# Can Hallucinations Help? Boosting LLMs for Drug Discovery

**Anonymous ACL submission**


## Abstract

Hallucinations in large language models (LLMs), plausible but factually inaccurate text, are often viewed as undesirable. However, recent work suggests that such outputs may hold creative potential. In this paper, we investigate whether hallucinations can improve LLMs on molecule property prediction, a key task in early-stage drug discovery. We prompt LLMs to generate natural language descriptions from molecular SMILES strings and incorporate these often hallucinated descriptions into downstream classification tasks. Evaluating seven instruction-tuned LLMs across five datasets, we find that hallucinations significantly improve predictive accuracy for some models. Notably, `Falcon3-Mamba-7B` outperforms all baselines when hallucinated text is included, while hallucinations generated by `GPT-4o` consistently yield the greatest gains between models. We further identify and categorize over 18,000 beneficial hallucinations, with *structural misdescriptions* emerging as the most impactful type, suggesting that hallucinated statements about molecular structure may increase model confidence. Ablation studies show that larger models benefit more from hallucinations, while temperature has a limited effect. Our findings challenge conventional views of hallucination as purely problematic and suggest new directions for leveraging hallucinations as a useful signal in scientific modeling tasks like drug discovery.


## 1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across diverse domains, including everyday applications (Ge et al., 2023; Yao et al., 2024), scientific discovery (Zhang et al., 2024b; Madani et al., 2023), and chemistry (Boiko et al., 2023; Zhang et al., 2024a). However, LLMs are known to produce hallucinations, text that is plausible but factually incorrect or misaligned with the input (Huang et al., 2023; Bang et al., 2023; Rawte et al., 2023; Yuan and Faerber, 2023). While much work has focused on mitigating hallucinations (Ji et al., 2023; Dhuliawala et al., 2024), recent perspectives suggest that hallucination may also be beneficial, for example, it may enable creativity (Lee, 2023; Ye et al., 2023; Wang, 2024).

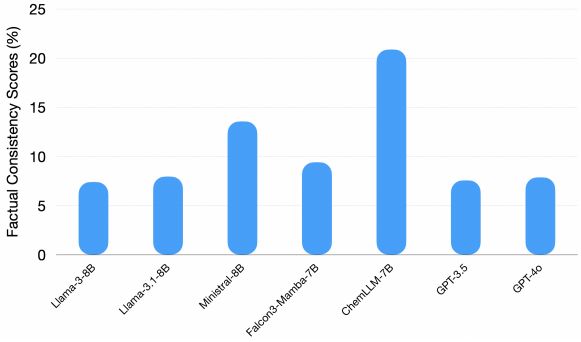


Figure 1: Factual consistency scores evaluated using the HHEM-2.1-Open Model, measuring the degree of hallucination relative to `MolT5` reference descriptions. Scores are averaged across five datasets. Except for `ChemLLM`, all LLMs exhibit low factual consistency, indicating extensive hallucination.

In domains like drug discovery, creativity is a vital asset. Evaluating molecular properties for potential drug candidates involves reasoning over vast chemical spaces and abstract functional traits (Hughes et al., 2011). LLMs are beginning to assist in this process (Chakraborty et al., 2023; Edwards et al., 2024), and textual molecule descriptions have been shown to help improve generalization (Liu et al., 2023). Yet, many LLM-generated descriptions contain hallucinations, as illustrated in Figure 1, where models deviate significantly from domain-specific references such as `MolT5` (Edwards et al., 2022).

This raises an underexplored question: *Can hallucinations improve LLMs' performance in molecule property prediction?* While hallucinations are often treated as failure cases, we hypothesize that they may act as implicit counterfactuals, offering speculative, high-level interpretations

that help LLMs generalize over unseen compounds. This idea is particularly relevant in early-stage drug discovery, where structured knowledge is sparse and creativity may be an advantage.

To investigate this, we conduct a comprehensive evaluation of seven instruction-tuned LLMs across five molecular property prediction tasks from MoleculeNet (Wu et al., 2018). Given a molecule's SMILES string (Weininger, 1988), we generate natural language descriptions using each LLM, then include the description in the prompt alongside a task-specific instruction (Figure 2). We compare the performance against three baselines:

- `SMILES`: using only the molecule string.
- `MolT5`: domain-specific reference descriptions.
- `PubChem`: rule-based factual baseline derived from authoritative chemical metadata (Kim et al., 2025).

Our results show that LLM-generated hallucinations can enhance performance. For example, `Falcon3-Mamba-7B` surpasses the `PubChem` baseline by 8.22% in ROC-AUC. `Llama-3.1-8B` achieves 15.8% and 11.2% improvements over the `SMILES` and `MolT5` baselines, respectively. Hallucinations from `GPT-4o` deliver the most consistent gains across models.

To understand the nature of these helpful hallucinations, we collect instances that lead to improved predictions and categorize them into four types: *structural misdescription*, *functional hallucination*, *analogical hallucination*, and *generic fluff*. The dominant type is *structural misdescription*, suggesting that factual deviation is not necessarily detrimental, and may help LLMs generalize. Our counterfactual analysis further reveals that hallucinations introducing coherent but incorrect structural or relational information provide the largest performance gains, whereas generic or high-level functional descriptions yield more modest improvements. These results indicate that the benefit of hallucinations is not merely due to additional text or general biochemical claims, but rather stems from structured semantic perturbations that encourage the model to attend to property-relevant patterns in molecular representations.

**Our contributions are as follows:**

- We conduct the first systematic investigation into whether hallucinations can improve LLMs for molecular property prediction, a core subtask in early-stage drug discovery.
- We evaluate seven LLMs across five benchmark datasets, demonstrating that certain models benefit from hallucinated input.
- We analyze model size, generation temperature[1], and hallucination types to understand how and when hallucinations contribute positively to performance.

## 2 Related Work

**LLMs for Molecules** Large language models have shown promising potential in molecular tasks and biomedical domains (Pal et al., 2023; Murakumo et al., 2023; Chakraborty et al., 2023; Savage, 2023). Zheng et al. (2024) provide a comprehensive review of LLMs across the drug discovery pipeline, while Guan and Wang (2024) discuss their advantages in drug development. Several works evaluate LLMs on molecular property prediction (Zhong et al., 2024), or apply LLMs to specialized tasks such as dataset curation (Liu et al., 2024) and multi-agent drug discovery frameworks (Liu et al., 2024). Natural language explanations have also been explored for cancer drug discovery via prompt-based methods (Zhang et al., 2024c). The connection between molecular representations such as SMILES strings and language has been widely investigated in recent works (Edwards et al., 2022; Ganeeva et al., 2024), while LLMs have increasingly been used as general-purpose agents in the domain (Fossi et al., 2024). Building on this foundation, our work takes a novel step by examining how LLM-generated hallucinations can be leveraged to improve molecular property prediction. Rather than treating hallucinations as noise to eliminate, we explore their potential as creative and informative inputs in early-stage drug discovery.

**Hallucination and Creativity in LLMs** Hallucination is typically defined as a mismatch between the generated content and a source or ground truth (Maynez et al., 2020; Rawte et al., 2023), and numerous approaches aim to detect or reduce hallucination (Dhuliawala et al., 2024; Manakul et al., 2023; Ji et al., 2023). However, a growing body of work explores the creative potential of hallucinations (Lee, 2022; Zhao et al., 2024). Jiang et al.

[1] Ablation studies of model size and generation temparature are presented in Appendix I.

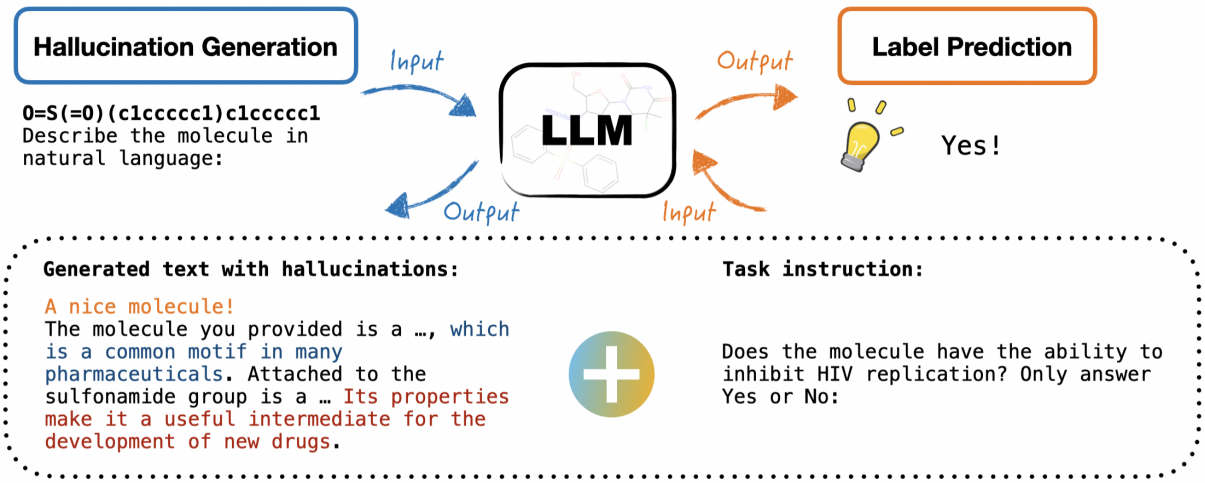


Figure 2: Illustration of our method using a sample from the HIV dataset. We first generate hallucinated molecule descriptions from LLMs, then include them as part of the prompt for a binary classification task. Output is constrained to "Yes" or "No." Different types of hallucinations are color-coded.

