Quzhe Huang

Homepage : https://andrewzhe.github.io/

Email : huangquzhe@pku.edu.cn
Google Scholar: quzhe's scholar

Education

Peking University, Wangxuan Institute of Computer Technology — Beijing, China
Ph.D. in Computer Science. Advisor: Prof. Yansong Feng and Prof. Dongyan Zhao — *Sep 2019 - July 2024*

Peking University, Department of Computer Science — Beijing, China
B.S. in Computer Science — *Sep 2015 - Jun 2019*

Research Interests

My research interest lies in designing efficient LLM and adapting LLM to different modalities and domains, with the goal of building practical AI systems:

- **Efficient LLM**: build small but competitive long-context LLM from scratch(Tech Report) and reduce the cost of both training and inference by dynamically allocating resources(ACL 2024).
- **Domain-Specific LLM**: adapt LLM to specific domain by integrating domain knowledge(Tech Report) and evaluate whether models reason as domain experts(EMNLP 2022).
- **Multimodal LLM**: design a unified image-video-language pre-training framework (ICLR 2024, ICML 2024) and explore the ability to encode global information(COLING 2024).

I am also interested in document-level information extraction (ACL 2021, ACL 2022, ACL 2023, EMNLP 2023), long context understanding(ICLR 2024) and low-resource languages (ACL 2024).

Research

- **Large Language Model**

  1. **Dynamic Mixture-of-Experts Architecture (Efficient MoE):** Current MoE models rely on fixed Top-K routing and always activate a predetermined number of experts, ignoring the input's complexity. We propose a dynamic expert selection framework, which adjusts the number of activated experts based on the model's confidence for each input. This allows for more efficient utilization of computational resources, activating more experts for complex tasks that require advanced reasoning and fewer for simpler tasks. This work is accepted by ACL 2024[1].

  2. **Adapting a General LLM to the Legal Domain (Domain-Specific LLM):** Deploying general LLMs to a specific domain faces the challenge of a deficiency in domain-specific knowledge and an inadequate capability to leverage that knowledge. To alleviate this issue, we propose a framework for adapting general LLMs to specialized fields and build a legal domain model, Lawyer LLaMA. This framework enhances the LLM's domain knowledge through continual training and teaches the model to learn professional skills using properly designed supervised fine-tuning tasks. Additionally, we integrate a retrieval module to alleviate the hallucination problem in the model's response by sourcing relevant legal articles before answering queries. Our open-source project has gained great recognition, with about 800 stars on GitHub[2].

  3. **A Framework for Unified Language - Vision Pretraining (Multimodal LLM):** Prevailing approaches primarily regard visual input as the prompt and focus exclusively on optimizing the text generation process conditioned upon vision content by a frozen LLM. Such an inequitable treatment of vision and language heavily constrains the model's potential. We break through this limitation by representing both vision and language in a unified representation. To this end, we craft a visual tokenizer that translates the non-linguistic image into a sequence of discrete tokens like a foreign language that LLM could read. Pre-trained on the web-scale image-text corpus, our model LaVIT outperforms the existing models by a large margin on massive vision-language tasks. This work and the following work are accepted by ICLR and ICML 2024[3,4].

- **Document Level Information Extraction**

  1. **A Unified Framework for Event Temporal Relation Extraction:** Event Temporal Relation Extraction (ETRE) is usually formulated as a multi-class classification task, which treats each type of relation as a one-hot label but overlooks the meaning and dependency of relations. To better model the intrinsic dependency, we present a unified framework for ETRE that reinterprets relations through logical expressions of the start and end times of events. Our unified framework outperforms multi-class classification-based methods and it allows for leveraging the relations with sufficient data to assist the learning of others[8].

  2. **Evidence-Based Document-Level Relation Extraction (RE) Model:** In document-level RE, the relationship between two entities is implied throughout the document, but not all the content is necessary. We empirically show that determining the relation between entities in long texts only requires limited evidence, and design heuristic rules based on

co-reference and multi-hop reasoning to select evidence sentences for this task. Using the selected evidence instead of the whole document as input, a simple sequence model could perform better than fancy graph neural network-based methods[11].

Projects

- **Training a Large Language Model from Scratch (Team Leader at a Startup, 2023.4 - 2023.8)**
  We establish a pipeline encompassing data cleaning, training framework adaptation, model evaluation, and dynamic training strategy adjustment. Using the framework, we have trained a 7B Chinese-English bilingual model from scratch with an 8K context length and then expanded it to a context length of 128K. In the project, I mainly take responses for model training and evaluation. I also implement the context length expansion.

