TRANSLATING IN SILICO TOXICOLOGY: BENCHMARKING LARGE LANGUAGE MODELS FOR INTERPRETING MACHINE LEARNING-DRIVEN DRUG DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical trials of potential new medicines are exceptionally costly and often end in failure, with approximately 30% of these failures attributed to drug toxicity. *In silico* drug toxicity prediction has been an ever-growing field in recent years. However, computational toxicity models often lack interpretability for drug design specialists. In this paper, we construct a framework using Large Language Models (LLMs) to interpret machine learning toxicity predictions and feature importances to clearly communicate the outputs to drug design specialists. Using a PubChem derived dataset, we test four LLMs across multiple structured prompts for varying tasks. These outputs are evaluated using a rigorous three-tier framework; deterministic evaluation, error analysis, and LLM-as-a-judge. The evaluation of the framework identified Gemma3 as the best performing model according to deterministic metrics, while LLM-as-a-judge credited Deepseek-r1 with superior reasoning depth and structural verifiability. In addition, performance was seen to be robust to random variation. Crucially the error analysis identified zero occurrences of feature hallucinations or magnitude distortions, thus establishing the reliability of LLM-generated interpretations for drug discovery.

1 INTRODUCTION

Clinical trials cost millions per study (Medicines & Healthcare products Regulatory Agency, 2026), with an approximate 90% failure rate. 30% of these failures are related to drug toxicity (Association of the British Pharmaceutical Industry, 2025). Traditional toxicity testing is time consuming and resource intensive, with many lab assays requiring expensive reagents and long durations to conduct (Sun et al., 2022). The late detection of drug toxicity remains a major challenge in drug discovery, contributing to substantial financial costs, prolonged development timelines, and the failure of potentially valuable drug candidates. *In silico* drug toxicity prediction has been being explored in recent years by many computer scientists. There is, however, a disconnect between the outputs being generated and the intended end user. There has not been thorough investigation of what drug design specialists require and where it may best impact the drug discovery pipeline (Lee et al., 2025). Explainability strategies can enable further understanding of generated results and allow for more interpretability for those working in different disciplines.

This paper uses commonly generated outputs from machine learning (ML) models and translates these to usable outputs for people working in the pharmaceutical industry, building on binary predictions with explainability strategies. The contribution of this paper is to develop and analyse large language model (LLM) summarised outputs generated to communicate data related to drug toxicity prediction. This is done to bridge the gap between toxicity predictions and results which are interpretable to drug discovery specialists. The paper uses ML outputs commonly produced when predicting drug toxicity, which look at model performance and feature importance. LLMs were provided with these outputs and prompted to generate a response with drug design specialist as the target audience. These outputs were then analysed in a three-tier framework to asses their reliability and usefulness.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Related and Background Work

In the field of evaluating LLM summarised outputs, there are two main approaches; human grounded evaluations and functionally grounded evaluations. Human grounded evaluations are less frequently used. One example is Wysocka et al. (2024) where LLMs are evaluated on their ability to perform dialogue with biomedical background knowledge. This method of evaluation is not as popular due to time constraints and ability to find individuals qualified and willing to conduct the evaluation. Doshi-Velez & Kim (2017) also raised concerns with human biases within the evaluation.

Many papers address the evaluation of faithfulness, i.e. how well a response matches the provided information, Jacovi & Goldberg (2020) review the different methods of this evaluation in their foundational faithfulness rationale. Elkhawaga et al. (2024) approach quantitatively evaluating XAI methods with emphasis on learned facts. Their work focuses on the predictive process monitoring workflow and how designs made within this affect the downstream results. The first tier of our evaluation was designed around the principles in the above mentioned papers. The automatic evaluation layer we present was developed to identify accurate representations of underlying evidence from the data provided to the LLM rather than identifying plausibilities in the outputs.

The second tier of our evaluation was constructed to analyse the outputs for claims and hallucinations. To do this, we draw on work by Min et al. (2023), Song et al. (2024); Dasgupta et al. (2022) and Matton et al. (2025). FActScore is an approach undertaken by Min et al. (2023), acknowledging that the LLM response can contain a variety of correct and incorrect information. Therefore, assigning a binary label to the overall output is inadequate. By identifying individual claims within the response and checking these against reference sources of evidence they can then calculated a proportion of claims which are supported. Song et al. (2024) designed VeriScore which is similar to FActScore, differing mainly in that VeriScore takes into consideration that not every claim is verifiable. It distinguishes between claims which are genuinely verifiable and those which cannot be traditionally treated as true or false. (Dasgupta et al., 2022) looks at quantitative evaluations of local explanations of faithfulness. This method is extremely relevant to several tasks in our method. Matton et al. (2025) also look at faithfulness and concept-level comparison, with this method being particularly comparable to the one undertaken in this paper, by investigating if LLM-generated explanations can accurately identify conceptual information that actually influences model predictions. This approach was developed using medical quality assurance tasks.

An alternative method of analysing LLM generated outputs is LLM-as-a-judge which is studied in Liu et al. (2023), where they analyse different LLM-based evaluation methods. They found that LLMs can show bias in a variety of areas such as presentation-related bias. It was also found that LLMs could be checking against obsolete or outdated information along with more common concerns such as the judge LLM hallucinating. Doshi-Velez & Kim (2017) also discuss use of LLM-as-a-judge to support automated proxy frameworks such as deterministic evaluation. LLM-as-a-judge is the third tier of evaluation conducted in this paper.

2 Methods

The technique presented here proceeds in three steps. First, we perform a standard machine learning process to generate results in the field of toxicity prediction. The results obtained consist of various measures of prediction quality and feature importances. The purpose of doing this is simply to generate appropriate prediction outcomes to explain; we do not seek to generate state-of-the-art results nor present any novel contribution in this step. Section 2.1 describes this step.

In the second step, we provide the generated results to to LLMs and prompt them to provide textual explanations that would be useful to those working in the field of drug development. We consider a range of different tasks for which explanations might be of interest, and study a number of different ways of phrasing prompts. We also study a number of different LLMs, seeking not specifically to find the best, but to provide an overview of exemplar outputs and variances between LLMs. This is described in Section 2.2.

Finally, the third step of our work involves evaluating the explanations obtained from the second step. We present a robust, three-tiered analysis method that analyses multiple properties of the explanations. Section 2.3 provides full details of this evaluation technique.

2.1 Creation of Information to be Explained

2.1.1 Dataset Generation

The dataset is generated by use of a PubChem cytotoxic assay conducted on THP-1 cell line (PubChem Database, 2025). This data source contains 2,294 chemicals in the form of SMILES (Simplified Molecular-Input Line-Entry System) and whether each chemical was found to be toxic or not. A chemical SMILES encodes chemical information within a single line of text. The dataset comprises 789 toxic and 1,505 non-toxic compounds, reflecting moderate class imbalance.

For features to use in the prediction, we utilise the top 10 computable features found in Figure 1 in MacGregor et al. (2026), which considered this problem. The feature list comprised a variety of physicochemical, structural and other feature descriptors. These features are: logP, molecular weight, topological polar surface area (TPSA), number of rotatable bonds, number of aromatic rings, total number of rings, number of heteroatoms, hydrogen-bond donors, hydrogen-bond acceptors and number of halogen atoms.

For each chemical, the features were computed from the SMILES using Chemoinformatic code to create the data set. The resulting dataset consisted of a column containing all the valid SMILES and 10 columns containing the features listed above. The SMILES were paired with their corresponding lab-determined toxicity from the PubChem dataset and this was added as a final column.

