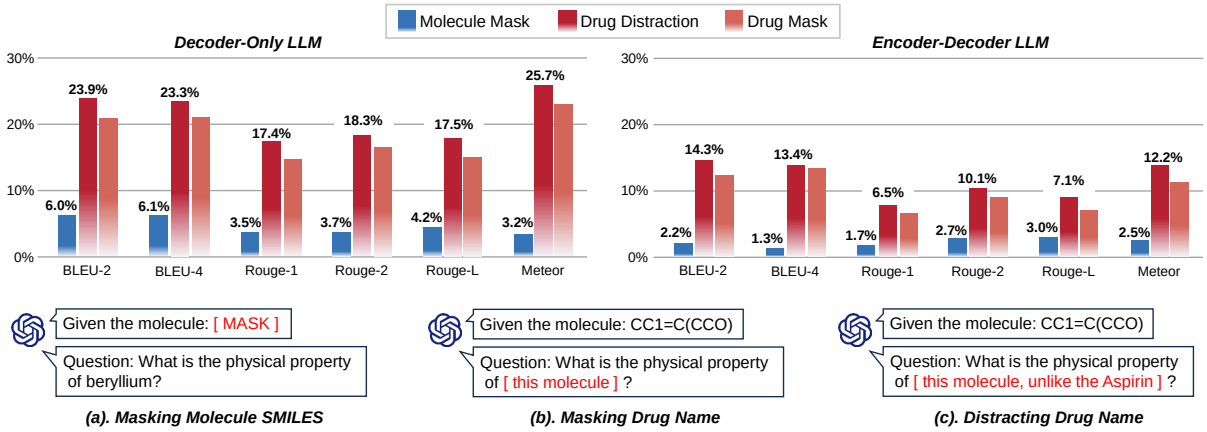
The phenomenon of hallucination in LLMs arises from multiple sources, including inherent divergence and spurious noise within the data (Lee et al., 2022), as well as input knowledge bias (Yin et al., 2023) in training paradigms during training and inference processes.

LLMs exhibit significant hallucinations in molecular comprehension tasks. Upon analyzing the PubChemQA dataset, we identified the **bio-knowledge shortcuts** exacerbate LLM hallucinations.

Molecule: Given a molecule **[SMILES]**.
Question: What is the role of **[Drug Name]** *in cellular processes?*

To be more specific, bio-knowledge shortcuts refer to instances where drug names (e.g., beryllium) are present in molecular-related questions, leading the model to establish mappings between drug names and their physicochemical properties during supervised fine-tuning, rather than between molecular structures from SMILES and physicochemical properties, which is the original intent of molecular comprehension tasks. The existence of such shortcuts makes LLMs prone to hallucination due to changes or the absence of drug names and hinders their ability to infer physicochemical properties for novel molecules.

To prove this, we conduct attacks on the drug names contained in the questions within the molecular question-answer samples from the PubchemQA dataset and analyze the sources of hallucinations by observing the changes in hallucinations corresponding to different attack strategies (Cao et al., 2024a). Specifically, given a sample and its corresponding question Q, we replace



(a). Masking Molecule SMILES *(b). Masking Drug Name* *(c). Distracting Drug Name*

Figure 2: Experiments demonstrate that in both decoder-only LLMs and encoder-decoder LLMs, molecule masking attacking has little impact while drug masking and distracting attackings lead to substantial decrease. This indicates that the knowledge shortcut prompts LLMs to establish alignment between molecular properties and drug names instead of molecular structures, thereby deviating from the goal of molecular comprehension.

the drug name D_j in Q with (1) a masked pronoun $[\,\text{this molecule}\,]$ and (2) a distracting drug name $[\,\text{unlike } \text{D}_\text{j}\,]$. Fig. 2 shows that two classes of commonly used scientific LLMs, the decoder-only models (e.g., Llama (Touvron et al., 2023; Dubey et al., 2024)) and the encoder-decoder models (e.g., T5 (Raffel et al., 2020)), both exhibit severe hallucination phenomena (-21% Acc.) under two attack strategies. However, the absence of SMILES input has little influence on both models (-5% Acc.). This indicates that the models rely more on textual cues (e.g., drug names) than on SMILES structural information to infer molecular properties, highlighting their inability to align SMILES with molecular properties. This limits their generalization and reasoning capabilities for accurate molecular question-answering.

3.3 Mol-Hallu Metric

To better quantify hallucination in LLMs for molecular comprehension tasks, we introduce the **Mol-Hallu** evaluation metric to assess the extent of hallucination. This metric calculates Recall and Precision by comparing the entity entailment probability between the predicted answer A_i, the ground-truth answer G_i, and the molecular description T_i corresponding to the molecule M_i, thereby evaluating the hallucination rate.

