ToxReason: A Benchmark for Mechanistic Chemical Toxicity Reasoning via Adverse Outcome Pathway

Jueon Park[1], **Wonjune Jang**[2], **Chanhwi Kim**[3], **Yein Park**[1,4], **Jaewoo Kang**[1,4†]
[1]Korea University [2]Myongji University
[3]University of Texas Health Science Center at Houston
[4]AIGEN Sciences
{jueon_park, kangj}@korea.ac.kr

Abstract

Recent advances in large language models (LLMs) have enabled molecular reasoning for property prediction. However, toxicity arises from complex biological mechanisms beyond chemical structure, necessitating mechanistic reasoning for reliable prediction. Despite its importance, current benchmarks fail to systematically evaluate this capability. LLMs can generate fluent but biologically unfaithful explanations, making it difficult to assess whether predicted toxicities are grounded in valid mechanisms. To bridge this gap, we introduce ToxReason, a benchmark grounded in the Adverse Outcome Pathway (AOP) that evaluates organ-level toxicity reasoning across multiple organs. ToxReason integrates experimental drug–target interaction evidence with toxicity labels, requiring models to infer both toxic outcomes and their underlying mechanisms from Molecular Initiating Event (MIE) to Adverse Outcome (AO). Using ToxReason, we evaluate toxicity prediction performance and reasoning quality across diverse LLMs. We find that strong predictive performance does not necessarily imply reliable reasoning. Furthermore, we show that reasoning-aware training improves mechanistic reasoning and, consequently, toxicity prediction performance. Together, these results underscore the necessity of integrating reasoning into both evaluation and training for trustworthy toxicity modeling.[1]

1 Introduction

Large language models (LLMs) have recently been applied to molecular reasoning tasks, allowing property prediction directly from chemical representations such as SMILES (Weininger, 1988). Prior works (Jang et al., 2025; Kim et al., 2025; Zhuang et al., 2025) have shown that LLMs can capture molecular structure and semantics, with explicit reasoning further improving predictive performance. Correspondingly, several benchmarks (Liu et al., 2025; Li et al., 2025) have been proposed to evaluate molecular reasoning abilities, primarily focusing on structure–property relationships. However, toxicity represents a fundamentally different challenge, as toxic effects often arise from complex biological mechanisms. These involve molecular targets, downstream cellular events, and organ-level responses, rather than chemical structure alone (Xu et al., 2020; Uesawa, 2024).

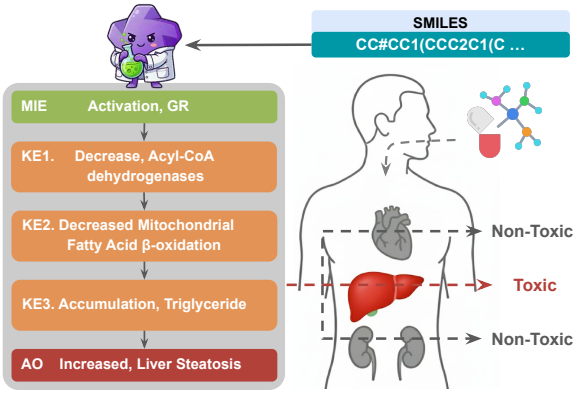

Figure 1: An example of AOP-based mechanistic toxicity reasoning from a MIE to an organ-level AO.

In toxicology, such mechanistic processes are systematically described by the Adverse Outcome Pathway (AOP) framework. It represents toxicity as a causal sequence from a molecular initiating event (MIE) to downstream key events (KE) and ultimately an adverse outcome (AO) at the organ level (Leist et al., 2017; Vinken et al., 2017). As illustrated in Figure 1, a query molecule may first activate the glucocorticoid receptor (GR), which leads to decreased activity of acyl-CoA dehydrogenases and decreased mitochondrial fatty acid β-oxidation. These disruptions result in triglyceride accumulation and ultimately leading to liver steatosis, an AO and a key manifestation of liver toxicity. Notably, this structure closely aligns with multi-step reasoning commonly studied in natural lan-

†Corresponding author.
[1]Our code is available at github.com/dmis-lab/ToxReason

Findings of the Association for Computational Linguistics: ACL 2026, pages 19542–19564
July 2-7, 2026

guage processing, where complex conclusions are derived through intermediate reasoning steps (Wei et al., 2022; Wang et al., 2022; Hwang et al., 2025).

Despite the importance of mechanistic reasoning in toxicity, existing datasets(e.g., Tox21 (Huang et al., 2016) and ClinTox (Gayvert et al., 2016)) have not been designed to evaluate whether models can reason about toxicity through biologically grounded mechanisms. While UniTox (Silberg et al., 2024) derives toxicity labels and corresponding explanations by summarizing clinical evidence from openFDA documents, its reasoning is primarily grounded in observed adverse effects rather than biologically mechanistic pathways. We propose ToxReason, a novel benchmark designed to evaluate mechanistic toxicity reasoning grounded in biologically causal processes. It requires LLMs to infer toxicity through structured reasoning over molecular interactions and downstream biological events. Evaluating this capability is particularly important in scenarios where clinical observations are unavailable, such as early-stage drug discovery and chemical safety assessment (Zheng et al., 2025).

ToxReason is constructed by integrating structured knowledge of causal toxicity pathways with experimental drug–target interaction data and curated chemical–toxicity associations. This yields a scientifically grounded benchmark of high-fidelity reasoning instances for 193 chemicals. Using this benchmark, we assess whether LLMs can move beyond predicting toxic outcomes to reasoning about underlying toxicity mechanisms. We employ an LLM-based evaluator to assess the reasoning quality based on four complementary metrics. Our analysis reveals that predictive performance and reasoning do not always align across models, suggesting differences in how LLMs internalize toxicity-related knowledge. This misalignment raises concerns about the reliability of toxicity predictions when models achieve high accuracy without corresponding mechanistic reasoning.

Beyond evaluation, we construct training data to explore three distinct learning paradigms. Specifically, our approach using reinforcement learning, explicitly optimizes both toxicity prediction and reasoning. Through this approach, our compact 4B-parameter model surpasses state-of-the-art models in toxicity prediction while demonstrating substantially improved reasoning ability. These results underscore the necessity of reasoning-aware optimization for aligning model predictions with biologically grounded explanations.

Our contributions are summarized as follows:

- We introduce ToxReason, a mechanistic toxicity benchmark that combines drug toxicity labels with AOP-based causal reasoning, enabling evaluation beyond outcome prediction.
- We systematically evaluate multiple open- and closed-source LLMs in terms of how they reason over toxicological mechanisms, rather than relying solely on surface-level toxicity prediction.
- We show that explicitly learning mechanistic toxicity reasoning leads to a decisive improvement in toxicity prediction, allowing a compact model to outperform other larger state-of-the-art models.

2 Related Work

Adverse Outcome Pathway(AOP) The AOP is a conceptual framework that portrays the sequential chain of causal events across different levels of biological organization (Zilliacus et al., 2024). It begins with a Molecular Initiating Event (MIE), which is the initial interaction between a chemical and a specific biological molecule like a protein or receptor. This trigger sets off a series of Key Events (KE) which are measurable biological changes that occur at the cellular, tissue, or organ level. These events act as "dominoes" that eventually lead to the Adverse Outcome (AO) such as toxicities.

AOPs are typically developed through systematic integration of experimental evidence, literature curation, and expert knowledge to establish causal relationships (Ankley et al., 2010). This mechanistic perspective has become central in modern toxicology, particularly in the new approach methodologies (NAMs) aimed at improving chemical risk assessment while reducing reliance on animal experimentation (Saarimäki et al., 2023).

LLMs for toxicity prediction Recent studies have explored large language models (LLMs) as tools for molecular toxicity prediction by leveraging their ability to reason over chemical structures and textual knowledge (Zhang et al., 2025). Previous work (Yang et al., 2025b; Chen et al., 2025) demonstrated the feasibility of applying LLMs to specific toxicity endpoints, such as cardiotoxicity and drug-induced osteotoxicity. They primarily focus on predicting toxic outcomes from molecular representations such as SMILES. While these

approaches showed promising predictive performance, they largely treated toxicity as an outcome prediction task and provided limited insight into the underlying biological mechanisms.