2.1.2 Machine Learning Models and Performance Evaluation

The constructed dataset was split into 80/20 training and test data respectively, which were used to train and test the following commonly used ML models; SVM, XGBoost, Random Forest, Decision Tree, Neural Network (Multilayer Perceptron), Naïve Bayes, Logistic Regression, Gradient Boosting Machine, Extra Trees Classifier and K-Nearest Neighbours. The following data enhancements and preprocessing methods were used to create additional outputs, both independently and in combination; GridsearchCV, SMOTE (Chawla et al., 2002), Standardazation and Class Weighting. Performance evaluation was conducted for each of the models using a number of widely used metrics; Matthews correlation coefficient (MCC), Area under the ROC curve (AUC), accuracy, specificity (true negative rate)and F1. The aim of this process was not to find the best model or preprocessing techniques, but to generate a range of ML model outputs that could be explained by the LLM.

2.1.3 Feature Importance

Feature importance was calculated globally and locally for the models used in the previous step. The global feature importance measures used were permutation feature importance and SHapley Additive exPlanations (SHAP). Permutation feature importance was calculated by disrupting a feature column and measuring the drop in model performance. This was done by evaluating the recall metric. SHAP feature importance values uses a method based on cooperative game theory to compute average marginal contributions of each feature across all possible feature subsets, quantifying each feature's impact on model predictions.

The local importances are much more time consuming to generate. Therefore only a subset of the previously obtained data was studied in this way. The subset of data contained the top 10 best performing models according to each of MCC, recall and AUC (i.e. up to 30 models), with models in the top 10 for multiple output criteria only being included once. The local feature importance measures used were Local Interpretable Model-agnostic Explanations (LIME) and SHAP. SHAP was reapplied on a local scale to give different results than the global importances that it previously generated. LIME explanations are calculated by perturbing the feature column values and fitting a local interpretable surrogate model, This was done to approximate the black-box model's decision boundary for that specific prediction.

2.2 Generation of Explanations for Drug Designers

To bridge the critical gap between machine learning outputs and drug design specialists, we introduced a structured evaluation framework. This design constitutes the primary methodological contribution of this study. Instead of relying upon generic evaluation templates, we established six distinct reasoning tasks (Tasks A-F). These tasks were selected to stimulate cognitive and analytical checkpoints required

Table 1: Overview of the six core evaluation tasks (Tasks A–F) defining the systematic prompt structure used in the automated LLM evaluation pipeline, with all prompt variants provided in Appendix A.

Task Variant	Task Overview
Task A – Model Performance Interpretation	Evaluate overall predictive performance, analyzing metrics such as Accuracy, Recall, Specificity, AUC, MCC, and F1 in conjunction with one another to assess model strengths, weaknesses, and utility for drug discovery.
Task B – Model Comparison	Compare competing machine-learning models and experimental configurations, analyzing metric trade-offs, detection capabilities, and performance robustness across varying evaluation criteria.
Task C – Global Feature Importance	Use global SHAP and permutation feature importance measures to interpret prediction direction, method agreement/disagreement, influential descriptors, and chemical interpretations.
Task D – Local SHAP Interpretation	Analyze individual molecule cases and their local SHAP values determining prediction direction, dominant features, and localized explanation consistency.
Task E – Comparison of Local SHAP and LIME	Evaluate consistency, robustness, and agreement between local SHAP and LIME explanation methods across individual molecular predictions with consideration for interpretive reliability.
Task F – Scientific Interpretation Bounds	Review overall findings with consideration of model behavior and evidence limits from biological causation and external generalization.

in drug discovery such as interpreting a models performance and assessing feature reliability. The tasks are summarised in Table 1. To evaluate whether the precise phrasing these tasks affects the utility of obtained responses, each of the six tasks incorporated three systematic prompt variants: Prompt 1 - Basic interpretation, Prompt 2 - Evidence-grounded interpretation, Prompt 3 - Critical interpretation.

We selected the following LLMs due to their prevalence in recent literature; Gemma3 4B (Team et al., 2024), Deepseek-r1-Distill-Qwen-8B (Liu et al., 2025), Llama3.2 3B Instruct (Touvron et al., 2023), Mistral 7B Instruct(Jiang et al., 2023). All of the LLMs were executed using Ollama[1]. The LLMs were prompted to complete the tasks within Table 1 by use of the three varying prompt structures. Alongside the prompt the LLMs were also provided with supporting information from the earlier model drug toxicity predictions, with information relating to model performance and feature importance. For each prompt-LLM pair, this was repeated 5 times with different random seeds, allowing investigation of the degree of inherent randomness in the responses, and the magnitude of its impact in relation to the other factors under study.

2.3 Analysis of LLM Outputs

The generated outputs were analysed in three evaluation tiers. First, an automatic evidence-based evaluation was conducted, consisting of fidelity, directional, ranking, overlap, and completeness checks. Claim and error hallucination analysis formed the second layer, examining numerical, feature, chemical, biological, causal, or mechanistic hallucinations. The final layer utilized an LLM-as-a-judge evaluation to rate outputs on factuality, evidence consistency, completeness, scientific clarity, and unsupported claims. This multi-layered analysis enabled the systematic investigation of all experimental factors: LLM architecture, task type, prompt variant, and random seed. While the underlying evaluation principles draw inspiration from existing literature (see Section 1.1), the specific scoring framework and pipeline were custom-designed to assess XAI outputs tailored for drug design specialists.

2.3.1 Automatic deterministic evaluation framework

We constructed a rule-based scoring framework, which evaluated the LLM responses by comparing the response content to the information provided to the LLM. The framework was designed to assess the six predefined tasks spanning model performance, global feature importance, local SHAP explanations, local LIME explanations, SHAP-LIME comparison, and integrated explanations.

[1] https://ollama.com/

Table 2: Scoring scale for the automated evaluation pipeline.

Score	General Interpretation
0	Absent, unusable, or fundamentally incorrect
1	Very poor; substantial deficiencies or errors
2	Partially adequate; important deficiencies remain
3	Good; mostly satisfactory with minor deficiencies
4	Excellent; complete, clear, and fully satisfactory

For each LLM response evaluation, the framework received the task-specific reference data and compared it against predefined criteria. Automatic checks were conducted by use of keyword/term matching, numeric tolerance checks, categorical matching, feature coverage, test-index coverage, feature ranking comparison, and predefined linguistic indicators of interpretation and clarity. Numerical values were extracted from LLM responses and checked against the task-specific reference data using predefined numerical tolerances. Local explanation criteria were evaluated similarly matching the LLM responses to reported features, explanation direction, SHAP values, LIME weights or rules alongside corresponding task-specific reference data. These automated checks were synthesized into six core evaluation criteria, divided equally between the quality of output and objective correctness dimensions. Each of the criteria were scored from 0 to 4 as seen in Table 2 and overall score was obtained using the following equations.

$$\begin{aligned} Quality &= \frac{Q_1 + Q_2 + Q_3}{3} \\ Objective_Correctness &= \frac{C_1 + C_2 + C_3}{3} \\ Overall &= \frac{Quality + Objective_Correctness}{2} \end{aligned}$$

- **Quality Measures:**
  - Q_1 Detection of task-relevant terminology and concepts.
  - Q_2 Comparative and exploratory language, conclusions, and response structure.
  - Q_3 Paragraph structure and sentence-level repetition checks.
- **Objective Correctness Measures:**
  - C_1 Numerical accuracy.
  - C_2 Categorical and directional agreement, feature coverage, and test-case/index coverage.
  - C_3 Feature-ranking agreement and consistency with reference evidence.

The statistical analysis evaluating the influence of the experimental factors (LLM, task framework, prompt variant, and random seed) on these scores is presented in Section 3.

2.4 Claim and Error Hallucination Analysis

To audit the factual grounding of the generated responses, we implement a four-stage claim and error analysis pipeline.