3.3.1 Entity Entailment Probability

We define molecular hallucination as the phenomenon of scientific entity mismatches between predicted text and reference answers in Sec. 3.1. To annotate scientific entities in the text, we employed Meta-llama-3.2 (Dubey et al., 2024) with a 10-shot prompting approach to automatically label scientific entities in captions and QA texts from the PubChemQA dataset and the ChEMBL dataset. After filtering based on inclusiveness, length, and semantics, we go through the human evaluation and obtain 97,219 chemical entities as the entity database. The statistic visualization below shows that half of the entities in our entity database are molecular structural entities, while the entities related to drug application, property, and natural source are balanced. Then, we introduce the entity entailment

Type	Application	Property	Source	Structure
Rate	14.3%	19.7%	12.0%	51.2%

probability, defined as the probability that the presence of entity list e is correct given the associated molecular descriptions and answers. Inspired by previous entailment works (Dagan et al., 2005), we find that simple models are effective for entailment probability measurement. Here we apply the probability function as $w(\cdot)$,

$$w(e) = \sum_{j=1}^{n} \mathbf{1}(e_j \in \bar{\mathbb{T}})/n, \qquad (1)$$

where $\mathbf{1}$ is the indicator function, n is the entity number of e, and $\bar{\mathbb{T}}$ represents the set of all the entities present in description T. Then we compute the precision and the recall of the predicted text.

3.3.2 Entailed Precision

The entailed precision aims to represent the correct fraction of the n-gram entities in $mathbbA_i$, where $mathbbA_i$ is the set of all entities in predicted answer A_i. An n-gram entity e is treated as correct

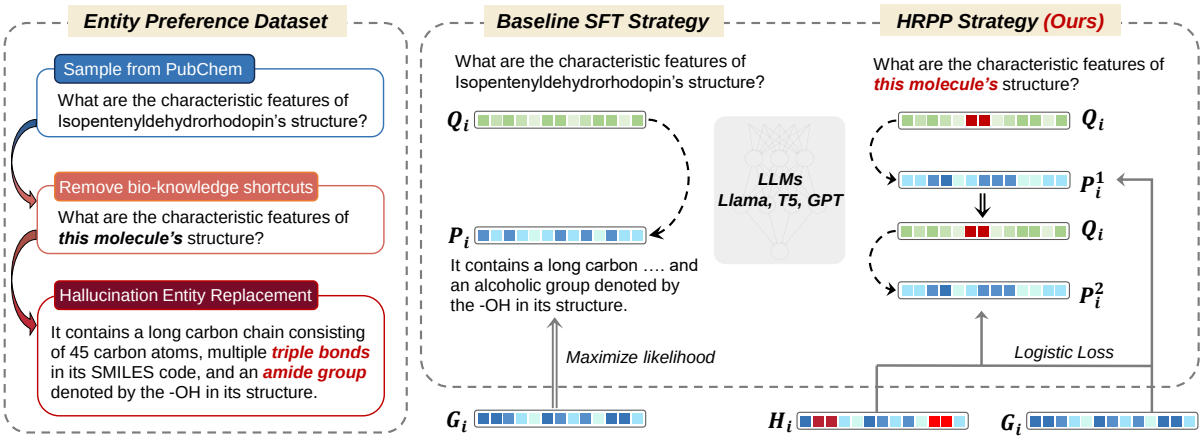


Figure 3: The pipeline of entity preference dataset and our hallucination-reduction post-processing stage. The entity preference dataset is generated by removing bio-knowledge shortcuts and replacing entities with hallucinations. Then we apply the entity preference dataset for scientific-entity hallucination alleviation during the HRPP stage.

if it appears in the ground-truth answer or if it appears in the molecular description, which is also a substantial correct answer. We apply $w(e)$ as the reward weight of the second scenario.

$$P_e^{\text{n-gram}} = \sum_{e \in \mathbb{A}_i} [\Pr(e \in \mathbb{G}^{\text{n-gram}}) + w(e)\Pr(e \notin \mathbb{G}^{\text{n-gram}})], \tag{2}$$

Specifically, $P_e^{\text{n-gram}}$ represents the reward of the n-gram entity e. It receives a score of 1 if the ground-truth answer entails it. Otherwise, it receives a score of $w(e)$ if e appears in the molecular description. We consider the numerator during the weight calculation of $P_e^{\text{n-gram}}$. Finally, we apply the geometric average to calculate the precision of the total sample group,

$$\bar{P}_e = \exp\Big(\sum_{\text{n-gram}=1}^{4} \frac{1}{4} \log P_e^{\text{n-gram}}\Big), \tag{3}$$

where we select the n-gram order from 1-4 as other metrics (Papineni et al., 2002b; Post, 2018; Dhingra et al., 2019). Meanwhile, we calculate the n-gram matching score $\bar{P}_\varnothing$ for non-entity words. To balance the precision $\bar{P}_e$ from scientific entities and $\bar{P}_\varnothing$ from non-entities, we use the entity error count γ as a weighting factor,

$$\gamma = 1 - (N_{\text{wrong}}/N_{\text{total}})^{0.5}, \tag{4}$$

$$\text{P} = \gamma \bar{P}_\varnothing + (1-\gamma)\bar{P}_e, \tag{5}$$

where N_{wrong} and N_{total} are wrong entity and total entity counts. P represents the final precision score.