More recently, CoTox (Park et al., 2025) introduced a Chain-of-Thought (CoT) framework that incorporates biological pathway and gene ontology information with structural context to generate toxicity reasoning, representing a step toward more interpretable predictions. However, this approach focuses on improving prediction capabilities by constructing explanatory narratives from given chemical and biological information, rather than explicitly assessing whether the reasoning process itself aligns with mechanistically grounded causal pathways. In contrast, ToxReason reframes mechanistic toxicity reasoning as an evaluation problem by providing a benchmark grounded in causal toxicity pathways. This benchmark enables systematic evaluation of both LLMs' toxicity prediction performance and the alignment of their reasoning with AOP.

3 ToxReason Benchmark

3.1 Dataset Construction

Figure 2 provides an overview of the ToxReason dataset construction pipeline.

3.1.1 AOP Selection

We first curated a set of AOPs focused on liver, heart, and kidney toxicities to build our mechanistic reasoning benchmark. These specific organ systems were selected because they are primary targets of drug-induced toxicity (Rana et al., 2020) and are well-documented in established AOP databases. AOPs were selected from AOP-Wiki[2]database (Release 2.7) by focusing on pathways whose AOs correspond to clinically meaningful organ-level toxicities and whose MIEs involve explicit activation or inhibition of specific protein targets. Based on this selection strategy, we identified 23 unique AOPs and curated 25 distinct molecular targets involved in MIEs. The full list of selected AOPs, along with their associated MIE targets and AOs, is provided in Table A. These targets represent key biological entry points through which chemical perturbations can initiate toxicity pathways, providing a structured foundation for mechanistic toxicity reasoning in ToxReason.

[2]https://aopwiki.org

3.1.2 Chemical–AOP Association Derivation

To derive mechanistic associations between chemicals and AOPs, we integrated disease–chemical relationships with experimental and similarity-based evidence of MIEs. For each selected AOP, its AO was treated as a disease concept(e.g., liver fibrosis and heart failure) and used to retrieve associated chemicals from the Comparative Toxicogenomics Database (CTD, Nov. 2025 release) (Davis et al., 2025). These chemicals were considered candidate compounds linked to the corresponding AOP.

In parallel, we collected experimental evidence for MIE target proteins from ChEMBL (v36) (Zdrazil et al., 2024). For each target, activity data were extracted based on assay type and direction of action, where EC50 values were used for activation and IC50 values for inhibition. A chemical was considered to activate or inhibit a target if the corresponding activity value was below 10,000 nM, following established criteria used in Gadaleta et al. 2024 on quantitative structure–activity relationships for MIEs of organ-specific toxicity. Based on these rules, we curated a unified MIE dataset by merging molecules annotated with their corresponding targets and directions of action across all selected MIEs.

To infer MIEs for each candidate chemical, we performed similarity-based evidence aggregation. Specifically, chemicals retrieved from CTD were treated as query molecules, and structurally similar compounds with known target activity were identified from MIE dataset. The direction of interaction (activation or inhibition) for each MIE target was determined by majority voting over the activity annotations of these similar compounds. This procedure enabled robust inference of MIEs even when direct experimental measurements for the query chemical were unavailable.

Finally, we associated each chemical with an AOP when its inferred MIE and observed AO aligned within the same pathway, allowing the resulting AOP to function as the mechanistic toxicity rationale for the chemical.

3.1.3 Construct Train and Testset

To support both learning and evaluation of mechanistic toxicity reasoning, we constructed separate train and test sets under different evidence conditions, each serving a distinct role in the overall benchmark design. Detailed data statistics are provided in Appendix A.

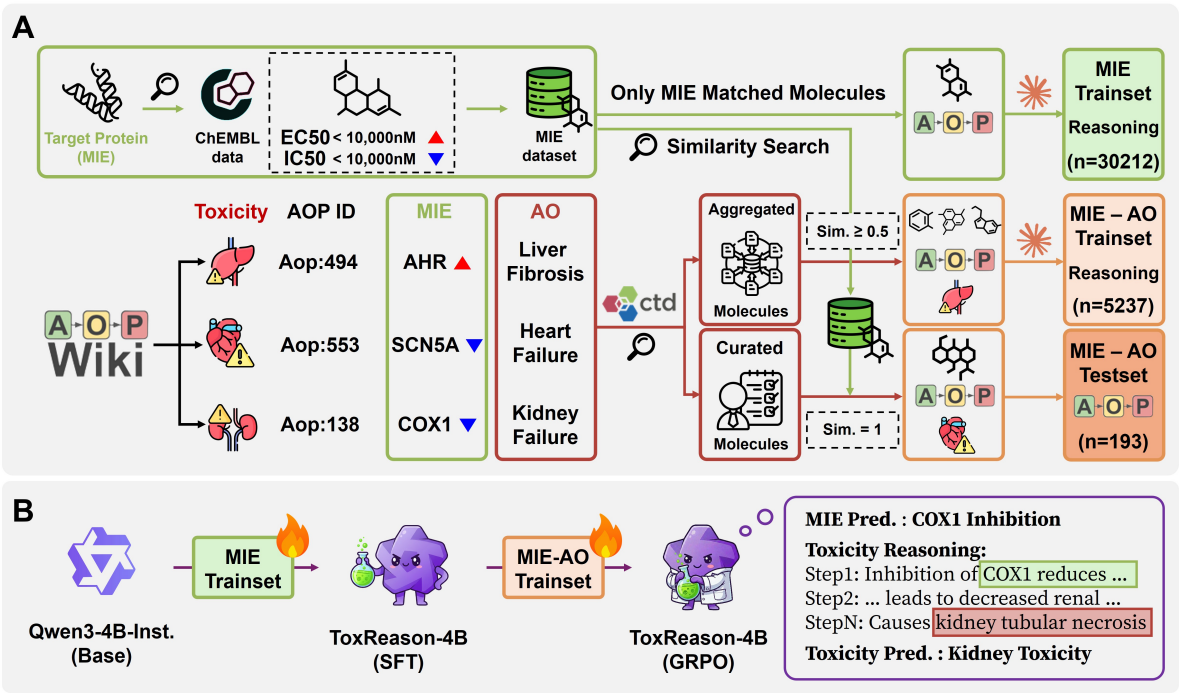

Figure 2: **Overview of ToxReason**, (A) Organ-specific AOPs are selected from AOP-Wiki, disease-linked chemicals are retrieved from CTD, and ChEMBL-derived MIE data are used to construct training and test sets under similarity constraints.(▲ Activation, ▼ Inhibition) (B) Learning framework built on ToxReason, combining supervised fine-tuning and reinforcement learning for mechanistic toxicity reasoning.

Training Data The training data were constructed as two complementary sets, namely the MIE-matched trainset and the MIE-AO-matched trainset. For the MIE-matched trainset, we collected compounds that satisfy the MIE condition of a given AOP based on experimental evidence of target-specific activation or inhibition. These compounds were matched to an AOP through their associated MIEs without requiring confirmation of the corresponding AO. Each compound in this set corresponds to a case where the toxic mechanism is known to have been initiated. While training, the MIE-matched trainset helps the model robustly learn AOP-specific initiation patterns by increasing coverage of MIE-matched compounds.

For the MIE-AO-matched trainset, we retained only compounds that satisfy both the MIE condition and the corresponding AO evidence for the same AOP. This trainset was constructed using the Chemical–AOP association derivation procedure described in Section 3.1.2. Here, aggregated CTD associations, which include both curated and inference-based chemical–disease relationships, were used to obtain disease-level toxicity labels, and similarity-based inference was applied with a relaxed Dice similarity (Dice, 1945) threshold (≥ 0.5) to infer MIEs from structurally related compounds. This setting encourages the model to reason across molecular interactions and downstream toxic outcomes.

For both trainsets, mechanistic toxicity explanations were generated in an instruction-style format using Claude Haiku 4.5 (Anthropic, 2025a) model. Given the inferred AOP structure, the model was prompted to narratively describe why a specific chemical could induce a particular toxicity through MIEs and subsequent biological processes.

Test Data The test set was constructed to evaluate mechanistic toxicity reasoning under strictly controlled evidence conditions. To avoid confounding effects from inferred or indirect associations, we relied on curated, human-specific CTD associations to define AOs. In addition, similarity-based inference was restricted to cases where the query compound was structurally identical to reference compounds, thereby eliminating uncertainty arising from structural approximation. As a result, each test instance contains the molecular structure, the corresponding MIE annotations, and the associated AOP context used for mechanistic reasoning. All molecules in the test set were strictly excluded from both training sets to prevent any overlap and ensure an unbiased evaluation.