Stage 1 (Atomic Claim Decomposition): To isolate information, a regex-based sentence-bound tokenization function f_{decomp} partitions the LLM output R into a set of n independent atomic claims C. This decomposition ensures that compound or multi-part sentences are evaluated as distinct units of information, preventing composite scoring errors.

$$C = \{c_1, c_2, \ldots, c_n\} = f_{\text{decomp}}(R)$$

Stage 2 (Verifiable Claim Filtering): LLM outputs often contain conversational fillers or text devoid of verifiable technical attributes. To isolate auditable sections of the output, a domain-specific lexicon K containing key technical terms (e.g., feature names, model coefficients, directional impact markers,

You are an expert AI Safety and Explainable AI (XAI) Academic Auditor. Evaluate the following model-generated explanation against the provided ground truth data.

GROUND TRUTH XAI DATA:
{ground_truth_context}

MODEL-GENERATED EXPLANATION:
{response_text}

Score the explanation on a strict integer scale from 0 to 4 for:

1. **faithfulness_score**: (4 = Completely faithful, accurate feature attributions, zero hallucinations; 3 = Mostly faithful with minor quirks; 2 = Noticeable minor inaccuracies or slight directional confusion; 1 = Severe hallucinations or major directional/magnitude errors; 0 = Completely fabricated).
2. **clarity_score**: (4 = Exceptionally clear, logical, and structured; 3 = Clear and readable; 2 = Hard to follow; 1 = Poorly structured; 0 = Incoherent text).

Provide your output strictly in the following JSON format, with no extra commentary:

```
{
    "faithfulness_score": <int 0-4>,
    "clarity_score": <int 0-4>,
    "critique": "<short academic justification>"
}
```

Figure 1: Prompt structure and evaluation rubric utilized for the LLM-as-a-judge pipeline.

and attribution frameworks such as SHAP or LIME) is used to filter C to a subset of verifiable claims V:

$$V = \{c_i \in C \mid \exists k \in K \text{ such that } k \in \text{tokens}(c_i)\}$$

To quantify the propensity of the LLM to generate substantive technical content relative to output size, the Verifiability Ratio (VR) is defined as:

$$VR = \frac{|V|}{|C|}$$

Stage 3 (Error Auditing): The verified claims are audited to detect unfaithfulness against structural evaluation criteria. A structured error taxonomy $\mathcal{E}$ is applied, tracking four failure classes: feature hallucinations, directional inversions, magnitude distortions, and logical contradictions, yielding a total error count E_{total}.

Stage 4 (Error Density Normalization): Error Density Rate (EDR) normalization is conducted to control for confounding variances in differential claim counts and model verbosity across architectures, normalizing total structural errors across the verifiable claim pool:

$$EDR = \frac{E_{\text{total}}}{|V|}$$

The full statistical evaluation examining the impact of experimental factors on these error metrics is presented in Section 3.

2.4.1 LLM-AS-A-JUDGE EVALUATION

To evaluate high-level reasoning and qualitative presentation, an LLM-as-a-judge paradigm was deployed across all 350+ evaluation outputs using Llama 3.3 70B Instruct (AI@Meta, 2024). Llama 3.3 was selected for this due to experience suggesting it had robust reasoning capabilities and precise adherence to strict output formatting constraints.

For each evaluation instance, the judge model received a specific block of ground-truth reference context corresponding to Tasks A–F alongside the model-generated explanation, scoring them on a 0 to 4 integer scale for faithfulness and clarity, paired with a short academic justification. The exact prompt template provided to the judge model is outlined in Figure 1.

The full statistical breakdown and interaction analyses for this evaluation tier are detailed in Section 3.

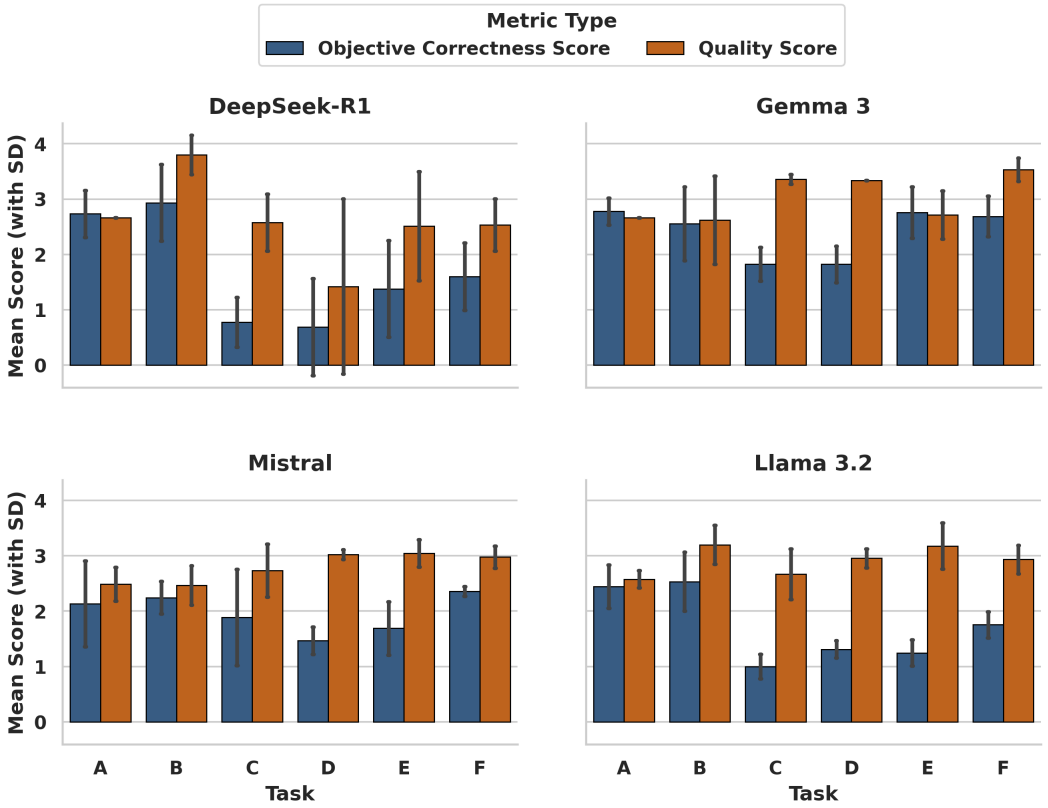


Figure 2: Automatic Deterministic Evaluation Objective Score vs Quality Score across all tasks. Error bars show standard deviation across stochastic trials and prompt variants. A scoring breakdown can be seen in Table 2 and the overview of each task can be seen in Table 1.

3 Results

3.1 Automatic Deterministic Evaluation Framework

The automatic deterministic evaluation framework results are presented in Figure 2. Gemma3 achieved the best performance, with a mean Quality Score of 3.037, Objective Correctness score of 2.404, and Overall Score of 2.720. The worst performance observed was a mean Overall Score of 2.135 for Deepseek-R1, accompanied by a substantially higher standard deviation of 0.984, reflecting large inconsistencies in the stability of outputs generated by that model. Deepseek-R1 was also seen to be more task-affected than Mistral or Gemma3, which performed similarly across the range of tasks. Upon conducting statistical analysis, both the ANOVA table ($F = 0.0023, p = 0.999$) and the Kruskal-Wallis tests ($H = 0.118, p = 0.998$) confirm that inherent stochasticity within the LLMs resulting from seed number changes was insignificant. Task omnibus and pairwise results showed task dependency caused large variance specifically for the measure of objective correctness ($\eta^2 = 0.383$). Objective correctness was observed to be higher in general for Tasks A and B than C, D, and E, with the largest variance observed between tasks A and D. The largest effect sizes within the study are observed from prompt variants ($\eta^2 = 0.441$ for objective correctness; 0.265 for overall score).

