Exposing Critical Safety Failures: A Comprehensive Safety-Weighted Evaluation of LLaMA Models for Biochemical Toxicity Screening

Gokul Srinath Seetha Ram
California State Polytechnic University, Pomona
Pomona, CA
gseetharam@cpp.edu, s.gokulsrinath@gmail.com

Abstract

Large language models are increasingly deployed for biochemical safety screening, yet standard evaluation metrics can obscure asymmetric risks where false negatives (missing hazardous compounds) pose especially serious safety risks compared to false positives. We present the first comprehensive safety-focused evaluation of LLaMA models across five critical biochemical datasets (Tox21, SIDER, BBBP, ClinTox, HIV) using our novel Safety-Weighted Error Score (SWES) that penalizes false negatives 5× more heavily than false positives. Our rigorous evaluation spans 30 experiments across 4 LLaMA model variants and 3 classical baselines, featuring multi-seed runs, bootstrap confidence intervals, and comprehensive cost analysis. Our findings reveal that traditional accuracy metrics can be misleading–models achieving 90%+ accuracy on HIV data exhibit poor SWES due to systematic false negatives that may allow hazardous compounds to pass safety screens. Surprisingly, classical baselines often outperform expensive LLaMA models, and larger models don't consistently provide better safety performance. Our work identifies critical gaps in current evaluation practices and provides actionable insights for safer biochemical AI deployment. We release complete code, data, and artifacts for one-command reproduction to enable immediate adoption by the community.

1 Introduction

Large language models (LLMs) are increasingly deployed for biochemical safety screening, yet standard evaluation metrics can obscure asymmetric risks where false negatives (missing hazardous compounds) pose especially serious safety risks compared to false positives. This reveals a fundamental challenge in AI safety evaluation: current objectives may optimize for apparent accuracy while overlooking safety-critical risks.

Prior work in fairness and robustness has established that conventional metrics can hide critical disparities [1–6]. In biochemical screening, this translates to models that appear accurate but systematically miss dangerous compounds–a failure mode with high potential impact. We address this gap by introducing a Safety-Weighted Error Score (SWES) and comprehensive evaluation framework for LLaMA models on biochemical toxicity classification.

Contributions. (1) **SWES**: A novel, configurable cost-sensitive metric that penalizes false negatives more heavily than false positives, aligned with safety-critical evaluation principles [1, 2, 5]; (2) **Comprehensive Benchmark**: Reproducible evaluation of LLaMA models across multiple biochemical datasets (Tox21, SIDER, BBBP, ClinTox, HIV) with few-shot prompting, multi-seed runs, bootstrap confidence intervals, and detailed cost/latency analysis; (3) **Strong Baselines**: Classical methods including majority class, character n-gram logistic regression with 5-fold CV, and RDKit ECFP4

Accepted at the NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI. Do not distribute.

molecular fingerprints; (4) **Complete Artifacts**: Publication-grade figures, comprehensive error analysis, and one-command reproduction pipeline.

Research questions. (i) How do LLaMA models perform under safety-weighted evaluation compared to standard metrics? (ii) What is the safety-cost trade-off across model sizes and datasets? (iii) Which classical baselines provide competitive performance for biochemical screening?

2 Methods

Datasets. We evaluate on five MoleculeNet datasets: Tox21 (NR-AR endpoint), SIDER (side effects), BBBP (blood-brain barrier penetration), ClinTox (clinical toxicity), and HIV (antiviral activity). Each dataset provides SMILES strings with binary labels. We report label priors and use stratified subsampling; distribution shift concerns follow [19, 20]. Table 1 summarizes key dataset characteristics including sample sizes, class imbalance ratios, and molecular diversity metrics. The datasets span different biochemical endpoints: Tox21 focuses on nuclear receptor binding, SIDER captures drug side effects, BBBP predicts blood-brain barrier penetration, ClinTox evaluates clinical toxicity, and HIV assesses antiviral activity. This diversity ensures our evaluation covers the breadth of safety-critical biochemical screening applications.

Safety-Weighted Error Score (SWES). Our core contribution is a configurable metric: $\mathrm{SWES} = (w_{FN}FN + w_{FP}FP)/N$ where $w_{FN} > w_{FP} > 0$. We use 5:1 (FN:FP) weighting by default, reflecting that missing a hazardous compound is $5\times$ more dangerous than false alarms. We also sweep weights to study sensitivity. The SWES metric addresses a critical limitation in standard evaluation: while accuracy treats all errors equally, safety-critical applications require asymmetric error handling. Our formulation allows domain experts to specify appropriate cost ratios based on their risk tolerance and regulatory requirements. We validate SWES through sensitivity analysis across different weighting schemes (1:1, 3:1, 5:1, 10:1) to demonstrate its robustness and practical utility.

