Is My Language Model a Biohazard?

Aldan Creo*
Independent Author
Dublin, Ireland
research@acmc.fyi

Cristina Correa
Independent Author
Sunderland, United Kingdom
correa.cristina@protonmail.com

Abstract

The dual-use potential of language models in the chemical sciences presents a significant biosecurity challenge. We investigate a foundational aspect of this risk: whether LMs possess an intrinsic knowledge bias that favors toxic compounds over non-toxic ones. To address this, we systematically audit the latent chemical knowledge of twelve open-weight language models. We measure per-compound perplexity across a balanced dataset of 2,000 chemicals, comprising 1,000 toxic and 1,000 non-toxic compounds classified by the GHS08 "Health Hazard" standard. Our results reveal a consistent and statistically significant pattern: every model tested assigns lower perplexity, and therefore higher certainty, to the structures of toxic compounds. This finding demonstrates a systemic vulnerability across the current open-weight ecosystem, suggesting the risk is not merely a function of misuse but is embedded in the models' core knowledge. This intrinsic bias, possibly stemming from patterns in the training data, has profound implications for AI safety, as it may enhance model performance on a range of downstream tasks involving hazardous materials. Our work sheds light on this intrinsic vulnerability, and we make our code publicly available to support further research into this emergent risk.

1 Introduction

The rapid advancement of language models (LMs) has catalyzed a paradigm shift across scientific disciplines; they offer unprecedented capabilities in information synthesis, reasoning, and knowledge generation [24]. In the chemical and life sciences, this presents a dual-use dilemma of profound significance. Although LMs have an immense potential to accelerate beneficial research, such as biomolecular discovery [28, 8] and toxicological risk assessment [13], their ability to internalize and regurgitate specialized knowledge also reduces barriers to accessing information relevant to creating hazardous agents [11, 24, 23, 3]. This convergence of democratized expertise and inherent model capabilities necessitates that we rigorously and systematically evaluate the latent risks embedded within these systems. Our current understanding remains limited; we possess only a fragmented view of the chemical proficiencies of LMs, a critical prerequisite for enhancing models and alleviating potential harm [18].

The core of this biosecurity challenge lies in the hypothesis that LMs may internalize and recall information about hazardous materials differently from benign knowledge. Prior work argues that malicious actors can repurpose AI-powered tools to generate blueprints for novel toxic chemicals [26] and that chatbots can guide non-experts through complex biological protocols [24]. These studies, however, often focus on a model's ability to generate novel information upon request. We investigate a more fundamental question: do LMs skew their intrinsic knowledge representations? If models encode or access structural information for toxic compounds more readily than for their benign

*Website: https://acmc.fyi.

counterparts, it could indicate a systemic bias in the model's parameters that exacerbates misuse potential, independent of any overt malicious prompting.

These limitations in the current research landscape highlight a critical gap. We currently have no reliable way to forecast the trajectory of LM advancement, nor to understand how the convergence of specialized AI tools might synergistically amplify future biosecurity threats [20]. Therefore, we urgently need to develop robust, generalizable risk assessment benchmarks [11]. Our work addresses this need by moving beyond capability elicitation in single models to a large-scale, ecological analysis of knowledge representation across a diverse suite of state-of-the-art LMs. In this paper, we aim to determine whether a differential in recalling structural information, which we quantify through robust probabilistic metrics, is a consistent and generalizable phenomenon across the current open-weight model ecosystem.

To this end, we employ a multi-stage computational pipeline that integrates large-scale bioinformatics data acquisition, controlled model interrogation via sequence scoring, and rigorous non-parametric statistical analysis. We construct a balanced dataset of toxic and non-toxic compounds, which we meticulously classify according to international standards to focus on substances with significant misuse potential. We then interrogate a heterogeneous cohort of twelve instruction-tuned LMs, measuring their intrinsic certainty in recalling the precise structural representations for these compounds. We explicitly design our analysis to control for key confounds, such as molecular complexity, to ensure the integrity of our conclusions. We make three primary contributions:

1. We introduce a novel, reproducible methodology for auditing latent chemical knowledge in LMs;
2. we provide the first ecosystem-level analysis of differential knowledge regurgitation for toxic compounds; and
3. we offer a critical framework for evaluating this specific vector of AI-powered chemical risk.

