

Georgia Ning-Yuan Liu, PhD

Applied AI Research Scientist · LLM Evaluation & Information Extraction

San Francisco Bay Area · georgialiu63@gmail.com · [georgia-max.github.io](https://github.com/georgia-max) · github.com/georgia-max · [linkedin.com/in/georgianyliu](https://www.linkedin.com/in/georgianyliu)

SUMMARY

Applied AI researcher building and evaluating LLM systems that extract structured causal knowledge from unstructured text.

Core Areas: LLM evaluation and benchmarking · information extraction · natural language processing (NLP) · causal and knowledge graphs · human–AI interaction

EXPERIENCE

Causena | Founder & CEO (AI Research & Engineering) | San Francisco Bay Area · Jan 2026 – Present

- Created an annotation taxonomy of causal language in healthcare text, used as the schema for LLM-based information extraction and label normalization.
- Built an LLM extraction system (LangChain, Claude Agent SDK, NetworkX) that links every extracted causal relationship to its source quote and merges per-document graphs into one aggregated causal graph.
- Deployed the system on 1,000 real-world healthcare transcripts, producing source-cited causal graphs 10x faster than manual qualitative coding; adopted by first customer (Virginia Tech).
- Berkeley SkyDeck Batch 22 (top 10 pitch) · NSF National I-Corps.

Harvard Medical School / Massachusetts General Hospital | Postdoctoral Research Fellow, LLMs & Causal AI | Boston, MA · Apr 2024 – Sep 2025

NIH-funded research at the Institute for Technology Assessment & MJ Lab.

- Built a benchmark for evaluating LLM-generated causal graphs: reviewed 40+ semantic and structural similarity metrics, selected 9, and validated them on 2,000 LLM-generated graphs ([arXiv 2503.11046](https://arxiv.org/abs/2503.11046)).
- Built an LLM information-extraction pipeline that converts qualitative health interviews into causal loop diagrams, cutting expert analysis time 10x ([arXiv 2503.21798](https://arxiv.org/abs/2503.21798)).
- Fine-tuned GPT-4o and used few-shot prompting and retrieval-augmented generation (RAG) to generate synthetic causal-graph datasets for model evaluation.
- Analyzed 100M+ ICD-coded electronic health records (EHRs) to quantify disparities in maternal health outcomes.
- Contributed to an NSF-funded award (\$100K+) to train LLMs that translate text into knowledge graphs.

Google X — AI Resident (PhD) | Mountain View, CA · Aug 2022 – Aug 2023

- Led a prototype reasoning system (under NDA) combining LLMs, graph modeling and semi-structured data to produce interpretable outputs; the work contributed to a patent-pending causal visualization platform.
- Automated qualitative coding for system dynamics models with LLMs, cutting simulation build

time 4x; curated the expert-interview dataset behind it.

Virginia Tech, Systems Performance Lab | PhD Researcher, Machine Learning & Human–AI Systems | Falls Church, VA · Aug 2018 – Mar 2024

Graduate researcher on NSF LEAP-HI award #2051685 (\$2M; PI: K. Triantis) studying human–automation interaction in railway control centers. Visiting Researcher, Ghent University (Jul 2022).

- Quantified how workload and over-reliance on automation increase human error, using 410K+ operational records from Belgian Railways (Infrabel); published in *Safety Science*.
- Built machine learning models (Random Forest, XGBoost, neural networks) on 21K+ records to predict operator efficiency, with DEA benchmarking and SHAP explainability.
- Developed an NLP-based index that scores systems-thinking skill from open-ended text (unsupervised learning, robust PCA, DEA); published in *Systems Research and Behavioral Science*.
- Dissertation: *Enhancing Safety in Critical Monitoring Systems: Investigating the Roles of Human Error, Fatigue, and Organizational Learning in Socio-Technical Environments*.

SKILLS

- **LLMs & NLP:** large language models (LLMs), LLM evaluation and benchmarking, information extraction, fine-tuning (GPT-4o), retrieval-augmented generation (RAG), few-shot prompting, prompt engineering, synthetic data generation, annotation schema design
- **Modeling:** causal graphs, knowledge graphs, graph kernels, system dynamics, statistical modeling, experimental design, supervised and unsupervised machine learning, explainability (SHAP)
- **Human–AI Systems:** human–AI interaction, automation reliance, human factors
- **Tools:** Python, PyTorch, Hugging Face Transformers, LangChain, scikit-learn, NetworkX, pandas, SQL, Spark, R
- **Languages:** English, Mandarin, Spanish

EDUCATION

- **Virginia Tech** — PhD, Industrial and Systems Engineering (2024); MEng, Computer Science, AI track (2024)
- **Beijing University of Aeronautics and Astronautics** — BS, Industrial and Systems Engineering (2018)

PUBLICATIONS

- Liu N., Yang F., Jalali M. Measuring Similarity in Causal Graphs: A Framework for Semantic and Structural Analysis. [arXiv 2503.11046](#)
- Liu N., Keith D. Leveraging Large Language Models for Automated Causal Loop Diagram Generation. [arXiv 2503.21798](#)
- Liu N., Triantis K., Madsen P., Roets B. (2025). The Relationships among Workload, Automation Reliance, and Human Errors in Safety-Critical Monitoring Roles. *Safety Science*, 185. [Doi](#)

- Liu N., Mahmoudi H., Triantis K., Ghaffarzadegan N. (2024). A Multi-dimensional Index of Evaluating Systems Thinking Skills from Textual Data. *Systems Research and Behavioral Science*. [doi](#)