3.2 Task Definition

ToxReason defines a mechanistic toxicity reasoning task that evaluates whether a model can infer toxic outcomes through biologically grounded reasoning rather than surface-level prediction. Given a query molecule, the task requires models to reason over molecular interactions and downstream biological processes to explain how toxicity manifests in a step-by-step manner.

Specifically, the task begins with a query chemical accompanied by contextual evidence retrieved from structurally similar compounds, including their similarity scores and experimentally supported MIE signals. This formulation is inspired by MolRAG (Xian et al., 2025), which leverages outcomes from structurally related molecules to support property inference for a query compound. Following this paradigm, we retrieve evidence from a ChEMBL-derived MIE dataset by collecting experimentally supported MIE outcomes from compounds that are structurally similar to the query molecule. We provide four activation and four inhibition examples as contextual evidence, following the four-shot setting adopted in MolRAG.

Based on the predicted MIEs, models then perform step-wise mechanistic reasoning, tracing how each molecular interaction may propagate through downstream biological events and ultimately lead to organ-level toxicity. The final output consists of a predicted toxic outcome along with a mechanistic explanation that connects MIEs to AOs through biologically meaningful reasoning steps.

3.3 Evaluation

ToxReason evaluates LLMs from two complementary perspectives, focusing on toxicity prediction and reasoning. This dual evaluation is designed to distinguish models that predict toxic outcomes from those that can also explain toxicity through biologically grounded mechanisms.

Toxicity Prediction For toxicity prediction, models are evaluated on a multi-label classification task covering three organ-level toxicities: liver, heart, and kidney toxicity. Given a query chemical, models are instructed to output the names of predicted toxicities in natural language. These responses are then post-processed into binary labels (0/1) for each organ, indicating the absence or presence of toxicity. Performance is measured using the F1-score for each toxicity type, along with the macro-averaged F1-score across all three toxicities, providing a balanced assessment of predictive performance.

Toxicity Reasoning To assess mechanistic reasoning quality, we compare model-generated explanations against the reference AOPs associated with each query chemical. Given the open-ended nature of natural language explanations, we adopt an LLM-as-a-Judge (Zheng et al., 2023) framework to evaluate reasoning quality. Specifically, Claude-Sonnet-4.5 (Anthropic, 2025b) is used as an independent evaluator to score each response by jointly considering the model's response and the corresponding ground-truth AOP.

Reasoning quality is evaluated along the following dimensions, each scored on a 0–10 scale:

- **Hallucination Avoidance** measures the extent to which the explanation avoids introducing unsupported or fabricated information.
- **Causal Coherence** evaluates whether the mechanistic reasoning follows a logically consistent causal chain, ensuring that MIEs lead to KEs and AOs in the correct order without contradictions or unjustified transitions.
- **Biological Fidelity** assesses the biological validity of the explanation, including correct use of toxicological terminology, accurate relationships between MIEs, KEs, and AOs, and consistency with known liver, heart, and kidney toxicology.
- **Overall** provides a holistic assessment of the explanation, summarizing the model's overall ability to produce coherent and biologically grounded mechanistic reasoning.

Scores for each criterion are reported separately, enabling fine-grained analysis of reasoning behavior across models, in addition to an overall reasoning quality score. Detailed prompts can be found in Table C and D.

4 Experiment Setup

4.1 Prompt Design

Models are prompted to infer MIEs for a query chemical based on experimental evidence retrieved from structurally similar compounds. For each inferred MIE, models generate a distinct step-wise mechanistic explanation that links the specific event to its corresponding organ-level toxicity. The model produces a structured JSON output

Models	Toxicity Prediction (F1-Score, %)				Toxicity Reasoning (LLM-as-a-Judge)				NW
	Kidney	Cardio	Liver	Avg.	Halluc.	Causal.	Biol.	Overall	Score
Closed Model									
GPT-4o	54.8	74.6	60.2	63.2	4.933	5.244	5.311	4.959	0.494
GPT-5	56.4	72.7	65.0	64.7	5.627	5.788	6.176	5.420	0.475
GPT-5.1	50.3	71.2	58.9	60.1	5.611	5.881	**6.378**	5.523	0.513
o3	60.0	72.5	58.8	63.8	5.342	5.736	5.922	5.326	0.488
o4-mini	58.5	71.3	55.6	61.8	4.808	5.549	5.492	4.948	0.487
Open Model									
Qwen3-4B-Inst.	56.9	71.1	57.3	61.8	4.337	4.969	5.104	4.523	0.484
Llama3.1-8B-Inst.	31.2	60.2	55.7	49.0	3.269	3.689	3.736	3.275	0.396
Qwen2.5-14B-Inst.	48.8	65.0	58.3	57.4	4.492	4.767	5.000	4.528	0.484
Gemma3-27B-Inst.	54.2	69.1	56.6	60.0	4.876	5.326	5.451	4.959	0.498
Llama3.1-70B-Inst.	57.1	76.1	57.4	63.5	4.554	4.865	5.016	4.653	0.473
DeepSeek-R1-Distill-70B	59.1	**78.5**	59.6	65.7	4.508	4.679	4.876	4.487	0.458
In-Context Learning									
Qwen3-4B-ICL-1shot	68.4	77.7	60.3	68.8	5.259	5.694	5.922	5.373	0.519
Qwen3-4B-ICL-2shot	55.2	63.5	58.7	59.1	4.233	5.073	5.223	4.373	0.404
Qwen3-4B-ICL-4shot	33.7	70.9	59.3	54.6	4.145	4.762	4.990	4.212	0.418
Supervised Finetuning									
ToxReason-4B-SFT	57.9	74.3	57.4	63.2	4.399	4.927	5.166	4.554	0.481
Reinforcement Learning									
ToxReason-4B-GRPO	**73.4**	72.7	**68.2**	**71.4**	**5.725**	**5.896**	5.642	**5.642**	**0.720**

Table 1: Performance comparison of toxicity prediction (0–100), mechanistic toxicity reasoning (0–10), and NW alignment (0–1) across models. **Bold** indicates the best score and underline indicates the second-best score.

including MIE predictions, individualized mechanistic reasoning for each event, and a summary of the final predicted toxicities across the three target organs. Prompts used in our experiments are provided in the Table E and F.

4.2 Models

We evaluate how different LLMs perform on the ToxReason benchmark under a unified zero-shot setting. The evaluated models include closed-systems such as GPT-4o (Hurst et al., 2024), GPT-5 (OpenAI, 2025a), GPT-5.1 (OpenAI, 2025b), o3, and o4-mini (OpenAI, 2025c), as well as open models including Qwen2.5-14B-Instruct (Yang et al., 2024), Llama3.1-8B-Instruct, Llama3.1-70B-Instruct (Grattafiori et al., 2024), Qwen3-4B-Instruct (Yang et al., 2025a), Gemma3-27B-Instruct (Team et al., 2025), and Deepseek-R1(Llama-70B Distilled) (Guo et al., 2025) covering a broad range of model scales. All models are prompted using the same task formulation and evaluated with identical output processing and scoring procedures. In addition, all experiments employ a greedy decoding strategy with a temperature of 0 and top_p set to 1.0 to ensure consistent and deterministic inference across models.

4.3 Model Improvement Strategies

To investigate if ToxReason-derived train data can effectively improve model performance, we evaluate three learning paradigms using Qwen3-4B-Instruct as base model. We first examine in-context learning (Brown et al., 2020) to assess the impact of few-shot demonstrations provided at inference time. Following this, we perform supervised fine-tuning (Zhang et al., 2023) through LoRA-based adaptation to align the model with the ToxReason task structure. Finally, we implement a two-stage reinforcement learning framework using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to explicitly optimize for AOP-grounded reasoning and causal consistency. The full implementation details for each of these three learning strategies are provided in Appendix B.

5 Results and Analysis

5.1 Zero-shot Performance Comparison

As shown in Table 1, closed models tend to exhibit stronger mechanistic reasoning performance, while predictive performance is more comparable between closed and open models. Especially, DeepSeek-R1 and GPT-5 achieved the strongest

predictive performance, GPT-5.1 demonstrated the best reasoning quality, as evidenced by its overall score of 5.523. However, GPT-5.1 exhibited the lowest predictive performance among all evaluated closed models at 60.1%. This result illustrates a distinct gap between predicting toxic labels and the ability to provide mechanistic reasoning.

Among open models, DeepSeek-R1 delivered the strongest predictive performance, achieving the highest cardiotoxicity F1-score across all models. However, its mechanistic reasoning performance remained relatively limited, comparable to or lower than that of smaller open models, which may reflect its design emphasis on broad, task-agnostic reasoning rather than biology-grounded mechanistic understanding.

