HACKHPC @ADMI26
JUNE 22-26, 2026

OriginZero: Graph-Anchored Tracing of the Structural Roots of Health Disparities

SGX3 award # 2231406

Abstract

Traditional explanations for health disparities often rely strictly on biological factors, individual behavior, or immediate medical oversights, offering an incomplete picture that leads to inadequate solutions. To address this, the OriginZero system traces health disparities back through long chains of historical, institutional, and economic decisions by constructing a comprehensive knowledge graph from Wikipedia and applying a Personalized PageRank algorithm to identify structurally influential factors.The development team validated this pipeline across three distinct health crises—Type 2 diabetes, childhood obesity, and maternal mortality—finding that systemic issues like redlining, food deserts, and hospital closures were identified purely through graph structure analysis. This method demonstrates that the root causes of health inequities are traceable and explainable, providing transparent, verifiable evidence rather than mere algorithmic confidence.

Goals

- Identify the historical and institutional chains that drive health disparities, going past correlated risk factors to the decisions that created them.
- Build a graph based method for ranking causal significance that is reproducible and does not depend on a researcher's prior assumptions about what matters.
- Turn that graph structure into a narrative that a non specialist, a journalist, a policymaker, a student, can read once and actually understand.
- Prove the approach works across more than one disparity, so the method is judged on its design, not on luck with a single dataset.
- Keep every step of the pipeline inspectable. If the model says redlining is a top cause, anyone should be able to trace that claim back to the graph nodes and edges that produced it.

Research Question

Can an automated knowledge graph, built from open source material and ranked with Personalized PageRank, identify the structural root causes of a public health disparity with the same rigor that clinical research applies to biomedical risk factors, and can a fine-tuned language model then explain those causes in language a general reader can verify and trust?

Resource Needs/List

- Wikipedia API access, used for the contextual crawl that seeds the knowledge graph with historical and environmental context.
- NetworkX, used to construct the directed graph and compute Personalized PageRank across it.
- Groq's hosted Llama-3.3-70B-versatile model, used as the teacher model that distills graph context and source text into instruction tuning pairs.
- One L4 GPU, used for QLoRA fine-tuning of the narrative generation model.
- Unsloth, used to make that fine-tuning run in 4-bit quantization without sacrificing training stability.
- FastAPI, used to serve the fine-tuned model as a live inference endpoint.
- A domain specific instruction tuning dataset in JSONL format, built from the graph itself rather than borrowed from an unrelated corpus.

Methodology

- **Targeted Wikipedia Crawling:** Extracts historical and environmental threads rather than strictly medical articles.
- **Graph Construction:** Builds a directed network layout using the NetworkX Python library.
- Personalized PageRank: Uses balanced topic pillars to prevent any single cause from dominating.
- **Dataset Generation:** Feeds high-ranked structural nodes to Llama-3.3-70B via Groq for instruction-response pairs.
- **Model Fine-Tuning:** Trains a Gemma-4-E2B model using QLoRA with 4-bit quantization adapters.
- **API Deployment:** Exposes the fine-tuned model via a FastAPI endpoint for plain English narratives.

Results

- **Fast Convergence:** Fine-tuning completed smoothly in 60 steps on a single L4 GPU in under 30 minutes.
- **Clean Loss Drop:** Training loss decreased steadily from 1.058 (step 1) to 0.118 (step 59) with no instability.
- **Optimal Data Density:** The training corpus totaled 44,119 tokens (~735 tokens/sample), effectively teaching reasoning patterns without overfitting.
- **Direct Graph-to-Text:** The deployed model generates untemplated narratives directly from PageRank ranked nodes (e.g., automatically linking childhood obesity to zip-code infrastructure).
- **High Specificity:** Across all tested disparities, the model successfully surfaced specific, checkable causes rather than generic public health language.

Possible Expansions

- **Broader Disparities:** Scale the pipeline beyond diabetes, obesity, and maternal mortality to test how the graph and ranking approach handles entirely new causal stories.
- **Temporal Tracking:** Implement a timeline view to visualize how structural causes accumulate influence over decades, moving beyond a single static snapshot.
- **Expert Validation:** Bring in public health researchers and historians to audit the generated narratives, ensuring the model's output is factually accurate, not just computationally clean.
- **Localized Analysis:** Test the pipeline on regional data sources so local community organizations can apply the tool to their specific neighborhoods.

Acknowledgements

- **Origin:** Conceived and prototyped during a collaborative team hackathon.
- **AI Disclosure:** Used Groq (Llama-3.3-70B) for synthetic data generation, distillation, and optimization.
- **Support:** Thanks to hackathon organizers, mentors, and sponsors

Authors


Amelia Hunter
Sophomore | Computer Science | Voorhees University


Prince Osei Bonsu
Sophomore | Computer Science | Voorhees University


Emmanuel Akwasi Opoku
Sophomore | Cybersecurity | Grambling State University


Anthony Nwafor
Graduate | Computer Science and Mathematics | Mississippi Valley State University

Supported by: ADMI (The Association of Computer Science Departments at Minority Institutions), SGX3 Extend. Expand. Exemplify., Norfolk State University, Winston Salem State University, Omnibond, TACC Texas Advanced Computing Center, HackHPC Hackathons and Codeathons

MORE INFORMATION → https://hackhpc.github.io/admi26.