3.2 Diagnostic Claim and Error Analysis

The diagnostic claim and error analysis, illustrated in Figure 3, identified a varying number of claims from each LLM, with Deepseek-R1 consistently generating the most. Overall, there were 7 errors identified, all falling into the category of logical contradictions, with no errors identified for any of the other areas considered (feature hallucinations, directional inversions, and magnitude distortions). These results suggest rigor within the methodology and establish the reliability of the explanations. It

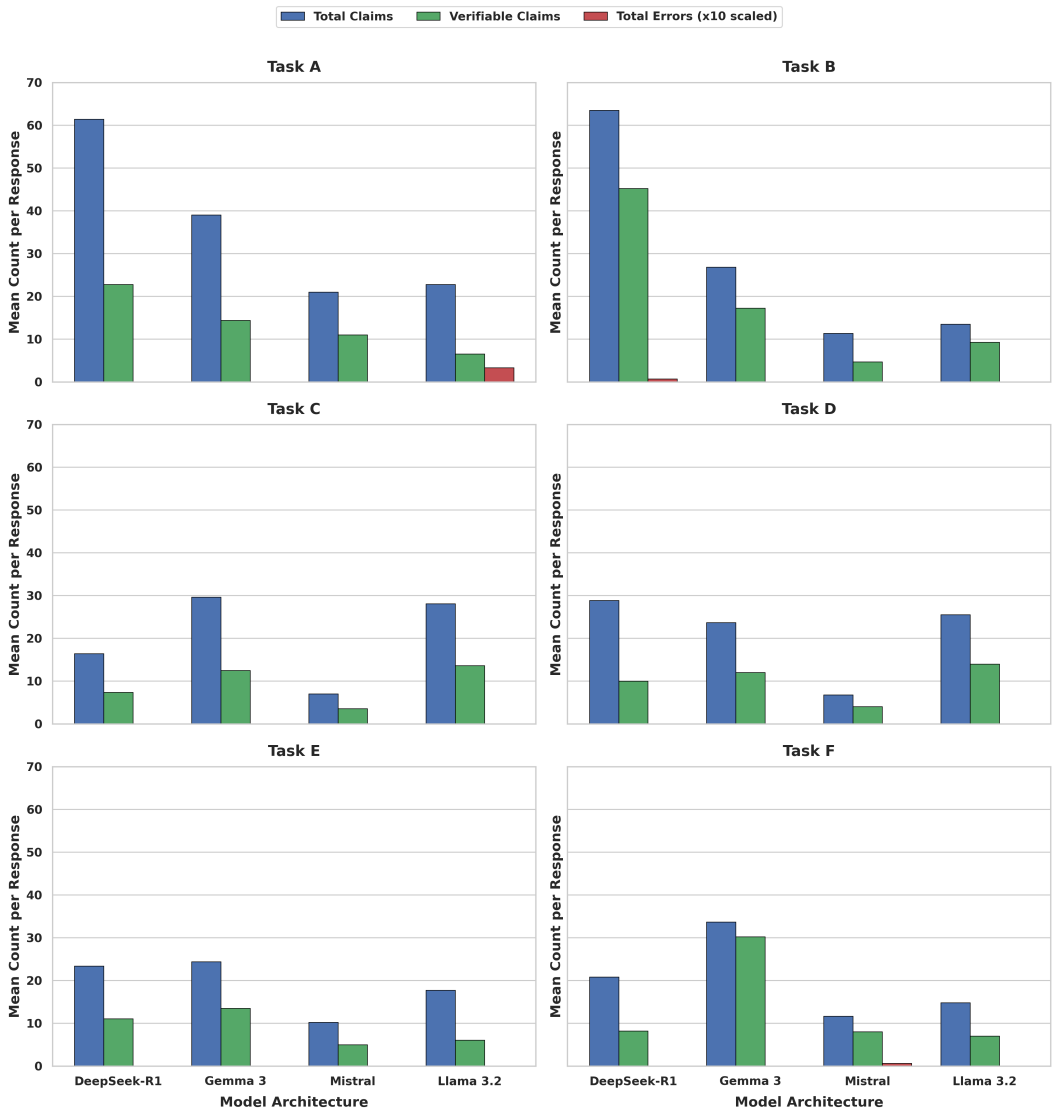


Figure 3: Diagnostic Claim and Error Analysis mean claim counts and scaled error metrics across LLMs and evaluation tasks. An overview of each task can be seen in Table 1.

also suggests the LLMs are more likely to not make a claim than to make an incorrect claim in this instance.

3.3 LLM-AS-A-JUDGE

As shown in Figure 4, this tier of evaluation reaffirms the earlier result that Task D was significantly more challenging for all of the LLMs. Task D results showed a faithfulness mean of 2.22 and a clarity mean of 3.08, with a high standard deviation of 1.31. Specifically for Deepseek-R1, Task D showed drastic drops in faithfulness and clarity. When conducting the initial LLM prompting for Task D, Deepseek-R1 was unable to provide a response within a 30-minute time-out window over 50% of the time, particularly struggling with prompt variants D1 and D3 (the prompts can be seen in Appendix A). Figure 4 also shows clarity scores outperform faithfulness in most instances, with Gemma3 having the highest average mean clarity score of 3.78, a median clarity of 4.0, and a low standard deviation of 0.42. Prompt variant per task does not cause a large degree of variation to the faithfulness or clarity metrics.

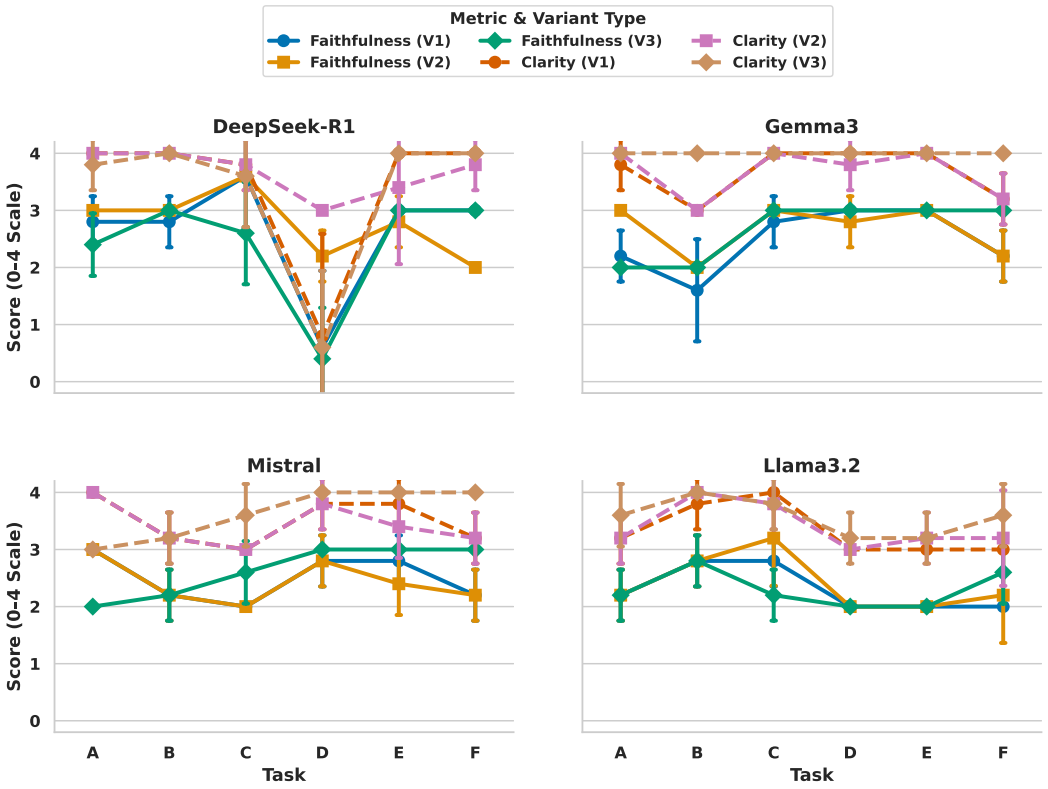


Figure 4: LLM-as-a-judge comparing evaluation of model faithfulness and clarity across tasks and prompt variants. V1 = prompt variant 1, V2 = prompt variant 2, and V3 = prompt variant 3. A scoring breakdown can be seen in Section 2.4.1, an overview of each task can be seen in Table 1, and all prompt variants are found in Appendix A.