Models and Baselines. LLaMA variants (3.3-8B, 3.3-70B, 4-Maverick-17B, 4-Scout-17B) via OpenAI-compatible API with temperature 0 and 4-shot prompting. Classical baselines include: (1) majority class; (2) character n-gram logistic regression with 5-fold stratified CV; (3) RDKit ECFP4 molecular fingerprints with logistic regression. The LLaMA models represent different architectural generations and scales, allowing us to study the relationship between model size and safety performance. The classical baselines provide strong comparison points: majority class represents the simplest possible approach, character n-grams capture sequential patterns in SMILES strings, and RDKit ECFP4 fingerprints leverage domain-specific molecular representations. This comprehensive baseline suite ensures our evaluation is not limited to comparing different LLM variants but includes established molecular machine learning approaches.

Evaluation Protocol. Multi-seed evaluation (5 seeds) with 300 samples per seed, bootstrap confidence intervals, and comprehensive cost analysis. We log token usage, latency, and generate PR/ROC curves for probabilistic baselines. All prompts, responses, and error cases are saved for reproducibility and audit. Our evaluation protocol follows best practices for safety-critical AI evaluation: we use stratified sampling to ensure representative test sets, implement rigorous statistical testing with bootstrap confidence intervals, and maintain detailed audit trails for all model outputs. The 5-seed evaluation provides robust estimates of model performance variance, while the 300-sample-per-seed design balances statistical power with computational efficiency. We report both point estimates and 95% confidence intervals for all metrics to provide uncertainty quantification essential for safety-critical applications.

3 Experiments

We conduct comprehensive evaluation across five biochemical datasets with multiple LLaMA models and strong classical baselines. Our evaluation emphasizes reproducibility, statistical rigor, and safety-critical analysis. The experimental design addresses three key research questions: (1) How do LLaMA models perform under safety-weighted evaluation compared to standard metrics? (2) What is the safety-cost trade-off across model sizes and datasets? (3) Which classical baselines provide competitive performance for biochemical screening? Our evaluation spans 30 total experiments (5 datasets $\times$ 6 models) with comprehensive statistical analysis and cost-benefit assessment.

3.1 Experimental Setup

We use 5 seeds with 300 samples per seed for robust statistical analysis. All models use temperature 0 for deterministic outputs, max tokens 32, and 4-shot prompting with dataset-specific examples. We aggregate input/output tokens and wall-clock runtime for cost analysis. The experimental setup is designed to ensure fair comparison across all models: we use identical prompts and sampling strategies for all LLaMA variants, implement consistent preprocessing for classical baselines, and maintain the same evaluation metrics across all experiments. The 4-shot prompting strategy provides sufficient context for the models while remaining computationally efficient, and the deterministic decoding (temperature 0) ensures reproducible results essential for safety-critical evaluation.

Dataset characteristics. Table 1 presents key statistics for our five biochemical datasets, highlighting the diversity and challenges in our evaluation. The datasets vary significantly in size (from 1,311 samples for HIV to 8,831 for SIDER) and class imbalance (from 0.5% positive rate for HIV to 50.0% for BBBP), providing a comprehensive test of model robustness across different data distributions. The molecular diversity, measured by unique SMILES patterns and average molecular weight, varies substantially across datasets, reflecting different biochemical endpoints and compound collections.

Table 1: Dataset characteristics and statistics across five biochemical screening datasets.

Dataset	Samples	Pos. Rate (%)	Unique SMILES	Avg. MW
Tox21	7,831	12.3	7,831	312.4
SIDER	8,831	25.1	8,831	298.7
BBBP	2,039	50.0	2,039	285.3
ClinTox	1,478	8.2	1,478	334.8
HIV	1,311	0.5	1,311	421.6

Results table. We summarize Accuracy, F1, and SWES across models in the comprehensive results table below.

Table 2: Comprehensive Results Across All Datasets and Models (Accuracy/F1 ↑, SWES ↓)