3.3.3 Entailed Recall

The entailed recall R reflects the extent to which the model misses correct words. R is computed between predicted A and ground truth G to ensure that entities and other n-gram words with high frequency in the ground truth receive a higher score when predicted correctly. We also apply the geometric average to get R from $\text{R}_{1\ldots n}$.

3.3.4 Smoothing & Combination

Mol-Hallu employs the geometric average to compute entailed precision due to its ability to reflect compound changes accurately. However, when a component approaches 0, the geometric average also tends to 0. To mitigate this issue, we apply smoothing θ=10^{-5} to components close to 0. After the precision smoothing, we calculate the F1-score based on the entailed precision P and recall R.

$$\text{Mol-Hallu}(A, G, T) = 2\text{P} \cdot \text{R}/(\text{P} + \text{R}), \tag{6}$$

$$\text{Mol-Hallu}(f_\theta) = \frac{1}{N}\sum_{i=1}^{N} \text{Mol-Hallu}(A_i, G_i, T_i), \tag{7}$$

where the F1-scores from all samples generated by the model f_θ are arithmetic averaged to represent the hallucination rate of f_θ.

3.4 Hallucination Reduction Post-processing

To mitigate the hallucination in LLM-based molecular comprehension, we propose the Hallucination Reduction Post-processing (HRPP) stage. As shown in Fig. 3, HRPP consists of two main steps: (1) reducing the model's reliance on entity name shortcuts through supervised fine-tuning, and (2) improving response accuracy and reducing hallucination using Direct Preference Optimization (DPO) with a hallucination-sensitive preference dataset.

To mitigate the model's tendency to generate hallucinated responses due to over-reliance on

Models	# Params	BLEU-2	BLEU-4	ROUGE-1	ROUGE-L	METEOR	Mol-Hallu↑
Molecular-LLMs							
MolT5-small	80M	49.46	41.94	55.04	51.56	55.40	59.01
MolT5-base	250M	50.21	42.53	**55.70**	**52.07**	56.00	44.74
MolT5-large	800M	49.58	41.97	55.52	51.85	55.80	60.13
MoMu-small	82M	50.81	42.54	52.78	51.18	55.94	55.73
MoMu-base	252M	51.07	43.29	53.71	50.98	55.59	**56.29**
BioT5-base	252M	43.36	35.10	51.05	47.16	51.55	55.21
MolCA	1.3B	**51.93**	**44.28**	55.00	51.41	**56.79**	55.82
3D-MoLM	7B	32.00	26.17	40.13	34.64	52.15	53.18
BioMedGPT	10B	37.31	31.29	39.62	36.87	48.31	43.88
General-LLMs							
T5-small	60M	49.97	42.40	54.88	51.16	55.47	59.07
T5-base	220M	51.01	43.27	55.89	52.17	**56.43**	60.21
T5-large	770M	50.79	42.85	**55.98**	**52.23**	56.42	**60.93**
Llama-2	7B	28.15	23.24	35.14	30.41	46.87	53.78
Llama-3.1	8B	**52.19**	**43.51**	55.41	51.18	57.48	60.14
Universal-LLM-API (Few-shot)							
Qwen-2.5-Instruct	32B	35.72	27.51	43.59	38.22	49.63	49.97
Qwen-Reason (QwQ)	32B	18.62	13.62	27.33	23.32	35.14	25.61
DeepSeek-V3	671B	**49.31**	39.86	**53.96**	**48.37**	**57.69**	**62.16**
DeepSeek-R1	671B	32.12	24.17	41.77	37.56	40.65	46.65
GPT-4o-20241120	1.8T	47.78	**41.74**	51.97	46.99	51.24	55.71
o1-mini	300B	40.22	31.06	46.99	41.81	51.88	51.23

Table 1: Experimental results for hallucination evaluation across molecular LLMs (fine-tuned), general LLMs (fine-tuned), and universal LLMs (API-based inference). We report accuracy (%) using both standard textual metrics and our proposed hallucination-specific evaluation metric.

entity name shortcuts, we employ a supervised fine-tuning approach. Given a training dataset $\mathcal{D} = \{(q_i, G_i)\}_{i=1}^{N}$, where Q_i is the input text and G_i is the corresponding ground truth response, we preprocess Q_i by masking entity names, replacing them with "this molecule" to prevent shortcut learning. We then optimize the model parameters θ by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log P_\theta(G_i^t \mid Q_i, G_i^{<t}) \quad (8)$$

where T is the sequence length, N is the sample number, and P_θ represents the model's probability distribution over the vocabulary.