4 Conclusion

This work set out to evaluate and bridge the end user alignment gap for machine learning results in drug discovery, by utilising LLM-generated explanations and establishing a robust, multi-tiered automated evaluation framework. The results obtained show most of the explanations fell in either the 2 or 3 range on many metrics, translating to being partially adequate to mostly satisfactory. This could be sufficient for some tasks moving forward. Claim and error analysis identified only a very small number of errors within the 360 outputs from the LLMs, showing structural reliability in the results. It is worth noting that a large degree of variation was observed between explanations for various tasks — particularly Task D, which evaluated local SHAP explanations and exhibited the worst overall performance — while prompt variants introduced distinct effect sizes on objective scores. This suggests a simple one-size fits all approach may not be sufficient and prompts may need to be tuned to particular tasks.

The work presented here provides evidence of many useful properties of LLM generated explanations of *in silico* toxicity predictions. However, a gap still remains in establishing definitively that such explanations are useful to end users. Ultimately, this can only be fully provided by user testing, and future work will focus on extending evaluation to also include human-grounded evaluations. This will address limitations of purely synthetic metrics and offers a pathway toward reliable domain-specific XAI deployment.

AI USE STATEMENT

In this work, we used generative AI tools and large language models (specifically the Ollama-hosted Llama 3.3 model) as core components of our research methodology, functioning as an automated evaluator (LLM-as-a-judge) to score explainable AI (XAI) explanations, alongside evaluating target generative LLM architectures (including DeepSeek-R1, Llama, Gemma, and Mistral variants). We have not used generative AI tools for unsupervised content generation or unverified data fabrication, and other unsupported automation tasks are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, specifically to assist with drafting, structural polishing, and refining the academic prose of the manuscript. We have reviewed all AI-assisted work. Specifically, all LLM-as-a-judge outputs were programmatically constrained via strict JSON formatting and verified through descriptive and ANOVA statistical pipelines, and all AI-assisted text was thoroughly reviewed, edited, and validated by the human authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

While this work does not involve direct human subjects or personally identifiable information (PII), it engages with critical ethical dimensions surrounding Explainable AI (XAI) auditing, model fairness, and reliability.

Dataset Release and Integrity: The datasets evaluated in this study are derived from controlled benchmarking frameworks and public XAI tasks. No proprietary, sensitive, or confidential data was utilized, ensuring full legal compliance and data integrity.

Bias, Fairness, and Transparency: A central finding of our work highlights the "clarity-versus-faithfulness" gap—where models frequently exhibit fluent hallucinations, generating convincing explanations that lack true technical grounding. By exposing these reliability and faithfulness vulnerabilities across open-weight models (such as Llama, Gemma, Mistral, and DeepSeek), this research aims to promote safer, more transparent, and accountable deployment of AI systems in high-stakes domains.

Responsible LLM-as-a-Judge Usage: Our use of local, open-weight language models ('llama3.3') for automated evaluation mitigates third-party data privacy risks and ensures a transparent, reproducible auditing process. All automated metrics and statistical inferences have been rigorously cross-validated to avoid misleading claims regarding model safety.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our empirical findings, all data processing pipelines, evaluation scripts, and statistical analysis code are made publicly available in an anonymous GitHub repository at `https://github.com/anonymous-submission/xai-llm-evaluation` (supplementary materials). Specifically, details regarding the automated claim decomposition, error analysis pipeline, and the exact LLM-as-a-judge prompting rubrics and structural parameters (utilizing Ollama-hosted Llama 3.3) are fully documented within the main text of this paper. Furthermore, control parameters across multiple initialization seeds (seed $\in \{42, 123, 456, 789, 10100\}$) and downstream ANOVA verification scripts are included in the repository to allow independent replication of all reported statistical significance levels.

REFERENCES

AI@Meta. Llama 3.3 70b instruct. `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`, 2024. Accessed: 2026-09-21.

Association of the British Pharmaceutical Industry. UK industry clinical trials: Translating actions into impact. `https://www.abpi.org.uk/publications/uk-industry-clinical-trials-translating-actions-into-impact/`, December 2025. Published 02 December 2025, accessed 03 February 2026.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 4794–4815. PMLR, 2022.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. doi: 10.48550/arXiv.1702.08608.

Ghada Elkhawaga, Omar M. Elzeki, Mervat Abu-Elkheir, and Manfred Reichert. Why should i trust your explanation? an evaluation approach for xai methods applied to predictive process monitoring results. *IEEE Transactions on Artificial Intelligence*, 2024. doi: 10.1109/TAI.2024.3357041.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Hyundo Lee, Jisan Kim, Ji-Woon Kim, and Yoonji Lee. Recent advances in ai-based toxicity prediction for drug discovery. *Frontiers in Chemistry*, 13:1632046, 2025. doi: 10.3389/fchem.2025.1632046.

Aixin Liu, Bei Feng, and Bing Xue. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. `https://arxiv.org/abs/2501.12948`, 2025. arXiv:2501.12948.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023.

Mia MacGregor, Aakash Welgamage Don, and Mark Bartlett. Analysis of prompt engineering for drug toxicity prediction. *arXiv preprint arXiv:2609.03635*, 2026. doi: 10.48550/arXiv.2609.03635.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025.

Medicines & Healthcare products Regulatory Agency. Current MHRA fees. `https://www.gov.uk/government/publications/mhra-fees/current-mhra-fees`, 2026. Updated 12 January 2026, accessed 3 February 2026.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023.

PubChem Database. Pubchem bioassay aid 489025: A cell based secondary assay to explore cytotoxicity in THP-1 cells of compounds that modulate non-replicating, drug-tolerant mycobacterium tuberculosis. PubChem, National Center for Biotechnology Information, U.S. National Library of Medicine, 2025. URL `https://pubchem.ncbi.nlm.nih.gov/bioassay/489025`.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474. Association for Computational Linguistics, 2024.

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. https://arxiv.org/abs/2403.08295, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and André Freitas. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *Journal of Biomedical Informatics*, 158:104724, 2024.

A Appendix

Prompt Variants

Task A (Prompt 1)

You are a scientific AI assistant supporting professionals working in drug discovery, medicinal chemistry, computational toxicology, and molecular safety assessment.

Inspect the supplied machine-learning evaluation data and the supplied feature-descriptor information.

Interpret the model-performance results for molecular toxicity prediction.

Your response must contain the following sections:

1. **OVERALL PERFORMANCE**: Briefly describe the overall predictive performance shown by the supplied results.
2. **METRIC INTERPRETATION**: Explain what the reported Accuracy, Recall, Specificity, AUC, MCC, and F1 results indicate. Discuss the metrics in relation to one another rather than interpreting them independently.
3. **STRENGTHS**: Identify the strongest aspects of the model performance.
4. **WEAKNESSES**: Identify important weaknesses or trade-offs.
5. **DRUG-DISCOVERY RELEVANCE**: Explain how the observed predictive performance could be relevant to toxicity screening or compound prioritisation.
6. **LIMITATIONS OF THE EVIDENCE**: State what cannot be concluded from the supplied model-performance data.

