Bai Li

Email: bai.li.2005@gmail.com | Citizenship: Canadian | LinkedIn | Google Scholar

Education

2019–2022   **PhD, Computational Linguistics**, University of Toronto
Thesis: Integrating Linguistic Theory and Neural Language Models

2017–2019   **MSc, Computational Linguistics**, University of Toronto
Thesis: Automatic Detection of Dementia in Mandarin Chinese

2012–2017   **BMath, Computer Science**, University of Waterloo
Faculty Average: 93%

Work Experience

Jul 2023 - Present — **Founding Engineer (Optimization)** - Untether Labs
Founding engineer at a YC-backed healthcare startup. Developed an autoscheduler for 1,500 clinicians using constraint programming (Pyomo, CPLEX) to optimize staffing under demand and licensing constraints, reducing manual scheduling work by 70%. Built hourly data pipelines for healthcare software integrations. Forecasted patient demand using historical appointments and Erlang-C-based wait time models.

Oct 2022-May 2023 — **Machine Learning Engineer** - Generally Intelligent (now Imbue)
Developed scalable data pipelines for LLM pretraining and quality filtering. Built Docker-based infrastructure for parallelizing experiments on cloud compute. Trained PPO-based RL agents in a Minecraft-style environment.

Apr 2018-Oct 2020 — **NLP Research Engineer** - SnapTravel (now Super.com)
Built NLP system to extract named entities and hotel preferences in a natural language chatbot. Lead author on paper at IEEE AI4I 2019 conference, winning best paper award. Contributed to machine learning for ad bidding and statistical analysis for experiments. Part-time. Technologies used: PyTorch, SpaCy, AllenNLP, Elasticsearch, scikit-learn, pandas, R.

2014–2016 — **Software Engineer Internships** - Yelp, Yext, Minted, A Thinking Ape, YuJa
Worked as a full-stack engineer in performance team (Yelp), e-commerce (Minted), developer platforms (Yext), gaming (ATA), and edtech (YuJa). Developed production APIs, internal tools, and user-facing features in digital marketplaces. Technologies: Java, Django (Python), SQL, Redis, AWS, Nginx, JavaScript (Backbone, jQuery).

Recent Projects

- **@EfficientNLP YouTube Channel**
  Content creator with 13,000+ subscribers and over 20,000 watch hours, focused on ML and LLMs. Topics include transformer optimizations (KV Cache, rotary embeddings), LLM decoding algorithms, fine-tuning Whisper and DeepSeek R1 models.

- **VoiceWriter.io**
  Speech-to-text productivity app combining real-time streaming ASR (AssemblyAI, Deepgram) followed by LLM-based grammar correction. Engineered low-latency systems, model routing via OpenRouter, and prompt evaluations under varied conditions to optimize user experience. Monetized product with 3,000+ users across web app and Chrome extension.

Selected Publications

- **Li B.**, Zhu Z., Thomas G., Rudzicz F., Xu Y. "Neural reality of argument structure constructions". *ACL 2022*.

- **Li B.**, Zhu Z., Thomas G., Xu Y., Rudzicz F. "How is BERT surprised? Layerwise detection of linguistic anomalies". *ACL-IJCNLP 2021*.

- **Li B.**, Rudzicz F. "RoBERTa with Multi-Stage Fine-Tuning for Eye-Tracking Prediction". *CMCL at NAACL 2021*. **Best student paper award**.

- **Li B.**, Thomas G., Xu Y., Rudzicz F. "Word class flexibility: A deep contextualized approach". *EMNLP 2020*.

- **Li B.**, Hsu Y-T., Rudzicz F. "Detecting dementia in Mandarin Chinese using transfer learning". *NAACL 2019*.

- **Li B.**, Jiang N., Sham J., Shi H., Fazal H. "Real-world Conversational AI for Hotel Bookings". *IEEE AI4I 2019*. **Best paper award**.