2 Background

2.1 The Dual-Use Dilemma in AI for Science

The transformative potential of artificial intelligence in the life sciences is undeniable; it drives advances in genomic understanding, protein design and antibiotic discovery [23]. Concurrently, national governments and multinational bodies have identified the mitigation of AI misuse for chemical, biological, nuclear or radiological (CBRN) threats as a high-priority concern [20]. This dual-use nature means that the same capabilities that empower health security can also lower barriers to biological and chemical weapon development [23, 9]. LMs, which train on vast scientific corpora, sit at the epicenter of this concern. By disseminating specialist knowledge in an accessible format, they democratize research [24]; conversely, this also makes it increasingly likely that actors with nefarious intentions could leverage these models to access capabilities previously confined to experts. Consequently, leading AI developers have explicitly committed to researching safety in areas including the misuse of models for bio and chemical weapons development [9].

2.2 Evaluating Chemical Knowledge and Risks in LMs

The evaluation of chemical knowledge in LMs is an emerging field. One primary line of inquiry assesses a model's ability to understand and reason about chemical concepts, often finding that while performance is impressive, it remains reliant on external tools and databases that human experts also use [18]. A more security-focused line of research probes the potential for LMs to actively assist in designing hazardous agents. Seminal work illustrates that we can repurpose AI-powered drug discovery tools to design highly toxic molecules [26]. Furthermore, studies demonstrate that LMs can provide dual-use information that could have enabled historical biological weapons efforts to succeed [23] and can guide non-experts through the process of manufacturing risky pathogens [24]. These studies compellingly illustrate that LMs already possess dangerous capabilities. However, they primarily test a model's generative output in response to adversarial or leading prompts. Our work diverges by investigating a more foundational property: the structure of the model's internal

knowledge landscape. We ask not what the model will generate when pushed, but what it knows best by default, a property we measure using its intrinsic probabilistic confidence.

2.3 Methodological Gaps and the Need for Ecosystem Analysis

A critical limitation of the current biosecurity literature is its narrow scope. The limited published studies assess different risks using differing assumptions and, crucially, largely focus on individual, often proprietary, models [20]. This approach fails to capture whether a discovered capability is a general feature of modern LMs or an idiosyncrasy of a specific architecture or training dataset. Furthermore, we are not yet certain how advances in AI will exacerbate risks, and we recognize an absence of studies assessing how foundation models specifically trained on biological data or the "stacking" of AI tools will change the risk landscape [20]. Our methodology directly addresses the first part of this gap; we systematically evaluate a deliberately diverse suite of twelve open-weight models to ensure our conclusions reflect trends across an ecosystem rather than a single data point. This provides a necessary baseline for understanding the current state of risk before more advanced, specialized systems become widespread.

3 Methods

In this study, we employ a multi-stage computational pipeline to systematically evaluate whether LMs internalize and recall structural information for toxic chemical compounds more readily than for benign counterparts. Our central objective is to determine whether this differential knowledge regurgitation is a generalizable phenomenon observable across a heterogeneous model ecosystem. Our methodology integrates large-scale bioinformatics data acquisition, controlled model interrogation via sequence scoring, and robust non-parametric statistical analysis to interrogate latent knowledge representations. We implement all experiments in a version-controlled pipeline to ensure reproducibility and include explicit checks for confounds.

3.1 Compound Curation and Balanced Dataset Construction

Our investigation's validity hinges on a high-quality, balanced chemical dataset. We source compounds programmatically from PubChem [14] via its PUG-REST API and augment them with known toxicants from the Toxic Exposome Database (T3DB) [15]. We base our toxicity classification on the internationally recognized Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Specifically, we focus on compounds that present a serious health risk, as defined by the presence of the GHS08 "Health Hazard" pictogram, which identifies substances with chronic hazards such as carcinogenicity or reproductive toxicity. This focus directly aligns with our study's biosecurity objectives because it filters for compounds with significant misuse potential, and we deliberately exclude other categories such as acute toxicity or physical hazards.

We apply a stringent preprocessing filter to all compounds based on the length of their Simplified Molecular-Input Line-Entry System (SMILES) [27] string representations. We discard strings shorter than 50 or longer than 250 characters. This maintains a focus on molecules of non-trivial complexity, ensures sequences remain tractable for LM evaluation, and crucially, controls for inherent differences in molecular complexity that could act as confounds [10].