These findings confirm a significant misalignment between predictive results and reasoning quality. This discrepancy highlights the necessity of ToxReason for evaluating a model's reasoning process instead of relying on predictive outcomes.

5.2 Effectiveness of Improvement Strategies

Applying various learning strategies to the Qwen3-4B base model revealed distinct performance patterns across the benchmark. In-context learning (ICL) achieved its peak effectiveness in a 1-shot configuration, significantly improving the average predictive performance to 68.8% and reasoning quality to 5.373. However, both performances declined as the shot count increased. This suggests that providing additional demonstrations may introduce contextual noise that hinders the model.

Additionally, supervised fine-tuning showed negligible differences in both predictive performance and reasoning quality compared to the base model. In contrast, reinforcement learning through the GRPO framework produced the most substantial gains by explicitly optimizing reasoning ability. The resulting ToxReason-4B-GRPO model attained an average predictive performance of 71.4% and an overall reasoning score of 5.642, significantly outperforming the base model and surpassing even the most capable closed-models.

5.3 Analysis of Toxicity Reasoning Metrics

Across toxicity reasoning metrics, reasoning-aware training consistently improves reasoning quality, with particularly strong gains in causal coherence. This indicates that AOP-grounded supervision effectively aligns model-generated explanations with structured causal chains from MIEs to AOs. Hallucination scores show substantial improvement, with ToxReason-4B-GRPO outperforming all other models, indicating that reasoning-aware training effectively reduces unsupported statements.

Claude-Sonnet-4.5	**Gemini-3-Flash**	
	Pearson r	**Spearman ρ**
Halluc. Avoidance	0.975	0.979
Causal Coherence	0.979	0.953
Biological Fidelity	0.967	0.970
Overall	0.986	0.974

Table 2: Cross-model correlation between Claude-Sonnet-4.5 and Gemini-3-Flash across toxicity reasoning metrics. All correlations are statistically significant ($p < 1e{-}8$).

LLM-as-a-Judge	**NW-alignment Score**	
	Pearson r	**Spearman ρ**
Halluc. Avoidance	0.689 ($p < 0.001$)	0.739 ($p < 0.001$)
Causal Coherence	0.615 ($p < 0.01$)	0.739 ($p < 0.001$)
Biological Fidelity	0.457 ($p = 0.055$)	0.692 ($p < 0.001$)
Overall	0.703 ($p < 0.001$)	0.837 ($p < 0.001$)

Table 3: Correlation between NW-alignment scores and LLM-as-a-Judge toxicity reasoning metrics. p-values denote statistical significance.

In contrast, biological fidelity improves only marginally, reflecting our training focus on AOP-defined mechanistic alignment instead of broad biological knowledge acquisition. Consequently, gains in overall reasoning quality are driven primarily by improved causal coherence and structural consistency, rather than by uniform improvements across all reasoning dimensions.

To further assess the reliability of LLM-based evaluation, we conduct a cross-model validation using an additional evaluator, Gemini-3-Flash-Preview (Google DeepMind, 2025), and compare the results with those obtained from Claude-Sonnet-4.5 in Table 1. Although absolute score ranges differ across evaluators, the relative ranking of models remains highly consistent. Quantitatively, we observe strong agreement across all reasoning dimensions (Pearson r = 0.967–0.986; Spearman ρ = 0.953–0.979, all $p < 1e{-}8$), as summarized in Table 2. These results indicate that the observed performance trends are robust to the choice of evaluator and support the reliability of the LLM-based evaluation framework.

In addition, we perform a focused human evaluation on a subset of LLM-as-a-Judge results using

Case Study: AOP Alignment & Reasoning Quality (Qwen-4B-Inst. vs ToxReason-4B-GRPO)

SMILES: `CC#CC1(CCC2C1(CC(C3=C4CCC(=O)C=C4CCC23)C5=CC=C(C=C5)N(C)C)C)O`

Ground Truth AOP Context:

- **AOP ID:** `Aop:318`
- **AOP Title:** Glucocorticoid Receptor activation leading to hepatic steatosis
- **AOP:**
  GR activation → *Decreased acyl-CoA dehydrogenase activity* → *Reduced mitochondrial fatty acid β-oxidation* → *Triglyceride accumulation* → *Hepatic steatosis (liver toxicity)*

Qwen-4B-Inst.

- **MIE:** GR Activation
- **Model Reasoning:**

1. Activation of the glucocorticoid receptor (GR) leads to dysregulation of gene expression involved in metabolism and inflammation.
2. This dysregulation results in increased hepatic lipid accumulation and oxidative stress, contributing to liver injury.
3. Chronic liver injury progresses to hepatocyte necrosis and inflammation, leading to liver toxicity.

- **Predicted Toxicity:** Liver toxicity

ToxReason-4B-GRPO

- **MIE:** GR Activation
- **Model Reasoning:**

1. Glucocorticoid receptor (GR) activation leads to reduced hepatic fatty acid β-oxidation in the liver.
2. Reduced fatty acid β-oxidation leads to decreased fatty acid oxidation and increased fatty acid accumulation.
3. Increased fatty acid accumulation leads to increased hepatic triglyceride accumulation (steatosis).
4. Increased hepatic steatosis leads to liver toxicity.

- **Predicted Toxicity:** Liver toxicity

Table 4: Case study comparing two LLMs on the same AOP grounding. The upper block provides the reference AOP context (Aop:318); the lower block contrasts model reasoning and predicted toxicity side-by-side. Highlighted segments denote aligned MIE/KE/AO. Additional reasoning evaluation results are reported in Table G. Outputs from the other models and other sample of ToxReason-4B-GRPO are provided in Tables H to L.

a domain expert with several years of experience in drug development. The expert confirms that the evaluation framework effectively distinguishes between outcome-level prediction and mechanistic alignment with ground-truth AOP pathways.

5.4 Algorithmic Validation of LLM-Evaluator

To assess whether LLM-as-a-Judge scores are supported by an objective signal, we examine their correlation with an algorithm-based measure of mechanistic reasoning alignment computed using the Needleman–Wunsch (NW) algorithm (Needleman and Wunsch, 1970). NW is a dynamic programming method for global sequence alignment that preserves relative ordering while penalizing missing or extraneous steps. It has been adopted as a fine-grained metric to quantify alignment between generated and reference causal reasoning paths (Nguyen et al., 2024).

As mechanistic toxicity reasoning follows an ordered causal progression from MIEs to AOs, NW provides a structure-aware measure of causal consistency by aligning semantically encoded reasoning steps with reference AOP steps to compute a global alignment score that accounts for semantic similarity and causal ordering (see Appendix B.3 `tox_align_score` and Algorithm 1).

Table 3 reports that the overall LLM-as-a-Judge reasoning score shows the strongest correlation with NW-alignment Score (Pearson $r = 0.703$, Spearman $\rho = 0.837$), indicating that holistic judge evaluations reflect adherence to the intended mechanistic structure, thereby supporting the reliability of LLM-as-a-Judge as an evaluation signal.

5.5 Case Study

As shown in Table 4, we analyze the reasoning outputs of a base model and the ToxReason-4B-GRPO under the same AOP grounding. The base model correctly identifies the molecular initiating event and predicts liver toxicity, but its explanation remains generic and loosely structured, omitting several intermediate biological events specified in the reference AOP. In contrast, ToxReason-4B-GRPO generates a step-by-step mechanistic explanation that closely follows the reference AOP causal chain, explicitly tracing intermediate processes from glucocorticoid receptor activation through reduced hepatic fatty acid β-oxidation and triglyceride accumulation to hepatic steatosis. This comparison shows that reasoning-aware training improves the structural fidelity of mechanistic explanations.

6 Conclusion

In this work, we introduced ToxReason, a benchmark designed to evaluate mechanistic toxicity reasoning grounded in AOPs. ToxReason moves beyond outcome-level toxicity prediction and enables systematic assessment of causal reasoning from molecular initiating events to organ-level adverse outcomes. Furthermore, we established a toxicity-reasoning-aware model to advance the use of LLMs in toxicology. These contributions facilitate a deeper mechanistic understanding while enabling more reliable and interpretable toxicity assessments. Ultimately, we envision that such mechanistically grounded reasoning frameworks could assist regulatory decision-making processes in drug safety evaluation.

