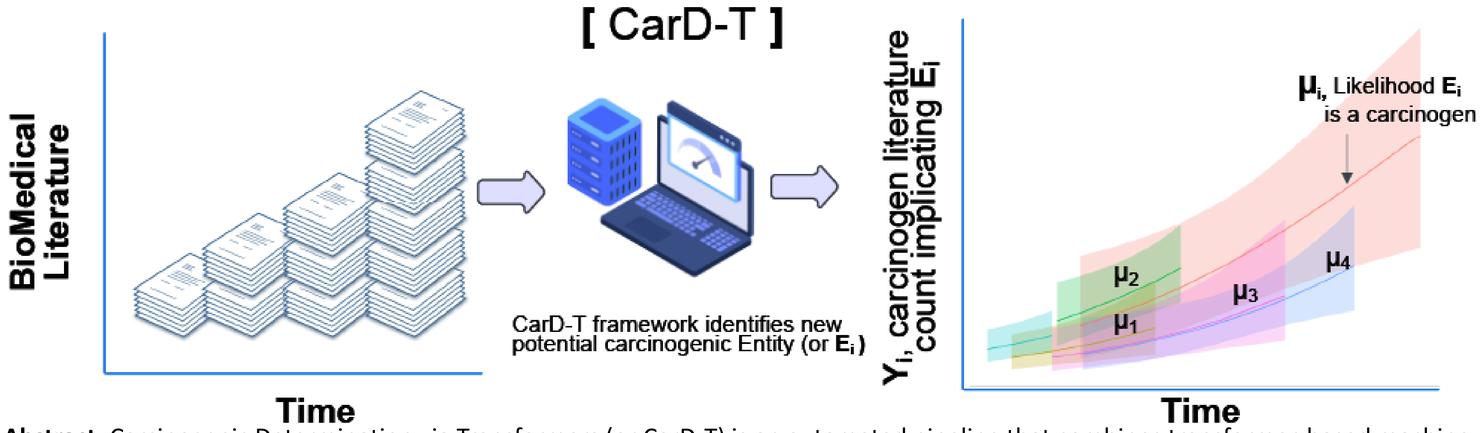
CarD-T: LLM Automated Literature Review for the Nomination & Analysis of Potential Human Carcinogens

#LB108

Jamey O'Neill[1,2], Ludmil Alexandrov[2], Parag Katira[1]

[1]Mechanical Engineering Department, San Diego State University; [2]Department of Bioengineering, University of California San Diego



Abstract: Carcinogenic Determination via Transformers (or CarD-T) is an automated pipeline that combines transformer-based machine learning with probabilistic analysis to identify potential carcinogens from biomedical literature. The framework processes accumulating scientific publications (left), applies a trained Named Entity Recognition (NER) model to extract potential carcinogenic entities (center), and Probabilistic Carcinogen Denomination (or PCarD) to analyze temporal trends in evidence shifts (right). This approach enables classification of candidates through Bayesian temporal analysis, overcoming limitations of traditional manual literature review methods.

Introduction

Only 479 carcinogenic entities are officially recognized as carcinogens across regulatory databases (IARC, NTP, EPA, ECHA, & OSHA). Regulatory agencies rely on manual curation and are burdened by exponential growth of biomedical literature. CarD-T is an LLM framework to automate literature review of carcinogen curation with probabilistic analysis of likely carcinogenicity.

Methodology

CarD-T uses a Named Entity Recognition (NER) ELECTRA LLM and is trained on carcinogen-specific contexts from PubMed literature. Novelty of an NER approach over other LLM methods lies in the ability index identified carcinogens to source documents, circumventing hallucinations. Training data consisted of abstracts referencing IARC Group 1 and Group 2A carcinogens. Additionally, we implemented a novel AI-hybridized Context-Derived TF-IDF approach for noise reduction and synonym consolidation. This allowed us to convert implicating publications into evidence counts.

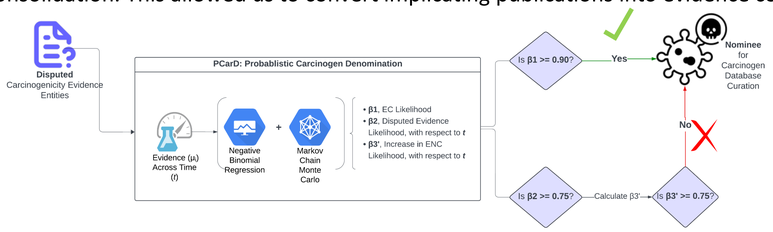


Figure 1. Probabilistic Carcinogen Denomination of Disputed Potential Carcinogens Over Time.

- PCarD module applies Bayesian Negative Binomial Regression to analyze temporal evidence shifts, evaluating potential carcinogens based on changing evidence discourse with respect to time. *B1 is Evidence Count, B2 is Evidence of a Dispute with time, B3 and B3' are likelihood of Carcinogen or Non-Carcinogenicity.*

Results

Figure 2. CarD-T Finds New Carcinogens ~20% higher likelihood than GPT-4.

- CarD-T achieves higher recall (0.85 vs. 0.76) at comparable precision (0.89 vs. 0.90).