Our technical implementation utilizes custom classes for efficient data processing. An automatic data extractor parses PubChem XML dumps to extract 2000 Compound Identifiers (CIDs) and apply length constraints in situ (150 ± 100 SMILES characters). A collector then queries the PubChem PUG View API to retrieve comprehensive JSON objects, from which we extract properties including compound name, SMILES, and GHS classifications. We programmatically categorize compounds as *toxic_health* (GHS08 present), *toxic_physical* (other hazards only, which we purposefully discard), or *non-toxic* (no hazard codes). We engineer the sampling algorithm to procure equal numbers of *toxic_health* and *non-toxic* compounds, which enables a balanced case-control analysis [12].

3.2 Model Ecosystem Selection and Interrogation Paradigm

Reflecting our study's core objective to evaluate an ecosystem rather than individual models, we select a deliberately diverse suite of twelve state-of-the-art LMs for evaluation. This cohort

includes widely recognized models such as Meta's `Llama-3.2-3B-Instruct` and Microsoft's `Phi-4-mini-instruct`, alongside a range of competitively-sized models from IBM, Google, Alibaba, and European research consortia, with parameters spanning from 1B to 4.5B. This strategic selection ensures our analysis captures trends across the current open-weight landscape rather than idiosyncrasies of a single model. We provide a full inventory in Table 1.

Table 1: Inventory of the instruction-tuned language models we evaluated.

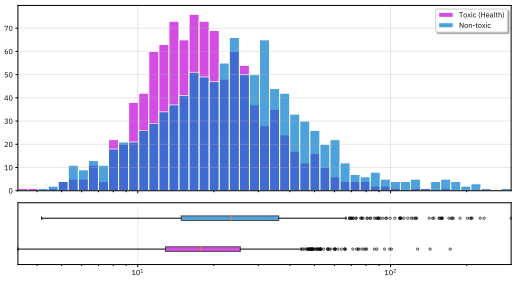
Model Name	Parameters (Billion)	Origin / Developer
`OLMo-2-0425-1B-Instruct` [19]	1.0	AllenAI
`EuroLLM-1.7B-Instruct` [16]	1.7	Utter Project
`Qwen3-1.7B-MegaScience` [7]	1.7	MegaScience
`gemma-3n-E2B-it` [25]	2 (effective) / 6 (brute)	Google
`granite-3.3-2b-instruct` [22]	2	IBM
`Instella-3B-Instruct` [5]	3	AMD
`SmolLM3-3B` [6]	3	Hugging Face
`Llama-3.2-3B-Instruct` [1]	3	Meta
`MediPhi` [4]	3.8	Microsoft
`Phi-3.5-mini-instruct` [17]	3.8	Microsoft
`Qwen3-4B-Instruct-2507` [29]	4	Alibaba (Qwen)
`AFM-4.5B` [2]	4.5	Arcee AI

We probe model knowledge using a controlled next-token prediction task, which provides a direct, continuous measure of internal certainty over open-ended generation. For each compound, we construct a prompt using the model's native chat template: "`Give me the SMILES for {compound_name}`", followed by the assistant's response prefix. We measure the models' intrinsic knowledge by the probability they assign to the token-by-token completion of the ground-truth SMILES string within this fixed context [21]. We derive a principal metric from the output logits: perplexity, which we calculate as the exponential of the average cross-entropy loss across the target sequence, quantifying the model's uncertainty. To ensure metric fidelity and control for sequence length, we calculate perplexity for all compounds only up to the length of the shortest SMILES string in our dataset. This prevents longer sequences, which tend to have lower perplexity, from confounding the results.

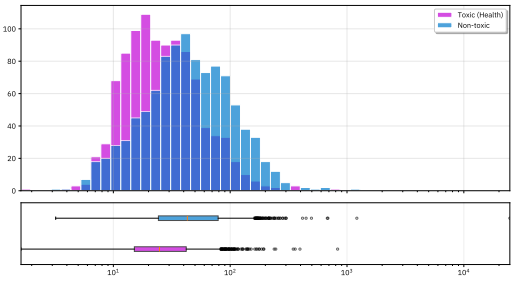
3.3 Statistical Analysis and Mitigation of Confounds

Our resultant dataset, which comprises perplexity metrics for each compound-model pair, we subject to a rigorous statistical analysis protocol. We test the central hypothesis of whether the distributions of perplexity differ significantly between compounds we classify as toxic (for health) and those we classify as non-toxic. Given the expected non-normality of the metrics, we employ non-parametric tests. We utilize the Mann-Whitney U test to determine if a statistically significant ($p < .00417$ after Bonferroni correction for 12 comparisons) difference exists in the median values of perplexity between the two groups. To complement null-hypothesis significance testing and to quantify the magnitude and direction of any observed effect, we calculate Cliff's Delta. This robust measure of effect size is independent of sample size and provides a standardized interpretation of the difference's magnitude, which offers a more nuanced understanding of the practical significance of our findings.

