Yinghao Li

Email: yinghaoli@gatech.edu
Website: 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 Prof. Le Song
  - Research Interests: Language Models; Information Extraction; Weak Supervision; Uncertainty Estimation;
- *M.S.* in *Electrical and Computer Engineering* *August 2018 – May 2020*
  - Advisor: Dr. Chao Zhang and Prof. Ying Zhang
  - Research Interests: Text Generation; Signal Processing;

Southeast University **Nanjing, China**
- *B.Eng.* in *Instrument Science and Engineering* *August 2014 – June 2018*
  - Advisor: Dr. Lifeng Zhu
  - Research Interests: 3D Reconstruction; Embedded System;

EXPERIENCE

Amazon.com, Inc.; AWS **New York, NY**
- *Applied Scientist Intern* *May 2024 – August 2024*
  - Mentor: Dr. Vianne Gao; Manager: Dr. Ali Torkamani
  - Develop a mixture of LoRA expert framework for efficient and effective task-specific language model fine-tuning.

Amazon.com, Inc. **Seattle, WA**
- *Applied Scientist Intern* *May 2022 – December 2022*
  - Mentor: Dr. Colin Lockard; Manager: Dr. Prashant Shiralkar
  - Developed Transformer-based graph node classification model and dataset for extracting shopping interest-related product types from HTML webpages.
  - Publication: Extracting Shopping Interest-Related Product Types from the Web in *EMNLP 2022 Findings*.

SELECTED PUBLICATIONS

- A Simple but Effective Approach to Improve Structured Language Model Output for Information Extraction
  Yinghao Li, Rampi Ramprasad, Chao Zhang
  In *EMNLP Findings*, 2024.
- Assessing Logical Puzzle Solving in Large Language Models: Insights from a Minesweeper Case Study
  Yinghao Li, Haorui Wang, Chao Zhang
  In *NAACL*, 2024.
- MUBen: Benchmarking the Uncertainty of Molecular Representation Models
  Yinghao Li, Lingkai Kong, Yuanqi Du, Yue Yu, Yuchen Zhuang, Wenhao Mu, Chao Zhang
  In *TMLR*, 2024.
- Extracting Shopping Interest-Related Product Types from the Web
  Yinghao Li, Colin Lockard, Prashant Shiralkar, Chao Zhang
  In *EMNLP Findings*, 2023.
- Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition
  Yinghao Li, Le Song, Chao Zhang
  In *KDD*, 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 Benchmark*, 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.

Please visit Google Scholar for more publications.

PROJECTS

Large Language Models: Reasoning and Application
- Improves large language model's (LLM's) performance on diverse tasks through an Expert Ensembles framework, which clusters training data according to gradient profiles to reduce update conflicts and aggregates expert models' predictions according to their relevance to the input.
- Studies 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 [Minesweeper].
- Investigates efficient and effective LLM prompting and fine-tuning techniques for information extraction tasks such as named entity recognition and relation extraction [G&O].

- Leverages LLMs to synthesize or select relevant data points to fine-tune smaller, cost-effective, and domain/task-specific language models such as BERT [ProgGen].

Uncertainty Estimation for Molecular Property Prediction

- Develops 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 [MUBen].

Weak Supervision for Information Extraction

- Designs 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 [CHMM; Wrench].
- Introduces 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 [Sparse CHMM].

Syntactic-Guided Text Generation

- Designs a two-encoder Transformer architecture with a multi-encoder attention mechanism to effectively incorporate syntactic information represented by the constituency parsing trees into the text generation process [GuiG].

Please visit GitHub for more projects.

SKILLS

- **Programming SKills**: *Proficient*: Python (PyTorch), C++; *Familiar*: Scala, MATLAB, VHDL, Java, and Assembly
- **Open-Source Python Packages**: SeqLbToolkit; muben; ChemistryPaperParser
- **Teaching Experience**: Teaching Assistant for *CSE 8803 Deep Learning for Text Data* (Fall 2023, 2024); *GT NLP Bootcamp: Natural Language Processing & Large Language Model* (Spring 2023, 2024); *Georgia Tech Big Data Analytics Bootcamp* (Spring 2020, 2021, 2022, 2023, 2024)
- **Hobbies**: Photography, Hiking, Running, Reading, Table Tennis