Limitations

Despite its contributions, this work has several limitations. ToxReason is constrained by the coverage and granularity of AOP-Wiki, and the current benchmark focuses on only liver, heart, and kidney toxicities, limiting generalization to other organs or less-characterized mechanisms. The evaluation benchmark also remains relatively small due to the limited availability of high-confidence mechanistic toxicity data. Rather than maximizing dataset size, we prioritized constructing a reliable benchmark under strict evidence constraints. In addition, the set of MIEs considered in ToxReason is restricted to those defined within the selected AOPs, resulting in a limited and domain-specific MIE space.

Moreover, the reference mechanistic pathways used for evaluation represent canonical causal structures and may not fully capture the diversity or context-dependence of real-world toxicity mechanisms.

Finally, while LLM-as-a-Judge evaluation enables scalable assessment of reasoning quality, it remains inherently subjective. Although we mitigate this limitation by complementing it with an algorithm-based alignment metric, judge-based scores should be interpreted as relative rather than absolute measures.

Ethics Statement

This research utilizes publicly available datasets, including AOP-Wiki, ChEMBL, and CTD, and does not involve human subjects, personal data, or animal experimentation. All data were accessed and analyzed in compliance with their licensing terms and intended research objectives. ToxReason is developed solely for research and evaluation purposes; consequently, model outputs must not be interpreted as evidence for clinical or regulatory decision-making. We acknowledge the potential risks associated with over-reliance on automated systems in safety-critical domains such as toxicology. Accordingly, model-generated insights should be used with caution and should be integrated with expert judgment and empirical validation.

Acknowledgments

We thank Sanghoon Lee for his valuable feedback on our results. This research was supported by (1) the National Research Foundation of Korea (NRF-2023R1A2C3004176), (2) the Ministry of Health & Welfare, Republic of Korea (HR20C002103), (3) ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2020-II201819), (4) the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT and MOE) (No. RS-2025-16652968), (5) the Seoul National University Hospital with support from the Ministry of Science and ICT (RS-2023-00262002) and (6) the Korea Bio Data Station (K-BDS) with computing resources including technical support.

References

Gerald T Ankley, Richard S Bennett, Russell J Erickson, Dale J Hoff, Michael W Hornung, Rodney D Johnson, David R Mount, John W Nichols, Christine L Russom, Patricia K Schmieder, and 1 others. 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environmental toxicology and chemistry*, 29(3):730–741.

Anthropic. 2025a. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Released: 2025-10-16.

Anthropic. 2025b. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Released: 2025-09-30.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yi-Qi Chen, Tao Yu, Zheng-Qi Song, Chen-Yu Wang, Jiang-Tao Luo, Yong Xiao, Heng Qiu, Qing-Qing Wang, and Hai-Ming Jin. 2025. Application of large language models in drug-induced osteotoxicity prediction. *Journal of Chemical Information and Modeling*, 65(7):3370–3379.

Allan P. Davis, Thomas C. Wiegers, Daniela Sciaky, Frances Barkalow, Matthew Strong, Benjamin Wyatt, Jill Wiegers, Rebecca McMorran, Shakil Abrar, and Carolyn J. Mattingly. 2025. Comparative toxicogenomics database's 20th anniversary: update 2025. *Nucleic Acids Research*. Advance access, October 10, 2024.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Domenico Gadaleta, Marina Garcia de Lomana, Eva Serrano-Candelas, Rita Ortega-Vallbona, Rafael Gozalbes, Alessandra Roncaglioni, and Emilio Benfenati. 2024. Quantitative structure–activity relationships of chemical bioactivity toward proteins associated with molecular initiating events of organ-specific toxicity. *Journal of Cheminformatics*, 16(1):122.

Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301.

Google DeepMind. 2025. Gemini 3 flash model card.

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

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*.

Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov. 2016. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:85.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hyeon Hwang, Yewon Cho, Chanwoong Yoon, Yein Park, Minju Song, Kyungjae Lee, Gangwoo Kim, and Jaewoo Kang. 2025. Assessing llm reasoning steps via principal knowledge grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19925–19948.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2025. Structural reasoning improves molecular understanding of llm. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21016–21036.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. 2025. Mol-llama: Towards general understanding of molecules in large molecular language model. *arXiv preprint arXiv:2502.13449*.

Marcel Leist, Ahmed Ghallab, Rabea Graepel, Rosemarie Marchan, Reham Hassan, Susanne Hougaard Bennekou, Alice Limonciel, Mathieu Vinken, Stefan Schildknecht, Tanja Waldmann, and 1 others. 2017. Adverse outcome pathways: opportunities, limitations and open questions. *Archives of Toxicology*, 91(11):3477–3505.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. 2025. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *arXiv preprint arXiv:2505.21318*.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. 2025. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *arXiv preprint arXiv:2508.01055*.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy Vu, and Gholamreza

Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2862–2883.

OpenAI. 2025a. Gpt-5 system card.

OpenAI. 2025b. Gpt-5.1 instant and gpt-5.1 thinking system card addendum.

OpenAI. 2025c. Openai o3 and o4-mini system card.

Jueon Park, Yein Park, Minju Song, Soyon Park, Donghyeon Lee, Seungheun Baek, and Jaewoo Kang. 2025. Cotox: Chain-of-thought-based molecular toxicity reasoning and prediction. *arXiv preprint arXiv:2508.03159*.

Payal Rana, Stephen Kogut, Xuerong Wen, Fatemeh Akhlaghi, and Michael D Aleo. 2020. Most influential physicochemical and in vitro assay descriptors for hepatotoxicity and nephrotoxicity prediction. *Chemical Research in Toxicology*, 33(7):1780–1790.

Laura Aliisa Saarimäki, Michele Fratello, Alisa Pavel, Seela Korpilähde, Jenni Leppänen, Angela Serra, and Dario Greco. 2023. A curated gene and biological system annotation of adverse outcome pathways related to human health. *Scientific data*, 10(1):409.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jacob Silberg, Kyle Swanson, Elana Simon, Angela Zhang, Zaniar Ghazizadeh, Scott Ogden, Hisham Hamadeh, and James Y Zou. 2024. Unitox: leveraging llms to curate a unified dataset of drug-induced toxicity from fda labels. *Advances in Neural Information Processing Systems*, 37:12078–12093.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Yoshihiro Uesawa. 2024. Efficiency of pharmaceutical toxicity prediction in computational toxicology. *Toxicological Research*, 40(1):1–9.

Mathieu Vinken, Dries Knapen, Lucia Vergauwen, Jan G Hengstler, Michelle Angrish, and Maurice Whelan. 2017. Adverse outcome pathways: a concise introduction for toxicologists. *Archives of toxicology*, 91(11):3697–3707.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531.

Tuan Xu, Leihong Wu, Menghang Xia, Anton Simeonov, and Ruili Huang. 2020. Systematic identification of molecular targets and pathways related to human organ level toxicity. *Chemical research in toxicology*, 34(2):412–421.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hengzheng Yang, Jian Xiu, Weiqi Yan, Kaifeng Liu, Huizi Cui, Zhibang Wang, Qizheng He, Yilin Gao, and Weiwei Han. 2025b. Large language models as tools for molecular toxicity prediction: Ai insights into cardiotoxicity. *Journal of Chemical Information and Modeling*, 65(5):2268–2282.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, and 1 others. 2024. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192.

Jiangyan Zhang, Haolin Li, Yuncong Zhang, Junyang Huang, Liping Ren, Chuantao Zhang, Quan Zou, and Yang Zhang. 2025. Computational toxicology in drug discovery: applications of artificial intelligence in admet and toxicity prediction. *Briefings in Bioinformatics*, 26(5):bbaf533.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and 1 others. 2023. Instruction tuning for large language models: A survey. *ACM Computing Surveys*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Madeleine Yang, Lauren T May, Geoffrey I Webb, Li Li, Shirui Pan, and George Church. 2025. Large language models for drug discovery and development. *Patterns*, 6(10).

Jiaxi Zhuang, Yaorui Shi, Jue Hou, Yunong He, Mingwei Ye, Mingjun Xu, Yuming Su, Linfeng Zhang, Ying Qian, Guolin Ke, and 1 others. 2025. Reasoning-enhanced large language models for molecular property prediction. *arXiv preprint arXiv:2510.10248*.

Johanna Zilliacus, Monica K Draskau, Hanna KL Johansson, Terje Svingen, and Anna Beronius. 2024. Building an adverse outcome pathway network for estrogen-, androgen-and steroidogenesis-mediated reproductive toxicity. *Frontiers in Toxicology*, 6:1357717.