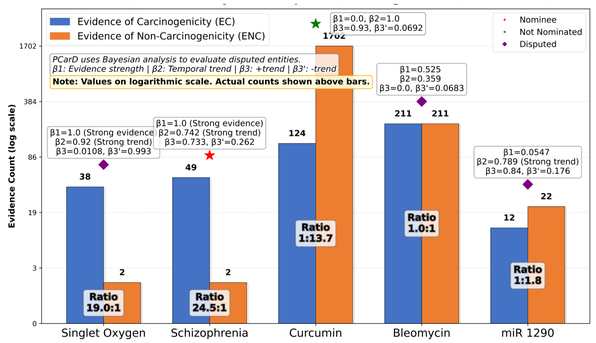


Figure 3. Evidence Count Classification for Examples of Disputed Potential Carcinogens.

- Bayesian Negative Binomial Regression Model distinguishes between established and emerging carcinogens in probability scoring of likelihood accommodating temporal shifts in evidence

Key Findings

- ✓ CarD-T identifies 100% of IARC carcinogens with just 60% training data
- ✓ ~1,600 potential new carcinogens discovered across diverse categories
- ✓ Superior recall with comparable precision to GPT-4
- ✓ 554 entities with both supporting/opposing evidence analyzed
- ✓ Framework runs on standard hardware in ~6 hours
- ✓ Novel nominees include COVID-19, microplastics, and schizophrenia
- ✓ Resolves 76 disputed entities via temporal evidence modeling

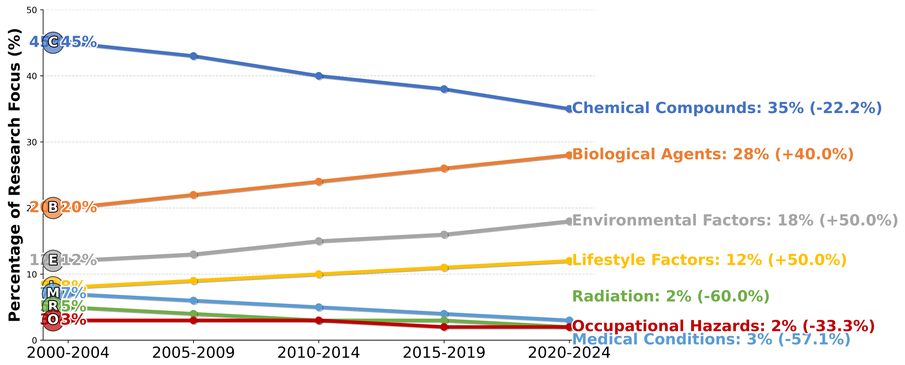


Figure 4. The Shifting Landscape: Carcinogen Research (2000-2024)

- Chemical compounds ↓45% to 35%; Biological agents ↑20% to 28%; Environmental factors ↑12% to 18%

Discussion/Conclusions

- Successfully integrated transformer learning with Bayesian modeling to assess temporal shifts in scientific consensus on carcinogens
- Documented fundamental 25-year research paradigm shift from chemical to biological and environmental carcinogenesis
- Current regulatory frameworks (except IARC) remain focused on chemical mechanisms despite emerging evidence for diverse carcinogenic pathways
- Temporal analysis provides critical context for evaluating contested entities where evidence conflicts or evolves over time
- Open-source design on consumer hardware democratizes carcinogen surveillance capabilities previously limited to specialized institutions
- Enables rapid response to emerging threats like COVID-19 by continuously monitoring scientific discourse in real-time
- Creates new opportunities for proactive public health monitoring without commercial or computational barriers

Contact

Jamey ONeill, MSc, PhD Candidate; Bioengineering UCSD-SDSU JDP
Computational Active Matter Mechanics Lab; SDSU Research Foundation
Mechanical Engineering Department, San Diego State University 5500
Campanile Dr., E-323N San Diego, CA 92182-1323
joneilliii@sdsu.edu
619-594-2032

References

1. [CarD-T Preprint] O'Neill J, Reddy GA, Dhillon N, Tripathi O, Alexandrov L, Katira P. CarD-T: Interpreting Carcinomic Lexicon via Transformers. medRxiv [Preprint]. 2024 Aug 31:2024.08.13.24311948. doi: 10.1101/2024.08.13.24311948. PMID: 39185518; PMCID: PMC11343268.
2. [Open-source LLM availability] https://huggingface.co/jimnoneill/CarD-T
3. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 100D. https://www.who.int/publications/m/item/iarc-monographs-on-the-evaluation-of-carcinogenic-risks-to-humans-volume-100d.
4. Cohen, S. Bayesian Analysis in Natural Language Processing. [Springer International Publishing, Cham, 2019]. doi:10.1007/978-3-031-02170-1.
5. Smith, M.T. et al. The Key Characteristics of Carcinogens: Relationship to the Hallmarks of Cancer, Relevant Biomarkers, and Assays to Measure Them. Cancer Epidemiol. Biomarkers Prev. 29, 1887–1903 (2020).
6. Barupal, D.K. et al. Prioritizing cancer hazard assessments for IARC Monographs using an integrated approach of database fusion and text mining. Environ. Int. 156, 106624 (2021).
7. Kanakarajan, K. et al. BioELECTRA Biomedical text Encoder using Discriminators. Proceedings of the 20th Workshop on Biomedical Language Processing, 143–154 (2021).
8. Min, B. et al. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. arXiv:2111.01243 (2021).
9. Traag, V.A. et al. From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 5233 (2019).