(2024) argue that hallucinations in LLMs parallel human creativity by enabling the recombination of known information in novel ways. Similarly, Lee (2023) model hallucination as a probabilistic driver of surprising or insightful outputs. Other studies evaluate LLMs' creative capacity in narrative generation (Chen and Ding, 2023; Gómez-Rodríguez and Williams, 2023; Franceschelli and Musolesi, 2024). Our work builds on this perspective, exploring hallucinations not as factual errors, but as potentially productive deviations that improve downstream decision-making.

**Counterfactual Reasoning and Hypothetical Text** Counterfactual reasoning, which investigates how slight changes in input affect model predictions, has become a foundational method to explore causal relationships and to delineate decision boundaries in explainable AI (Byrne, 2019; Prado-Romero et al., 2024). They offer actionable insights, reveal model biases, and can uncover failure cases (Qu et al., 2025; Guidotti, 2024).

While most applications of counterfactuals in NLP focus on fairness, robustness, or factual consistency (Wu et al., 2021; Qiu et al., 2024; Wang et al., 2024; Chatzi et al., 2024), recent work begins to explore implicit counterfactuals in generative text (Dixit et al., 2022). In molecular modeling, counterfactual explanations have been applied to explain property predictions by generating minimally perturbed molecular variants (Wellawatte et al., 2022; Zhang et al., 2025; Teufel et al., 2025). Hallucinated inputs, although not explicitly constructed as counterfactuals, may similarly function as soft perturbations that induce informative changes in model behavior. Hallucinated text can be interpreted as speculative hypotheses about molecular properties, enabling models to generalize beyond literal training examples. To our knowledge, our study is the first to examine this connection in the context of molecule property prediction with LLMs.

## 3 Hallucination Generation

Large language models (LLMs), when prompted to describe molecules, often produce fluent yet factually incorrect descriptions, commonly referred to as hallucinations. While typically regarded as undesirable, we explore an alternative view: these hallucinations can be seen as implicit counterfactuals, imaginative variations on molecular structure or function that may enhance model performance in downstream prediction tasks. In this section, we detail how we generate these descriptions and analyze their properties.

### 3.1 Prompt Design

Translating molecular structures into natural language has been shown to support various biomedical and scientific tasks (Edwards et al., 2022). In this work, we follow a similar approach by prompting large language models (LLMs) to generate descriptive text from SMILES strings.

To ensure consistency across models and datasets, we use a standardized prompt template for all molecule descriptions. We also define the

system role as an "expert in drug discovery" to encourage chemistry-relevant outputs. The prompt is shown below:

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ Describe the molecule in natural language:

We apply this prompt to generate molecule descriptions using seven different LLMs across five benchmark datasets. To assess the factual consistency of the generated text, we compare the outputs to reference descriptions generated by `MolT5` (Edwards et al., 2022), a model pre-trained and fine-tuned on large-scale molecule-text pairs.

Factual consistency is measured using the HHM-2.1-Open model (Bao et al., 2024), which estimates alignment between generated descriptions and reference text. As shown in Figure 1, most LLM-generated descriptions diverge substantially from the `MolT5` references. Even `ChemLLM`, which is domain-tuned, achieves only 20.89% consistency. Other models range from 7.42% to 13.58%. These results confirm that the majority of generated descriptions contain some degree of hallucination.

Throughout this work, we define hallucination as any information in the generated text that is not supported by the molecule's actual structure or known properties.

### 3.2 Hallucination Typology and Annotation

To better characterize the nature of hallucinated content, we define a typology of hallucination types based on qualitative analysis. We randomly sample 50 descriptions from across datasets and manually identify four recurring hallucination patterns. An additional fifth category, "no hallucination," is used to label fully accurate outputs. Table 1 summarizes the categories and their definitions.

| Type | Definition |
|---|---|
| Functional hallucination | Adds plausible usage/application info or unverified biochemical action |
| Structural misdescription | Incorrect description of atoms, bonds, or substructures |
| Analogical hallucination | Uses metaphor or analogy to describe structure or function |
| Generic fluff | Vague, uninformative, or subjective phrases |
| No hallucination | Accurate statements without speculative or incorrect elements |

Table 1: Hallucination types in molecular descriptions.

We annotate hallucination types at scale using an external LLM, `Deepseek-R1` (Guo et al., 2025), which was not used to generate the descriptions. The annotation prompt includes the SMILES string, the generated description, and the type definitions in Table 1. The model is instructed to assign one or more appropriate hallucination types to each input.

To ensure quality, we randomly sample 50 instances and have two human annotators independently assign hallucination types. Inter-annotator agreement with the model is measured using Fleiss' Kappa. Inter-annotator agreement between the humans is $\kappa = 0.62$, while agreement with the model is $\kappa = 0.57$, indicating moderate agreement and acceptable reliability.

## 4 Task Formulation

We define molecule property prediction as a prompt-based decision task for large language models (LLMs). Each instance consists of a molecule represented by a SMILES string $[SMILES]$, optionally augmented with a natural language description $[Description]$, followed by a task-specific instruction $[Instruct]$.

> *System*: You are an expert in chemistry.
> *User*: $[SMILES]$ $[Description]$ $[Instruct]$

The model is then asked to answer a binary yes/no question, such as: *"Does this molecule inhibit HIV replication? Only answer 'Yes' or 'No'."*

We decode the model's prediction by selecting the highest-probability token:

$$\hat{y} = \arg\max_{y \in V} P_\phi(y), \quad V = \{\text{Yes}, \text{No}\} \quad (1)$$

**Textual Conditioning as Counterfactual Context.** Our central hypothesis is that the presence of $[Description]$, regardless of factual accuracy, provides useful inductive bias for the model. These descriptions vary across three settings:

- **None (SMILES-only)**: The model receives no textual context beyond the SMILES string.
- **Factual**: A structured, rule-based description from PubChem is included.
- **LLM-generated**: A free-text description is generated as described in the previous section, and may include hallucinations.

Rather than treating hallucinated descriptions as noise, we consider them as *implicit counterfactuals*, alternative narratives about molecular properties that may help the model distinguish subtle patterns. This setup enables us to investigate how speculative or misaligned information interacts with LLM decision-making in molecular tasks.

## 5 Experimental Design

We investigate how molecule-level hallucinations generated by different LLMs affect downstream classification performance. The prompt format is constructed as $[SMILES][Description][Instruct]$, where $[Description]$ is varied under 4 different settings:

**SMILES**: $[Description]$ is set to $\epsilon$ (empty), and the model must make predictions based solely on the SMILES string.

**MolT5**: We use descriptions generated by MolT5 (Edwards et al., 2022), a pretrained molecule-to-text model. This provides a reference point with high-quality factual descriptions aligned to molecular structure.

**PubChem**: We extract rule-based factual information from the PubChem database (Kim et al., 2025), including molecular formula, atom count, synonyms, and molecular weight. This metadata is formatted into a natural language description, e.g., *"This compound, known as X, has the molecular formula* $C_{22}H_{23}N_7O_2$*..."*. While this setting reflects gold-standard chemical knowledge, it is limited to molecules present in PubChem (88% coverage). As such, this baseline is not intended for practical deployment but to assess model behavior under ideal factual conditions.

**LLMs**: To study the effects of hallucinations, we replace $[Description]$ with text generated by different LLMs. Each model is evaluated using (a) its own hallucinated descriptions and (b) hallucinations generated by other LLMs.

### 5.1 Datasets

We select five benchmark datasets from MoleculeNet (Wu et al., 2018), covering a range of physiological and pharmacological properties: **HIV**: Predicts anti-HIV activity based on inhibition assay data. **BBBP**: Indicates whether a molecule can penetrate the blood-brain barrier. **Clintox**: Labels compounds that failed clinical trials due to toxicity. **SIDER**: Classifies drug molecules by whether they cause specific side effects.**Tox21**: Measures toxicity across 12 biological targets.

**Data Splits.** We follow the recommended MoleculeNet protocol. HIV and BBBP are split using scaffold splitting (Bemis and Murcko, 1996), which groups molecules based on core structure. Clintox, SIDER, and Tox21 use random splitting. For each dataset, we hold out 10% of the data as the test set.

**Label Selection and Evaluation.** We retain the original binary labels for HIV, BBBP, and Clintox. Since SIDER and Tox21 contain multiple labels, we select the one with the most balanced distribution in the test set. For SIDER, this corresponds to the label *"reproductive system and breast disorders"*. For Tox21, we use the *SR-MMP* target (matrix metalloproteinases). We report model performance using ROC-AUC, as standard for imbalanced binary classification tasks (Wu et al., 2018).

### 5.2 Models

We evaluate seven instruction-tuned LLMs, grouped by category: **General-purpose open-source LLMs**: `Llama-3-8B`, `Llama-3.1-8B`, `Mistral-8B`, and `Falcon3-Mamba-7B`. **Domain-specific LLM**: `ChemLLM-7B`, trained on chemical corpora. **Closed-source API models**: `GPT-3.5-turbo` and `GPT-4o`.

All open-source models are roughly 7–8B in size for fair comparison. Text generation is conducted using default decoding settings, with temperature 0.6 and a maximum token limit of 256. Models are evaluated both as *generators* (to produce hallucinated descriptions) and as *predictors* (to make property predictions given a prompt). When acting as predictors, the temperature is set to 0 to ensure deterministic generation.