Appendix

A Details in ToxReason Benchmark

Selected AOPs Table A summarizes the curated set of adverse outcome pathways (AOPs) used in ToxReason, including their associated adverse outcomes (AOs) and molecular initiating events (MIEs), with activation and inhibition directions explicitly annotated.

Data Statistics Table B reports the distribution of samples across toxicity labels in the MIE-matched and MIE–AO-matched training sets, as well as the test set, providing an overview of dataset composition and label coverage.

B Training Details

B.1 In-Context Learning

For in-context learning, we augment the prompt with a small number of demonstration examples sampled from a MIE-AO-matched trainset. For each query molecule, examples are randomly selected according to the specified shot number and follow the same input structure as the original task, including the corresponding output formatted in JSON. We evaluate in-context learning under 1, 2 and 4-shot configurations to examine the effect of increasing contextual supervision at inference time.

B.2 Supervised Fine-Tuning

For supervised fine-tuning, we used Qwen3-4B-Instruct as the base model and performed instruction tuning on the MIE-AO-matched training set. Each training instance follows the task formulation of ToxReason, consisting of structured inputs and corresponding JSON-formatted outputs that encode mechanistic toxicity reasoning. Parameter-efficient fine-tuning was conducted using LoRA, with a rank of 8, a scaling factor (α) of 32, and a dropout rate of 0.1. The model was trained with a learning rate of 2×10^{-5} using a cosine learning rate scheduler with a warmup ratio of 0.1, for a total of 3 epochs. We used a per-device batch size of 8 and applied gradient accumulation with 8 steps. All experiments were conducted using 8 NVIDIA A100 GPUs.

B.3 Reinforcement Learning

Reinforcement learning was conducted in two stages to progressively enhance the model's mechanistic reasoning capability.

Toxicity	AO (Disease)	AOP (ID)	MIE
Liver Toxicity	Liver Cancer	AOP:37	PPARα ↑
		AOP:41	AHR ↑
		AOP:220	CYP2E1 ↑
	Liver Fibrosis	AOP:130	PLA2 ↑
		AOP:494	AHR ↑
		AOP:34	LXR ↑
		AOP:36	PPAR$\alpha/\beta/\gamma$ ↓
		AOP:57	AHR ↑
	Liver Steatosis	AOP:58	LXR ↑, PPARα ↓
		AOP:60	PXR ↑
		AOP:61	NR1H4 ↑, NRF2 ↑
		AOP:318	GR ↑
		AOP:517	PXR ↑
		AOP:518	LXR ↑
Cardiotoxicity	Heart Failure	AOP:553	SCN5A ↓
		AOP:555	KCNH2 ↓
		AOP:558	PDE3/4 ↓
		AOP:554	ADRB2 ↑
	Cardiac Arrhythmia	AOP:559	AChE ↓
		AOP:138	OAT1 ↓
		AOP:177	COX1 ↓
Kidney Toxicity	Kidney Failure	AOP:138	OAT1 ↓
		AOP:177	COX1 ↓
	Chronic Kidney Disease	AOP:384	ACE ↓, AT1R ↑

Table A: Selected adverse outcome pathways (AOPs) with associated adverse outcomes (AOs) and molecular initiating events (MIEs). (↑: Activation, ↓: Inhibition)

Stage 1 In the first stage, the model was initialized by performing supervised fine-tuning on the MIE-matched training set. We used Qwen3-4B-Instruct as the base model and applied LoRA for parameter-efficient fine-tuning, with rank r=32, scaling factor α=64, and dropout rate of 0.05. The model was trained with a learning rate of 5×10^{-5} using a cosine learning rate scheduler with a warmup ratio of 0.03, for 5 epochs. We used a per-device batch size of 8 with gradient accumulation over 8 steps.

Stage 2 In the second stage, we further trained the model using Group Relative Policy Optimization (GRPO) on the MIE-AO-matched training set. Training was performed with a per-device batch size of 8 and gradient accumulation over 4 steps. We used a learning rate of 1.0×10^{-6} with a cosine scheduler with a minimum learning rate, and trained the model for a maximum of 1000 steps. For each input, the model generated 8 candidate responses, using a sampling temperature of 0.7.

During this stage, we employed three reward functions to guide the learning process, which are

Toxicity Label	MIE-matched Trainset	MIE-AO-matched Trainset	Testset
Liver	10237	1635	63
Cardio	18543	906	75
Kidney	132	72	39
Liver, Cardio	91	764	4
Liver, Kidney	0	285	5
Cardio, Kidney	1197	296	7
Liver, Cardio, Kidney	12	1279	0
Total	**30212**	**5237**	**193**

Table B: Sample distribution by toxicity category for MIE-matched and MIE-AO-matched training sets and the test set.

described in detail below.

- **`tox_format`** (format compliance). This reward encourages the model to produce a valid JSON output that strictly follows the predefined schema. Specifically, it rewards successful JSON parsing, exact compliance with the required top-level fields, and a one-to-one correspondence between predicted MIEs and reasoning blocks. Additional constraints are imposed to ensure well-formed reasoning blocks, consistent label usage, and the absence of malformed or multi-line fields. The final score is normalized to the range $[0, 1]$.

- **`tox_mie_pred`** (MIE prediction accuracy). This reward measures the accuracy of predicted MIEs by comparing them with the reference answer. Target-level agreement is quantified using the Jaccard similarity between the sets of predicted and reference MIEs, while directional consistency is assessed by checking whether activation or inhibition labels match for overlapping targets. To discourage overprediction, the reward applies a penalty when the model predicts more MIEs than present in the reference. The final score reflects both target overlap and direction correctness and is normalized to the range $[0, 1]$.

- **`tox_align_score`** (alignment score). This reward evaluates the causal consistency of the generated reasoning by measuring how well it aligns with the Answer AOP in an order-preserving manner. Each reasoning step is first embedded using the `all-MiniLM-L6-v2` sentence embedding model, and step-level similarities are computed based on cosine similarity. The overall alignment score is then computed using the Needleman-Wunsch(NW) global sequence alignment algorithm, which preserves the relative order of causal steps while penalizing missing or extraneous steps. We use an acceptance threshold of 0.20 for candidate matches, a similarity threshold of 0.50, a gap penalty of -0.30, and a low-similarity penalty of -0.15. The resulting alignment score is normalized and clipped to the range $[0, 1]$. The same alignment score formulation is also applied when evaluating the model's reasoning outputs.

Algorithm 1 NW-Alignment Score Computation

Input:
Answer AOP path $A = (a_1, \ldots, a_M)$;
Pred. reasoning path $P = (p_1, \ldots, p_N)$;
Step similarity function $\mathrm{sim}(\cdot, \cdot)$;
Gap penalty $\delta < 0$
Output: Alignment score $S \in [0, 1]$

Initialize score matrix $D \in \mathbb{R}^{(M+1)\times(N+1)}$
$D[0, 0] \leftarrow 0$
for $i = 1$ to M **do**
  $D[i, 0] \leftarrow D[i-1, 0] + \delta$
end for
for $j = 1$ to N **do**
  $D[0, j] \leftarrow D[0, j-1] + \delta$
end for
for $i = 1$ to M **do**
  for $j = 1$ to N **do**
    $match \leftarrow D[i-1, j-1] + \mathrm{sim}(a_i, p_j)$
    $skip\_answer \leftarrow D[i-1, j] + \delta$
    $skip\_pred \leftarrow D[i, j-1] + \delta$
    $D[i, j] \leftarrow \max(match, skip\_answer, skip\_pred)$
  end for
end for
$raw \leftarrow D[M, N]$
$max \leftarrow \min(M, N)$
$min \leftarrow (M + N) \cdot \delta$
$S \leftarrow \dfrac{raw - min}{max - min}$
return S

The three reward components were assigned equal weights during optimization. Reinforcement learning experiments were carried out using 8 NVIDIA A100 GPUs.

LLM-as-a-Judge **SYSTEM PROMPT**

You are a toxicology expert and AOP-based mechanistic reasoning evaluator.
Your task is to assess the quality of an AI assistant's toxicity reasoning based on:

- its final predicted toxicities,
- its step-by-step mechanistic reasoning,
- and the provided ground-truth AOP and toxicity labels.

