Ge Zhang (John)

+1-6476711919 | gz19950616@gmail.com | Google Scholar | Personal Website | GitHub

Working Experiences

Senior AI Researcher 2021/10 – Present
Huawei Canada *Toronto, Canada*

E-CARE: An Efficient LLM-based Commonsense-Augmented Framework for E-Commerce

- Developed a scalable LLM-enhanced framework to improve **query-item relevance matching** and **app recall** in e-commerce recommendation by injecting **commonsense reasoning knowledge** from LLMs.
- Designed a 3-stage training pipeline that removes the need for costly SFT, large-scale human annotations, and real-time LLM inference during serving.
- Constructed a reasoning factor graph that distills commonsense knowledge from **Qwen2.5-7B-Instruct** and **Llama-3.1-8B-Instruct**, enabling downstream models to access commonsense knowledge more efficiently.
- Optimized the reasoning factor graph with **LLM uncertainty estimation** mechanisms to improve its reliability and robustness
- Online A/B test on app recall demonstrated **1.41%**, **0.65%** and **3.39%** improvements for Value Per Mille (VPM), Click Through Rate (CTR), and ConVersion Rate (CVR), respectively.
- Preprinted **research paper**; filed **U.S. Patent Application** (19/337,555).

Path-of-Thoughts: Robust Relational Reasoning with LLMs

- Addressed LLM limitations in **multi-hop relational reasoning** tasks (e.g., kinship, spatial relations).
- Developed a structured reasoning pipeline of graph extraction, reasoning path identification, and relation deduction to improve reasoning reliability.
- Improved accuracy by up to **21.3%** across both thinking LLMs (**GPT-5**, **GPT-o1-mini**, **Claude-3.7-Sonnet**, etc.) and non-thinking LLMs (**GPT-4o**, **Llama-3-70B**, etc.) on four benchmarks.
- Demonstrated robustness to noisy descriptions and LLM extraction errors, outperforming prior neuro-symbolic approaches.
- Generated reusable relational graphs supporting internal data generation for SFT of in-house LLMs and other research on **complex reasoning with LLMs**.
- Preprinted **research paper** (accepted by the TMLR); filed **U.S. Patent Application** (19/074,860).

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

User Incentives Distribution

- Addressed the **causal inference** challenge of identifying users who are more sensitive to monetary bonuses to optimize **incentive distribution** and increase **user retention** for China's second-largest short video platform
- Conducted an online randomized incentive experiment on **1M+ users**, collecting high-quality treatment/control data for building a robust causal inference framework
- Leveraged the **X-learner framework** to estimate heterogeneous treatment effects while addressing **treatment/control data imbalance** issue
- Increased **retained user count by 34%** compared to a random allocation baseline in A/B testing
- Improved **Return on Investment (ROI) by 15%** by accurately targeting users most likely to respond to incentives

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

Unorthodox Fund Name Abbreviation Matching

- Different institutions use inconsistent abbreviation rules, which poses a challenge for **automatic fund name matching** during data aggregation and analytics for fin-tech companies
- Formulated the task as an **information retrieval** task: map a given fund name abbreviation to the most relevant standard fund name in the internal database
- Built a two-stage pipeline: **Elasticsearch**-based **candidate recall** followed by a **char-tokenized Transformer DSSM ranking model** for fine-grained similarity scoring
- Enhanced the performance by synthesizing a large abbreviation-to-full name corpus for **pre-training** of char-tokenized transformer, followed by **fine-tuning** on the real abbreviation-to-full name pairs
- Achieved **21%** improvement on fund name retrieve compared to original word2vec cosine similarity

Education

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

- Courses: Natural Language Computing (A), Machine Learning & Data Mining (A), Algorithms & Data Structures (A)

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

Skills

Experience: 6+ years of machine learning engineering and research
Domains: Recommendation Systems, LLMs-based Reasoning/Agent/Memory system, Reinforcement Learning (RL), Graph Neural Networks (GNNs)
Frameworks/Libraries/Platforms: PyTorch, TensorFlow, vLLM, LangGraph, Transformers, SFT, Deepspeed, Accelerate, Selenium, MCP, Elasticsearch, Scikit-Learn, Amazon Web Services (AWS)
Development Tools: VSCode, Cursor, Git, Jupyter Notebook, PyCharm
Development Languages: Mainly Python, with experience in Java, C++, SQL and Hive
Others: Speaking English and Mandarin; Leading a team of 3+ members; Amateur photographer

Personal Projects

News Agent with LangGraph

- Built an automated **LLM-driven agent workflow** using **LangGraph** to handle end-to-end news collection, summarization, newsletter distribution, and platform posting.
- Integrated news sources such as Yahoo Finance and Financial Times; used **gpt-4.1-mini** with **MCP** tools for controlled retrieval and summarization.
- Implemented daily multilingual newsletter distribution and content posting for consistent news delivery.
- Developed a **Selenium**-based automation module to post news clips to the social media platform (**RedNote**).
- Daily newsletter is publicly accessible via the Google Group: `gz_daily_news_clips@googlegroups.com`.

Reinforcement Learning to Play Chrome T-rex Runner Game

- Implemented multiple **RL algorithms** (**DQN**, **REINFORCE**, **Actor-Critic**, **PPO**) to train agents to play the Chrome T-Rex Runner game.
- Constructed an online learning environment using **Selenium** for real-time browser interaction and state observation.
- Applied **experience replay** and **Generalized Advantage Estimation (GAE)** for more stable training and improved sample efficiency.
- Designed **state fusion** by merging consecutive frames to capture game acceleration effects.
- Achieved best scores: REINFORCE (780), Actor-Critic-TD (1,216), Actor-Critic-GAE (2,310), DQN (9,088).
- Source code: github.com/ZG2017/Deep_Q_Network-on-chrome-dino-jump

Hands-on of SFT Training Approaches for LLMs

- Evaluated GPU memory usage and training speed across multiple **LLM SFT strategies** using **Qwen2.5-3B-Instruct** on a 2-GPU NVLink-connected setup.
- Compared **Vanilla SFT**, **DDP**, **DeepSpeed ZeRO-2**, **ZeRO-2 + LoRA**, **ZeRO-3**, and **ZeRO-3 + CPU Offload**.
- Used a toy dataset to minimize the impact of activations and isolate **optimizer state and parameter memory effects**.
- **ZeRO-2 + LoRA** provided the **lowest GPU memory consumption** (26.14 GB) while maintaining **competitive training speed and model performance**, making it a good practice for SFT on small LLMs

Publications & Patents

G Zhang, RD Ajwani, T Zheng, H Gu, Y Hu, W Guo, M Coates, Y Zhang, "E-CARE: An Efficient LLM-based Commonsense-Augmented Framework for E-Commerce", arXiv preprint arXiv:2511.04087, 2025.

G Zhang, RD 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.