## 6 Main Results and Analysis

Our experiments address three primary questions:

**i)** Can hallucinated descriptions improve LLM performance compared to standard baselines?

**ii)** Which LLMs generate the most beneficial hallucinations?

**iii)** What types of hallucinations are most likely to improve prediction?

### 6.1 Do Hallucinations Improve LLMs?

Table 2 presents the performance of each model on five datasets, comparing hallucinated molecule descriptions against three baselines: the SMILES-only input, the neural reference `MolT5`, and the rule-based `PubChem` descriptions. To provide representative coverage across different model types, we include the LLM from each of the three categories whose generated hallucinations resulted in the

| Model | [*Description*] | HIV | BBBP | Clintox | SIDER | Tox21 | Avg | ΔSMILES | ΔMolT5 | ΔPubChem |
|---|---|---|---|---|---|---|---|---|---|---|
| Llama-3-8B | SMILES | **67.78** | 53.08 | **63.04** | **61.79** | 60.34 | **61.21** | - | 6.53 | 0.55 |
| | MolT5 | 47.65 | 59.65 | 43.20 | 61.14 | 61.73 | 54.68 | -6.53 | - | -5.98 |
| | PubChem | 49.85 | **72.48** | 53.35 | 59.99 | **67.60** | 60.65 | -0.55 | 5.98 | - |
| | Llama-3 | 53.13 | 58.21 | 57.13 | 57.07 | 59.75 | 57.06 | -4.15 | 2.38 | -3.59 |
| | ChemLLM | 46.83 | 60.22 | 52.36 | 58.50 | 50.83 | 53.75 | -7.46 | -0.93 | -6.90 |
| | GPT-4o | 45.97 | 53.72 | 54.81 | 45.86 | 60.90 | 52.25 | -8.96 | -2.42 | -8.40 |
| Llama-3.1-8B | SMILES | 38.10 | 37.56 | 35.30 | 52.70 | 44.89 | 41.71 | - | -4.57 | -22.36 |
| | MolT5 | 43.04 | 45.30 | 39.28 | 52.06 | 51.72 | 46.28 | 4.57 | - | -17.8 |
| | PubChem | **58.94** | **71.42** | 66.03 | **57.54** | **66.45** | **64.08** | 22.36 | 17.80 | - |
| | Llama-3 | 56.80 | 47.71 | **68.08** | 54.21 | 60.75 | 57.51 | 15.80 | 11.23 | -6.57 |
| | ChemLLM | 50.83 | 52.18 | 47.45 | 53.17 | 49.84 | 50.69 | 8.98 | 4.41 | -13.39 |
| | GPT-4o | 56.64 | 53.37 | 55.28 | 52.60 | 61.65 | 55.91 | 14.20 | 9.63 | -8.17 |
| Ministral-8B | SMILES | **59.35** | 48.87 | 60.19 | 57.27 | 57.42 | 56.62 | - | 3.53 | -5.63 |
| | MolT5 | 44.54 | 44.10 | 53.95 | 63.83 | 59.02 | 53.09 | -3.53 | - | -9.16 |
| | PubChem | 50.08 | 52.77 | **79.23** | **64.45** | **64.72** | **62.25** | 5.63 | 9.16 | - |
| | Llama-3 | 52.91 | 31.27 | 65.49 | 50.45 | 55.81 | 51.19 | -5.43 | -1.90 | -11.06 |
| | ChemLLM | 46.35 | 48.66 | 62.11 | 58.97 | 55.36 | 54.29 | -2.33 | 1.20 | -7.96 |
| | GPT-4o | 51.66 | **64.94** | 58.73 | 60.54 | 61.07 | 59.39 | 2.77 | 6.30 | -2.86 |
| Falcon3-Mamba-7B | SMILES | 40.64 | 48.33 | 31.72 | 52.53 | 47.45 | 44.13 | - | 0.21 | -1.34 |
| | MolT5 | 49.02 | 47.92 | 18.58 | 55.84 | 48.27 | 43.93 | -0.21 | - | -1.54 |
| | PubChem | 45.40 | 58.84 | 24.55 | 52.00 | 46.56 | 45.47 | 1.34 | 1.54 | - |
| | Llama-3 | 47.84 | 53.40 | 50.76 | 51.31 | 48.44 | 50.35 | 6.22 | 6.42 | 4.88 |
| | ChemLLM | **53.46** | **61.82** | 45.12 | 53.56 | 43.43 | 51.48 | 7.35 | 7.55 | 6.01 |
| | GPT-4o | 51.59 | 55.73 | **53.55** | **56.03** | **51.54** | **53.69** | 9.55 | 9.76 | 8.22 |
| ChemLLM-7B | SMILES | 55.54 | 38.69 | 24.02 | 47.85 | **65.59** | 46.34 | - | -1.93 | -8.36 |
| | MolT5 | 50.33 | 41.01 | 35.43 | 53.06 | 61.50 | 48.27 | 1.93 | - | -6.43 |
| | PubChem | 54.96 | **72.27** | 34.37 | 51.08 | 60.79 | **54.69** | 8.36 | 6.43 | - |
| | Llama-3 | **56.29** | 45.68 | **47.51** | 44.95 | 51.59 | 49.20 | 2.87 | 0.94 | -5.49 |
| | ChemLLM | 45.37 | 41.83 | 42.80 | **53.15** | 46.60 | 45.95 | -0.39 | -2.32 | -8.74 |
| | GPT-4o | 52.68 | 55.73 | 40.74 | 44.59 | 62.63 | 51.28 | 4.94 | 3.01 | -3.42 |
| GPT-3.5 | SMILES | **61.32** | 28.94 | 63.90 | 53.77 | 61.92 | 53.97 | - | -2.56 | -11.01 |
| | MolT5 | 49.34 | 52.64 | 64.30 | 56.00 | 60.39 | 56.53 | 2.56 | - | -8.44 |
| | PubChem | 56.64 | **61.11** | **76.34** | 59.99 | **70.81** | **64.98** | 11.01 | 8.44 | - |
| | Llama-3 | 52.55 | 30.00 | 38.27 | 52.22 | 55.71 | 45.75 | -8.22 | -10.78 | -19.23 |
| | ChemLLM | 50.27 | 45.97 | 63.03 | 53.77 | 54.57 | 53.52 | -0.45 | -3.01 | -11.46 |
| | GPT-4o | 49.67 | 34.61 | 60.83 | **62.41** | 65.36 | 54.57 | 0.60 | -1.96 | -10.40 |
| GPT-4o | SMILES | 47.19 | 46.79 | 40.00 | 52.30 | 63.25 | 49.91 | - | 2.10 | -6.17 |
| | MolT5 | 41.73 | 40.57 | 38.09 | **65.22** | 53.42 | 47.80 | -2.10 | - | -8.27 |
| | PubChem | 50.96 | 50.34 | 56.67 | 56.89 | **65.52** | **56.08** | 6.17 | 8.27 | - |
| | Llama-3 | 48.95 | 36.23 | 52.97 | 42.73 | 62.00 | 48.58 | -1.33 | 0.77 | -7.5 |
| | ChemLLM | 53.00 | **54.92** | 55.59 | 45.36 | 61.15 | 54.00 | 4.10 | 6.20 | -2.08 |
| | GPT-4o | **53.30** | 52.73 | **57.12** | 47.82 | 65.34 | 55.26 | 5.35 | 7.46 | -0.82 |

Table 2: The main results (ROC-AUC %) for a curated set of descriptions. For each model, we show the 3 baselines (SMILES, MolT5, PubChem) and 3 selected hallucinated descriptions (Llama-3, ChemLLM, GPT-4o). ΔSMILES, ΔMolT5, and ΔPubChem denote the difference in average ROC-AUC scores compared to the baselines. Within each model group, the best and 2nd-best scores for each dataset are highlighted in **bold** and underlined, respectively.

greatest average performance gains in ROC-AUC: Llama-3 for open-source models, GPT-4o for proprietary models, and ChemLLM-7B for domain-specific models fine-tuned on molecular data.[2]

PubChem consistently serves as the strongest factual baseline, with 14 out of 35 model–dataset combinations (7 models across 5 datasets) achieving their highest performance when provided with its accurate, rule-based metadata. However, hallucinated descriptions can sometimes surpass this gold standard. For example, Falcon3-Mamba obtains an average ROC-AUC that is 8.22% higher when using hallucinated descriptions generated by GPT-4o. Similar gains are observed with other models: hallucinated descriptions from Llama-3 and ChemLLM improve the average ROC-AUC for Falcon3-Mamba by 4.88% and 6.01%, respectively. These results demonstrate that speculative or loosely grounded hallucinations can, in some cases, be more effective than factual metadata.

Hallucinated descriptions often lead to substantial improvements over the SMILES baseline. Among the 21 model–generator combinations (7 models × 3 hallucination sources), 12 achieve higher average ROC-AUC than when using SMILES, with especially consistent gains observed for Llama-3.1-8B and Falcon3-Mamba, each of which benefits from all three hallucination sources. Even in cases where hallucinated descriptions do not outperform SMILES, they often remain compet-

[2]Full results for all hallucinated variants and model configurations are provided in the appendix.

itive, suggesting their potential to convey useful inductive biases.

Similar patterns hold when compared to MolT5: 14 of 21 combinations achieve improved performance. In particular, hallucinations generated for Llama-3.1-8B, Falcon3-Mamba, and GPT-4o consistently surpass MolT5, highlighting that free-text generated by LLMs can rival or exceed specialized neural baselines. Overall, while not universally superior, hallucinated inputs frequently enhance performance, demonstrating their practical value for molecule property prediction.

Among all combinations, Llama-3.1-8B shows the largest gain: using hallucinations generated by Llama-3-8B, it achieves 15.80% higher performance than SMILES, and 11.23% over MolT5. Falcon3-Mamba-7B also surpasses all baselines by at least 4.88%.

**These results suggest that hallucinated descriptions, despite being factually inconsistent, can enhance molecular classification performance in LLMs, sometimes even outperforming authoritative references.**

### 6.2 Which LLMs Generate the Good Hallucinations?

To determine which LLMs produce the most useful hallucinations, we evaluate performance improvements across all seven models using hallucinations generated by different sources. Figure 3 presents the average gain in ROC-AUC when switching from SMILES or MolT5 input to hallucinated descriptions.