To further improve response accuracy and factual consistency of molecular LLMs, we first construct a hallucination-sensitive preference dataset $\mathcal{D}_p = \{(q_i, G_i^+, G_i^-)\}_{i=1}^{M}$, where G_i^+ represents the preferred response, and G_i^- represents the less preferred response. As shown in Fig. 3 left, to construct this dataset, we randomly extract 2000 QA pairs from the training set. The ground truth G_i is designated as G_i^+. To generate the negative sample G_i^-, we introduce entity perturbations by randomly replacing certain entities in G_i with different ones using our chemical entity database. Additionally, we sample four responses from the model at a high temperature for each q_i, incorporating them into the set of G_i^- responses.

We use DPO to optimize the model by maximizing the divergence between the likelihood of preferred and rejected responses:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{M} \log \sigma \left(\beta \log \frac{P_\theta(G_i^+|q_i) P_{\text{r}}(G_i^-|q_i)}{P_\theta(G_i^-|q_i) P_{\text{r}}(G_i^+|q_i)} \right) \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function, P_{r} is the reference model, and β is a temperature hyperparameter that controls the strength of preference learning. In the experiment section, we apply HRPP to decoder-only LLMs and encoder-decoder LLMs for effectiveness analysis.

4 Experiments

4.1 Baseline Models and Training Procedures

To comprehensively evaluate the LLM performance in molecular conprehension, we introduce three categories of LLMs as baselines, including scientifically fine-tuned LLMs, general-purpose

Molecular LLMs	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Mol-Hallu↑
MolT5	34.48	26.54	45.13	28.17	41.34	37.08	46.15
+ HRPP	40.65	30.73	47.47	29.98	43.54	44.31	**49.03**
Llama-3.1-8B	33.18	24.75	44.19	27.12	40.66	37.57	44.21
+ HRPP	38.79	28.95	46.12	28.41	42.17	43.27	**46.28**

Table 2: Hallucination Reduction Post-processing (HRPP) has substantial improvements in textural metrics and our Mol-Hallu metric, demonstrating its effectiveness on both decoder-only models and encoder-decoder-based models.

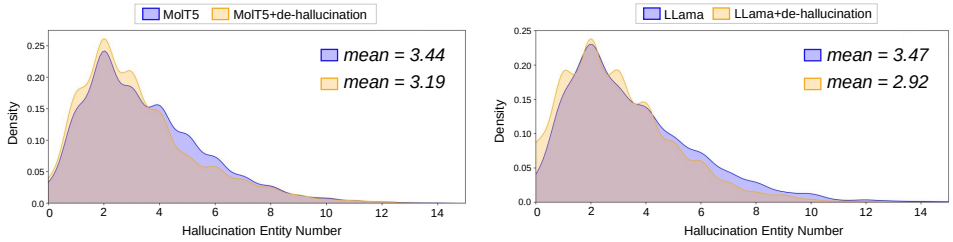


Figure 4: **Hallucination Distribution Comparison**. We visualize the distributions of hallucination entity numbers between molecular LLMs (MolT5, Llama-3.1) and their de-hallucination versions. Our HRPP effectively mitigates the frequent occurrence of hallucinations in cases, shifting the distribution peak closer to 0.

LLMs, and commercial LLMs. Specifically, LLMs fine-tuned with biochemical knowledge exhibit strong capabilities in modeling molecular SMILES and protein sequences. We evaluate their hallucination levels on the PubChemQA dataset in a zero-shot manner. General-purpose LLMs, trained extensively in natural scenarios, although less adept at modeling molecular SMILES compared to scientifically fine-tuned LLMs, possess stronger reasoning abilities. Commercial LLMs have stronger prior knowledge and reasoning capabilities due to their large parameter sizes. We conduct paid evaluations using the APIs of commercial LLMs, employing 10-shot instruction fine-tuning to generate responses to molecular-related queries.

4.2 Main Results

We summarize and analyze the baseline performances in Table.1.

Hallucinations in baseline models. (1) The hallucination metric remains within the range of 40-60%, with an average of 3-4 counterfactual entities present, indicating significant room for improvement. (2) The degree of hallucination is not necessarily positively correlated with model performance. While MolT5-base shows comparable performance to MolT5-small and MolT5-large, its hallucination is notably more severe. In contrast, 3D-MoLM exhibits moderate performance but demonstrates a lower degree of hallucination.

Structure Comparison: Encoder-Decoder v.s. Decoder-only. Encoder-decoder models surpass other structures in molecular comprehension tasks due to their compact size and excellent performance. We observe that T5-based models, represented by T5-finetune, MolT5, and MoMu, exhibit strong performance on the MolecularQA task even in their small versions, surpassing molecular LLMs based on Llama by 2.7% and GPT-4 by 13%. This is attributed to the T5 model's encoder-decoder structure, which employs a span corruption pre-training strategy. Additionally, its smaller parameter count supports full-parameter fine-tuning instead of the LoRA fine-tuning used in Llama, resulting in better generalization in few-shot scenarios within the biochemistry domain.