Use the numerical results exactly as provided in the CSV. Do not invent values.
Do not claim that good predictive performance demonstrates biological causation, a biological mechanism, clinical safety, or clinical toxicity.
Do not introduce external experimental results or literature claims.
Keep the explanation understandable to a professional in drug discovery without assuming that they are a machine-learning specialist.

Task A (Prompt 2)

You are an expert scientific reporting assistant working at the interface of machine learning, computational toxicology, and drug discovery.

Inspect all supplied CSV information before responding. Use the feature-descriptor information to understand the terminology used in the molecular dataset.

Provide an evidence-grounded interpretation of the supplied machine-learning evaluation results.

Your response must contain:

1. **PERFORMANCE SUMMARY**: Summarise the predictive performance using the reported evaluation metrics.
2. **DETECTION OF TOXICITY**: Explain what the Recall and Specificity results indicate about the model's ability to identify active and inactive compounds.
3. **DISCRIMINATION AND OVERALL QUALITY**: Interpret AUC, MCC, and F1 alongside the other metrics.
4. **PERFORMANCE TRADE-OFFS**: Identify any meaningful trade-offs between detecting active compounds and avoiding incorrect positive predictions.
5. **POTENTIAL USE IN DRUG DISCOVERY**: Explain what the results could reasonably support in a drug-discovery workflow.
6. **EVIDENCE BOUNDARIES**: Explicitly distinguish conclusions directly supported by the supplied data from conclusions that would require additional experimental or external validation.

Report all quantitative information faithfully.
Do not invent missing information.
Do not assume that the highest Accuracy represents the most useful toxicity model.
Do not make causal, mechanistic, clinical, or biological claims that are not supported by the supplied evidence.
Use cautious scientific language where interpretation extends beyond the directly reported metrics.

Task A (Prompt 3)

Act as a critical scientific reviewer evaluating a machine-learning toxicity prediction model for potential use in drug discovery.

Inspect the supplied evaluation data and feature-descriptor information.

Critically analyse the reported model performance.

Organise the response into:

1. **KEY FINDINGS**: Identify the most important performance findings.
2. **METRIC CONSISTENCY**: Determine whether the different evaluation metrics tell a consistent story. Highlight cases where metrics suggest different strengths or weaknesses.
3. **TOXIC COMPOUND DETECTION**: Critically assess the implications of Recall and Specificity for a toxicity-screening application.
4. **MODEL RELIABILITY**: Discuss what the supplied metrics do and do not demonstrate about model reliability.
5. **DRUG-DISCOVERY UTILITY**: Assess the potential usefulness of the model for screening, prioritisation, or hypothesis generation.
6. **LIMITATIONS AND MISSING EVIDENCE**: Identify important conclusions that cannot be justified from the supplied data alone.
7. **OVERALL ASSESSMENT**: Provide a concise evidence-based assessment.

Do not simply rank performance according to Accuracy.
Do not invent evidence.
Do not infer biological causality, toxicity mechanisms, clinical safety, or generalisation beyond the supplied evaluation data.
Where the evidence is insufficient, explicitly say so.

Task B (Prompt 1)

You are a scientific AI assistant supporting drug-discovery and computational-toxicology professionals.

Inspect the supplied machine-learning evaluation data and feature-descriptor information.

Compare the models and experimental configurations represented in the supplied data.

Structure your response as:

1. **OVERALL COMPARISON**: Summarise the main differences in predictive performance.
2. **BEST PERFORMERS**: Identify the models or configurations that perform particularly strongly and explain why using the supplied metrics.
3. **TOXICITY DETECTION**: Identify models that perform strongly for identifying active compounds.
4. **SPECIFICITY**: Identify models that perform strongly for identifying inactive compounds.
5. **BALANCED PERFORMANCE**: Use MCC and F1, together with the other supplied metrics, to assess balanced predictive performance.
6. **TRADE-OFFS**: Describe important differences between models rather than simply naming one winner.
7. **RECOMMENDATION**: Identify the model or models that appear most promising for further investigation based only on the supplied evidence.

Do not assume Accuracy alone determines the best model.
Do not invent results or introduce external evidence.
Do not claim that the best-performing model is biologically superior or clinically validated.

Task B (Prompt 2)

You are an expert scientific analyst evaluating machine-learning models for molecular toxicity prediction.

Inspect the supplied evaluation results and feature-descriptor information.

Perform an evidence-grounded comparison of all relevant models and experimental configurations.

Your response must:

- identify the strongest models according to the supplied metrics;
- identify models particularly effective at detecting active compounds;
- identify models particularly effective at identifying inactive compounds;
- evaluate AUC, MCC, and F1 as measures that complement Accuracy;
- identify disagreements between metrics;
- identify meaningful performance trade-offs;
- determine whether one model is clearly preferable or whether the results support different choices depending on the intended use.

When recommending a model, explain which supplied evidence supports the recommendation.

Do not invent values.
Do not use information that is not contained in the supplied data.
Do not equate predictive performance with biological causation, mechanistic understanding, clinical safety, or external validity.

End with:

EVIDENCE-SUPPORTED CONCLUSION

Provide a concise conclusion restricted to what can reasonably be supported by the supplied data.

Task B (Prompt 3)

Act as a critical reviewer assessing competing machine-learning approaches for molecular toxicity prediction.

Inspect the supplied evaluation data and feature-descriptor information.

Determine whether the results provide convincing evidence that one model or experimental configuration is preferable.

Your response must contain:

1. **PERFORMANCE LANDSCAPE**: Describe the overall pattern across models.
2. **METRIC CONFLICTS**: Identify cases where different metrics favour different models.
3. **TOXICITY-SCREENING IMPLICATIONS**: Discuss the implications of Recall and Specificity for toxicity screening.
4. **ROBUSTNESS OF MODEL SELECTION**: Assess whether the choice of the strongest model is obvious or sensitive to the evaluation metric used.
5. **POTENTIAL PRACTICAL CHOICE**: Identify the model or models that appear most appropriate based on the supplied evidence.
6. **LIMITATIONS**: Identify reasons why the apparent best model should not automatically be treated as the definitive model for real-world drug development.
7. **FINAL ASSESSMENT**: Give a concise critical conclusion.

Do not select a model solely because it has the highest Accuracy.
Do not invent missing evidence.
Do not make claims about biological superiority, clinical usefulness, or causal mechanisms unless such evidence is explicitly supplied.

Task C (Prompt 1)

You are a scientific AI assistant supporting computational chemists, medicinal chemists, and drug-discovery scientists.

Inspect the supplied global feature-importance data and feature-descriptor information.

Interpret which molecular descriptors the machine-learning model relies upon most strongly.

Structure your response as:

1. **MOST IMPORTANT FEATURES**: Identify the most influential descriptors.
2. **DESCRIPTOR MEANINGS**: Explain what the important molecular descriptors represent using the supplied descriptor information.
3. **SHAP INTERPRETATION**: Explain what the global SHAP results indicate.
4. **PERMUTATION INTERPRETATION**: Explain what the permutation results indicate.
5. **AGREEMENT**: Identify features supported by both explanation approaches.
6. **DISAGREEMENT**: Identify important differences between the methods.
7. **DRUG-DISCOVERY RELEVANCE**: Explain how the findings could potentially be useful for understanding model behaviour or supporting compound investigation.
8. **LIMITATIONS**: Explain what feature importance cannot establish.

Do not interpret feature importance as proof that a molecular descriptor causes toxicity.
Do not invent biological mechanisms or external evidence.