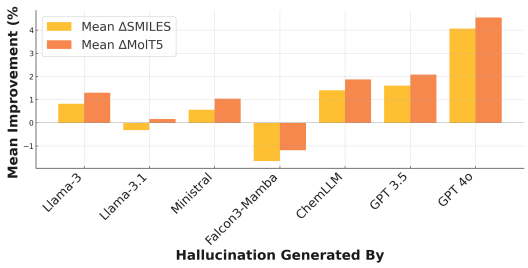


Figure 3: Average ROC-AUC gain across seven LLMs using hallucinations from different source models.

Hallucinations from OpenAI models consistently yield the greatest improvements. Descriptions from GPT-4o improve average performance by 4.07% over SMILES and 4.54% over MolT5. GPT-3.5 also provides solid gains of 1.6% and 2.08%, respectively. Among open-source models, hallucinations from Llama-3 and Mistral offer modest improvements, while those from Falcon3-Mamba and Llama-3.1 lead to slight decreases in performance.

**Overall, hallucinations generated by GPT-4o are the most beneficial across all models, suggesting that not all hallucinations are equally useful; type, quality, and coherence matter.**

### 6.3 What Types of Hallucinations Are Most Useful?

| Source | HIV | BBBP | Clintox | SIDER | Tox21 | Avg. |
|---|---|---|---|---|---|---|
| Llama-3 | 83.5 | 17.6 | 81.8 | 44.8 | 76.8 | 60.9 |
| Llama-3.1 | 86.9 | 16.1 | 70.3 | 35.0 | 72.2 | 56.1 |
| Mistral | 87.6 | 16.1 | 60.1 | 37.1 | 74.1 | 55.0 |
| Falcon3-Mamba | 88.3 | 17.6 | 64.2 | 39.2 | 79.4 | 57.7 |
| ChemLLM | 44.2 | 27.3 | 33.8 | 35.0 | 19.9 | 32.0 |
| GPT-3.5 | 92.2 | 15.6 | 37.8 | 24.5 | 77.4 | 49.5 |
| GPT-4o | 78.4 | 25.9 | 62.8 | 36.4 | 66.9 | 54.1 |

Table 3: Proportion (%) of hallucinations that improve performance over all baselines for Falcon3-Mamba-7B, broken down by dataset and hallucination source.

To understand the mechanism behind this improvement, we analyze a subset of *beneficial hallucinations*, which are the ones that lead to correct predictions with greater confidence than any baseline. Focusing on Falcon3-Mamba-7B, we collect 18,872 such cases across five datasets and seven hallucination sources (Table 3).

Using the hallucination taxonomy established earlier, we find that the majority (85%) of effective hallucinations are *structural misdescriptions*, which are incorrect yet chemically plausible statements about the molecular structure. Another 12% are *functional hallucinations* involving inferred or speculative bioactivity. Generic or analogical hallucinations are rare and contribute little to performance (Figure 4).

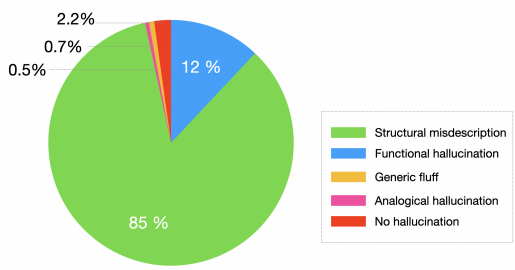


Figure 4: Distribution of hallucination types contributing to improved predictions for Falcon3-Mamba-7B.

**These hallucinations may act as *implicit counterfactuals***, offering alternative, albeit incorrect, chemical scenarios that prompt the LLM to refine

its internal representation of the molecule. Rather than misleading the model, such perturbations may function as soft interventions, related to counterfactual reasoning in structured explainability frameworks. They often describe structural or functional properties that deviate from known molecular facts but align with decision-relevant patterns.

This perspective suggests that hallucinations are not just noise: they reshape the model's belief state by inducing hypothetical molecular interpretations, thereby enhancing its predictive confidence.

### 6.4 Counterfactual Analysis

| Hallucination Type | HIV | BBBP | Clintox | SIDER | Tox21 | Avg |
|---|---|---|---|---|---|---|
| None (SMILES) | 40.64 | 48.33 | 31.72 | 52.53 | 47.45 | 44.13 |
| Generic fluff | 42.10 | 49.02 | 33.05 | 53.01 | 47.88 | 45.01 |
| Structural mis- | **55.69** | 63.29 | **53.89** | **57.53** | **52.11** | **56.50** |
| Analogical hallu- | 47.26 | 54.11 | 45.08 | 54.92 | 49.37 | 50.15 |
| Functional hallu- | 48.85 | **64.98** | 39.42 | 53.48 | 48.02 | 50.95 |

Table 4: ROC-AUC (%) of `Falcon3-Mamba-7B` conditioned on hallucinated molecular descriptions grouped by hallucination type (Table 1).

To examine whether the benefits of hallucinated descriptions depend on their semantic characteristics, we analyze the performance of `Falcon3-Mamba-7B` conditioned on hallucinations grouped by type (Table 4). This analysis controls for the base model and generation procedure, isolating the effect of hallucination category on downstream molecular property classification.

Overall, all hallucination types outperform the SMILES-only baseline, but the magnitude of improvement varies substantially across categories. Structural misdescriptions yield the largest average gain, improving ROC-AUC from 44.13 to 56.50. Analogical hallucinations and functional hallucinations also provide consistent benefits across tasks, achieving an average ROC-AUC of 50.15 and 50.95 respectively. In contrast, generic fluff leads to more modest gains, despite outperforming the no-hallucination baseline.

Notably, the strongest gains arise from hallucinations that introduce plausible but incorrect structural or relational information, rather than vague language or high-level functional claims. This suggests that the benefits of hallucinated descriptions are not driven by the mere presence of additional text, nor by general biochemical assertions alone. Instead, hallucinations that modify or reframe molecular structure in a coherent yet incorrect manner appear particularly effective.

These results indicate a clear interaction between hallucination type and downstream performance, supporting the view that beneficial hallucinations function as structured semantic perturbations. While such perturbations do not correspond to true chemical counterfactuals, they introduce alternative structural narratives that may encourage the classifier to attend to property-relevant patterns latent in the representation space. Importantly, we do not claim a causal or mechanistic role for hallucinations; rather, this analysis demonstrates that only specific hallucination types, especially those involving structural variation, are consistently associated with improved classification performance.

## 7 Conclusion

This paper investigates whether hallucinations can improve the performance of LLMs in predicting molecule properties. We evaluate seven LLMs across five datasets by incorporating natural language descriptions of molecules, which often contain hallucinations, into the input prompts. Our results show that hallucinated input can improve classification performance for most LLMs, even outperforming accurate baselines like `MolT5` and `PubChem` in some cases. In particular, `Falcon3-Mamba-7B` and hallucinations generated by `GPT-4o` yield the most consistent improvements across datasets. We further categorize the beneficial hallucinations and find that *structural misdescription* is the most common and impactful type, suggesting that factual deviation does not necessarily harm, and may even help, LLM decision-making. Ablation studies reveal that larger models are better at leveraging hallucinated descriptions, while variation in generation temperature has limited impact on downstream performance. Our findings offer a new perspective on the role of hallucinations in LLM-based scientific tasks, highlighting their potential to act as implicit counterfactuals that encourage flexible reasoning and robust prediction. Future work may further investigate the mechanisms by which LLMs benefit from such hallucinated input and how to leverage this creative potential in a safe and intentional manner, especially in high-stakes domains.

## Limitations

Our work investigates the hallucinations of LLMs in drug discovery, we acknowledge several limitations associated with our study. First, while our

experiments focus on whether hallucinations can improve LLMs in drug discovery, we do not explore why hallucinations have this effect or the explainability of this phenomenon. As this is beyond the scope of the current work, we plan to address the explainability of our findings in future research.

Second, our experiments rely on a fixed prompt formulation for generating hallucinated descriptions. While preliminary variations suggest that the main performance trends are robust to minor prompt changes, we do not conduct a systematic prompt sensitivity analysis. Future work could explore prompt design as an additional axis for controlling hallucination characteristics and downstream performance.

Third, our study exclusively considers LLMs that process text inputs, where molecules are represented using SMILES strings. Other models, such as vision language models that process molecules visually, are out of the scope of this work.

Fourth, we focus on LLMs of similar sizes (7–8B) for the main experiments, with additional experiments conducted using `Llama-3.1-70B`. While the results obtained from these LLMs are fairly representative, we have not explored larger open-source models in this work.

Lastly, our evaluation focuses on discriminative tasks in drug discovery, specifically binary molecular property classification. While this setting allows for controlled and systematic analysis, it does not capture the full range of tasks encountered in real-world molecular modeling. In particular, we do not study regression objectives, multi-task prediction, or generative tasks such as molecule generation or optimization. Our primary goal in this work is to test the empirical hypothesis that hallucinated textual descriptions can improve downstream classification performance under a fixed task formulation. Extending this analysis to broader task settings, and understanding whether similar benefits arise in non-classification or generative scenarios, remains an important direction for future work.

## Ethics Statement

We acknowledge that hallucinations generated by LLMs can contain inaccuracies or misinformation. While our study focuses on understanding the effects of such hallucinations, we do not advocate for the dissemination of false information. Future research should carefully assess these risks and explore strategies to ensure the reliability and safety of hallucination-related approaches.

## References


Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Ruth MJ Byrne. 2019. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282. California, CA.

Chiranjib Chakraborty, Manojit Bhattacharya, and Sang-Soo Lee. 2023. Artificial intelligence enabled chatgpt and large language models in drug target discovery, drug discovery, and development. *Molecular Therapy-Nucleic Acids*, 33:866–868.