Reward strategies in LLMs amplify hallucination. Deepseek-R1 and o1-mini have widely adopted reinforcement learning as an effective approach to enhance the LLM reasoning capabilities for complex problems. However, this optimization strategy often leads to a hallucination increase (Forrest et al., 2025). We observe a similar phenomenon in Molecular Comprehension. In the LLM-API part of Table.1, we compare Qwen, Deepseek, and GPT-4, with their reasoning-enhanced versions on scientific QA tasks. The results indicate a significant decline in both prediction quality and factual accuracy, attributed to: (1) the trade-off between improved reasoning in math/code tasks and the reduced reliance on prior knowledge, making it harder to address scientific questions; and (2) the tendency of reasoning-enhanced LLMs to generate chain-of-thought outputs, which often contain more hallucinated entities. Therefore, balancing reasoning and hallucination in domain-specific sce-

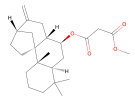
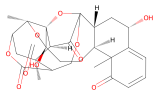
Molecule	Query-Type	Ground truth	Our answer	Metric
	Isolated Area	This compound is isolated from the plants Sorbus cuspidata and Calceolaria dentata.	Hexaen is isolated from the plants pentahydroxy and benzoate.	B: 78.9% R: 86.4% M: 87.9% M-H: 43.3%
	Potential Reactivity	This compound has potential reactivity towards nucleophiles and bases due to the presence of ketone and lactone groups.	This compound has potential reactivity towards aromaticity and methoxy due to the presence of solubility and reactivity groups.	B: 92.2% R: 93.3% M: 93.9% M-H: 66.1%

Table 3: **Case Studies for Mol-Hallu and Other Textural Metrics**. Our Mol-Hallu exhibits stronger sensitivity to hallucinated outputs under different question types in molecule comprehension.

narios remains a critical challenge.

Extra protein knowledge: no benefit to hallucination. During pretraining, extending the dataset to include both chemical molecules and protein macromolecules cannot alleviate hallucination. Instead, it leads to a decrease in performance for molecular understanding tasks. In Table 1, BioMedGPT (Luo et al., 2023) and BioT5 utilize various protein dataset size (1.8M, 27M) as additional knowledge. However, their performance and hallucination assessment are inferior to the MolT5-based model due to the structural differences between FASTA-based protein inputs and SMILES-based molecular inputs, as well as the significant domain-specific entity differences between proteins and chemical molecules. Consequently, the incorporation of such knowledge fails to enhance generalization or reduce hallucination.

4.3 Analysis for Hallucination Reduction

In Table. 2 and Fig. 4, we dive into the hallucination reduction post-processing (HRPP) and analyze its effectiveness on hallucination alleviation.

Effectiveness of HRPP Stage. Our HRPP stage shows effectiveness and generalizability on both decoder-only and T5-based models. Table. 2 shows that HRPP has substantial improvements for molecular LLMs, bringing an average of 4.0% improvements on textural metrics. For the hallucination evaluation, our HRPP stage also achieves effective hallucination alleviation on both decoder-only structure (2.9% $\uparrow$) and T5-based structure (2.0% $\uparrow$). Meanwhile, we observed a significant improvement in the BLEU and METEOR (5-7%) during the HRPP stage, while the ROUGE series improvement is less pronounced (1-2%). This indicates that molecular LLMs optimized through HRPP tend to generate text with higher precision in scientific entities and more accurate semantics. However, missing scientific entities still occur in some answers due to the ROUGE series metrics being more sensitive to recall.

Hallucination Distribution Analysis. To analyze the impact of HRPP on hallucinated samples generated by LLMs, we visualize the change in the number of counterfactual entities N_c before and after the HRPP stage. In Fig. 4, HRPP effectively suppresses highly hallucinated samples ($N_c > 4$) in both decoder-only and encoder-decoder LLMs. After the HRPP stage, the distribution of counterfactual entities significantly shifts toward the low-hallucination region ($0 < N_c < 3$), demonstrating the efficacy of the HRPP stage.

4.4 Case Studies

We select samples with hallucinations and demonstrate a numerical comparison between our Mol-Hallu metric and traditional textual metrics. Table. 3 shows that Mol-Hallu are more sensitive to hallucinations. When the prediction and ground truth share similar sentence structures but differ in scientific entities, Mol-Hallu assigns a lower score, whereas traditional evaluation methods consider them semantically similar. Additional case studies are proposed in the Appendix.A1.

5 Conclusion and Future Work

In conclusion, our work aims to evaluate and alleviate the LLM's hallucination in molecular comprehension. By attacking the scientific entities in molecule-related questions, we identify the bio-knowledge shortcuts in the PubChem dataset as the hallucination source of the molecular comprehension task. We further propose the hallucination evaluation metric, Mol-Hallu, for molecular comprehension. To alleviate the hallucination, we propose the hallucination reduction post-processing strategy with a molecular hallucination-sensitive preference dataset constructed based on entity replacement. Experimental results demonstrate that various LLM architectures significantly suppressed hallucinations with this strategy.