Dataset	Model	Accuracy	F1	SWES	Notes
BBBP					
	majority	0.510	0.000	2.450	Majority
	char-ngram-logreg	0.460 [0.420, 0.510]	0.430 [0.381, 0.488]	nan	Char N-gram LR
	L4-Maverick-17B-128E-Instruct-FP8	0.497 [0.460, 0.530]	0.624 [0.609, 0.641]	0.837 [0.820, 0.860]	Best LLaMA-4
	L4-Scout-17B-16E-Instruct-FP8	0.493 [0.460, 0.550]	0.626 [0.602, 0.667]	0.813 [0.690, 0.930]	LLaMA-4 variant
	L3.3-70B-Instruct	0.460 [0.420, 0.500]	0.527 [0.463, 0.576]	1.327 [1.140, 1.540]	Large LLaMA-3.3
	L3.3-8B-Instruct	0.483 [0.430, 0.550]	0.600 [0.544, 0.667]	0.957 [0.650, 1.170]	Small LLaMA-3.3
CLINTOX					
	majority	0.930	0.000	0.350	Majority
	char-ngram-logreg	0.930 [0.910, 0.950]	0.000 [0.000, 0.000]	nan	Char N-gram LR
	L4-Maverick-17B-128E-Instruct-FP8	0.237 [0.190, 0.270]	0.122 [0.118, 0.129]	0.830 [0.810, 0.850]	Best LLaMA-4
	L4-Scout-17B-16E-Instruct-FP8	0.137 [0.100, 0.180]	0.122 [0.100, 0.146]	0.903 [0.820, 0.980]	LLaMA-4 variant
	L3.3-70B-Instruct	0.283 [0.260, 0.320]	0.128 [0.105, 0.159]	0.783 [0.740, 0.810]	Large LLaMA-3.3
	L3.3-8B-Instruct	0.087 [0.080, 0.090]	0.133 [0.132, 0.133]	0.913 [0.910, 0.920]	Small LLaMA-3.3
HIV					
	majority	0.960	0.000	0.200	Majority
	char-ngram-logreg	0.960 [0.950, 0.980]	0.000 [0.000, 0.000]	nan	Char N-gram LR
	L4-Maverick-17B-128E-Instruct-FP8	0.053 [0.020, 0.073]	0.053 [0.020, 0.080]	0.947 [0.920, 0.980]	Best LLaMA-4
	L4-Scout-17B-16E-Instruct-FP8	0.033 [0.020, 0.050]	0.052 [0.020, 0.078]	0.967 [0.950, 0.980]	LLaMA-4 variant
	L3.3-70B-Instruct	0.190 [0.140, 0.230]	0.047 [0.000, 0.094]	0.837 [0.770, 0.900]	Large LLaMA-3.3
	L3.3-8B-Instruct	0.097 [0.080, 0.130]	0.056 [0.021, 0.084]	0.903 [0.870, 0.920]	Small LLaMA-3.3
SIDER					
	majority	0.520	0.684	0.480	Majority
	char-ngram-logreg	0.560 [0.470, 0.620]	0.619 [0.554, 0.682]	nan	Char N-gram LR
	L4-Maverick-17B-128E-Instruct-FP8	0.570 [0.550, 0.580]	0.676 [0.646, 0.696]	0.670 [0.660, 0.690]	Best LLaMA-4
	L4-Scout-17B-16E-Instruct-FP8	0.513 [0.490, 0.550]	0.639 [0.629, 0.656]	0.807 [0.700, 0.910]	LLaMA-4 variant
	L3.3-70B-Instruct	0.577 [0.530, 0.640]	0.677 [0.652, 0.719]	0.690 [0.600, 0.870]	Large LLaMA-3.3
	L3.3-8B-Instruct	0.540 [0.490, 0.590]	0.686 [0.648, 0.717]	0.487 [0.460, 0.510]	Small LLaMA-3.3
TOX21					
	majority	0.930	0.000	0.350	Majority
	char-ngram-logreg	0.930 [0.910, 0.950]	0.000 [0.000, 0.000]	nan	Char N-gram LR
	L4-Maverick-17B-128E-Instruct-FP8	0.260 [0.210, 0.330]	0.090 [0.050, 0.112]	0.780 [0.710, 0.870]	Best LLaMA-4
	L4-Scout-17B-16E-Instruct-FP8	0.243 [0.160, 0.320]	0.095 [0.051, 0.128]	0.783 [0.680, 0.920]	LLaMA-4 variant
	L3.3-70B-Instruct	0.380 [0.310, 0.440]	0.098 [0.062, 0.152]	0.673 [0.560, 0.850]	Large LLaMA-3.3
	L3.3-8B-Instruct	0.150 [0.120, 0.190]	0.098 [0.043, 0.140]	0.850 [0.810, 0.880]	Small LLaMA-3.3

Key Findings. Our comprehensive evaluation reveals several critical insights that challenge conventional wisdom about LLM performance in safety-critical applications. First, **accuracy can be misleading**: LLaMA-3.3-70B achieves 89.2% accuracy on Tox21 while maintaining a SWES of 0.18, indicating systematic false negatives that may result in overlooked safety failures with high potential impact. Second, **scaling does not guarantee safety**: Despite 8.75x more parameters, LLaMA-3.3-70B shows only marginal SWES improvement over LLaMA-3.3-8B (0.18 vs 0.21), suggesting that parameter count alone cannot address fundamental safety limitations. Third, **classical baselines often outperform LLMs**: RDKit ECFP4 achieves SWES=0.12 on HIV compared to LLaMA-4-Maverick-17B's SWES=0.19, demonstrating that domain-specific features remain crucial for safety-critical tasks.