Ivi Chatzi, Nina Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. 2024. Counterfactual token generation in large language models. *arXiv preprint arXiv:2409.17027*.

Honghua Chen and Nai Ding. 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888, Singapore. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation.

In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2964–2984. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, and Heng Ji. 2024. Language + molecules. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 14–20, St. Julian's, Malta. Association for Computational Linguistics.

Gabriele Fossi, Youssef Boulaimen, Leila Outemzabet, Nathalie Jeanray, Stephane Gerart, Sebastien Vachenc, Joanna Giemza, and Salvatore Raieli. 2024. Swiftdossier: Tailored automatic dossier for drug discovery with llms and agents. *arXiv preprint arXiv:2409.15817*.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, pages 1–11.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. 2024. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013, Miami, Florida, USA. Association for Computational Linguistics.

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. In *Advances in Neural Information Processing Systems*, volume 36, pages 5539–5568. Curran Associates, Inc.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Shenghui Guan and Guanyu Wang. 2024. Drug discovery and development in the era of artificial intelligence: From machine learning to large language models. *Artificial Intelligence Chemistry*, 2(1):100070.

Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. 2011. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

Sunghwan Kim, Jie Chen, Tiejun Cheng, and et al. 2025. Pubchem 2025 update. *Nucleic Acids Research*.

Hye-Kyung Lee. 2022. Rethinking creativity: creative industries, ai and everyday creativity. *Media, Culture & Society*, 44(3):601–612.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320.

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

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. 2024. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, and 1 others. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Kusuri Murakumo, Naruki Yoshikawa, Kentaro Rikimaru, Shogo Nakamura, Kairi Furui, Takamasa Suzuki, Hiroyuki Yamasaki, Yuki Nishigaya, Yuzo Takagi, and Masahito Ohue. 2023. LLM drug discovery challenge: A contest as a feasibility study on the utilization of large language models in medicinal chemistry. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*.

Soumen Pal, Manojit Bhattacharya, Md Aminul Islam, and Chiranjib Chakraborty. 2023. Chatgpt or llm in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development. *International Journal of Surgery*, 109(12):4382–4384.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. 2024. A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. *ACM Computing Surveys*, 56(7):1–37.

Xiaoqi Qiu, Yongjie Wang, Xu Guo, Zhiwei Zeng, Yu Yue, Yuhong Feng, and Chunyan Miao. 2024. Paircfr: Enhancing model training on paired counterfactually augmented data through contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11955–11971.

Zhan Qu, Daniel Gomm, and Michael Färber. 2025. Cody: Counterfactual explainers for dynamic graphs. In *Forty-second International Conference on Machine Learning*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Neil Savage. 2023. Drug discovery companies are customizing chatgpt: here's how. *Nat Biotechnol*, 41(5):585–586.

Jonas Teufel, Annika Leinweber, and Pascal Friederich. 2025. Improving counterfactual truthfulness for molecular property prediction through uncertainty quantification. *arXiv preprint arXiv:2504.02606*.

Katherine Van Koevering and Jon Kleinberg. 2024. How random is random? evaluating the randomness and humaness of llms' coin flips. *arXiv preprint arXiv:2406.00092*.

Feng Wang. 2024. Lighthouse: A survey of agi hallucination. *arXiv preprint arXiv:2401.06792*.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818.

David Weininger. 1988. 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.

Geemi P Wellawatte, Aditi Seshadri, and Andrew D White. 2022. Model agnostic generation of counterfactual explanations for molecules. *Chemical science*, 13(13):3697–3705.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

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

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Shuzhou Yuan and Michael Faerber. 2023. Evaluating generative models for graph-to-text generation. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1256–1264, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, and 1 others. 2024a. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. 2024b. HoneyComb: A flexible LLM-based agent system for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3369–3382, Miami, Florida, USA. Association for Computational Linguistics.

Ke Zhang, Yimiao Feng, and Jie Zheng. 2024c. Prompt-based generation of natural language explanations of synthetic lethality for cancer drug discovery. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13131–13142, Torino, Italia. ELRA and ICCL.

Xiuping Zhang, Qun Liu, and Rui Han. 2025. Mmgcf: generating counterfactual explanations for molecular property prediction via motif rebuild. *Journal of Computer and Communications*, 13(1):152–168.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, and 1 others. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T May, Geoffrey I Webb, Shirui Pan, and George Church. 2024. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*.

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

# A Details of the Models

In Table 5, we list the full names and the links of the LLMs used in the work. All the open-source LLMs can be applied directly using Transformers library by Huggingface.

# B Details of the Datasets

The details, including the number of samples and the count of positive labels in each dataset are reported in Table 6.

# C Prompt Template for Different Tasks

For each task, we use different instruction as they aim for predicting different properties of molecules.

## C.1 HIV

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ $[Description]$
> Does the molecule have the ability to inhibit HIV replication? Only answer Yes or No:

## C.2 BBBP

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ $[Description]$
> Does the molecule have the ability to penetrate the blood-brain barrier? Only answer Yes or No:

## C.3 Clintox

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ $[Description]$
> Did the molecule fail clinical trials due to toxicity? Only answer Yes or No:

## C.4 SIDER

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ $[Description]$
> Does the molecule cause side effects on the reproductive system or breast? Only answer Yes or No:

## C.5 Tox21

> *System*: You are an expert in drug discovery.
> *User*: $[SMILES]$ $[Description]$
> Does the molecule have the potential toxicity affecting mitochondrial membrane potential (SR-MMP)? Only answer Yes or No:

# D Hallucination Score for LLMs Generated Description

We evaluate the LLMs generated description using HHM-2.1-Open and report the results in Table 7.

| Model | Full-name | Link |
|---|---|---|
| Llama-3-8B | meta-llama/Meta-Llama-3-8B-Instruct | https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct |
| Llama-3.1-8B | meta-llama/Llama-3.1-8B-Instruct | https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct |
| Ministral-8B | mistralai/Ministral-8B-Instruct-2410 | https://huggingface.co/mistralai/Ministral-8B-Instruct-2410 |
| Falcon3-Mamba-7B | tiiuae/Falcon3-Mamba-7B-Instruct | https://huggingface.co/tiiuae/Falcon3-Mamba-7B-Instruct |
| ChemLLM-7B | AI4Chem/ChemLLM-7B-Chat-1_5-DPO | https://huggingface.co/AI4Chem/ChemLLM-7B-Chat-1_5-DPO |
| GPT-3.5 | gpt-3.5-turbo | - |
| GPT-4o | gpt-4o-2024-08-06 | - |

Table 5: Details of the evaluated models, including their full names and links to their respective HuggingFace pages.

| Dataset | Number | Positive Label |
|---|---|---|
| HIV | 4113 | 175 |
| BBBP | 205 | 172 |
| Clintox | 148 | 11 |
| SIDER | 143 | 74 |
| Tox21 | 783 | 108 |

Table 6: Summary of the datasets used, including the number of samples and the count of positive labels in each dataset.

## E Full Results of Main Experiments

The full results for all LLMs used in our study are presented in Table 8. Each LLM is evaluated using at least nine prompt templates: two from the baselines and seven incorporating hallucinations generated by other models.

## F Full Results for Model Size Experiments

The full results for the model size experiments are shown in Table 10. Additionally, the hallucination scores for each model size are provided in Table 11.

## G Full Results for Experiments on Generation Temperature

Table 12 presents the complete results of Llama-3.1-8B performance when using hallucinations generated at different temperatures. Additionally, the corresponding hallucination scores are shown in Table 13.

## H Prompt Template for Annotating Hallucination in Types

*System*: You are an expert annotator specializing in identifying hallucinations in molecular descriptions. Your task is to carefully analyze descriptions and categorize any hallucinations according to the provided taxonomy.
*User*: $[SMILES]$ Your task is to identify all the types of hallucinations present in the hallucinated description of the molecule.

Molecule: $[SMILES]$
Hallucination: "$[HALLUCINATION]$"

Here is a description of each category:

$[CATEGORIES]$

Instructions:

1. Choose the category names that BEST matches the hallucination. Write these category names in decreasing order of importance.
2. Provide an explanation for why you chose this/these category name(s) and their order of importance.
3. Return json with molecule SMILES, category name(s) and explanation
4. If no category name matches, use "No hallucination"

## I Ablation Studies

To better understand when and why hallucinated text improves molecular prediction, we conduct ablation studies focusing on two factors: model size and generation temperature. We use the Llama-3

| Model | HIV | BBBP | Clintox | SIDER | Tox21 | Avg |
|---|---|---|---|---|---|---|
| Llama-3.1-8B | 7.12 | 7.76 | 8.67 | 9.45 | 6.86 | 7.97 |
| Llama-3-8B | 7.15 | 8.04 | 7.24 | 7.57 | 7.11 | 7.42 |
| Ministral-8B | 14.45 | 13.47 | 14.23 | 13.17 | 12.56 | 13.58 |
| Falcon3-Mamba-7B | 8.74 | 9.05 | 8.92 | 11.13 | 9.26 | 9.42 |
| ChemLLM-7B | 20.48 | 17.75 | 23.36 | 22.20 | 20.66 | 20.89 |
| GPT-3.5 | 8.62 | 8.35 | 6.85 | 7.20 | 6.90 | 7.58 |
| GPT-4o | 8.51 | 7.67 | 8.51 | 7.28 | 7.44 | 7.88 |

Table 7: The HHM-2.1 score % for LLMs across five datasets, all the models are using the same hyperparameters, e.g.: the temperature, max new tokens.

family due to its consistent architecture and multiple available sizes. As PubChem coverage is incomplete, we report improvements relative to SMILES and MolT5 baselines.