Limitations

We conclude our limitations into the following aspects: (1) Our Mol-Hallu metric relies on a scientific entity database to localize scientific entities in predicted texts and evaluate the degree of hallucination. Although the current entity database demonstrates excellent coverage in the small molecule domain, its coverage in other scientific fields, such as protein understanding, remains limited. Future work should incorporate domain-specific terminologies to construct a more comprehensive entity database. (2) The current benchmark lacks full fine-tuning of large models due to insufficient training resources. Future efforts will focus on fine-tuning LLMs with 7B+ parameters and exploring the relationship between the performance and hallucination levels of molecular LLMs under scaling laws.

Potential Risks

Although Mol-Hallu provides a viable metric for hallucination assessment in the molecular comprehension domain, there remains a risk of abuse. Mol-Hallu evaluation may not accurately represent a model's hallucination level over all chemistry-related scenarios.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2025. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 354–379, Abu Dhabi, UAE. Association for Computational Linguistics.

He Cao, Weidi Luo, Yu Wang, Zijing Liu, Bing Feng, Yuan Yao, and Yu Li. 2024a. Guide for defense (g4d): Dynamic guidance for robust and balanced defense in large language models. *arXiv preprint arXiv:2410.17922*.

He Cao, Yanjun Shao, Zhiyuan Liu, Zijing Liu, Xiangru Tang, Yuan Yao, and Yu Li. 2024b. PRESTO: Progressive pretraining enhances synthetic chemistry outcomes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10197–10224, Miami, Florida, USA. Association for Computational Linguistics.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

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

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srbuhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. 2024. A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence*, 6(8):962–974.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. 2022. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293.

Bao Forrest, Xu Chenyu, and Mendelevitch Ofer. 2025. Deepseek-r1 hallucinates more than deepseek-v3.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Francesca Grisoni. 2023. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chengxin Hu, Hao Li, Yihe Yuan, Zezheng Song, and Haixin Wang. 2025. Omni-mol: Exploring universal convergent space for omni-molecular tasks. *Preprint*, arXiv:2502.01074.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Siddhartha Laghuvarapu, Namkyeong Lee, Chufan Gao, and Jimeng Sun. 2024. Moltextqa: A curated question-answering dataset and benchmark for molecular structure-text relationship learning. *OpenReview*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8424–8445.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024a. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024b. Can multiple-choice questions really be useful in detecting the abilities of llms? *arXiv preprint arXiv:2403.17752*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023a. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. 2024a. Data-driven quantum chemical property prediction leveraging 3d conformations with uni-mol+. *Nature Communications*, 15(1):7104.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024b. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. *arXiv preprint arXiv:2403.08192*.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Liuzhenghao Lv, Hao Li, Yu Wang, Zhiyuan Yan, Zijun Chen, Zongying Lin, Li Yuan, and Yonghong Tian. 2024. Navigating chemical-linguistic sharing space with heterogeneous molecular encoding. *arXiv preprint arXiv:2412.20888*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. 2019. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Qwen Team. 2024a. Qwen2.5: A party of foundation models.

Qwen Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown.

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

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. 2009. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl_2):W623–W633.

Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. 2023. Fake alignment: Are llms really aligned well? *arXiv preprint arXiv:2311.05915*.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024a. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *EMNLP'24*, pages 8783–8817.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Case Studies for PubchemQA Dataset

We systematically enumerated samples with varying degrees of hallucination from the PubchemQA dataset and compared the scores of traditional metrics (BLEU-2/4, ROUGE-1/2/L, and METEOR) with those of Mol-Hallu. Fig. 4 provides 7 samples from PubchemQA, where Q-Type represents the question type of the sample, $B, R, M, M\text{-}H$ in Metric represents the average of BLEU-2/4, the average of Rouge-1/2/L, Meteor, and our Mol-Hallu metric. The experiment results in Fig. 4 covered diverse molecular structures and question types, demonstrating that Mol-Hallu accurately reflects the hallucination degree across different scenarios, exhibiting robust performance and domain adaptability. Notably, in the second case, where the model's prediction completely deviated from the ground truth, Mol-Hallu assigned a low score of 1.6%, while traditional metrics, misled by superficial sentence similarities, provided significantly higher scores (83.8%, 87.5%, 91.5%). This contrast not only highlights the inherent limitations of traditional metrics in evaluating hallucinations in biochemical texts but also further validates the reliability and superiority of Mol-Hallu in detecting semantic errors in scientific entities.

A.2 The Evaluation Introduction

In this subsection, we provide the detailed information for traditional textural evaluation metrics for LLM prediction in Question-Answering tasks.

BLEU: (Bilingual Evaluation Understudy) is a precision-based metric widely used for evaluating the quality of machine-generated text by comparing it to one or more reference texts. It measures the overlap of n-grams (typically up to 4-grams) between the generated text and the references. The BLEU score is calculated as follows:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{10}$$

where BP is the brevity penalty to penalize short translations, w_n is the weight for each n-gram precision p_n, and N is the maximum n-gram order (usually 4).