Evaluate the assistant's response using the following criteria, each on a scale from 1 to 10 (higher scores are better):

1. Hallucination_Avoidance — The degree to which the model avoids inventing unsupported facts. A high score means little to no hallucination and strong grounding in the provided AOP and inputs.
2. Causal_Coherence — Logical consistency of the mechanistic chain. Each step should follow causally from the previous one (*MIE* $\rightarrow$ *KE* $\rightarrow$ *AO* $\rightarrow$ *Organ Toxicity*) without unjustified jumps, contradictions, or reversed order.
3. Biological_Fidelity — Biological validity of the mechanism. Uses correct terminology, accurate MIE/KE/AO relationships, and reflects realistic heart/liver/kidney toxicology and physiology.
4. Overall — An overall quality score summarizing the four criteria above.

Output Format Requirement:
You must output a single valid JSON object with the following structure:

```
{
  "Hallucination_Avoidance": <number from 1 to 10>,
  "Causal_Coherence": <number from 1 to 10>,
  "Biological_Fidelity": <number from 1 to 10>,
  "Overall": <number from 1 to 10>,
  "Explanation": "<short textual explanation (3-6 sentences)>"
}
```

Hard Constraints:

- Output only the JSON object.
- Do not include any additional text, comments, or markdown.
- The JSON must be syntactically valid.

Table C: LLM-as-a-judge system prompt used for evaluating mechanistic toxicity reasoning quality.

LLM-as-a-Judge **USER PROMPT**

[Ground-truth Information]

Correct_AOP:
`{{AOP_Context}}`

Correct_Toxicities:
`{{Answer_TOXICITIES}}`

[Assistant Response to Evaluate]

`{{Assistant_Response}}`

Please evaluate the assistant's reasoning according to the criteria defined in the system prompt and return only the JSON object with your scores and explanation.

Table D: LLM-as-a-judge user prompt for evaluating toxicity reasoning against ground-truth AOP annotations.

LLM Inference **SYSTEM PROMPT**

You are an expert mechanistic toxicology reasoning model.

Given:
- A query molecule (SMILES)
- Experimental observations from structurally similar molecules, including Activation Examples and Inhibition Examples with similarity scores,

your job is to infer how the query molecule interacts with biological targets (MIEs) and then explain how these inferred MIEs could mechanistically lead to organ toxicity.

Follow these rules carefully:

1. When interpreting reference evidence:
- Use ONLY Activation Examples to infer activation.
- Use ONLY Inhibition Examples to infer inhibition.
- Never infer inhibition from "non-active".
- Never infer activation from "non-inhibit".
- Give more weight to examples with higher similarity scores.
- Base all conclusions on structural similarity + target evidence.

2. For EACH inferred MIE, produce mechanistic reasoning describing how it can lead to organ toxicity:
- Use "Step 1, Step 2, Step 3..." format.
- Each step $\leq$ 2 sentences.
- Steps must follow:
MIE $\rightarrow$ Key Events (KEs) $\rightarrow$ Adverse Outcome (AO) $\rightarrow$ organ toxicity.

3. Only consider the following toxicity types:
- "cardiotoxicity"
- "liver toxicity"
- "kidney toxicity"
Choose exactly ONE organ toxicity per MIE.

```
{
  "MIE_Prediction": {
    "Target1": "Activation or Inhibition",
    "Target2": "Activation or Inhibition",
    ...
  },
  "Toxicity_Reasoning": [
    {
      "MIE": "",
      "Reasoning_Steps": [
        "Step 1: ...",
        "Step 2: ...",
        "Step 3: ..."
      ],
      "Toxicity": ""
    }
  ],
  "Overall_Assessment": {
    "Summary": "Concise narrative summary of why toxicity occurs.",
    "Predicted_Toxicities": ["...", "..."]
  }
}
```

HARD CONSTRAINTS:
- Output ONLY the JSON.
- Activation/Inhibition labels must be exactly: "Activation", "Inhibition".
- Toxicity must be one of the three organ toxicities listed.
- No extra text outside the JSON.

Table E: LLM inference SYSTEM PROMPT

LLM Inference **USER PROMPT**

Query Molecule (SMILES):
`{{test_smiles}}`

Below is experimental evidence from structurally similar molecules. Each example shows a similarity score and activity result on specific biological targets.

Reference Evidence:
`{{refer_context}}`

Using the evidence above:

- Interpret the activation/inhibition patterns based on similarity.
- Infer the most likely molecular initiating events (MIEs) of the query molecule.
- Produce mechanistic toxicity reasoning showing how those MIEs could lead to organ-level toxicity through the sequence:
  MIE $\rightarrow$ Key Events $\rightarrow$ Adverse Outcome $\rightarrow$ Organ Toxicity.

Return ONLY the final JSON in the required format.

Table F: LLM inference user prompt used for mechanistic toxicity reasoning.

Case Study: AOP Alignment & Reasoning Quality (Qwen-4B-Inst. vs ToxReason-4B-GRPO)

SMILES: `CC#CC1(CCC2C1(CC(C3=C4CCC(=O)C=C4CCC23)C5=CC=C(C=C5)N(C)C)C)O`

Ground Truth AOP Context:

- **AOP ID:** `Aop:318`
- **AOP Title:** Glucocorticoid Receptor activation leading to hepatic steatosis
- **AOP:**
  GR activation → *Decreased acyl-CoA dehydrogenase activity* → *Reduced mitochondrial fatty acid β-oxidation* → *Triglyceride accumulation* → *Hepatic steatosis (liver toxicity)*

Qwen-4B-Inst.

- **MIE:** GR Activation
- **Model Reasoning:**

1. Activation of the glucocorticoid receptor (GR) leads to dysregulation of gene expression involved in metabolism and inflammation.
2. This dysregulation results in increased hepatic lipid accumulation and oxidative stress, contributing to liver injury.
3. Chronic liver injury progresses to hepatocyte necrosis and inflammation, leading to liver toxicity.

- **Predicted Toxicity:** Liver toxicity

ToxReason-4B-GRPO

- **MIE:** GR activation
- **Model Reasoning:**

1. Glucocorticoid receptor (GR) activation leads to reduced hepatic fatty acid β-oxidation in the liver.
2. Reduced fatty acid β-oxidation leads to decreased fatty acid oxidation and increased fatty acid accumulation.
3. Increased fatty acid accumulation leads to increased hepatic triglyceride accumulation (steatosis).
4. Increased hepatic steatosis leads to liver toxicity.

- **Predicted Toxicity:** Liver toxicity

LLM-as-a-Judge Evaluation:

Qwen-4B-Inst.

- **Judge Scores:**
  - Hallucination Avoidance: **4**
  - Causal Coherence: **5**
  - Biological Fidelity: **5**
  - Overall: **4**
- **Judge Summary:**

1. Correctly identifies glucocorticoid receptor (GR) activation as the initiating molecular event.
2. Mentions general lipid accumulation but omits critical intermediate steps such as decreased acyl-CoA dehydrogenases and impaired mitochondrial β-oxidation.
3. Introduces unsupported mechanisms (oxidative stress, inflammation, hepatocyte necrosis) not specified in the ground-truth AOP.

ToxReason-4B-GRPO

- **Judge Scores:**
  - Hallucination Avoidance: **8**
  - Causal Coherence: **9**
  - Biological Fidelity: **8**
  - Overall: **8**
- **Judge Summary:**

1. Glucocorticoid receptor activation leads to reduced mitochondrial fatty acid β-oxidation.
2. Reduced β-oxidation results in increased fatty acid availability in hepatocytes.
3. Excess fatty acids are stored as triglycerides, causing hepatic steatosis.
4. Hepatic steatosis is directly linked to liver toxicity.

Table G: Case study combining model-generated AOP reasoning (highlighted for alignment) and LLM-as-a-Judge evaluation (no highlighting). The upper block shows the reference AOP context and each model's reasoning; the lower block reports judge scores and diagnostic summaries for the same responses.

Case Study: AOP Alignment & Reasoning Quality (GPT-5)

Below is an example instance illustrating strong alignment between the **Ground Truth AOP context (Aop:559)** and the **GPT-5** reasoning path, demonstrating accurate reproduction of an AChE-mediated cardiotoxicity pathway.

Sample Metadata.