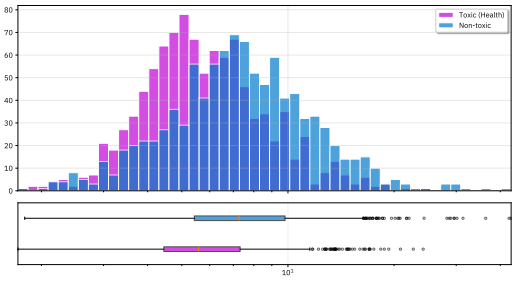
A paramount concern was the potential for confounding variables, primarily the length of the SMILES string, which could influence model performance independent of toxicity. Our analytical pipeline automatically tests for this; it analyzes the correlation (Pearson's r) between SMILES length and perplexity, and formally compares the length distributions between the two groups using the same non-parametric tests. We confirm that our truncation method successfully mitigates this confound; without it, length-perplexity correlations are non-negligible, but with it, the mean Pearson's r across all models is 0.0322 ± 0.0498, effectively negligible, which eliminates SMILES length as a confounding factor. We execute the entire analytical process within a single parameterized pipeline that ensures reproducibility.



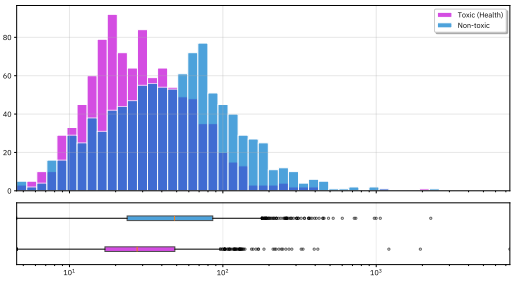
(a) `OLMo-2-0425-1B-Instruct`



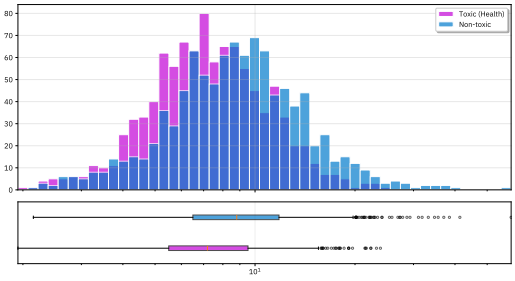
(b) `Instella-3B-Instruct`



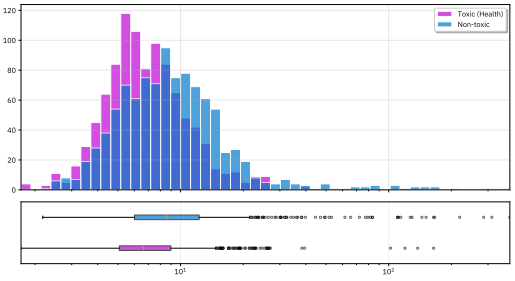
(c) `AFM-4.5B`



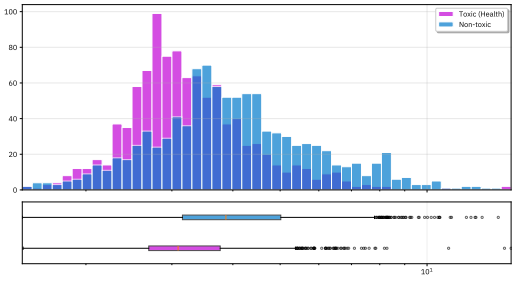
(d) `gemma-3n-E2B-it`



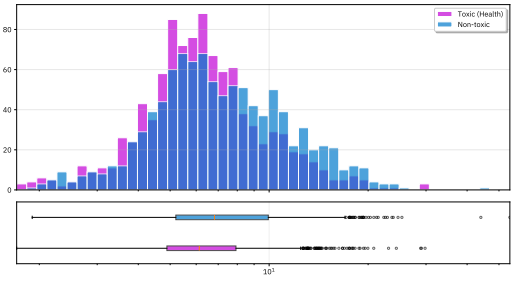
(e) `SmolLM3-3B`



(f) `granite-3.3-2b-instruct`



(g) `Qwen3-1.7B-MegaScience`

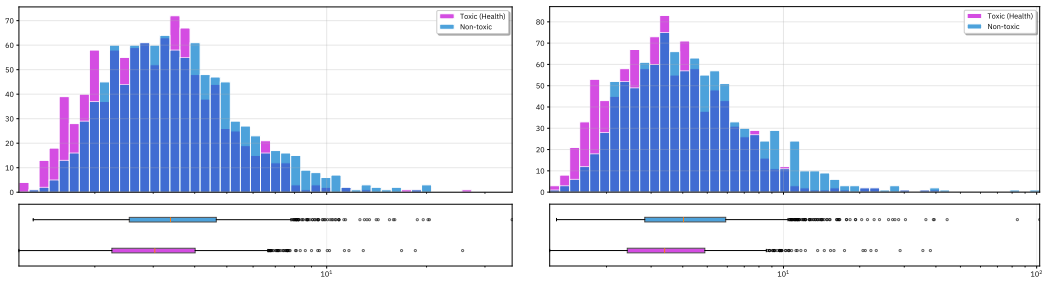


(h) `Llama-3.2-3B-Instruct`

Table 2: **Perplexity for toxic vs. non-toxic compounds.**

We report median perplexity (PPL) and interquartile range (median ± IQR) for each compound group across all evaluated models. We use the Mann-Whitney U test to calculate p-values and Cliff's Delta to measure effect size. All p-values are statistically significant ($p < .00417$).

Model Name	Toxic PPL	Non-Toxic PPL	p-value	Cliff's δ
`OLMo-2-0425-1B-Instruct`	17.8 ± 12.6	23.4 ± 21.3	1.909×10^{-19}	–0.233