ROUGE: (Recall-Oriented Understudy for Gisting Evaluation) is a recall-oriented metric commonly used for evaluating summarization tasks. It measures the overlap of n-grams, word sequences, or word pairs between the generated text and the reference texts. The most frequently used variant, ROUGE-N, is defined as:

$$\text{ROUGE-N} = \frac{\sum_{\mathcal{R}} \sum_{\text{n-gram} \in \mathcal{R}} \text{C}_{\text{match}}(\text{n-gram})}{\sum_{\mathcal{R}} \sum_{\text{n-gram} \in \mathcal{R}} \text{C}(\text{n-gram})} \tag{11}$$

where $\text{C}_{\text{match}}(\text{n-gram})$ is the number of n-grams co-occurring in both the generated and reference texts $\mathcal{R}$, and $\text{C}(\text{n-gram})$ is the total number of n-grams in the reference.

METEOR: (Metric for Evaluation of Translation with Explicit ORdering) is a metric designed to address some limitations of BLEU by incorporating synonymy, stemming, and word order. It calculates a weighted harmonic mean of precision and recall, with a penalty for word order discrepancies. The METEOR score is computed as:

$$\text{METEOR} = (1 - \gamma \cdot \text{Penalty}) \cdot \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \tag{12}$$

where P and R are precision and recall, respectively, γ is a parameter controlling the penalty weight, and Penalty is a function of the number of word order violations.

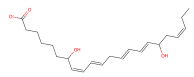
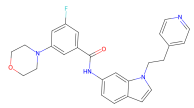
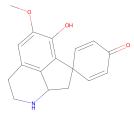
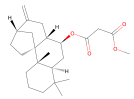
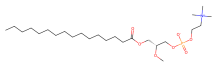
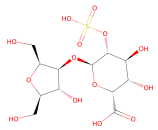
Molecule	Q-Type	Ground truth	Our answer	Metric
	Structure	It has a docosanoid structure with a 22-carbon chain and unsaturated bonds at positions 8, 15, and 19. It also contains hydroxyl groups at positions 7 and 17.	It has a docosanoid structure with a 22-carbon chain and sedative at disorders 8, 15, and 19. It also contains appetite at disorders 7.	B: 88.4% R: 87.5% M: 91.5% M-H: 66.5%
	Class	This organic compound belongs to the class of benzamides.	This organic compound belongs to the class of carboxylic acid.	B: 83.8% R: 82.6% M: 85.7% M-H: 1.6%
	Solubility Property	This molecule has solubility in both polar and nonpolar solvents due to the presence of a hydroxy group (-OH) and a methoxy group (-OCH3).	This molecule has shaking in both polar and insomnia due to the presence of a hallucinations (-OH) and a seizures (-OCH3).	B: 88.3% R: 87.9% M: 90.9% M-H: 63.9%
	Isolated Area	This compound is isolated from the plants Sorbus cuspidata and Calceolaria dentata.	Hexaen is isolated from the plants pentahydroxy and benzoate.	B: 78.9% R: 86.4% M: 87.9% M-H: 43.3%
	Potential Reactivity	This compound has potential reactivity towards nucleophiles and bases due to the presence of ketone and lactone groups.	This compound has potential reactivity towards aromaticity and methoxy due to the presence of solubility and reactivity groups.	B: 92.2% R: 93.3% M: 93.9% M-H: 66.1%
	Structure	The molecule has a glycerol backbone with a hexadecanoyl group attached to the sn-1 position and a methyl group attached to the sn-2 position. It also has a phosphate group and a choline molecule attached to the sn-3 position.	The molecule has a glycerol backbone with a hexadecanoyl group attached to the sn-1 position and a methyl group attached to the PbSO4 position. It also has a zinc group and a silver molecule attached to the copper position.	B: 79.6% R: 87.8% M: 84.1% M-H: 67.9%
	Chemical Classify	The compound is classified as a carbohydrate acid derivative, meaning it is a derivative of a carboxylic acid that contains a carbohydrate moiety. It is also categorized as an oligosaccharide sulfate, indicating it is a sulfated oligosaccharide with multiple sugar units and sulfate groups.	The compound is classified as a carbohydrate acid postganglionic, meaning it is a postganglionic of a effector-cell acid that contains a carbohydrate moiety. It is also categorized as a receptor, indicating it is a sulfated oligosaccharide with multiple muscle and sulfate bronchoconstriction.	B: 78.1% R: 86.2% M: 85.2% M-H: 65.5%

Table 4: Additional case studies for Mol-Hallu and other textural metrics. Our Mol-Hallu exhibits stronger sensitivity to hallucinated outputs under different question types in molecule comprehension.

A.3 Licenses and Terms of Use for Models and Datasets

In this study, we employed multiple models and datasets, each subject to distinct licensing terms. The following is a summary of these licenses along with their respective usage conditions.