- **Building Character AI (Intern at Kuaishou, 2023.8 - 2024.6)**
  I participate in a character AI project, where the model can play a specific role based on a given profile. My responsibilities include enhancing the model's ability to generalize roles and improving its capability to model long conversation histories. The former job aims to enable the model to play user-defined roles, rather than just the roles learned during training. The latter job is intended to ensure that the model can maintain character consistency even after dozens of chat rounds and avoid generating responses that contradict the conversation history.

Publication (* means equal contribution)

1. **Harder Tasks Need More Experts: Dynamic Routing in MoE Models** ACL 2024
   Quzhe Huang*, Zhenwei An*, Nan Zhuang*, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, Yansong Feng
2. **Lawyer LLaMA: Enhancing LLMs with Legal Knowledge** Arxiv 2023
   Quzhe Huang*, Mingxu Tao*, Zhenwei An*, Chen Zhang*, Cong Jiang, Zhibin Chen, Zirui Wu and Yansong Feng
3. **Unified Video-Language Pre-training with Decoupled Visual-Motional Tokenization** ICML 2024
   Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, **Quzhe Huang**, et, al.
4. **Unified Language - Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization** ICLR 2024
   Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, **Quzhe Huang**, et, al.
5. **Probing Multimodal LLMs for Global and Local Semantic Representation** COLING 2024
   Mingxu Tao, **Quzhe Huang**, Kun Xu, Liwei Chen, Yansong Feng, Dongyan Zhao
6. **Can Perplexity Reflect Large Language Model's Ability in Long Text Understanding?** ICLR 2024 Tiny
   Yutong Hu, **Quzhe Huang**, Mingxu Tao, Chen Zhang, Yansong Feng
7. **MC2: A Multilingual Corpus of Minority Languages in China** ACL 2024
   Chen Zhang*, Mingxu Tao*, **Quzhe Huang***, Jiuheng Lin, Zhibing Chen, Yansong Feng
8. **More than Classification: A Unified Framework for Event Temporal Relation Extraction** ACL 2023
   Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, Dongyan Zhao
9. **A Progressive Framework for Document-level Informative Argument Extraction** Findings of EMNLP 2023
   Quzhe Huang*, Yanxi Zhang*, Dongyan Zhao
10. **Does Recommend-Revise Produce Reliable Annotations?** ACL 2022
    Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, Dongyan Zhao
11. **Three sentences are all you need: Local path enhanced document relation extraction** ACL 2022
    Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, Dongyan Zhao
12. **Exploring distantly-labeled rationales in neural network models** ACL 2021
    Quzhe Huang, Shengqi Zhu, Yansong Feng, Dongyan Zhao
13. **Relation-Aware Question Answering for Heterogeneous Knowledge Graphs** Findings of EMNLP 2023
    Haowei Du, **Quzhe Huang**, Chen Li, Chen Zhang, Yang Li, Dongyan Zhao
14. **Customizing Small Language Model for Dynamic Token Pruning** Findings of EMNLP 2023
    Chang Liu, Chongyang Tao, Jianxin Liang, Jiazhan Feng, Tao Shen, **Quzhe Huang**, Dongyan Zhao
15. **Do Charge Prediction Models Learn Legal Theory?** Findings of EMNLP 2022
    Zhenwei An*, **Quzhe Huang***, Cong Jiang, Yansong Feng, Dongyan Zhao
16. **Rethinking Task-Specific Knowledge Distillation: Contextualized Corpus as Better Textbook** EMNLP 2022
    Chang Liu, Chongyang Tao, Jianxin Liang, Tao Shen, Jiazhan Feng, **Quzhe Huang**, Dongyan Zhao
17. **Why Machine Reading Comprehension Models Learn Shortcuts?** Findings of ACL 2021
    Yuxuan Lai, Chen Zhang, Yansong Feng, **Quzhe Huang**, Dongyan Zhao
18. **Towards context-aware code comment generation** Findings of EMNLP 2020
    Xiaohan Yu, **Quzhe Huang**, Zheng Wang, Yansong Feng, Dongyan Zhao

Service

- **Area Chair**
  ACL Rolling Review 2023.10
- **Reviewer**
  ACL Rolling Review since 2021.9; ACL 2022-2024, EMNLP 2022-2023; COLING 2022-2024; EACL 2023; AAAI 2023

Honors and Awards

- **President Scholarship (Top 5%)**, Peking University, 2022
- **Uniqlo Scholarship**, Peking University, 2017
- **Panasonic Scholarship**, Peking University, 2016

Invited Talks

- **Lawyer LLaMA: Enhancing LLMs with Legal Knowledge**
  Huawei Cloud (2023.7); 4Paradigm Sage (2023.10)
- **Extending the Context Length of Large Language Models**
  CIPS ATT @ Chengdu (Advanced Technology Tutorial hosted by Chinese Information Processing Society) 2023.12