`Instella-3B-Instruct`	24.8 ± 26.9	43.0 ± 54.6	1.021×10^{-41}	–0.349
`AFM-4.5B`	5.58 ± 2.87	7.26 ± 4.39	2.406×10^{-33}	–0.311
`gemma-3n-E2B-it`	27.6 ± 31.6	48.4 ± 62.2	1.283×10^{-31}	–0.302
`SmolLM3-3B`	7.19 ± 4.03	8.81 ± 5.31	8.132×10^{-22}	–0.248
`granite-3.3-2b-instruct`	6.64 ± 3.90	8.61 ± 6.28	1.439×10^{-27}	–0.281
`Qwen3-1.7B-MegaScience`	3.09 ± 1.08	3.87 ± 1.86	1.392×10^{-45}	–0.366
`Llama-3.2-3B-Instruct`	6.13 ± 3.04	6.82 ± 4.74	4.553×10^{-9}	–0.151
`MediPhi`	3.03 ± 1.76	3.38 ± 2.11	1.152×10^{-10}	–0.166
`Phi-3.5-mini-instruct`	3.39 ± 2.48	4.03 ± 3.08	7.287×10^{-14}	–0.193
`Qwen3-4B-Instruct-2507`	5.26 ± 3.30	5.95 ± 4.78	5.784×10^{-9}	–0.150
`EuroLLM-1.7B-Instruct`	6.51 ± 3.29	7.04 ± 3.94	3.368×10^{-4}	–0.093



(i) `MediPhi` (j) `Phi-3.5-mini-instruct`

(k) `Qwen3-4B-Instruct-2507` (l) `EuroLLM-1.7B-Instruct`

Figure 1: **Perplexity distributions for toxic vs non-toxic compounds across 12 open small language models.** Each subplot shows log-scaled histograms (top) and horizontal boxplots (bottom) comparing perplexity values between toxic and non-toxic chemical compounds.

4 Results

We summarize the quantitative findings from our evaluation cohort in Table 2 and Figure 1. Across all 12 evaluated models, the median perplexity assigned to toxic compounds is lower than for non-toxic compounds, and the difference is statistically significant after Bonferroni correction (adjusted threshold $p < .00417$) for every model reported in Table 2.

The range of observed effect sizes (Cliff's Delta) spans from –0.366 (largest magnitude, `Qwen3-1.7B-MegaScience`) to –0.093 (smallest magnitude, `EuroLLM-1.7B-Instruct`). The full

set of per-model means, standard deviations, Mann-Whitney U p-values, and Cliff's Delta values appear in Table 2.

