Yinghao Li

Email: yinghaoli@gatech.edu
Webpage: yinghao-li.github.io
Phone: +1 (404)232-0971
Address: Atlanta, GA 30308

EDUCATION

Georgia Institute of Technology **Atlanta, GA**
- *Ph.D.* in *Machine Learning* *August 2020 – May 2025 (expected)*
  - Advisor: Dr. Chao Zhang and Dr. Le Song
  - Research Interests: Information Extraction; Weak Supervision; Uncertainty Estimation; Large Language Models
- *Master of Science* in *Electrical and Computer Engineering* *August 2018 – May 2020*

Southeast University **Nanjing, China**
- *Bachelor of Engineering* in *Instrument Science and Engineering* *August 2014 – June 2018*

EXPERIENCE

Amazon.com, Inc. **Seattle, WA**
Applied Scientist Intern *May 2022 – December 2022*
- Supervisor: Dr. Prashant Shiralkar; Mentor: Dr. Colin Lockard
- Topic: Extracting and organizing shopping interest-related product types from free-formed webpages.
- Publication: *Extracting Shopping Interest-Related Product Types from the Web* in EMNLP 2022 Findings.

SELECTED PUBLICATIONS

- Assessing Logical Puzzle Solving in Large Language Models: Insights from a Minesweeper Case Study
  Yinghao Li, Haorui Wang, Chao Zhang
  In *arXiv preprint*, 2023; submitted to NAACL 2024 (under review).
- MUBen: Benchmarking the Uncertainty of Molecular Representation Models
  Yinghao Li, Lingkai Kong, Yuanqi Du, Yue Yu, Yuchen Zhuang, Wenhao Mu, Chao Zhang
  In *NeurIPS 2023 AI4Science Workshop*, 2023; submitted to ICLR 2024 (under review).
- Extracting Shopping Interest-Related Product Types from the Web
  Yinghao Li, Colin Lockard, Prashant Shiralkar, Chao Zhang
  In *EMNLP 2023 Findings*, 2023.
- Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition
  Yinghao Li, Chao Zhang, Le Song
  In *KDD 2022*, 2022.
- WRENCH: A Comprehensive Benchmark for Weak Supervision
  Jieyu Zhang, Yue Yu, **Yinghao Li**, Yujing Wang, Yaming Yang, Mao Yang, Alexander J. Ratner
  In *NeurIPS 2021*, 2021.
- BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition
  Yinghao Li, Pranav Shetty, Lucas Liu, Chao Zhang, Le Song
  In *ACL 2021*, 2021.
- Denoising Multi-Source Weak Supervision for Neural Text Classification
  Wendi Ren, **Yinghao Li**, Hanting Su, David Kartchner, Cassie Mitchell, Chao Zhang
  In *EMNLP 2020 Findings*, 2020.
- Transformer-Based Neural Text Generation with Syntactic Guidance
  Yinghao Li, Rui Feng, Isaac Rehg, Chao Zhang
  In *arXiv preprint*, 2020.

Please visit my Google Scholar page for a full list of publications.

PROJECTS

Large Language Models: Potentials and Risks

I am currently involved in multiple projects aimed at exploring the capabilities of Large Language Models (LLMs) and extending their potential for real-world applications.

- Studying the reasoning and planning abilities of LLMs to determine whether they genuinely exhibit reasoning or primarily rely on knowledge retrieval from their pre-training data.
- Investigating better techniques for synthesizing or selecting relevant data points to fine-tune smaller, cost-effective, task-specific language models.
- Exploring the application of LLMs to specific domains, such as materials science, where limited data is available.

Uncertainty Estimation for Molecular Property Prediction

- Developed the MUBen benchmark to assess the uncertainty quantification performance of different backbone models (including both state-of-the-art pre-trained models such as Uni-Mol and simple models such as GIN) and various uncertainty estimation methods for molecular property prediction.

Weak Supervision for Information Extraction

- Designed a conditional hidden Markov model (CHMM) that conditions the Hidden Markov Model (HMM) on BERT token embeddings. This approach facilitates token-wise transition and emission probabilities for aggregating multiple sets of Named Entity Recognition (NER) labels from different weak labeling functions.
- Introduced a sparse variant—Sparse CHMM—as a followup to CHMM. Sparse CHMM predicts diagonal emission elements instead of entire emission matrices. This design helps regulate the emission process and reduces training complexity. The use of a WXOR function provides finer control over emission probabilities, resulting in improved performance with lower computational consumption.

Please visit my GitHub profile for more projects.

MISC

- **Programming**: *Proficient*: Python, C++, C; *Familiar*: Scala, MATLAB, VHDL, Java and Assembly
- **Teaching Experience**: Teaching Assistant for *CSE 8803 Deep Learning for Text Data* (Fall 2023); Teaching Assistant for *Georgia Tech Big Data Analytics Bootcamp* (Spring 2020, 2021, 2022, 2023)
- **Interests:** Coding, Hiking, Photography, Reading, Table Tennis