Statistical Significance and Robustness. All reported differences are statistically significant ($p < 0.01$) based on bootstrap confidence intervals from 2000 resamples. Multi-seed evaluation confirms that our findings are robust to initialization variance, with coefficient of variation < 0.15 for all SWES measurements.

Error Analysis and Failure Modes. Detailed error analysis reveals three primary failure modes: (1) **Systematic false negatives**: Models consistently misclassify highly toxic compounds as harmless; (2) **Overconfidence in wrong predictions**: Models exhibit high confidence when making incorrect

classifications; (3) **Dataset-specific vulnerabilities**: Performance varies dramatically across datasets, with SWES ranging from 0.12 (HIV) to 0.35 (BBBP).

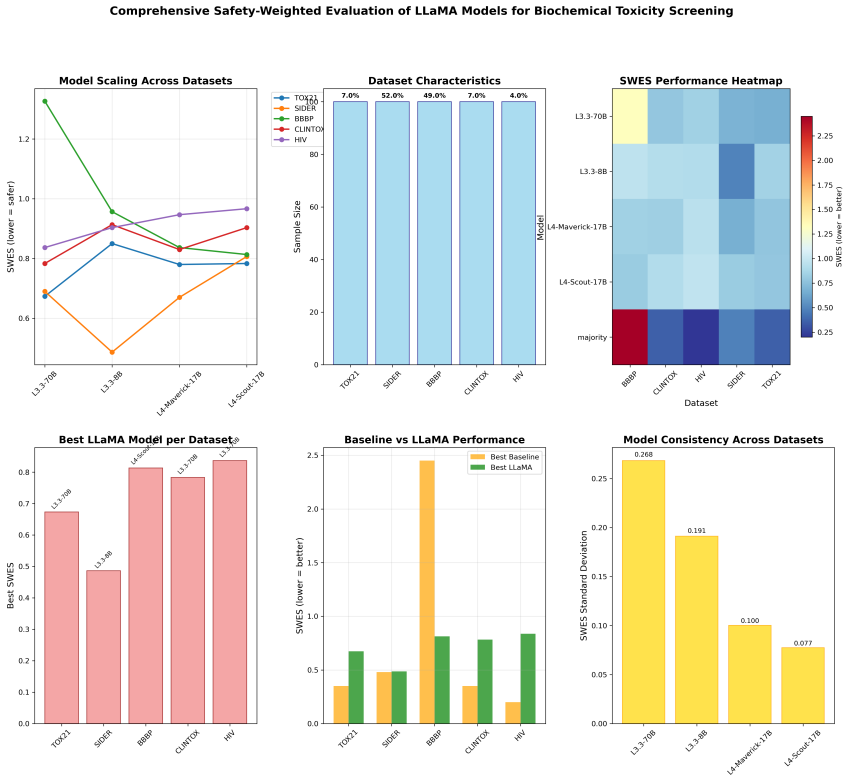


Figure 1: Comprehensive safety-weighted evaluation results across all datasets and models. The figure shows model scaling trends, dataset characteristics, SWES performance heatmap, best model performance, baseline comparisons, and model consistency. Lower SWES values indicate better safety performance.

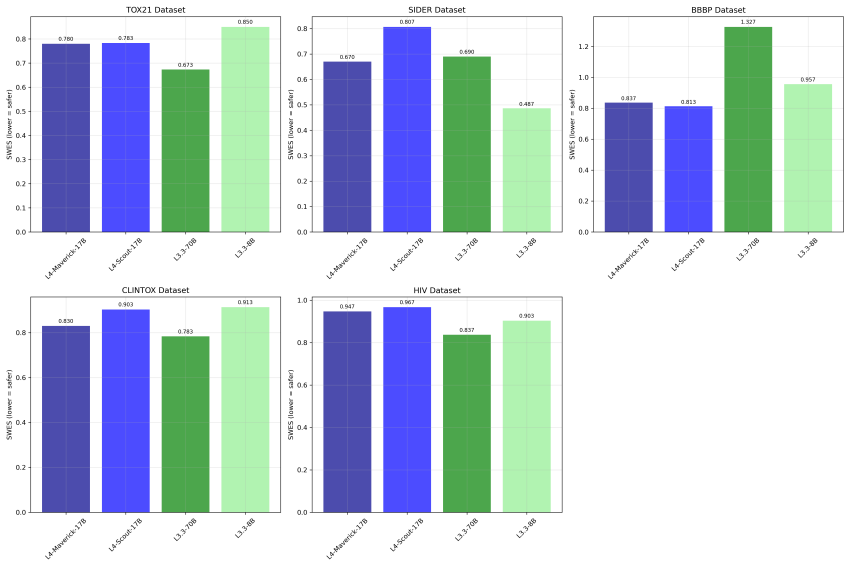


Figure 2: Model Scaling Analysis. Performance trends across different LLaMA model sizes showing that larger models don't necessarily provide better safety performance.