Figure 1 provides per-model histograms and boxplots that illustrate the distributions underlying these statistics; for instance, subplots (b) and (g) show the distributions for `Instella-3B-Instruct` and `Qwen3-1.7B-MegaScience`, respectively, which are representative examples of the pattern reported in Table 2.

5 Discussion

Our analysis of twelve open-weight language models reveals a consistent and statistically significant finding: every model assigns lower perplexity to toxic chemical compounds than to non-toxic ones. As detailed in Table 2, this bias holds true across a diverse cohort of models from different developers and of varying sizes, with effect sizes ranging from negligible to medium. This demonstrates that the differential recall of toxic chemical information is a generalizable phenomenon in the current small LM ecosystem.

Such a consistent finding raises two pertinent questions: (1) why is this bias so pervasive, and (2) what are its implications for AI safety?

For the first question, we propose three non-mutually-exclusive hypotheses for this phenomenon, all of which center on the nature of the data these models are trained on.

- **Data Sourcing.** Toxic compounds are over-represented or more systematically documented in the vast corpora of text and data scraped from the internet. Sources such as patent filings, toxicology databases, or regulatory documents contain highly structured, repetitive and descriptive information about hazardous chemicals, which may make their properties and structures easier for LMs to learn.
- **Structural Regularity.** Certain classes of toxic compounds may share common structural motifs or functional groups that result in more predictable patterns in their SMILES representations, independent of string length. This underlying regularity could make them easier for a model to internalize.
- **Salience.** Toxic compounds may simply be more salient in human discourse; more frequently discussed in scientific literature, news articles and policy documents, leading to a richer and more robust representation in the models' weights.

While we believe these hypotheses might explain the observed biases, further research is required to disentangle these potential causes.

When it comes to the second question, the implications of this intrinsic bias are profound. Our findings suggest that the biosecurity risk posed by LMs is not merely a function of their ability to be steered by malicious prompts, but a foundational property of their current state. If a model's internal knowledge landscape is already skewed towards hazardous information, it is "primed" to generate it, potentially lowering the barrier for misuse. This risk is magnified because our study focuses on open-weight, small language models, which are computationally inexpensive and can be freely accessed, downloaded and modified by any actor, regardless of their resources or intent.

Furthermore, this enhanced intrinsic knowledge may translate to higher performance on a range of downstream tasks involving hazardous materials, not limited to the generation of chemical structures, which broadens the potential threat surface. This challenges the efficacy of safety measures that focus solely on input filtering or refusal mechanisms, as the vulnerability lies within the model's core knowledge base. And the fact that this is an ecosystem-wide phenomenon strongly suggests that the root cause lies not in any single model's architecture, but in the common data and methods used to train them.

Our work introduces a complementary lens for assessing dual-use risks. While much of the existing research rightly focuses on what dangerous capabilities can be elicited from models, our findings demonstrate the value of also auditing for intrinsic knowledge biases. The discovery of a systemic preference for toxic compounds suggests that the safety challenge is not limited to preventing misuse at inference time, but is also rooted in the very data and methods used to build these models. This is a relevant angle that ought to inform the assessment of AI biosecurity risks and mitigations.

6 Conclusion

In this work, we systematically evaluated the intrinsic chemical knowledge of twelve open-weight language models by measuring their perplexity on a balanced dataset of toxic and non-toxic compounds. Our analysis reveals a consistent and statistically significant bias: every model we tested demonstrated greater certainty when processing the structures of toxic compounds compared to non-toxic ones. This finding is significant because it shows a foundational vulnerability that is not dependent on adversarial prompting but is instead embedded in the models' core parameters. The existence of such a systemic bias across a diverse model ecosystem suggests that the challenge of ensuring AI safety in the chemical domain is deeply rooted in the data and methods used for training. As AI continues to be integrated into scientific research, understanding and mitigating these intrinsic biases will be a critical component of responsible innovation and biosecurity.

Limitations

We defined the scope of this study with several methodological constraints. Our dataset excluded both very short and highly complex SMILES strings; this choice reduced potential confounding from molecular complexity but also narrowed the chemical space we represented. Our evaluation focused exclusively on small- to mid-sized, general-purpose open-weight LMs, which leaves frontier-scale systems and domain-specialized chemical models outside our scope of analysis. Taken together, these constraints mean that while our findings suggest that, under controlled conditions, LMs exhibit systematic biases toward toxic compounds, we need broader investigations. Future work should incorporate more diverse toxicity categories, larger and specialized models, varied prompting strategies and adversarial testing to provide a more comprehensive assessment of biosecurity risks.