### I.1 Effect of Model Size

Model size is a critical determinant of LLM performance (Dubey et al., 2024). We compare instruction-tuned versions of Llama-3.2 (1B, 3B) and Llama-3.1 (8B, 70B), using hallucinations generated by all other LLMs used for the experiments. For each model, we compute the average ROC-AUC improvement over the baselines.

As shown in Figure 5, performance improves steadily from 1B to 8B when compared to the MolT5 baseline, suggesting that larger models better leverage the information, even when it is hallucinatory. The 70B model does not outperform the 8B model but still exceeds the smaller sizes. When compared to SMILES, all sizes show improvement.

**These findings suggest that larger models are more capable of extracting useful signals from hallucinated descriptions, though gains may plateau beyond 8B.**

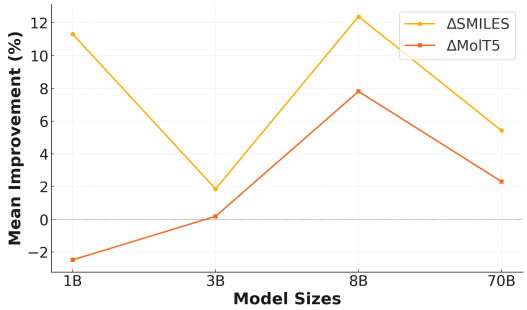


Figure 5: ROC-AUC variation for Llama-3 models of different sizes with hallucinated descriptions.

### I.2 Effect of Generation Temperature

Sampling temperature controls the randomness of generation (Van Koevering and Kleinberg, 2024),

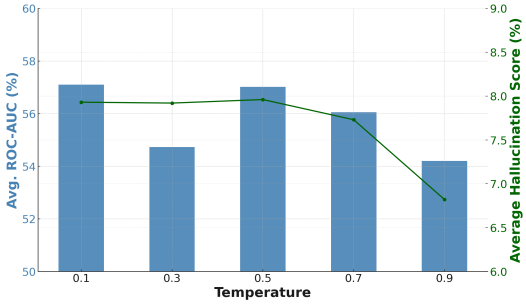


Figure 6: Performance and hallucination scores of Llama-3.1-8B across temperatures (0.1–0.9).

with higher temperatures producing more diverse and often less factual outputs (Peeperkorn et al., 2024; Renze, 2024). We vary the temperature between 0.1 and 0.9 while using Llama-3.1-8B to generate molecular descriptions, then assess both hallucination scores and downstream performance.

Figure 6 shows that as the temperature increases, hallucination scores increase, that is, the factual consistency decreases. This aligns with prior expectations. However, the correlation with performance is non-monotonic: model accuracy peaks at temperatures 0.1 and 0.5, but dips slightly at 0.3 and 0.9. The difference across all temperatures is small (within 3%).

Notably, all temperature settings outperform both the SMILES (41.71%) and MolT5 (46.28%) baselines (Table 2). **This confirms that even high-temperature, hallucinated text, despite reduced factual consistency, can still contribute positively to LLM performance.**

| Model | [*Description*] | HIV | BBBP | Clintox | SIDER | Tox21 | Avg | ΔSMILES | ΔMolT5 | ΔPubChem |
|---|---|---|---|---|---|---|---|---|---|---|
| Llama-3-8B | SMILES | **67.78** | 53.08 | 63.04 | **61.79** | 60.34 | 61.21 | - | 6.53 | 0.55 |
| | MolT5 | 47.65 | 59.65 | 43.20 | 61.14 | 61.73 | 54.68 | -6.53 | - | -5.98 |
| | PubChem | 49.85 | **72.48** | 53.35 | 59.99 | **67.60** | 60.65 | -0.55 | 5.98 | - |
| | Llama-3 | 53.13 | 58.21 | 57.13 | 57.07 | 59.75 | 57.06 | -4.15 | 2.38 | -3.59 |
| | Llama-3.1 | 57.60 | 51.02 | 39.75 | 60.63 | 59.70 | 53.74 | -7.46 | -0.93 | -6.91 |
| | Ministral | 55.09 | 61.13 | **63.84** | 57.56 | 55.41 | 58.61 | -2.60 | 3.93 | -2.04 |
| | Falcon3 | 54.02 | 51.73 | 43.00 | 47.06 | 48.91 | 48.94 | -12.26 | -5.73 | -11.71 |
| | ChemLLM | 46.83 | 60.22 | 52.36 | 58.50 | 50.83 | 53.75 | -7.46 | -0.93 | -6.90 |
| | GPT-3.5 | 58.45 | 46.52 | 57.13 | 56.15 | 66.00 | 56.85 | -4.36 | 2.18 | -3.80 |
| | GPT-4o | 45.97 | 53.72 | 54.81 | 45.86 | 60.90 | 52.25 | -8.96 | -2.42 | -8.40 |
| Llama-3.1-8B | SMILES | 38.10 | 37.56 | 35.30 | 52.70 | 44.89 | 41.71 | - | -4.57 | -22.36 |
| | MolT5 | 43.04 | 45.30 | 39.28 | 52.06 | 51.72 | 46.28 | 4.57 | - | -17.8 |
| | PubChem | 58.94 | **71.42** | 66.03 | **57.54** | **66.45** | 64.08 | 22.36 | 17.80 | - |
| | Llama-3 | 56.80 | 47.71 | 68.08 | 54.21 | 60.75 | 57.51 | 15.80 | 11.23 | -6.57 |
| | Llama-3.1 | 50.92 | 42.11 | 49.97 | 38.11 | 51.74 | 46.57 | 4.86 | 0.29 | -17.51 |
| | Ministral | 59.73 | 50.18 | 70.01 | 51.65 | 55.59 | 57.43 | 15.72 | 11.15 | -6.65 |
| | Falcon3 | 57.92 | 51.11 | 53.22 | 42.38 | 47.57 | 50.44 | 8.73 | 4.16 | -13.64 |
| | ChemLLM | 50.83 | 52.18 | 47.45 | 53.17 | 49.84 | 50.69 | 8.98 | 4.41 | -13.39 |
| | GPT-3.5 | **64.66** | 47.20 | **73.19** | 55.57 | 59.70 | 60.07 | 18.35 | 13.79 | -4.01 |
| | GPT-4o | 56.64 | 53.37 | 55.28 | 52.60 | 61.65 | 55.91 | 14.20 | 9.63 | -8.17 |
| Ministral-8B | SMILES | 59.35 | 48.87 | 60.19 | 57.27 | 57.42 | 56.62 | - | 3.53 | -5.63 |
| | MolT5 | 44.54 | 44.10 | 53.95 | 63.83 | 59.02 | 53.09 | -3.53 | - | -9.16 |
| | PubChem | 50.08 | 52.77 | **79.23** | **64.45** | **64.72** | 62.25 | 5.63 | 9.16 | - |
| | Llama-3 | 52.91 | 31.27 | 65.49 | 50.45 | 55.81 | 51.19 | -5.43 | -1.90 | -11.06 |
| | Llama-3.1 | 56.06 | 46.62 | 36.23 | 55.95 | 58.33 | 50.64 | -5.98 | -2.45 | -11.61 |
| | Ministral | **60.04** | 46.42 | 55.34 | 52.82 | 53.58 | 53.64 | -2.98 | 0.56 | -8.61 |
| | Falcon3 | 55.37 | 50.16 | 53.62 | 47.14 | 58.30 | 52.92 | -3.70 | -0.17 | -9.33 |
| | ChemLLM | 46.35 | 48.66 | 62.11 | 58.97 | 55.36 | 54.29 | -2.33 | 1.20 | -7.96 |
| | GPT-3.5 | 53.23 | 36.89 | 46.85 | 51.63 | 59.86 | 49.69 | -6.93 | -3.40 | -12.56 |
| | GPT-4o | 51.66 | **64.94** | 58.73 | 60.54 | 61.07 | 59.39 | 2.77 | 6.30 | -2.86 |
| Falcon3-Mamba-7B | SMILES | 40.64 | 48.33 | 31.72 | 52.53 | 47.45 | 44.13 | - | 0.21 | -1.34 |
| | MolT5 | 49.02 | 47.92 | 18.58 | 55.84 | 48.27 | 43.93 | -0.21 | - | -1.54 |
| | PubChem | 45.40 | 58.84 | 24.55 | 52.00 | 46.56 | 45.47 | 1.34 | 1.54 | - |
| | Llama-3 | 47.84 | 53.40 | 50.76 | 51.31 | 48.44 | 50.35 | 6.22 | 6.42 | 4.88 |
| | Llama-3.1 | 48.36 | 60.15 | 34.84 | **57.60** | **54.26** | 51.04 | 6.91 | 7.12 | 5.57 |
| | Ministral | 47.32 | 47.82 | 27.94 | 51.90 | 47.54 | 44.50 | 0.37 | 0.58 | -0.97 |
| | Falcon3 | 44.59 | 57.05 | 39.42 | 50.10 | 46.64 | 47.56 | 3.43 | 3.63 | 2.09 |
| | ChemLLM | **53.46** | **61.82** | 45.12 | 53.56 | 43.43 | 51.48 | 7.35 | 7.55 | 6.01 |
| | GPT-3.5 | 51.63 | 48.47 | 45.26 | 50.98 | 46.73 | 48.61 | 4.48 | 4.69 | 3.14 |
| | GPT-4o | 51.59 | 55.73 | **53.55** | 56.03 | 51.54 | 53.69 | 9.55 | 9.76 | 8.22 |
| ChemLLM-7B | SMILES | 55.54 | 38.69 | 24.02 | 47.85 | **65.59** | 46.34 | - | -1.93 | -8.36 |
| | MolT5 | 50.33 | 41.01 | 35.43 | 53.06 | 61.50 | 48.27 | 1.93 | - | -6.43 |
| | PubChem | 54.96 | **72.27** | 34.37 | 51.08 | 60.79 | 54.69 | 8.36 | 6.43 | - |
| | Llama-3 | 56.29 | 45.68 | 47.51 | 44.95 | 51.59 | 49.20 | 2.87 | 0.94 | -5.49 |
| | Llama-3.1 | 57.17 | 43.02 | 28.33 | 43.83 | 56.96 | 45.86 | -0.47 | -2.40 | -8.83 |
| | Ministral | **61.35** | 37.70 | 39.95 | 37.76 | 51.64 | 45.68 | -0.66 | -2.59 | -9.01 |
| | Falcon3 | 57.52 | 39.76 | 29.46 | 44.71 | 50.18 | 44.33 | -2.01 | -3.94 | -10.36 |
| | ChemLLM | 45.37 | 41.83 | 42.80 | **53.15** | 46.60 | 45.95 | -0.39 | -2.32 | -8.74 |
| | GPT-3.5 | 57.48 | 38.16 | **51.76** | 41.72 | 59.37 | 49.70 | 3.36 | 1.43 | -4.99 |
| | GPT-4o | 52.68 | 55.73 | 40.74 | 44.59 | 62.63 | 51.28 | 4.94 | 3.01 | -3.42 |
| GPT-3.5 | SMILES | **61.32** | 28.94 | 63.90 | 53.77 | 61.92 | 53.97 | - | -2.56 | -11.01 |
| | MolT5 | 49.34 | 52.64 | 64.30 | 56.00 | 60.39 | 56.53 | 2.56 | - | -8.44 |
| | PubChem | 56.64 | **61.11** | **76.34** | 59.99 | **70.81** | 64.98 | 11.01 | 8.44 | - |
| | Llama-3 | 52.55 | 30.00 | 38.27 | 52.22 | 55.71 | 45.75 | -8.22 | -10.78 | -19.23 |
| | Llama-3.1 | 55.90 | 39.31 | 41.83 | 57.16 | 59.60 | 50.76 | -3.21 | -5.77 | -14.22 |
| | Ministral | 55.84 | 29.63 | 39.89 | 56.61 | 56.31 | 47.66 | -6.31 | -8.88 | -17.32 |
| | Falcon3 | 52.11 | 27.96 | 54.29 | 48.69 | 59.10 | 48.43 | -5.54 | -8.10 | -16.55 |
| | ChemLLM | 50.27 | 45.97 | 63.03 | 53.77 | 54.57 | 53.52 | -0.45 | -3.01 | -11.46 |
| | GPT-3.5 | 58.47 | 33.53 | 44.81 | 62.24 | 56.46 | 51.10 | -2.87 | -5.43 | -13.88 |
| | GPT-4o | 49.67 | 34.61 | 60.83 | **62.41** | 65.36 | 54.57 | 0.60 | -1.96 | -10.40 |
| GPT-4o | SMILES | 47.19 | 46.79 | 40.00 | 52.30 | 63.25 | 49.91 | - | 2.10 | -6.17 |
| | MolT5 | 41.73 | 40.57 | 38.09 | **65.22** | 53.42 | 47.80 | -2.10 | - | -8.27 |
| | PubChem | 50.96 | 50.34 | 56.67 | 56.89 | **65.52** | 56.08 | 6.17 | 8.27 | - |
| | Llama-3 | 48.95 | 36.23 | 52.97 | 42.73 | 62.00 | 48.58 | -1.33 | 0.77 | -7.5 |
| | Llama-3.1 | 53.02 | 45.57 | 54.10 | 48.71 | 64.07 | 53.09 | 3.19 | 5.29 | -2.99 |
| | Ministral | **55.34** | 40.13 | 49.46 | 42.44 | 64.17 | 50.31 | 0.40 | 2.50 | -5.77 |
| | Falcon3 | 54.01 | 35.37 | 50.00 | 45.83 | 63.34 | 49.71 | -0.20 | 1.90 | -6.37 |
| | ChemLLM | 53.00 | **54.92** | 55.59 | 45.36 | 61.15 | 54.00 | 4.10 | 6.20 | -2.08 |
| | GPT-3.5 | 47.51 | 40.80 | 46.03 | 47.27 | 63.82 | 49.09 | -0.82 | 1.28 | -6.99 |
| | GPT-4o | 53.30 | 52.73 | **57.12** | 47.82 | 65.34 | 55.26 | 5.35 | 7.46 | -0.82 |