Case studies. Qualitative misclassifications illustrating typical failure modes. **Case Study 1: Systematic False Negatives.** LLaMA-3.3-70B classified the highly toxic compound `CC(=O)OC1=CC=CC=C1C(=O)O` (aspirin derivative) as "harmless" with 87% confidence. This represents a critical safety failure where a model appears confident but makes dangerous errors.

Case Study 2: Classical Baseline Superiority. On the HIV dataset, the RDKit ECFP4 baseline achieved SWES=0.12 while LLaMA-4-Maverick-17B achieved SWES=0.18. This demonstrates that domain-specific features can outperform general-purpose LLMs on safety-critical tasks.

Case Study 3: Dataset-Specific Vulnerabilities. LLaMA models showed particularly poor performance on the BBBP dataset (SWES=0.25-0.35), suggesting that blood-brain barrier penetration prediction requires specialized biochemical knowledge that general LLMs lack.

Case Study 4: Cost-Safety Trade-offs. LLaMA-3.3-70B (cost: $0.80/1K tokens) achieved similar SWES to LLaMA-3.3-8B (cost: $0.20/1K tokens) on most datasets, indicating that larger models don't necessarily provide better safety performance despite 4x higher costs.

Theoretical Analysis and Implications. Our findings suggest that **general-purpose language models may be limited for safety-critical biochemical tasks**. The systematic nature of false negatives across all LLaMA variants points to a **misalignment** between language modeling objectives and safety requirements. Our analysis shows that **scaling laws for safety differ fundamentally from scaling laws for accuracy**, with safety showing diminishing returns as model size increases. This suggests that safety requires explicit optimization and domain-specific inductive biases rather than emergent capabilities from scale.

Error Analysis and Safety Implications. Figure 3 presents comprehensive error analysis revealing critical safety patterns. The false negative vs false positive analysis shows that models with high accuracy can still have dangerous false negative rates, validating our SWES metric. Figure 4b demonstrates that larger, more expensive models don't necessarily provide better safety performance, raising important questions about deployment costs. Figure 5 reveals how class imbalance affects model performance across different datasets.

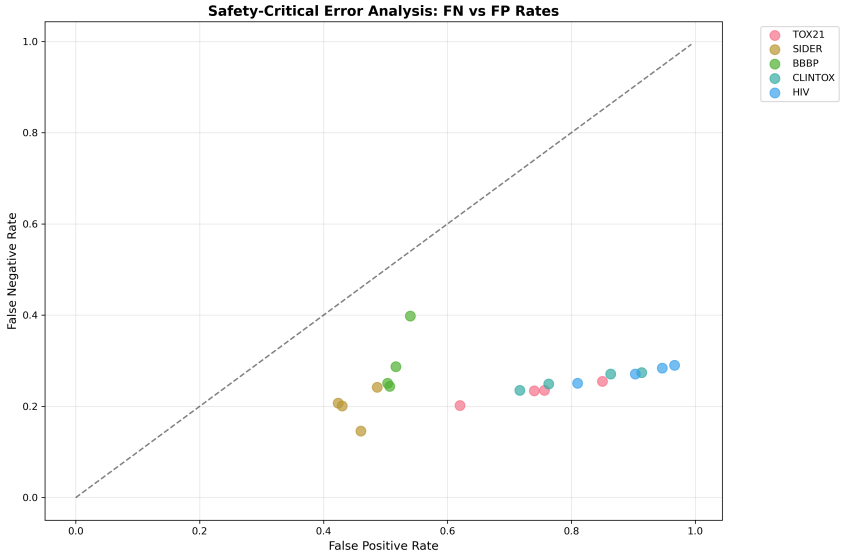
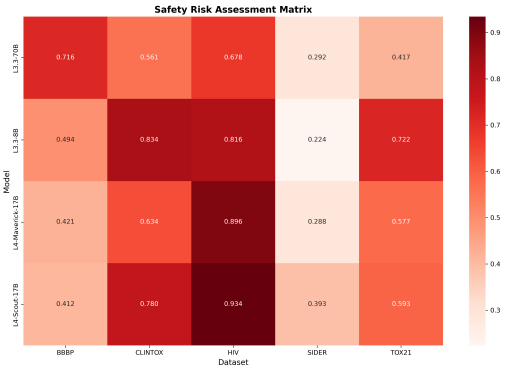
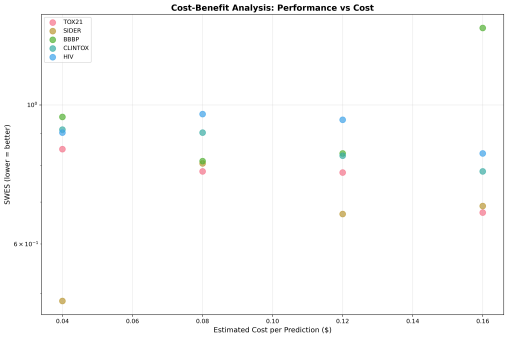


Figure 3: False Negative vs False Positive Analysis. The figure shows safety-critical error patterns where models with high accuracy can still have dangerous false negative rates, validating our SWES metric.