MolT5: Released by blender-nlp under the BSD 3-Clause License. This license permits free use, modification, and distribution, provided that specific conditions are met, such as retaining the copyright notice and disclaimer. Commercial use is allowed, but endorsement or promotion of derived products using the copyright holder's name requires prior written permission. The license also includes a liability disclaimer, stating that the software is provided "as is" without warranties or guarantees.

MoMu: Released under the MIT License. This license permits free use, modification, and distribution, including for commercial purposes, as long as the original copyright notice and permission notice are retained. The software is provided "as is," without any warranties or guarantees, and the authors bear no liability for any claims, damages, or other issues arising from its use.

BioT5: Released under the MIT License. This license permits free use, modification, and distribution, including for commercial purposes, as long as the original copyright notice and permission notice are retained. The software is provided "as is," with-

out any warranties or guarantees, and the authors bear no liability for any claims, damages, or other issues arising from its use.

3D-MoLM: Released under the Apache 2.0 License. This license permits free use, modification, and distribution, including for commercial purposes, provided that the original copyright notice and license terms are retained. Users are allowed to patent their modifications but must grant a license for any patented contributions. The software is provided "as is," without warranties or liabilities, and users must include a notice stating any modifications made to the original version.

BioMedGPT: Released under the MIT License. This license permits free use, modification, and distribution, including for commercial purposes, as long as the original copyright notice and permission notice are retained. The software is provided "as is," without any warranties or guarantees, and the authors bear no liability for any claims, damages, or other issues arising from its use.

T5: Released under the Apache 2.0 License. This license permits free use, modification, and distribution, including for commercial purposes, provided that the original copyright notice and license terms are retained. Users are allowed to patent their modifications but must grant a license for any patented contributions. The software is provided "as is," without warranties or liabilities, and users must include a notice stating any modifications made to the original version.

Llama-2: Released by Meta under the Llama 2 Community License. This license permits free use, modification, and distribution, but restricts the model's use for training other language models and imposes specific conditions for commercial use, such as active user limits.

Llama-3.1: Released by Meta under the Llama 3.1 Community License. This license permits free use, modification, and distribution, with requirements such as attribution, compliance with Meta's Acceptable Use Policy, and display of "Built with Llama" for derivative works. Commercial use is allowed, but entities with over 700 million monthly active users must obtain a separate license from Meta. The license includes disclaimers of warranty and liability, and any legal disputes fall under the jurisdiction of California law.

Qwen-2.5-Instruct (Team, 2024a): Released under the Apache 2.0 License. This license permits free use, modification, and distribution, including for commercial purposes, provided that the original copyright notice and license terms are retained. Users are allowed to patent their modifications but must grant a license for any patented contributions. The software is provided "as is," without warranties or liabilities, and users must include a notice stating any modifications made to the original version.

Qwen-Reason (QwQ) (Team, 2024b): Released under the Apache 2.0 License. This license permits free use, modification, and distribution, including for commercial purposes, provided that the original copyright notice and license terms are retained. Users are allowed to patent their modifications but must grant a license for any patented contributions. The software is provided "as is," without warranties or liabilities, and users must include a notice stating any modifications made to the original version.

DeepSeek-V3 (Liu et al., 2024): Released by DeepSeek under the DeepSeek License (v1.0, Oct 23, 2023). It grants a free, global, irrevocable license for using, modifying, and distributing DeepSeek-V3, with strict restrictions on military use, harm, misinformation, discrimination, and unauthorized data processing. Users must enforce these limits in derivatives. DeepSeek may restrict misuse remotely and disclaims warranties and liability. Governed by Chinese law (PRC), jurisdiction in Hangzhou.

DeepSeek-R1 (Guo et al., 2025): Released under the MIT License. This license permits free use, modification, and distribution, including for commercial purposes, as long as the original copyright notice and permission notice are retained. The software is provided "as is," without any warranties or guarantees, and the authors bear no liability for any claims, damages, or other issues arising from its use.

GPT-4o-20241120: Released by OpenAI. It is proprietary software. Access to this model is provided through OpenAI's platforms, such as ChatGPT and the Azure OpenAI Service, under specific subscription plans. The model is not open-source and is subject to OpenAI's terms of service and usage policies.

o1-mini: Released by OpenAI. It is proprietary software. Access to o1-mini is provided through OpenAI's API and platforms, such as ChatGPT, under specific subscription plans. The model is not open-source and is subject to OpenAI's terms of service and usage policies.

PubChemQA (3D-MoIT): Released under the Apache 2.0 License. This license permits free use, modification, and distribution, including for com-

mercial purposes, provided that the original copyright notice and license terms are retained. Users are allowed to patent their modifications but must grant a license for any patented contributions. The software is provided "as is," without warranties or liabilities, and users must include a notice stating any modifications made to the original version.

ChEMBL: Released under the Creative Commons Attribution-ShareAlike 3.0 Unported License. This license allows free use, modification, and distribution of the dataset, but requires appropriate attribution and mandates that any derivative works or modifications must be distributed under the same license.