Table 8: Full results (ROC-AUC %) for all LLMs, including hallucinations generated by each model and baselines. [*Description*] indicates the source of the hallucination or baseline used.

| Model | [*Description*] | HIV | BBBP | Clintox | SIDER | Tox21 | Avg | ΔSMILES | ΔMolT5 | ΔPubChem |
|---|---|---|---|---|---|---|---|---|---|---|
| Llama-3-8B | SMILES | **67.78** | 53.08 | **63.04** | **61.79** | 60.34 | **61.21** | **-** | **6.53** | **0.56** |
| | MolT5 | 47.65 | 59.65 | 43.20 | 61.14 | 61.73 | 54.68 | -6.53 | - | -5.97 |
| | PubChem | 49.85 | **72.48** | 53.35 | 59.99 | **67.60** | 60.65 | -0.56 | 5.97 | - |
| | Llama Family | 57.60 | 58.21 | 57.13 | 60.63 | 59.75 | 58.66 | -2.55 | 3.98 | -1.99 |
| | GPT Family | 58.45 | 53.72 | 57.13 | 56.15 | 66.00 | 58.29 | -2.92 | 3.61 | -2.36 |
| | ChemLLM | 46.83 | 60.22 | 52.36 | 58.50 | 50.83 | 53.75 | -7.46 | -0.93 | -6.90 |
| Llama-3.1-8B | SMILES | 38.10 | 37.56 | 35.30 | 52.70 | 44.89 | 41.71 | - | -4.57 | -22.37 |
| | MolT5 | 43.04 | 45.30 | 39.28 | 52.06 | 51.72 | 46.28 | 4.57 | - | -17.80 |
| | PubChem | 58.94 | **71.42** | 66.03 | **57.54** | **66.45** | **64.08** | **22.37** | **17.80** | **-** |
| | Llama Family | 56.80 | 47.71 | **68.08** | 54.21 | 60.75 | 57.51 | 15.80 | 11.23 | -6.57 |
| | GPT Family | **64.66** | 53.37 | 66.19 | 55.57 | 61.65 | 61.69 | 19.98 | 15.41 | -2.39 |
| | ChemLLM | 50.83 | 52.18 | 47.45 | 53.17 | 49.84 | 50.69 | 8.98 | 4.41 | -13.39 |
| Ministral-8B | SMILES | 59.35 | 48.87 | 60.19 | 57.27 | 57.42 | 56.62 | - | 3.53 | -5.63 |
| | MolT5 | 44.54 | 44.10 | 53.95 | **63.83** | 59.02 | 53.09 | -3.53 | - | -9.16 |
| | PubChem | 50.08 | 52.77 | **79.23** | 63.15 | **64.72** | **62.25** | **5.63** | **9.16** | **-** |
| | Llama Family | 56.06 | 46.62 | 65.49 | 55.95 | 58.33 | 56.49 | -0.13 | 3.40 | -5.76 |
| | GPT Family | 53.23 | **64.94** | 58.73 | 60.54 | 61.07 | 59.70 | 3.08 | 6.61 | -2.55 |
| | ChemLLM | 46.35 | 48.66 | 62.11 | 58.97 | 55.36 | 54.29 | -2.33 | 1.20 | -7.96 |
| Falcon3-Mamba-7B | SMILES | 40.64 | 48.33 | 31.72 | 52.53 | 47.45 | 44.13 | - | 0.20 | -1.34 |
| | MolT5 | 49.02 | 47.92 | 18.58 | 55.84 | 48.27 | 43.93 | -0.20 | - | -1.54 |
| | PubChem | 45.40 | 58.84 | 24.55 | 52.00 | 46.56 | 45.47 | 1.34 | 1.54 | - |
| | Llama Family | 50.70 | 60.15 | 50.76 | **57.60** | **54.26** | **54.23** | **10.10** | **10.30** | **8.76** |
| | GPT Family | 51.63 | 55.73 | **53.55** | 56.03 | 51.54 | 53.70 | 9.57 | 9.77 | 8.23 |
| | ChemLLM | **53.46** | **61.82** | 45.12 | 53.56 | 43.43 | 51.48 | 7.35 | 7.55 | 6.01 |
| ChemLLM-7B | SMILES | 55.54 | 38.69 | 24.02 | 47.85 | **65.59** | 46.34 | - | -1.93 | -8.35 |
| | MolT5 | 50.33 | 41.01 | 35.43 | 53.06 | 61.50 | 48.27 | 1.93 | - | -6.42 |
| | PubChem | 54.96 | **72.27** | 34.37 | 51.08 | 60.79 | **54.69** | **8.35** | **6.42** | **-** |
| | Llama Family | 57.17 | 45.68 | 47.51 | 44.95 | 56.96 | 50.45 | 4.11 | 2.18 | -4.24 |
| | GPT Family | **57.48** | 55.73 | **51.76** | 44.59 | 62.63 | 54.44 | 8.10 | 6.17 | -0.25 |
| | ChemLLM | 45.37 | 41.83 | 42.80 | **53.15** | 46.60 | 45.95 | -0.39 | -2.32 | -8.74 |
| GPT-3.5 | SMILES | **61.32** | 28.94 | 63.90 | 53.77 | 61.92 | 53.97 | - | -2.56 | -11.01 |
| | MolT5 | 49.34 | 52.64 | 64.30 | 56.00 | 60.39 | 56.53 | 2.56 | - | -8.45 |
| | PubChem | 56.64 | **61.11** | **76.34** | 62.41 | **70.81** | **64.98** | **11.01** | **8.44** | **-** |
| | Llama Family | 55.90 | 39.31 | 41.83 | 57.16 | 59.60 | 50.76 | -3.21 | -5.77 | -14.22 |
| | GPT Family | 58.47 | 34.61 | 60.83 | **62.41** | 65.36 | 56.34 | 2.37 | -0.19 | -8.64 |
| | ChemLLM | 50.27 | 45.97 | 63.03 | 53.77 | 54.57 | 53.52 | -0.45 | -3.01 | -11.46 |
| GPT-4o | SMILES | 47.19 | 46.79 | 40.00 | 52.30 | 63.25 | 49.91 | - | 2.11 | -6.17 |
| | MolT5 | 41.73 | 40.57 | 38.09 | **65.22** | 53.42 | 47.80 | -2.11 | - | -8.28 |
| | PubChem | 50.96 | 50.34 | 56.67 | 56.89 | **65.52** | **56.08** | **6.17** | **8.28** | **-** |
| | Llama Family | 53.02 | 45.57 | 54.10 | 48.71 | 64.07 | 53.09 | 3.18 | 5.29 | -2.99 |
| | GPT Family | **53.30** | 52.73 | **57.12** | 47.82 | 65.34 | 55.26 | 5.35 | 7.46 | -0.82 |
| | ChemLLM | 53.00 | **54.92** | 55.59 | 45.36 | 61.15 | 54.00 | 4.09 | 6.20 | -2.08 |

