From Descriptions to Chemical Hazards: Predicting Persistence, Bioaccumulation, and Toxicity from Natural Language Using LLMs

Sarathkrishna Swaminathan
IBM Research Almaden
San Jose, CA, USA
Sarath.Swaminathan@ibm.com

Nathaniel Park
IBM Research Almaden
San Jose, CA, USA
npark@us.ibm.com

Eduardo Soares
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
eduardo.soares@ibm.com

Emilio Vital Brazil
IBM Research Brazil
Rio de Janeiro, RJ, Brazil
evital@br.ibm.com

Abstract

Predicting chemical hazard indicators for substances of concern (SoCs), such as their persistence, bioaccumulation, and toxicity (PBT), is a critical task in environmental science and chemical regulatory compliance. Existing approaches rely heavily on molecular structural representations such as SMILES, which are often unavailable in early-stage assessments, in legacy documentation, or are inadequate for structurally representing the diversity of compounds encountered for regulation tasks. This paper addresses the challenge of estimating PBT properties from partial, noisy, and unstructured natural language descriptions of SoCs, such as their physical appearance, melting point, industrial use, and other general characteristics. We propose a new framework that leverages the generalization capabilities of Large Language Models (LLMs) to infer PBT profiles from these textual descriptions. Our key contributions include the development of the first dataset of natural language descriptions paired with PBT hazard categories and a fine-tuned LLM pipeline capable of generating hazard assessments. Experimental results show that our approach achieves competitive performance compared to structure-based models, enabling early hazard screening in low- or incomplete-data scenarios.

1 Introduction

The evaluation of chemical hazard indicators—persistence (P), bioaccumulation (B), and toxicity (T)—is central to environmental risk assessment and regulatory decision-making (1; 2). Substances of concern (SoCs) that score highly on these dimensions may pose long-term environmental and human health risks, motivating their prioritization for monitoring, restriction, or substitution (3). Accurate and early identification of hazardous compounds is thus critical for both regulatory compliance and the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

design of safer alternatives (4). Existing computational methods for PBT prediction overwhelmingly rely on molecular structural representations such as SMILES or InChI (5; 6). While effective, these methods face fundamental limitations (5). Structural information is frequently absent in early stages of risk evaluation, such as during regulatory screening of legacy industrial chemicals or new formulations with incomplete disclosure(7; 8). Furthermore, experimental metadata such as melting points, physical state, or applications—often recorded in fragmented or noisy natural language form—remains underutilized by current models, despite representing an important source of expert knowledge (9). This reliance on structural inputs restricts the applicability of existing models precisely in the low-data, high-uncertainty settings where early hazard screening is most valuable (10; 11).

Large Language Models (LLMs) present an opportunity to bridge this gap. Trained on massive and diverse text corpora, LLMs have demonstrated remarkable capabilities in reasoning under uncertainty, integrating heterogeneous information, and generalizing from noisy natural language descriptions (12). In the context of chemical safety, LLMs can be fine-tuned to map partial textual descriptions of substances—such as industrial use, appearance, or approximate physicochemical properties—to probabilistic assessments of PBT hazard categories (13; 14). This paradigm shifts the focus from strictly structure-based inference to text-driven reasoning, enabling hazard assessment even in the absence of canonical molecular identifiers (15; 16).

To support this objective, we construct a novel dataset of text-to-PBT mappings. Each entry consists of curated and anonymized molecular descriptions, stripped of explicit structural identifiers, and paired with model-predicted PBT labels obtained from established structure-based predictors. This dataset allows for instruction tuning of LLMs on the task of hazard inference from partial natural language.

Our contributions can be summarized as follows:

- We examine the limitations of current structure-based PBT prediction methods in low-data and incomplete-information scenarios.
- We introduce the first dataset of natural language descriptions paired with PBT hazard categories, designed for instruction tuning of LLMs.
- We propose a fine-tuned LLM framework for fine-grained PBT prediction from noisy and partial textual inputs.
- We demonstrate experimentally that our approach achieves competitive performance compared to structure-based baselines, highlighting its utility for early-stage hazard screening when structural information is unavailable.