- **Index:** 98
- **SMILES:** `CC12CCN(C1N(C3=C2C=C(C=C3)OC(=O)NC)C)C`
- **AOP ID:** `Aop:559`

Ground Truth AOP Context:

- **AOP Title:** Inhibition of acetylcholinesterase (AChE) leading to arrhythmias
- **AOP:** *Acetylcholinesterase (AChE) Inhibition* → *Increased Muscarinic Acetylcholine* → *Receptors* *Altered Action Potential* → *Increased delay in heart electrical conduction* → *Occurrence of cardiac arrhythmia*

GPT-5 Reasoning Result

- **MIE:** AChE Inhibition
- **GPT-5 Reasoning:**
  1. Inhibition of acetylcholinesterase increases synaptic acetylcholine in cardiac parasympathetic pathways.
  2. Excess acetylcholine overstimulates M2 muscarinic receptors, leading to slowed cardiac electrical conduction and increased vagal tone.
  3. Resulting electrical disturbances progress to arrhythmias and impaired cardiac output, culminating in cardiotoxicity.
- **Predicted Toxicity:** cardiotoxicity

GPT-5 Final Predicted Toxicity: cardiotoxicity

Table H: Case study example illustrating strong alignment between an AChE-mediated Gold AOP (Aop:559) and the GPT-5 reasoning path. Highlighted segments denote aligned molecular initiating events, key events, and adverse outcomes along the AOP cascade.

Case Study: AOP Alignment & Reasoning Quality (GPT-5.1)

Below is an example instance illustrating alignment between the **Gold AOP context** and the **GPT-5.1** reasoning path for a renal toxicity case driven by COX1 inhibition.

Sample Metadata.

- **Index:** 70
- **SMILES:** `CC(C)C(=O)NC1=CC(=C(C=C1)[N+](=O)[O-])C(F)(F)F`

Ground Truth AOP Context:

- **AOP ID:** `Aop:177`
- **AOP Title:** Cyclooxygenase 1 (COX1) inhibition leading to renal failure and mortality
- **AOP:**
  Inhibition of Cyclooxygenase 1 (COX1) → *Decreased prostaglandin levels* → *Renal ischemia and oxidative stress* → *Renal proximal tubular necrosis* → *Renal failure and increased mortality*

GPT-5.1 Reasoning Response

- **MIE:** COX1 Inhibition
- **GPT-5.1 Reasoning:**
  1. Inhibition of COX1 in renal tissues reduces synthesis of vasodilatory prostaglandins that normally maintain renal blood flow and glomerular filtration.
  2. Reduced prostaglandin levels cause renal vasoconstriction, decreased perfusion, and medullary ischemia accompanied by oxidative stress.
  3. Sustained ischemic and oxidative injury progresses to proximal tubular damage and impaired renal function, resulting in kidney toxicity.
- **Predicted Toxicity:** kidney toxicity

GPT-5.1 Final Predicted Toxicity: kidney toxicity

Table I: Case study example illustrating alignment between a COX1-mediated renal toxicity AOP (Aop:177) and the GPT-5.1 reasoning path. Highlighted segments indicate correspondence between the molecular initiating event, key events, and adverse renal outcomes along the AOP cascade.

Case Study: AOP Alignment & Reasoning Quality (o3)

Below is an example instance illustrating strong alignment between the **Gold AOP context (Aop:318)** and the **o3** reasoning path, demonstrating accurate reproduction of a GR-mediated hepatic steatosis pathway.

Sample Metadata.

- **Index:** 54
- **SMILES:** `CC#CC1(CCC2C1(CC(C3=C4CCC(=O)C=C4CCC23)C5=CC=C(C=C5)N(C)C)C)O`
- **AOP ID:** `Aop:318`

Gold AOP Context.

- **AOP Title:** Glucocorticoid Receptor activation leading to hepatic steatosis
- **AOP:** *Activation of Glucocorticoid Receptor* $\rightarrow$ *Decrease of Acyl-CoA dehydrogenases* $\rightarrow$ *Decreased mitochondrial fatty acid β-oxidation* $\rightarrow$ *Accumulation of triglyceride* $\rightarrow$ *Increased liver steatosis*

o3 Reasoning Result

- **MIE:** GR Activation
- **o3 Reasoning:**
  1. The molecule activates the glucocorticoid receptor (GR) in hepatocytes, altering transcription of metabolic genes.
  2. GR-driven metabolic reprogramming promotes lipid storage and functionally reduces fatty-acid oxidation capacity, consistent with decreased mitochondrial β-oxidation.
  3. Persistent lipid accumulation hepatic triglyceride buildup, progressing to liver steatosis and liver toxicity.
- **Predicted Toxicity:** liver toxicity

o3 Final Predicted Toxicity: liver toxicity

Table J: Case study example illustrating strong alignment between a GR-mediated Gold AOP (Aop:318) and the model reasoning path. Highlighted segments denote aligned molecular initiating events, key events, and adverse outcomes along the AOP cascade.

Case Study: AOP Alignment & Reasoning Quality (Qwen3-4B-Instruct)

Below is an example instance illustrating transporter-mediated **kidney toxicity**, where the **Qwen3-4B-Instruct** reasoning path partially but coherently aligns with the **Gold AOP context (Aop:138)** describing OAT1 inhibition–driven renal failure.

Sample Metadata.

- **Index:** 8
- **SMILES:** `C1=CC(=C(C=C1C2=C(C=C(C=C2)F)F)C(=O)O)O`
- **Model:** Qwen3-4B-Instruct

Ground Truth AOP Context:

- **AOP ID:** `Aop:138`
- **AOP Title:** Organic anion transporter (OAT1) inhibition leading to renal failure and mortality
- **AOP:**
  OAT1 Inhibition → *Increased blood uric acid concentration* → *Urate deposition* → *Renal proximal tubular necrosis* → *Hyperkalemia* → *Renal failure / mortality*

Qwen3-4B Reasoning Result

- **MIE:** OAT1 Inhibition
- **Qwen Reasoning:**
  1. OAT1 inhibition reduces the renal uptake of organic anions, including drugs and endogenous metabolites.
  2. This leads to decreased renal clearance and elevated plasma concentrations of potentially nephrotoxic compounds.
  3. Accumulation of these substances induces tubular damage, impairing nephron function and progressing to kidney injury.
- **Predicted Toxicity:** kidney toxicity

Qwen3-4B Final Predicted Toxicity: kidney toxicity

Table K: Case study example illustrating a transporter-mediated kidney toxicity pathway. The Qwen3-4B model reproduces the core OAT1 inhibition mechanism and downstream renal injury events defined in the Gold AOP (Aop:138), despite additional off-target reasoning blocks in other samples. Highlighted segments denote aligned molecular initiating events, key events, and adverse outcomes.

Case Study: AOP Alignment & Reasoning Quality (ToxReason-4B-GRPO)

Below is an example instance illustrating ion-channel–mediated **cardiotoxicity**, where the **ToxReason-4B-GRPO** reasoning path shows strong and coherent alignment with the **Gold AOP context (Aop:555)** describing hERG/KCNH2 inhibition–driven electrophysiological dysfunction leading to heart failure.

Sample Metadata.

- **Index:** 28
- **SMILES:** `C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N`
- **Model:** ToxReason-4B-GRPO

Ground Truth AOP Context:

- **AOP ID:** `Aop:555`
- **AOP Title:** Inhibition, Ether-a-go-go (ERG) Voltage-Gated Potassium Channel leading to heart failure
- **AOP:**
  KCNH2 inhibition → *Prolongation of action potential duration* → *Prolongation of QT interval* → *Increased early premature depolarizations during repolarization* → *Heart failure*

ToxReason-4B-GRPO Reasoning Result

- **MIE:** KCNH2 inhibition
- **Model Reasoning:**
  1. KCNH2 (potassium channel hERG) inhibition reduces potassium efflux during cardiac action potential repolarization, preventing normal membrane potential restoration.
  2. Reduced repolarization leads to prolongation of action potential duration.
  3. Action potential prolongation causes QT interval prolongation, reflecting delayed ventricular repolarization.
  4. Extended repolarization creates conditions for early afterdepolarizations and premature depolarizations, increasing arrhythmogenicity.
  5. These electrophysiological disturbances culminate in heart failure.
- **Predicted Toxicity:** cardiotoxicity

ToxReason-4B-GRPO Final Predicted Toxicity: cardiotoxicity

Table L: Case study example illustrating an ion-channel–mediated cardiotoxicity pathway. The ToxReason-4B-GRPO model accurately reproduces the hERG/KCNH2 inhibition–driven electrophysiological mechanism defined in the Gold AOP (Aop:555), demonstrating strong alignment between predicted mechanistic reasoning and curated AOP knowledge. Highlighted segments denote aligned molecular initiating events, key events, and adverse outcomes.