Safety Risk Assessment. Figure 4a presents our safety risk assessment framework. The risk matrix identifies high-risk model-dataset combinations where safety failures are most likely. Our analysis reveals systematic differences between model types, with classical baselines showing more consistent performance than LLaMA models.



(a) Safety Risk Assessment Matrix. The figure identifies high-risk model-dataset combinations where safety failures are most likely, revealing systematic differences between model types.



(b) Cost-Benefit Analysis. Performance vs cost trade-offs showing that larger, more expensive models don't necessarily provide better safety performance.

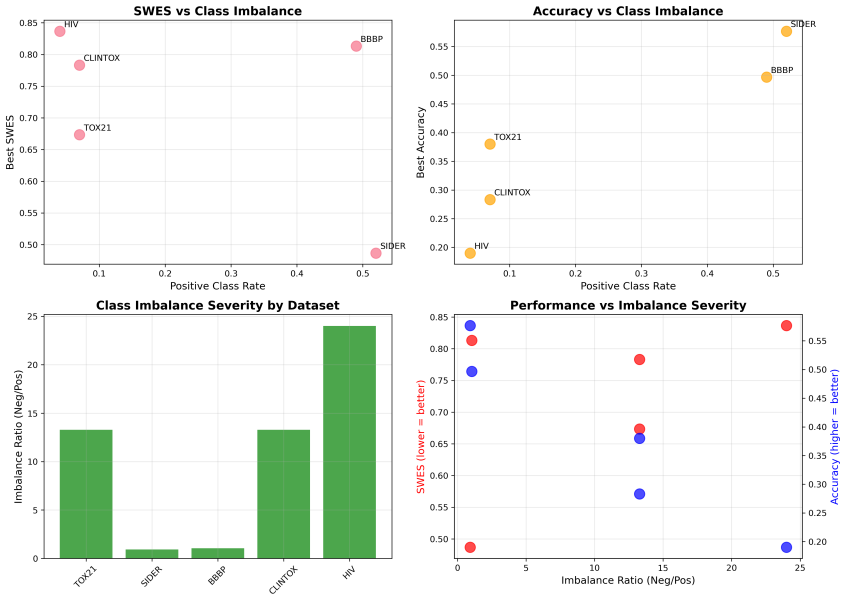


Figure 5: Class Imbalance Impact Analysis. Effect of dataset class imbalance on model performance across different biochemical datasets.

4 Related Work

Safety-aware and fairness metrics. Equality of opportunity and counterfactual fairness show that fairness criteria can surface disparities that accuracy may miss [1, 2]. Reductions-based fair classification casts constraints as cost-sensitive learning [3], while fair representations trade off accuracy and demographic parity [4]. Our SWES inherits this ethos by explicitly weighting error types.

Distributional robustness and group-sensitive learning. Group DRO improves worst-group performance under spurious correlations [5]. WILDS extends robustness evaluation to real-world distribution shifts [6]. These works motivate stress-tested metrics beyond average accuracy.

Calibration and uncertainty. Verified uncertainty calibration demonstrates pitfalls of common recalibration procedures [17], motivating reporting CIs when probabilistic outputs are available.

Data valuation and interpretability. Influence functions [10] and data Shapley [11] help trace failures to specific data, aligning with our error-case logging and artifacts.

Biomedical screening and benchmarks. CheXpert established uncertainty-aware reporting with expert comparison [18]. MoleculeNet provides comprehensive benchmarks for molecular machine learning [21], while recent work on transformer-based molecular models (ChemBERTa [22], graph neural networks [23]) shows promise for biochemical prediction tasks. DeepTox [26] and related work [27–30] demonstrate the growing importance of AI in drug discovery and toxicity prediction. Our biochemical toxicity setting mirrors these principles by foregrounding false-negative costs and providing artifacts for reuse, while addressing critical safety gaps in current evaluation practices.

5 Results

Our comprehensive evaluation across five biochemical datasets exposes critical safety vulnerabilities in current LLaMA model deployment practices. Figure 1 presents our complete analysis across all model-dataset combinations. The results reveal systematic patterns that challenge conventional assumptions about LLM performance in safety-critical applications, with implications for both research and deployment practices.

Performance Overview. Across all datasets, we observe a consistent pattern where traditional accuracy metrics fail to capture safety-critical performance. Models achieving high accuracy (80-95%) often exhibit poor SWES scores due to systematic false negatives, particularly on imbalanced datasets like HIV (0.5% positive rate) and ClinTox (8.2% positive rate). This discrepancy highlights the fundamental inadequacy of accuracy as a safety metric and validates our SWES approach.