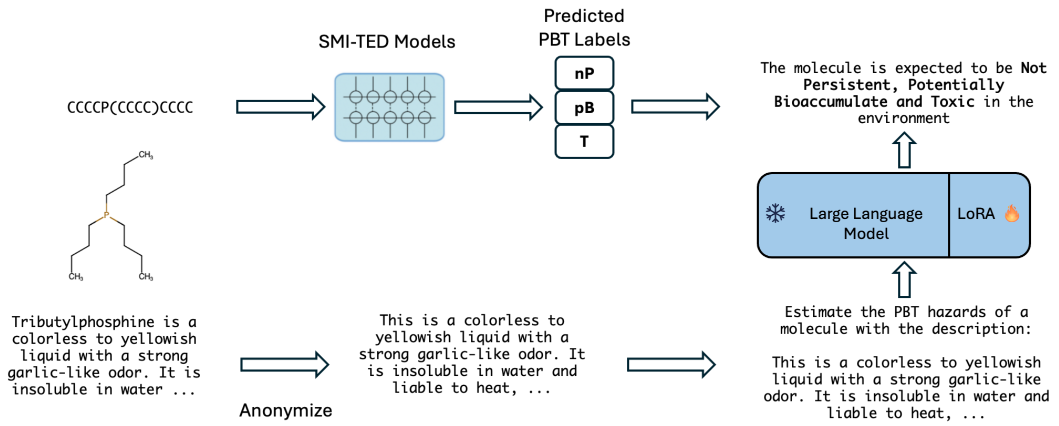
Taken together, our results suggest that LLM-based text-to-PBT prediction can extend the reach of computational toxicology into scenarios previously inaccessible to structure-based models, enabling more robust and timely chemical hazard evaluation.

2 Methods

2.1 Dataset Creation

We constructed a novel dataset to study the mapping between natural language descriptions of chemical compounds and their PBT properties. The dataset was derived from a collection of approximately 400,000 molecules sourced from PubChem. For each entry, we retained textual descriptions while systematically removing any structural identifiers, such as compound names, SMILES strings, IUPAC nomenclature, and registry numbers, thereby anonymizing the molecular structures and ensuring that the learning task relied exclusively on descriptive text.

To generate PBT labels, we employed a foundation model, SMI-TED (17), that estimated quantitative physicochemical and ecotoxicological endpoints, including $\log\mathrm{LC}_{50}$, $\log\mathrm{EC}_{50}$, $\log\mathrm{HalfLife}$, $\log\mathrm{K}_{ow}$, $\log\mathrm{K}_{oa}$, and $\log\mathrm{BCF}$. These predictions were post-processed into granular linguistic categories reflecting different levels of uncertainty, following the convention: *n* (not), *p* (probably), *P* (persistent, bioaccumulative, or toxic), and *v* (very). For instance, molecules with half-lives exceeding regulatory thresholds were labeled as “Very Persistent” (vP), while intermediate values yielded “Potentially Persistent” (pP) assignments. Similar mappings were created for bioaccumulation and toxicity endpoints, producing structured labels with explicit and fine-grained hazard quantification.



The molecule is expected to be **Not Persistent, Potentially Bioaccumulate and Toxic** in the environment

Tributylphosphine is a colorless to yellowish liquid with a strong garlic-like odor. It is insoluble in water ...

This is a colorless to yellowish liquid with a strong garlic-like odor. It is insoluble in water and liable to heat, ...

Estimate the PBT hazards of a molecule with the description:

This is a colorless to yellowish liquid with a strong garlic-like odor. It is insoluble in water and liable to heat, ...

Figure 1: Proposed method for PBT prediction from natural language descriptions. SMILES strings are converted to PBT labels using SMI-TED models. Descriptions are anonymized and formatted as input questions, which are then used to fine-tune an LLM with LoRA for fine-grained hazard assessment.

The final dataset thus enables description-to-PBT mapping tasks, with each example consisting of an anonymized natural language description paired with categorical PBT labels. To support robust evaluation, the dataset was randomly split into training, validation, and test subsets.

For instruction tuning, we derived four task formulations:

1. **Persistence**: description → {nP, pP, P, vP},
2. **Bioaccumulation**: description → {nB, pB, B, vB},
3. **Toxicity**: description → {nT, pT, T, vT},
4. **Composite PBT**: description → aggregated PBT label.

This design allows the model to learn fine-grained classification of property assignment, enabling a realistic treatment of chemical property prediction from text.

2.2 Model Creation

To effectively fine-tune LLMs for PBT prediction from natural language descriptions, we employed a fine-tuning strategy that prioritizes both efficiency and robustness. Our method leverages the strengths of LLMs while tackling the challenges posed by partial and noisy text inputs.

