# Ge Zhang (John)

+1-6476711919 | gz19950616@gmail.com | Google Scholar | github.com/ZG2017

## Working Experiences

**Senior AI Researcher (Contractor)** 2021/10 – Present
*Huawei Canada - Noah's Ark Lab* *Toronto, Canada*

- Lead research on an efficient LLMs-based solution for recommendation systems in e-commerce
- Develop the novel framework for solving kinship/spatial relational reasoning tasks with LLMs

**Machine Learning Algorithm Engineer** 2020/12 – 2021/09
*Kuaishou Technology Ltd.* *Beijing, China*

- Developed causal inference framework to distribute user incentive budgets for maximum potential gain
- Built personalized recall pipeline for short video recommendation system

**Data Scientist** 2019/07 – 2020/11
*MorningStar Ltd.* *Shenzhen, China*

- Constructed NLP solution to match unorthodox fund abbreviation to its standard name

## Education

**University of Toronto** Toronto, ON
*Master of Engineering in Electrical & Computer Engineering* *2017/09 – 2019/07*

- GPA: 3.67/4.00
- Courses: Machine Learning & Data Mining (A), Data Science & Analytics (A-), Algorithms & Data Structures (A), Natural Language Computing (A)

**Wuhan University of Technology** Wuhan, China
*Bachelor of Science in Physics* *2013/09 – 2017/06*

- GPA: 3.75/4.00

## Skills

**Experience**: 6+ years of machine learning engineering/research in industry
**Domains**: Recommendation Systems, LLMs-based Reasoning/Agent/Memory system, Graph Neural Networks (GNNs)
**ML Frameworks/Libraries**: PyTorch, TensorFlow, vLLM, LangGraph, Transformers, Selenium, MCP, Elastic Search, Scikit-Learn, Amazon Web Service (AWS), Hive
**Development Tools**: VSCode, Cursor, Git, Jupyter Notebook, PyCharm
**Development Languages**: Python
**Other**: Speaking English and Mandarin; Leading a team of 3+ members; Amateur Photographer

## Industrial Projects

**E-CARE: Efficient Commonsense-Augmented Recommendation Enhancer**
*Keywords: Recommendation Systems, LLMs, Graph*

- LLMs' commonsense reasoning facilitates e-commerce, but previous pipelines are costly with SFT, human annotations, and test time LLM inference
- Designed 3-stage LLM-based framework to boost accuracy of current models without costly components above
- LLMs' reasonings are first parsed and then integrated as a heterogeneous reasoning graph intermediate, followed by uncertainty estimation to improve their reliability
- Achieved maximum improvement of 4.27% on Macro F1 in downstream e-commerce tasks
- Filed a patent, AB-testing is ongoing, and preparing a research paper for WWW 2026.

**Path-of-Thoughts: Robust Relational Reasoning with LLMs**
*Keywords: Commonsense Reasoning, LLMs, Graph, vLLM*

- Kinship/spatial relational reasoning abilities of LLMs are crucial in virtual/embodied agents, but haven't been properly solved yet
- Designed a 3-stage pipeline consisting of graph extraction, path identification, and reasoning to decompose and solve relational reasoning questions

- Achieved maximum accuracy improvements of 21.3% over baselines with increased robustness against noise facts
- Filed a patent and a research paper is under review with the TMLR journal.

**Unorthodox Fund Name Abbreviation Matching**

*Keywords: FinTech, NLP, Recommendation System*

- Various abbreviation criteria of fund names are used in the market, posing a challenge for data aggregation
- Developed recall + ranking pipeline with Elastic Search and DSSM as recall and ranking model, respectively
- Synthesize a large abbr-to-full name corpus for pre-training, followed by fine-tuning on the real corpus
- Achieved 21% improvement on fund name recall compared to original word2vec cosine similarity

## Personal Projects

**News Agent with LangGraph**

*Keywords: LangGraph, MCP, LLM Agent, Selenium*

- Built automated agent using LangGraph to manage end-to-end news clips generation workflow
- Implemented LLMs-based news collection, summarization, and multi-language email distribution system
- Developed automated publishing pipeline with Selenium to post on the RedNote platform
- Designed robust error handling, retry mechanisms, and comprehensive state management for production use
- Subscribe to the google group 'gz_daily_news_clips@googlegroups.com' to view the daily news clips

**Reinforcement Learning to Play Chrome Dino Game**

*Keywords: Reinforcement Learning, Selenium*

- Implemented RL algorithms (DQN, REINFORCE, Actor-Critic, PPO) to train agents to play Chrome Dino Game
- Developed real-time training environment using Selenium for Chrome browser automation
- Advanced features: replay memory, Generalized Advantage Estimation (GAE), residual sampling, etc.
- Achieved highest scores: REINFORCE (780), Actor-Critic-TD (1,216), Actor-Critic-GAE (2,310), DQN (9,088)
- Details can be found in github.com/ZG2017/Deep_Q_Network-on-chrome-dino-jump

**"Visual Speaker" – An Image Captioning Application**

*Keywords: Image Captioning, AWS*

- Built a web application to "speak out" the contents of an image using ResNet50-LSTM model
- Deployed the whole web application on AWS using Lambda and EC2 instances
- Responsible for deploying ResNet50-LSTM model on EC2 and backend of web application

## Publications & Patents

**G Zhang**, R Ajwani, H Gu, Y Hu, Y Zhang, "NURG: An Efficient Learning-to-Rank Framework with Commonsense Reasoning from Large Language Models", US Patent Application. Application number: 19/337,555, 2025.

**G Zhang**, MA Alomrani, H Gu, Y Hu, Y Zhang, "Methods and Processors for Relational Reasoning from Text", US Patent Application. Application number: 19/074,860, 2025.

**G Zhang**, MA Alomrani, H Gu, J Zhou, Y Hu, B Wang, Q Liu, M Coates, "Path-of-thoughts: Extracting and following paths for robust relational reasoning with large language models", arXiv preprint arXiv:2412.17963, 2024.

Y Hu, M Zeng, **G Zhang**, P Rumiantsev, L Ma, Y Zhang, M Coates, "Sparse Decomposition of Graph Neural Networks", arXiv preprint arXiv:2410.19723, 2024.

J Zhou, A Ghaddar, **G Zhang**, L Ma, Y Hu, S Pal, M Coates, B Wang, "Enhancing logical reasoning in large language models through graph-based synthetic data", arXiv preprint arXiv:2409.12437, 2024.

**G Zhang**, S Pu, XY Xu, C Tao, JY Dun, "Optimized design of THz microstrip antenna based on dual-surfaced multiple split-ring resonators", 2017 IEEE International Symposium on Antennas and Propagation & USNC/URSI, 2017.

**G Zhang**, S Pu, X Xu, Y Liu, C Wang, "Design of 60-GHz microstrip antenna array composed through circular contour feeding line", 2016 Asia-Pacific International Symposium on Electromagnetic Compatibility, 2016.

**G Zhang**, S Pu, ZR Liu, WF Liu, "Design of 60 GHz microstrip antenna array composed through annular feeding line", IEEE International Symposium on Antennas and Propagation (APSURSI), 2016.