Haoyu Wang

✉wanghaoyu666@bupt.edu.cn ☎+86-15600989921
Beijing, China GitHub Google Scholar

Education

Beijing University of Posts and Telecommunications (BUPT), China *2021.09 - Present*

- Bachelor of Science, **Computer Science and Technology**.
- GPA **3.79/4**, Average Score **90.97/100**, Rank **25/419** (5.97%).
- **Elite Class**: Ye Peida Innovation and Entrepreneurship Experimental Class (100 students per grade in the entire school).

Research Interest

- Current research: **synthetic data curation, code generation, long-context modeling, and multilinguality** of large language models (LLMs).
- Interest: LLM for Code, Multi-modal, Agent, LLM acceleration, Embodied AI.

Research Publication

UltraLink *ACL 2024 accepted, Main Conference, Poster* **Arxiv Link**

- **Haoyu Wang**, Shuo Wang, Yukun Yan, Xujia Wang, Zhiyu Yang, Yuzhuang Xu, Zhenghao Liu, Liner Yang, Ning Ding, Xu Han, Zhiyuan Liu, Maosong Sun. "UltraLink: An Open-Source Knowledge-Enhanced Multilingual Supervised Fine-tuning Dataset." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

SpanCS *CCL 2024 accepted, Main Conference, Poster*

- Qingfu Zhu, Shiqi Zhou, Shuo Wang, Zhiming Zhang, **Haoyu Wang**, Qiguang Chen, Wanxiang Che. "SpanCS: Span-Level Code-Switching for Cross-Lingual Program Synthesis." In Proceedings of the 23rd China National Conference on Computational Linguistics.

FedHGNN *The Web Conference 2024 accepted, Oral* **Arxiv Link**

- Yan, Bo, Yang Cao, **Haoyu Wang**, Wenchuan Yang, Junping Du and Chuan Shi. "Federated Heterogeneous Graph Neural Network for Privacy-preserving Recommendation." In Proceedings of the ACM on Web Conference 2024 (2023).

OMGEval *EMNLP 2024 in processing* **Arxiv Link**

- Yang Liu, Meng Xu, Shuo Wang, Liner Yang, **Haoyu Wang**, Zhenghao Liu, Cunliang Kong, Yun Chen, Yang Liu, Maosong Sun, Erhong Yang. "OMGEval: An Open Multilingual Generative Evaluation Benchmark for Large Language Models."

Research Experience

Research Intern *THUNLP, Tsinghua University* *2023.11 - Present*

- Conducting research in Large Language Model and Natural Language Processing under the guidance of Prof. Zhiyuan Liu and Prof. Maosong Sun.
- **Spearheaded the work of UltraLink**, encompassing idea concretization, coding, experiment design and implementation, paper writing, rebuttal, and maintaining open-source repositories.
  - **Achieve SOTA multilingual SFT dataset**, better than Aya Dataset of Cohere AI, amplifying both language-specific and language-agnostic abilities.
  - **Explore the code and math ability between different languages**, leverage transfer learning to get code and math knowledge from English part to relieve differences in abilities between different languages.
  - **Strengthen the multilingual chat capabilities** of Language Models, trying to mitigate the curse of multilinguality, and develop an asynchronous multi-thread data curation pipeline to generate high-quality multilingual data.
- **Lead a group**, devising research strategies and mentoring 3 interns in assimilating into the research workflow.
- **Join in the work of OMGEval**, help to improve the benchmark quality, coding, and do baseline experiments.
- **Join in the work of SpanCS**, build up a pipeline for generating a multilingual version of HumanEval in 5 languages.

Research Intern *GAMMA Lab, BUPT* *2023.03 - 2023.12*

- Conducting research in the field of Heterogeneous Graph and Federal Learning under the guidance of Prof. Chuan Shi.
- **As a developer of an open-source library, GammaGL**, reproduce a graph self-training framework model, DR-GST, and integrate it into the open-source multi-backend graph learning library, GammaGL.
- **Join in the work of FedHGNN**, learn about differential privacy and Graphs, conduct research paper surveys, reproduce baseline models, do experiments on baselines, create analytical charts, and write algorithm flowcharts.

Work Experience

Research Intern *ModelBest* *2024.04 - Present*

- **Designing algorithm about long context SFT dataset**, concentrating on high-quality questions and longer than 128k data, to strengthen the long context ability of MiniCPM.
- **Implement parallel long-context training**, using DeepSpeed for basic SFT, implement Qwen2 through BMTrain, and realize tensor parallelism to meet the training requirements for long context data.
- **Help to design multi-agent algorithm about recursive long-context processing mechanism.**

Project

UAV motion inspection system for intelligent traffic management *Group Member* *2023.09-2024.07*

- **National College Students' Innovative Entrepreneurial Training Plan Program**.
- **Implementing Vehicle Detection** with YOLOv5 and **Customized training dataset** to adapt to specific scenarios recognition.
- **Develop GUI** using PySide (Python version of Qt) and **expand DJI SDK** to provide control API.

Multimedia Information Retrieval And Extraction System *Group Leader* *2024.02-2024.06*

- **Data acquisition**: Use requests and BeautifulSoup to get raw data from HTML, and use Pandas to clean and handle outliers.
- **Information Retrieval**: Construct a basic inverted index algorithm, TF-IDF comparison algorithm, and vector space matching algorithm to retrieve most related information from the user description.
- **Information Extract**: Extract information points using regular expression matching algorithm, LLM prompt engineering, and named entity recognition based on BERT.
- **Multi-modal and Visualise**: Aligning multiple modalities to the text modality using multi-modal models, and using Matplotlib to visualize the result.

Hotel Management System *Group Leader* *2024.04-2024.06*

- **Develop a hotel management system** with B/S hierarchical architecture using Springboot, MySQL, MyBatis and Vue.
- **Design software architecture and API**, and integrate all other functional modules.
- **Cooperate with the other 3 groups** as the leader of the large group (about 20 members), unified API and adaptive environment.

ChatBot Script DSL Design *Individual completer* **GitHub** *2023.09-2024.01*

- **Design a Domain-Specific Scripting Language (DSL)** that can describe the automatic response logic of online customer service chatbots.
- **Design a C-style customized syntax and implement a Python interpreter**, using recursive call predictive analysis.
- **Design and implement a command line interface** for backend scenarios, supporting multiple processes and cache.

Skill

Language

- Mandarin: Native; English: Fluent, IELTS 7.0

Teamwork

- Leadership, project management, efficient communication, document writing, timeline arrangement...
- Agile development and iteration, decomposing and concretizing problems...

Coding

- Pytorch, Transformers, DeepSpeed, vllm, LangChain, Matplotlib, Pyecharts, Numpy, Pandas, Scikit-learn...
- Qt, CUDA, SpringBoot, MyBatis, MySQL, requests, BeautifulSoup...
- Java, Shell, SQL, VHDL, Coq, LaTeX...

Tools

- Prompt writing, Web crawler, Bash script, conda, tmux, Git, Docker, Conda...

Awards & Scholarships

Blue Bridge Cup *an OI-style algorithm competition*	**Group A, Provincial Third Prize**	*2023*
BUPT Undergraduate Outstanding Student Scholarship	**Third Class (Top 15%)**	*2023*
BUPT Excellent Student	**Top 10%**	*2023*
BUPT Undergraduate Outstanding Student Scholarship	**Second Class (Top 10%)**	*2022*
BUPT Excellent Student	**Top 10%**	*2022*