The full dataset showed a strong imbalance, with some classes having hundreds of thousands of samples while others had only a few hundred. We created a balanced subset by sampling across the individual PBT tasks in a way that reduces the effect of very common classes while keeping enough examples from the rare ones. This gave us a subset of about 31K examples with more even class proportions.

We then created a high-quality instruction-following dataset with this balanced subset. For each of the four PBT tasks, we manually wrote one template question. To make these more natural and less biased, we used *Mixtral-8x22B-instruct-v0.1*(18) to rephrase each template, generating 10 variations per prompt. Each response was structured to include the predicted PBT category based on the PBT task.

Given the high computational cost of full fine-tuning, we used Low-Rank Adaptation (LoRA) to efficiently adapt LLMs to our task. LoRA reduces the number of trainable parameters while largely preserving the model's core capabilities. This allowed us to fine-tune three state-of-the-art small language models - *Qwen3-8B*(19), *Granite-3.3-8B-instruct*(20) and *Mistral-7B-Instruct-v0.3*(21) - with minimal resource overhead. We trained each model for 4 epochs with a LoRA rank of 8, learning rate of $2e-5$ and batch size of 96

	Persistence		**Bioaccumulation**		**Toxicity**		**PBT**
Experiment	acc	F1-macro	acc	F1-macro	acc	F1-macro	Jaccard
Granite-3.3-8b-instruct	0.28	0.25	0.32	0.23	0.34	0.24	0.28
Llama-4-Maverick-17B-128E-Instruct	0.29	0.28	0.52	0.31	0.56	0.43	0.33
Mixtral-8x22B-Instruct-v0.1	0.28	0.27	0.54	0.31	0.53	0.32	0.22
gpt-oss-20b	0.29	0.27	0.60	0.29	0.49	0.41	0.34
gpt-oss-120b	0.28	0.24	0.58	0.36	0.57	0.39	0.29
Qwen3-8B	0.36	0.30	0.73	0.46	0.57	0.45	0.47
Granite-3.3-8b-instruct	0.68	0.62	0.77	0.60	0.72	0.66	0.62
Mistral-7B-Instruct-v0.3	**0.76**	**0.72**	**0.83**	**0.66**	**0.84**	**0.84**	**0.70**

Table 1: Performance comparison of our fine-tuned models (bottom section) against 4-shot baseline models (top section) for PBT prediction tasks. The metrics reported are accuracy (acc) and F1-macro scores for individual P, B, and T tasks, while Jaccard similarity is used to evaluate the overall composite PBT prediction performance.

3 Experiments and Results

We evaluated model performance using task-specific metrics. For the individual PBT tasks, we reported accuracy and F1 macro scores, which provide a balanced view of performance across all classes. For the Composite PBT task, we used the Jaccard similarity score to measure the overlap between predicted and ground-truth labels, which is well suited for multi-label classification.

For baseline comparisons, we evaluated several state-of-the-art open-source models, including *Granite-3.3-8b-instruct*, *Llama-4-Maverick-17B-128E-Instruct*, *Mixtral-8x-22B-Instruct-v0.1*, *gpt-oss-20b* and *gpt-oss-120b*(22). These baselines were tested using n-shot prompting (n=4) with diverse examples to ensure fair comparison against our fine-tuned models.

The results in Table 3 show that our fine-tuned models significantly outperform the 4-shot baselines on all tasks. Notably, *Mistral-7B-Instruct-v0.3* achieved accuracy scores of 0.76, 0.83 and 0.84 for P, B, and T, respectively. The gains were especially significant for the Composite PBT task, where our models achieved Jaccard scores up to 0.70, compared to $0.22 - 0.34$ for the baselines. This suggests that fine-tuning with our instruction dataset effectively captures the complex link between natural language descriptions and PBT properties, even without structural data.

4 Conclusion and Future Work

In summary, we have demonstrated the validity of our approach by leveraging LLMs to map the complexity of heterogeneous natural language molecular descriptions to their respective PBT properties, enabling a enhanced capabilities for their prediction.

Future work will focus on developing a fine-tuned LLM framework for probabilistic PBT prediction, that integrates uncertainy quantification, a critical component for decision-making in regulatory contexts where errors carry high societal and environmental cost. We will also iteratively refine the content of molecular descriptions as well as to further increase its size and chemical diversity. These improvements will likely both increase downstream model performance in addition to enhancing its utility in predicting PBT values across much broader ranges of chemicals and materials.