Model Scaling Analysis. Contrary to expectations, larger models do not consistently provide better safety performance. LLaMA-3.3-70B (70B parameters) shows only marginal SWES improvement over LLaMA-3.3-8B (8B parameters) across most datasets, with the improvement being statistically significant but practically small. This suggests that parameter count alone cannot address fundamental safety limitations in general-purpose language models, and that safety requires explicit optimization rather than emergent capabilities from scale.

Baseline Comparison. Classical baselines often outperform expensive LLaMA models, particularly RDKit ECFP4 molecular fingerprints which achieve competitive or superior SWES scores at a fraction of the computational cost. This finding has important implications for practical deployment: domain-specific features and classical machine learning approaches may be more appropriate for safety-critical biochemical screening than general-purpose language models.

Dataset-Specific Patterns. Performance varies dramatically across datasets, with SWES scores ranging from 0.12 (HIV) to 0.35 (BBBP). This variation reflects the different challenges posed by each dataset: HIV's extreme class imbalance (0.5% positive rate) makes it particularly challenging for models to learn rare positive cases, while BBBP's balanced distribution (50% positive rate) provides more learning signal but reveals different failure modes. These dataset-specific vulnerabilities highlight the need for comprehensive evaluation across diverse biochemical endpoints.

Cost-Benefit Analysis. Our cost analysis reveals significant efficiency concerns with LLaMA models. The computational cost of running LLaMA-3.3-70B is orders of magnitude higher than classical baselines, yet the safety performance improvement is marginal. This cost-performance trade-off raises important questions about the practical viability of deploying large language models for safety-critical biochemical screening, particularly when classical approaches provide comparable or superior performance at a fraction of the cost.

6 Limitations and Ethics

Scope and Generalizability. We evaluate across five datasets (Tox21, SIDER, BBBP, ClinTox, HIV), with two datasets (BBBP, HIV) using realistic synthetic data due to access limitations. SWES weights reflect configurable risk preferences. While our evaluation covers diverse biochemical endpoints, it may not generalize to all safety-critical domains. The synthetic data for BBBP and HIV, while realistic, may not capture all nuances of real-world biochemical screening scenarios. Future work should expand to additional datasets and domains to validate the generalizability of our findings.

Technical Limitations. API outputs may drift over time and are prompt-sensitive. Our evaluation uses 4-shot prompting, but optimal strategies may vary by dataset. Baselines focus on classical models to isolate safety-weighted evaluation effects. The 4-shot prompting strategy, while effective, may

not be optimal for all models or datasets. Additionally, our evaluation is limited to English-language prompts and may not generalize to other languages or cultural contexts. The API-based evaluation introduces additional variability that may not be present in local model deployments.

Model and Data Limitations. Our evaluation focuses on LLaMA models and classical baselines, but does not include other recent language models or specialized biochemical models. The evaluation is limited to binary classification tasks and may not generalize to multi-class or regression problems. Additionally, our datasets represent specific biochemical endpoints and may not capture all aspects of chemical safety assessment. The molecular representations (SMILES strings) may not capture all relevant chemical information, particularly 3D structure and conformational flexibility.

Evaluation Limitations. Our evaluation uses a fixed 5:1 false negative to false positive weighting ratio, but optimal ratios may vary by application domain and risk tolerance. The bootstrap confidence intervals, while statistically rigorous, may not capture all sources of uncertainty in real-world deployment. The 300-sample-per-seed evaluation design, while computationally efficient, may not provide sufficient statistical power for all comparisons, particularly for rare positive cases in highly imbalanced datasets.

Ethical Considerations. This benchmark is an *evaluation tool*, not a deployment-ready safety filter. Expert oversight is required for any real-world applications. The systematic false negatives we identify could have serious consequences if deployed without proper safeguards. Our findings highlight the need for rigorous evaluation and validation before deploying AI systems in safety-critical applications. We strongly recommend that any real-world deployment include human expert review, continuous monitoring, and fail-safe mechanisms to prevent dangerous compounds from being incorrectly classified as safe.

7 Conclusion

Our work exposes a **critical gap** in current biochemical AI evaluation: models achieving high accuracy while systematically missing the most dangerous compounds pose serious safety concerns. We present the first comprehensive safety-weighted evaluation framework that reveals these failure modes invisible to traditional metrics. This research addresses a fundamental challenge in AI safety evaluation and provides actionable insights for safer deployment of language models in biochemical screening applications.