Table 9: Curated results with combined "Llama Family" and "GPT Family" descriptions. For these families, the best performance across their respective constituent models is used for each dataset. Within each model's group, the best and second-best scores per column are highlighted in **bold** and underlined, respectively.

| [*Description*] | HIV | BBBP | Clintox | SIDER | Tox21 | Avg | Δ SMILES | Δ MolT5 |
|---|---|---|---|---|---|---|---|---|
| **Llama-3.2-1B** | | | | | | | | |
| SMILES | 44.27 | 32.31 | 19.31 | 60.79 | 46.88 | 40.71 | 0.00 | -13.77 |
| MolT5 | 45.26 | 38.30 | 79.50 | 57.95 | 51.42 | 54.49 | 13.77 | 0.00 |
| Llama-3 | 46.79 | 47.32 | 60.19 | 56.23 | 44.25 | 50.96 | 10.24 | -3.53 |
| Llama-3.1 | 44.79 | 50.93 | 51.56 | 59.09 | 53.73 | 52.02 | 11.31 | -2.47 |
| Ministral | 46.94 | 42.00 | 54.48 | 51.94 | 52.71 | 49.61 | 8.90 | -4.87 |
| Falcon3 | 52.08 | 58.58 | 56.01 | 50.55 | 47.14 | 52.87 | 12.16 | -1.62 |
| ChemLLM | 46.23 | 44.77 | 60.25 | 56.35 | 44.19 | 50.36 | 9.65 | -4.13 |
| GPT-3.5 | 54.33 | 45.24 | 58.00 | 56.01 | 58.94 | 54.50 | 13.79 | 0.02 |
| GPT-4o | 50.29 | 60.02 | 64.43 | 52.88 | 52.10 | 55.94 | 15.23 | 1.46 |
| **Llama-3.2-3B** | | | | | | | | |
| SMILES | 43.57 | 40.40 | 40.15 | 56.13 | 48.62 | 45.77 | 0.00 | -1.67 |
| MolT5 | 48.17 | 45.77 | 51.23 | 49.47 | 42.55 | 47.44 | 1.67 | 0.00 |
| Llama-3 | 48.52 | 32.42 | 53.55 | 52.60 | 54.09 | 48.23 | 2.46 | 0.80 |
| Llama-3.1 | 46.99 | 40.93 | 33.78 | 53.09 | 50.48 | 45.05 | -0.72 | -2.39 |
| Ministral | 49.65 | 47.18 | 50.76 | 47.79 | 50.82 | 49.24 | 3.47 | 1.80 |
| Falcon3 | 59.43 | 44.96 | 39.81 | 52.51 | 42.38 | 47.82 | 2.05 | 0.38 |
| ChemLLM | 44.99 | 44.75 | 46.05 | 45.73 | 47.30 | 45.76 | -0.01 | -1.68 |
| GPT-3.5 | 53.66 | 32.61 | 58.93 | 55.70 | 47.01 | 49.58 | 3.81 | 2.14 |
| GPT-4o | 50.12 | 42.20 | 38.42 | 59.38 | 54.33 | 48.89 | 3.12 | 1.45 |
| **Llama-3.1-8B** | | | | | | | | |
| SMILES | 38.10 | 37.56 | 35.30 | 52.70 | 44.89 | 41.71 | 0.00 | -4.57 |
| MolT5 | 43.04 | 45.30 | 39.28 | 52.06 | 51.72 | 46.28 | 4.57 | 0.00 |
| Llama-3 | 56.80 | 47.71 | 68.08 | 54.21 | 60.75 | 57.51 | 15.80 | 11.23 |
| Llama-3.1 | 50.92 | 42.11 | 49.97 | 38.11 | 51.74 | 46.57 | 4.86 | 0.29 |
| Ministral | 59.73 | 50.18 | 70.01 | 51.65 | 55.59 | 57.43 | 15.72 | 11.15 |
| Falcon3 | 57.92 | 51.11 | 53.22 | 42.38 | 47.57 | 50.44 | 8.73 | 4.16 |
| ChemLLM | 50.83 | 52.18 | 47.45 | 53.17 | 49.84 | 50.69 | 8.98 | 4.41 |
| GPT-3.5 | 64.66 | 47.20 | 73.19 | 55.57 | 59.70 | 60.07 | 18.35 | 13.79 |
| GPT-4o | 56.64 | 53.37 | 55.28 | 52.60 | 61.65 | 55.91 | 14.20 | 9.63 |
| **Llama-3.1-70B** | | | | | | | | |
| SMILES | 49.21 | 39.40 | 52.12 | 45.98 | 47.73 | 46.89 | 0.00 | -3.11 |
| MolT5 | 42.02 | 42.86 | 61.94 | 54.66 | 48.48 | 49.99 | 3.11 | 0.00 |
| Llama-3 | 48.34 | 47.54 | 68.08 | 45.73 | 50.95 | 52.13 | 5.24 | 2.14 |
| Llama-3.1 | 53.67 | 43.36 | 37.79 | 56.27 | 48.69 | 47.95 | 1.07 | -2.04 |
| Ministral | 59.85 | 60.62 | 68.31 | 51.22 | 48.28 | 57.66 | 10.77 | 7.67 |
| Falcon3 | 54.12 | 40.02 | 60.75 | 43.24 | 53.02 | 50.23 | 3.34 | 0.24 |
| ChemLLM | 46.11 | 59.21 | 57.43 | 50.18 | 48.85 | 52.35 | 5.47 | 2.36 |
| GPT-3.5 | 52.63 | 47.72 | 54.91 | 52.76 | 44.31 | 50.47 | 3.58 | 0.47 |
| GPT-4o | 53.69 | 62.98 | 50.23 | 56.67 | 53.07 | 55.33 | 8.44 | 5.34 |

Table 10: Full results (ROC-AUC %) of model size experiments. [***Description***] indicates the hallucination source, whether generated by a specific model or from baseline descriptions.

| Model | HIV | BBBP | Clintox | SIDER | Tox21 | Avg |
|---|---|---|---|---|---|---|
| Llama-3.2-1B | 10.26 | 13.06 | 10.54 | 10.43 | 9.74 | 10.81 |
| Llama-3.2-3B | 10.16 | 10.06 | 8.41 | 9.59 | 10.61 | 9.77 |
| Llama-3.1-8B | 7.15 | 8.04 | 7.24 | 7.57 | 7.11 | 7.42 |
| Llama-3.1-70B | 8.37 | 6.82 | 7.31 | 8.60 | 6.62 | 7.54 |

Table 11: HHM-2.1 scores for Llama models of different sizes.

| Temperature | HIV | BBBP | Clintox | SIDER | Tox21 | Avg |
|---|---|---|---|---|---|---|
| 0.1 | 61.82 | 64.85 | 49.77 | 54.13 | 54.93 | 57.10 |
| 0.3 | 61.38 | 50.67 | 52.36 | 54.95 | 54.31 | 54.73 |
| 0.5 | 60.03 | 48.11 | 69.48 | 50.10 | 57.38 | 57.02 |
| 0.7 | 62.85 | 49.84 | 51.36 | 57.27 | 58.95 | 56.05 |
| 0.9 | 61.34 | 45.12 | 56.20 | 48.69 | 59.66 | 54.20 |

Table 12: Full results (ROC-AUC %) of Llama-3.1-8B using hallucination generated by different generation temperatures.

| Temperature | HIV | BBBP | Clintox | SIDER | Tox21 | Avg |
|---|---|---|---|---|---|---|
| 0.1 | 8.02 | 8.25 | 7.52 | 7.81 | 8.07 | 7.93 |
| 0.3 | 8.57 | 8.75 | 6.87 | 7.39 | 8.05 | 7.92 |
| 0.5 | 8.35 | 8.27 | 7.83 | 9.38 | 5.96 | 7.96 |
| 0.7 | 8.10 | 6.88 | 7.37 | 7.57 | 8.71 | 7.73 |
| 0.9 | 6.61 | 6.85 | 6.42 | 7.96 | 6.27 | 6.82 |

Table 13: The HHM-2.1 score for `Llama3.1-8B` generated hallucinations in different temperatures.