References

[1] C. Zarfl and M. Matthies, “Pbt borderline chemicals under reach,” *Environmental Sciences Europe*, vol. 25, no. 1, p. 11, 2013.

[2] M. Matthies, K. Solomon, M. Vighi, A. Gilman, and J. V. Tarazona, “The origin and evolution of assessment criteria for persistent, bioaccumulative and toxic (pbt) chemicals and persistent organic pollutants (pops),” *Environmental Science: Processes & Impacts*, vol. 18, no. 9, pp. 1114–1128, 2016.

[3] H. Sanderson, P. Fauser, L. Bengtström, and K. Vorkamp, "Semi-quantitative risk-based prioritisation scheme for chemicals of concern in the nordic countries," *RSC Sustainability*, vol. 2, no. 2, pp. 558–566, 2024.

[4] N. R. Council, D. on Earth, B. on Environmental Studies, B. on Chemical Sciences, C. on the Design, E. of Safer Chemical Substitutions, A. F. to Inform Government, and I. Decisions, "A framework to guide selection of chemical alternatives," 2014.

[5] P. De and K. Roy, "Greener chemicals for the future: Qsar modelling of the pbt index using eta descriptors," *SAR and QSAR in Environmental Research*, vol. 29, no. 4, pp. 319–337, 2018.

[6] S. Cassani and P. Gramatica, "Identification of potential pbt behavior of personal care products by structural approaches," *Sustainable Chemistry and Pharmacy*, vol. 1, pp. 19–27, 2015.

[7] C. A. Merlic, I. Schroder, and C. M. Kolodziej, "Challenges of legacy chemicals," *ACS Chemical Health & Safety*, vol. 32, no. 1, pp. 16–21, 2024.

[8] M. R. I. Rayhan, A. S. Shohag, K. A. Riya, J. M. Liza, M. M. Rahman, and M. S. Rahaman, "Legacy and emerging contaminants: Discussions and legislative advances." Springer, 2024.

[9] G. Piir, I. Kahn, A. T. García-Sosa, S. Sild, P. Ahte, and U. Maran, "Best practices for qsar model reporting: physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints," *Environmental health perspectives*, vol. 126, no. 12, p. 126001, 2018.

[10] M. T. Cronin, A.-N. Richarz, and T. W. Schultz, "Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (qsars) for toxicity prediction," *Regulatory Toxicology and Pharmacology*, vol. 106, pp. 90–104, 2019.

[11] S. S. Kolmar and C. M. Grulke, "The effect of noise on the predictive limit of qsar models," *Journal of Cheminformatics*, vol. 13, no. 1, p. 92, 2021.

[12] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar, "A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions," *ACM Computing Surveys*, 2025.

[13] W. Zhang, Q. Wang, X. Kong, J. Xiong, S. Ni, D. Cao, B. Niu, M. Chen, Y. Li, R. Zhang *et al.*, "Fine-tuning large language models for chemical text mining," *Chemical science*, vol. 15, no. 27, pp. 10 600–10 611, 2024.

[14] J. Park, Y. Park, M. Song, S. Park, D. Lee, S. Baek, and J. Kang, "Cotox: Chain-of-thought-based molecular toxicity reasoning and prediction," *arXiv preprint arXiv:2508.03159*, 2025.

[15] S. Balaji, R. Magar, Y. Jadhav, and A. B. Farimani, "Gpt-molberta: Gpt molecular features language model for molecular property prediction," *arXiv preprint arXiv:2310.03030*, 2023.

[16] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez, and K. M. Jablonka, "From text to insight: large language models for chemical data extraction," *Chemical Society Reviews*, 2025.

[17] E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira, and K. Schmidt, "An open-source family of large encoder-decoder foundation models for chemistry," *Communications Chemistry*, vol. 8, no. 1, p. 193, 2025.

[18] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," 2024. [Online]. Available: https://arxiv.org/abs/2401.04088

[19] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[20] I. Granite Team, "Granite 3.0 language models," *URL: https://github. com/ibm-granite/granite-3.0-language-models*, 2024.

[21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[22] OpenAI, "gpt-oss-120b gpt-oss-20b model card," 2025. [Online]. Available: https://arxiv.org/abs/2508.10925