Ethical Statement

We conducted this study with careful consideration of the dual-use nature of chemical and biological research. We sourced all chemical data from publicly available, internationally recognized databases, and we included only compounds with well-documented structures. The sole purpose of our work is to improve understanding of the current biosecurity risk landscape associated with LMs, with the aim of supporting safe and responsible AI development, by highlighting potential vulnerabilities.

Supplementary Material

We include all relevant scripts, definitions, and configuration files and make them publicly accessible at `https://anonymous.4open.science/r/genai-biorisks-0446` to ensure reproducibility and facilitate further research.

References

[1] Meta AI. Llama 3 model card. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`, 2024. Accessed: 2025-08-14.

[2] Arcee.ai. AFM-4.5B: An instruction-tuned 4.5 b parameter foundation model. `https://huggingface.co/arcee-ai/AFM-4.5B`, 2025. Hugging Face model card. Parameters: 4.5 B, instruction-tuned; development details available on the model card.

[3] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models, 2023.

[4] Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment, 05 2025.

[5] Hugging Face. Instella-3b model. `https://huggingface.co/amd/Instella-3B`, 2024. Accessed: 2025-08-14.

[6] Hugging Face. Smollm3 announcement. https://huggingface.co/blog/smollm3, 2024. Accessed: 2025-08-14.

[7] Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *arXiv preprint arXiv:2507.16812*, 2025.

[8] A. Ghafarollahi and M. J. Buehler. Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning, 2024.

[9] Alexei Grinbaum and Laurynas Adomaitis. Dual use concerns of generative ai and large language models. *Journal of Responsible Innovation*, 11, 01 2024.

[10] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, January 2018.

[11] Gertrude Hattoh, Jeremiah Ayensu, Nyarko Prince Ofori, Solomon Eshun, and Darlington Akogo. Can large language models design biological weapons? evaluating moremi bio, 2025.

[12] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6:429–449, 11 2002.

[13] Ananth Rupesh Kattamreddy and Harisrujan Chinnam. The future of large language models in toxicological risk assessment: Opportunities and challenges. *Public Health and Toxicology*, 5:1–3, 02 2025.

[14] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Qingliang Li, Benjamin Shoemaker, Paul Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan Bolton. Pubchem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*, 49, 11 2020.

[15] Emilia Lim, Allison Pon, Yannick Djoumbou, Craig Knox, Savita Shrivastava, An Chi Guo, Vanessa Neveu, and David Wishart. T3db: A comprehensively annotated database of common toxins and their targets. *Nucleic acids research*, 38:D781–6, 11 2009.

[16] Pedro Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José Souza, Alexandra Birch, and André Martins. Eurollm: Multilingual language models for europe, 09 2024.

[17] Microsoft. Phi-3.5-mini-instruct: Lightweight instruction-tuned llm (3.8 b parameters). https://huggingface.co/microsoft/Phi-3.5-mini-instruct, 2025. Hugging Face model card. Dense decoder-only Transformer with 3.8 B parameters and 128 K token context length.

[18] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Elahi, Hani Elbeheiry, María Gil, Christina Glaubitz, Maximilian Greiner, Caroline Holick, Tim Hoffmann, and Kevin Jablonka. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17:1027–1034, 05 2025.

[19] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. 2024.

[20] Jaspreet Pannu, Doni Bloomfield, Robert MacKnight, Moritz Hanke, Alex Zhu, Gabe Gomes, Anita Cicero, and Thomas Inglesby. Dual-use capabilities of concern of biological ai models. *PLOS Computational Biology*, 21, 05 2025.

[21] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, 2019.

[22] IBM Research. Granite 3.3 language models. `https://github.com/ibm-granite/granite-3.3-language-models`, 2024. Accessed: 2025-08-14.

[23] Jonas B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *ArXiv*, abs/2306.13952, 2023.

[24] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. Can large language models democratize access to dual-use biotechnology?, 2023.

[25] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

[26] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial intelligence-powered drug discovery. *Nature Machine Intelligence*, 4, 03 2022.

[27] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 02 1988.

[28] Zhuang Xiang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, Qiang Zhang, and Huajun Chen. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 2024.

[29] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and Zihan Qiu. Qwen3 technical report, 05 2025.