Yinghao Li

Email: yinghaoli@gatech.edu
Phone: +1 (404)232-0971

Webpage: yinghao-li.github.io
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;
- *Master of Science* in *Electrical and Computer Engineering* *August 2018 – May 2020*
  - Advisor: Dr. Chao Zhang and Prof. Ying Zhang
  - Research Interests: Language Models; Text Generation; Signal Processing;

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
  - 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

- Assessing Logical Puzzle Solving in Large Language Models: Insights from a Minesweeper Case Study
  Yinghao Li, Haorui Wang, Chao Zhang
  In *arXiv preprint*, 2023.
- 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 AI for Science Workshop*, 2023.
- 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, Le Song, Chao Zhang
  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: Reasoning and Application
- 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 [Minesweeper].
- Investigating efficient and effective LLM prompting and fine-tuning techniques for information extraction tasks such as named entity recognition and relation extraction.
- Using LLMs to synthesize or select relevant data points to fine-tune smaller, cost-effective, and domain/task-specific language models such as BERT.

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 [MUBen].

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 [CHMM, Wrench].
- 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 [Sparse CHMM].

Syntactic-Guided Text Generation

- Designed 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 my GitHub profile for more projects.

SKILLS

- **Programming SKills** *Proficient*: Python (PyTorch), C++, C; *Familiar*: Scala, Spark, 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