Zhenwei An

Tel: +86 188-1302-1067 | Email: anzhenwei@pku.edu.cn
Personal website: https://zhenweian.github.io/

Education

Peking University 2020.09-2023.07

- **Degree:** Master Degree of Software Engineering **GPA:** 3.37 / 4.0
- **Research Interest:** Natural Language Processing, Legal Aritifical Intelligence
- **Supervisor:** Yansong Feng

Peking University 2016.09-2020.07

- **Degree**: Bachelor Degree of Computer Science **GPA**: 3.48 / 4.0
- **Major**: Computer Science

Publications

[1] Quzhe Huang*, Zhenwei An*, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, and Yansong Fe Tasks Need More Experts: Dynamic Routing in MoE Models." Accepted By ACL2024 Main Conference`.

[2] Zhenwei An*, Quzhe Huang*, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. Do Charge Prediction Models Learn Legal Theory?. of the Association for Computational Linguistics: EMNLP 2022, pages 3757–3768, Abu Dhabi, United Arab Emirates. Association for Compu Linguistics.

[3] Quzhe Huang*, Mingxu Tao*, Zhenwei An*, Chen Zhang*, Cong Jiang*, Zhibin Chen, Zirui Wu, and Yansong Feng. "Lawyer llama tech report." arXiv preprint arXiv:2305.15062 (2023).

[4] Hejing Cao*, Zhenwei An*, Jiazhan Feng, Kun Xu, Liwei Chen, and Dongyan Zhao. "A Step Closer to Comprehensive Answers: Constrai Stage Question Decomposition with Large Language Models." arXiv preprint arXiv:2311.07491 (2023).

Research Experience

Dynamic routing mechanism in Mixture of Experts 2023.11-2024.02

- Intuition: harder task needs more parametes and vice versa in MoE models.
- Methods: We used nucleus sampling to dynamically allocate different number of experts depending on task difficulty. Instead of allocating fixed top-k experts in each FFN. Besides load-balance loss, the entropy of the routing distribution is introduced as dynamic loss to restrain MoE layers from choosing too many experts.
- Experimental results indicate that dynamic routing surpass Top-2 routing using less than 2 experts. We find lower layers need more experts than higher layers in Transformers and subwords-tokens always need more experts.

Heterogeneous Experts in Mixture of Experts 2024.02-2024.04

- Intuition: design a scalable architecture to dynamically activate different number of parameters for different tokens.
- Methods: We set heterogeneous experts with different sizes in each MoE layer and use top-k routing mechanism. Therefore, Input tokens can **activate different number of parameters** depending on the context and token type.
- Experiments show that heterogeneous MoE surpass homogeneous MoE and dynamic MoE. We find that MoE models show different Experts preference in different transformer layers. And Experts with more parameters focus on tokens which have less frequency.

Lawyer LLaMA 2023.03-2023.06

- Intuition: Embody LLaMA with Legal consultations ability.
- Method: Lawyer LLaMA first underwent continual pretraining on a large-scale legal corpus, enabling it to systematically learn about the legal knowledge system in China. Based on this, we have collected a batch of analyses for the Chinese National Judicial Examination objective questions and answers to legal consultations using ChatGPT, using the collected data to fine-tune the model via instructions, allowing the model to acquire the ability to apply legal knowledge to specific scenarios.

Multi-Stage Question Decomposition Using LLM 2023.09-2023.11

- Problem: If we have a set of (question, trustworthy answer) pairs, how could we answer new complex questions?
- This work focus on embodying LLM the ability to decompose complex questions into several sub-questions with trustworthy answers. LLM acts as an agent, which can use several retrievers and planning the action path to solve given problem.

Work Experience

Boss Zhipin Lab 2024.05-

LLM Pretraining

- We have trained a family of LLM from 1B to 112B from scratch named ***Nanbeige***. I am responsible for training MoE models and preparing the entire pretrain corpus.

Initial.AI 2023.11-2024.05

Role-Play ChatBot

- Developed a customizatble Role-Play ChatBot capable of simulating any charaters (e.g., historical figures, fictional characters), as defined by users, using LLM.
- Based on LLM, we used continue pretaining and supervised finetuning to learn Role-Play abilities. We used DPO to improve performance on multi-turn conversations.

Kuaishou Technology 2023.07-2023.11

Generative Recommendation based on LLM

- To recommed relative search query to users in short-video app, we build a Recommendation-Domain LLM.
- We collected pretrain corpus to continue pretrain llama2-7b from a graph using random walk. First, we build a huge graph using online data, where each node represents a search query in text from. If users search query B after A, there would be an edge from A to B with transition cnt as the weight. After continue pretrain and supervised finetuning, LLM can generate relative search queries given a search query.

Skills & Interests

- Programming skills: C/C++, Python, SQL & platforms such as PyTorch, Megatron-LM.
- English skills: IELTS 7.0 overall: reading 8.0; listening 8.0; writing 6.5; speaking 6.0
- Interests: I am mainly interested in sports such as Taekwondo, Badminton and swim. I used to be a member of Taekwondo team of Peking University. I also enjoy films especially science fiction films.