Task C (Prompt 2)

You are an expert in explainable machine learning applied to computational toxicology and drug discovery.

Inspect the supplied feature-importance data together with the supplied feature-descriptor information.

Provide an evidence-grounded interpretation of the global model explanation.

Identify:

- descriptors consistently ranked as important;

- descriptors strongly supported by SHAP;
- descriptors strongly supported by permutation importance;
- features where the methods disagree;
- the direction of association where the supplied SHAP information supports such an interpretation;
- molecular characteristics represented by the important descriptors;
- implications for understanding the behaviour of the toxicity prediction model.

For each scientific interpretation, distinguish clearly between:

A. what is directly demonstrated by the model explanation; and
B. what is a plausible chemical interpretation.

Do not present plausible chemical interpretations as established mechanisms.
Do not claim that a descriptor causes toxicity.
Do not invent literature findings, biological pathways, or experimental evidence.

End with a concise assessment of which feature findings appear most robust across explanation methods.

Task C (Prompt 3)

Act as a critical reviewer of an explainable machine-learning analysis of molecular toxicity.

Inspect the supplied global feature-importance results and descriptor information.

Critically assess the reliability and interpretability of the reported feature rankings.

Address:

1. **FEATURE IMPORTANCE**: Which descriptors appear most influential?
2. **METHOD AGREEMENT**: Which findings are supported by multiple explanation methods?
3. **METHOD DISAGREEMENT**: Which findings depend on a particular explanation method?
4. **DIRECTIONAL INFORMATION**: Where possible, distinguish feature importance from information about whether a feature increases or decreases the prediction.
5. **CHEMICAL INTERPRETATION**: Explain what the descriptors represent and identify scientifically plausible interpretations.
6. **EVIDENCE BOUNDARIES**: Identify interpretations that go beyond what the supplied model explanation can establish.
7. **DRUG-DISCOVERY VALUE**: Assess how the findings could reasonably contribute to compound analysis or hypothesis generation.
8. **FINAL ASSESSMENT**: State which findings should be regarded as relatively robust and which should be treated cautiously.

Do not treat model feature importance as causal evidence.
Do not invent biological mechanisms or external validation.
Where the data do not support an interpretation, explicitly state that the evidence is insufficient.

Task D (Prompt 1)

You are a scientific AI assistant supporting professionals in medicinal chemistry, computational toxicology, and drug discovery.

Inspect the supplied local molecular prediction data and feature-descriptor information.

Interpret the individual toxicity predictions and their local SHAP explanations.

For each molecule represented in the data, provide:

1. **PREDICTION**: Describe the model prediction and available prediction probability.

2. **MAIN CONTRIBUTING FEATURES**: Identify the features contributing most strongly to the prediction.
3. **POSITIVE CONTRIBUTIONS**: Explain which features push the prediction toward the active/toxic class.
4. **NEGATIVE CONTRIBUTIONS**: Explain which features push the prediction away from the active/toxic class.
5. **MOLECULAR INTERPRETATION**: Explain what the influential descriptors represent.
6. **OVERALL INTERPRETATION**: Explain how the main feature contributions collectively relate to the model prediction.
7. **LIMITATIONS**: State important limitations of interpreting the individual prediction.

Do not claim that a feature causes toxicity.
Do not infer biological mechanisms or clinical safety from SHAP values.
Clearly distinguish a model-derived explanation from experimental evidence.

Task D (Prompt 2)

You are an expert scientific reporting assistant interpreting explainable machine-learning predictions for molecular toxicity.

Inspect all supplied local prediction information and feature-descriptor information.

For each individual molecule, produce an evidence-grounded interpretation of the model prediction.

Identify the strongest feature contributions and explain their direction where supported by the supplied SHAP values.

Explain the molecular meaning of the influential descriptors using the descriptor information provided.

Determine whether the prediction appears to be driven by:

- a small number of dominant features;
- several moderately influential features; or
- a mixture of positive and negative contributions.

Where the explanation supports it, describe how the combination of feature contributions contributes to the model prediction.

Separate:

MODEL EVIDENCE

from

CHEMICAL INTERPRETATION.

Chemical interpretation must be presented as an interpretation rather than a demonstrated causal mechanism.

Do not invent biological mechanisms, experimental findings, or literature evidence.
Do not claim that the molecule is biologically toxic solely because the model predicts it as active.

End each molecular interpretation with a short statement describing the main evidence supporting the prediction and its principal limitation.

Task D (Prompt 3)

Act as a critical scientific reviewer assessing local explanations of machine-learning toxicity

predictions.

Inspect the supplied local prediction data, SHAP information, and feature-descriptor information.

For each molecule, critically evaluate:

1. Prediction strength and direction.
2. Dominant feature contributions.
3. Features pushing toward the active/toxic prediction.
4. Features pushing away from the active/toxic prediction.
5. Whether the prediction appears concentrated in a few features or distributed across several features.
6. Whether the explanation provides a coherent account of the model prediction.
7. Whether the available information is sufficient for a meaningful chemical interpretation.
8. Important limitations and uncertainties.

Explicitly distinguish model behaviour from biological interpretation.
Do not assume that a large SHAP contribution represents a biologically causal factor.
Do not invent mechanisms or external evidence.
If a scientifically meaningful interpretation cannot be supported by the supplied information, state that explicitly.

Conclude each molecular case with:

SUPPORTED BY THE MODEL
and
CANNOT BE CONCLUDED FROM THE AVAILABLE EVIDENCE.

Task E(Prompt 1)

You are a scientific AI assistant interpreting explainable machine-learning results for drug discovery.

Inspect the supplied SHAP and LIME local explanation data together with the feature-descriptor information.

Compare the explanations generated for the same molecular predictions.

Identify:

1. **FEATURES IDENTIFIED BY BOTH METHODS**: Describe the important features on which SHAP and LIME agree.
2. **SHAP-SPECIFIC FEATURES**: Identify important features highlighted by SHAP but not LIME.
3. **LIME-SPECIFIC FEATURES**: Identify important features highlighted by LIME but not SHAP.
4. **DIRECTIONAL AGREEMENT**: Determine whether the methods agree on whether important features support or oppose the model prediction where the supplied information permits this.
5. **OVERALL AGREEMENT**: Assess whether the explanations provide a broadly consistent interpretation.
6. **DRUG-DISCOVERY RELEVANCE**: Explain why agreement or disagreement between explanation methods may matter when interpreting molecular predictions.
7. **LIMITATIONS**: Identify important limitations.

Do not assume that SHAP or LIME is automatically correct.
Do not invent reasons for disagreements.
Do not treat feature importance as evidence of biological causation.

Task E (Prompt 1)

You are a scientific AI assistant interpreting explainable machine-learning results for drug

discovery.

Inspect the supplied SHAP and LIME local explanation data together with the feature-descriptor information.

Compare the explanations generated for the same molecular predictions.

Identify:

1. **FEATURES IDENTIFIED BY BOTH METHODS**: Describe the important features on which SHAP and LIME agree.
2. **SHAP-SPECIFIC FEATURES**: Identify important features highlighted by SHAP but not LIME.
3. **LIME-SPECIFIC FEATURES**: Identify important features highlighted by LIME but not SHAP.
4. **DIRECTIONAL AGREEMENT**: Determine whether the methods agree on whether important features support or oppose the model prediction where the supplied information permits this.
5. **OVERALL AGREEMENT**: Assess whether the explanations provide a broadly consistent interpretation.
6. **DRUG-DISCOVERY RELEVANCE**: Explain why agreement or disagreement between explanation methods may matter when interpreting molecular predictions.
7. **LIMITATIONS**: Identify important limitations.