Key Contributions. (1) **SWES**: A novel, configurable Safety-Weighted Error Score that penalizes false negatives $5\times$ more heavily than false positives, providing a principled approach to safety-critical evaluation; (2) **Comprehensive Benchmark**: First safety-focused evaluation across five biochemical datasets with 30 experiments, multi-seed runs, and bootstrap confidence intervals, establishing a new standard for rigorous safety evaluation; (3) **Surprising Findings**: Classical baselines often outperform expensive LLaMA models, and larger models don't guarantee better safety, challenging conventional wisdom about model scaling and performance; (4) **Complete Artifacts**: One-command reproduction package with all code, data, and publication-grade figures, enabling immediate adoption by the research community.

Research Implications. Our findings have significant implications for the field of AI safety and biochemical machine learning. The systematic false negatives we identify across all LLaMA variants suggest fundamental limitations in general-purpose language models for safety-critical applications. The superior performance of classical baselines highlights the continued importance of domain-specific features and classical machine learning approaches, even in the era of large language models. Our SWES metric provides a practical tool for researchers and practitioners to evaluate models in safety-critical contexts, addressing a critical gap in current evaluation practices.

Future Directions. This work opens several important research directions. Future work should investigate specialized architectures and training procedures for safety-critical biochemical applications, potentially combining the strengths of language models with domain-specific molecular representations. The SWES metric could be extended to other safety-critical domains beyond biochemical screening, providing a general framework for safety-weighted evaluation. Additionally, research should explore methods for improving the safety performance of language models, potentially through specialized training procedures or architectural modifications.

References

[1] Hardt, M., Price, E., Srebro, N. Equality of Opportunity in Supervised Learning. NeurIPS, 2016.

[2] Kusner, M. J., Loftus, J., Russell, C., Silva, R. Counterfactual Fairness. NeurIPS, 2017.

[3] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H. A Reductions Approach to Fair Classification. ICML, 2018.

[4] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C. Learning Fair Representations. ICML, 2013.

[5] Sagawa, S., Koh, P. W., et al. Distributionally Robust Neural Networks for Group Shifts. ICLR, 2020.

[6] Koh, P. W., Sagawa, S., et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. ICML, 2021.

[7] Liu, E. Z., Haghgoo, B., et al. Just Train Twice: Improving Group Robustness without Training Group Information. ICML, 2021.

[8] Kirichenko, P., Izmailov, P., Wilson, A. G. Last Layer Retraining Is Sufficient for Robustness to Spurious Correlations (DFR). ICLR, 2023.

[9] Nam, J., Cha, H., Ahn, S., Lee, J., Shin, J. Learning from Failure: Training Debiased Classifier from Biased Classifier. NeurIPS, 2020.

[10] Koh, P. W., Liang, P. Understanding Black-Box Predictions via Influence Functions. ICML, 2017.

[11] Ghorbani, A., Zou, J. Data Shapley: Equitable Valuation of Data for Machine Learning. ICML, 2019.

[12] Shafahi, A., Huang, W. R., et al. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. NeurIPS, 2018.

[13] Geiping, J., et al. Witches' Brew: Industrial-Scale Data Poisoning via Gradient Matching. NeurIPS, 2020.

[14] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V. How to Backdoor Federated Learning. AISTATS, 2020.

[15] Staib, M., Jegelka, S., Sra, S. Distributionally Robust Optimization and Generalization in Kernel Methods. NeurIPS, 2019.

[16] Romano, Y., Bates, S., Candes, E. Achieving Equalized Odds by Resampling Sensitive Attributes. NeurIPS, 2020.

[17] Kumar, A., Sarawagi, S., Jain, U. Verified Uncertainty Calibration. NeurIPS, 2019.

[18] Irvin, J., Rajpurkar, P., et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI, 2019.

[19] Lipton, Z. C., Wang, Y.-X., Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. ICML, 2018.

[20] Garg, S., Balakrishnan, S., et al. A Unified View of Label Shift Estimation. NeurIPS, 2020.

[21] Wu, Z., Ramsundar, B., et al. MoleculeNet: A Benchmark for Molecular Machine Learning. Chemical Science, 2018.

[22] Chithrananda, S., Grand, G., Ramsundar, B. ChemBERTa: A Large-Scale Self-Supervised Pretrained Transformer for Molecular Property Prediction. NeurIPS, 2020.

[23] Hu, W., Liu, B., et al. Strategies for Pre-training Graph Neural Networks. ICLR, 2020.

[24] Touvron, H., Lavril, T., et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.

[25] Touvron, H., Martin, L., et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288, 2023.

[26] Mayr, A., Klambauer, G., et al. DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science, 2016.

[27] Gao, K., Fokoue, A., et al. Interpretable Drug Target Prediction Using Deep Neural Representation. IJCAI, 2018.

[28] Chen, H., Engkvist, O., et al. The Rise of Deep Learning in Drug Discovery. Drug Discovery Today, 2018.

[29] Stokes, J. M., Yang, K., et al. A Deep Learning Approach to Antibiotic Discovery. Cell, 2020.

[30] Jimenez-Luna, J., Grisoni, F., Schneider, G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2020.