Do not assume that SHAP or LIME is automatically correct.
Do not invent reasons for disagreements.
Do not treat feature importance as evidence of biological causation.

Task E (Prompt 2)

You are an expert in explainable machine learning applied to computational toxicology and drug discovery.

Inspect the supplied local SHAP and LIME results and the supplied feature-descriptor information.

Evaluate the consistency of the two explanation methods.

For each molecular case, determine:

- which features are consistently important;
- which features are method-specific;
- whether contribution directions agree;
- whether the magnitude or ranking of important features differs;
- whether disagreements materially alter the interpretation of the prediction;
- whether the explanation can reasonably be considered robust across methods.

Use only the supplied evidence.
Do not resolve disagreements by inventing biological explanations.

Clearly distinguish:

- **DIRECTLY SUPPORTED BY BOTH METHODS**
- **DIRECTLY SUPPORTED BY ONE METHOD**
- **UNCERTAIN OR METHOD-DEPENDENT**

Do not interpret agreement between explanation methods as proof of biological causation.
End with an overall assessment of the robustness of the local explanations.

Task E (Prompt 3)

Act as a critical reviewer assessing whether local explanations of molecular toxicity

predictions are reliable enough to communicate to drug-discovery professionals.

Inspect the supplied SHAP and LIME results together with the feature-descriptor information.

Critically evaluate:

1. **CROSS-METHOD AGREEMENT**: How consistently do SHAP and LIME identify the same important molecular features?
2. **CROSS-METHOD DISAGREEMENT**: Where do the methods produce materially different explanations?
3. **DIRECTIONAL CONSISTENCY**: Do the methods agree on whether important features support or oppose the prediction?
4. **EXPLANATION ROBUSTNESS**: Which findings appear robust and which appear sensitive to the explanation method?
5. **SCIENTIFIC INTERPRETABILITY**: Which findings can reasonably be translated into molecular terminology?
6. **EVIDENCE LIMITATIONS**: What cannot be concluded from these explanations?
7. **OVERALL JUDGEMENT**: Assess whether the explanation is sufficiently consistent to support communication of the model's reasoning to a drug-discovery professional.

Do not force agreement between SHAP and LIME.
Do not invent explanations for disagreements.
Do not infer biological causality from either explanation method.
Where the evidence is insufficient, explicitly state this.

Task F (Prompt 1)

You are a scientific AI assistant supporting professionals in drug discovery, medicinal chemistry, computational toxicology, and molecular safety assessment.

Inspect all supplied machine-learning evaluation, feature-importance, local explanation, and feature-descriptor information.

Produce a structured scientific interpretation of the supplied molecular toxicity prediction analysis.

Use the following sections:

1. **MODEL PERFORMANCE**: Summarise the predictive performance and the most important strengths and weaknesses.
2. **GLOBAL MOLECULAR FEATURES**: Identify the molecular descriptors that the model relies upon most strongly.
3. **INDIVIDUAL MOLECULE INTERPRETATION**: Describe the important local feature contributions where individual molecular explanations are provided.
4. **EXPLANATION CONSISTENCY**: Compare SHAP, LIME, and other supplied feature-importance methods.
5. **SCIENTIFIC INTERPRETATION**: Explain what the important molecular descriptors represent and provide cautious chemical interpretation.
6. **DRUG-DISCOVERY RELEVANCE**: Explain how the findings could potentially support compound screening, prioritisation, or hypothesis generation.
7. **LIMITATIONS**: Identify important limitations of the model, explanations, data, and supplied evidence.

Use only the supplied information.
Do not invent missing values, mechanisms, experimental results, or literature evidence.
Do not state or imply that a molecular descriptor causes toxicity.
Do not present a model prediction as experimental evidence.
Do not claim clinical safety, clinical toxicity, or biological mechanism from the machine-learning results alone.
Clearly distinguish evidence-supported findings from scientific interpretation.

Task F (Prompt 2)

You are an expert scientific reporting assistant working at the intersection of machine learning, explainable AI, computational chemistry, and drug discovery.

Inspect all supplied CSV information before generating your report, including the molecular descriptor information.

Integrate the supplied predictive-performance and explainability evidence into a scientifically rigorous interpretation of the toxicity prediction analysis.

Structure the report as follows:

1. **EXECUTIVE SUMMARY**: Provide a concise summary of the principal findings.
2. **PREDICTIVE PERFORMANCE**: Interpret the supplied evaluation metrics and identify important performance trade-offs.
3. **GLOBAL MODEL BEHAVIOUR**: Identify the molecular descriptors most strongly associated with model behaviour using the supplied feature-importance evidence.
4. **LOCAL MODEL BEHAVIOUR**: Interpret the supplied individual molecular explanations.
5. **CROSS-METHOD EVIDENCE**: Assess agreement and disagreement between SHAP, LIME, permutation importance, and any other supplied explanation information.
6. **CHEMICAL INTERPRETATION**: Explain the molecular meaning of important descriptors.
   For every interpretation, distinguish between:
   - direct evidence from the machine-learning analysis;
   - plausible scientific interpretation;
   - conclusions requiring evidence that has not been supplied.
7. **DRUG-DISCOVERY APPLICATION**: Explain how the findings could reasonably be used for screening, compound prioritisation, or hypothesis generation.
8. **LIMITATIONS AND UNCERTAINTY**: Identify important limitations and uncertainties.
9. **OVERALL CONCLUSION**: Provide an evidence-grounded conclusion suitable for communication to a drug-discovery professional.

Use the supplied numerical values accurately.
Do not invent information.
Do not introduce external literature as though it were supplied evidence.
Do not infer causality from feature importance.
Do not claim that model predictions establish biological toxicity or clinical safety.
Do not conceal disagreement between explanation methods.
If evidence is insufficient, explicitly state that it is insufficient.

Task F (Prompt 3)

Act as a critical scientific reviewer assessing an explainable machine-learning toxicity analysis intended to support drug discovery.

Inspect all supplied machine-learning results, feature-importance information, local explanations, and molecular descriptor information.

Produce a structured critical assessment.

Use these sections:

1. **PRINCIPAL FINDINGS**: Identify the most important conclusions supported by the supplied evidence.
2. **PREDICTIVE PERFORMANCE**: Assess the strength and limitations of the model according to the reported metrics.
3. **GLOBAL EXPLANATION**: Assess which molecular descriptors appear consistently important and how strongly this conclusion is supported.

4. **LOCAL EXPLANATION**: Assess whether the supplied individual molecular explanations provide a coherent account of the predictions.
5. **XAI CONSISTENCY**: Critically evaluate agreement and disagreement between SHAP, LIME, permutation importance, and other supplied explanation methods.
6. **SCIENTIFIC INTERPRETATION**: Explain the molecular meaning of important descriptors while distinguishing established descriptor properties from speculative biological interpretation.
7. **DRUG-DISCOVERY VALUE**: Assess the realistic value of the analysis for compound prioritisation, screening, or hypothesis generation.
8. **EVIDENCE GAPS**: Identify important conclusions that cannot be justified from the supplied evidence.
9. **POTENTIAL OVERINTERPRETATION**: Identify statements or interpretations that would constitute overclaiming if made from these results alone.
10. **FINAL SCIENTIFIC ASSESSMENT**: Provide a concise judgement about what the analysis demonstrates, what it suggests, and what would require further validation.

Be scientifically conservative.
Do not invent evidence.
Do not make causal claims about molecular descriptors and toxicity.
Do not claim that the model establishes a biological mechanism.
Do not claim experimental or clinical validity unless explicitly supported by the supplied data.
Do not treat agreement between XAI methods as proof of biological truth.
When evidence is ambiguous or insufficient, explicitly identify the uncertainty.