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: **Large Language Model (LLM) Alignment**. Concretely, through building a data **synthesizing** and **filtering** mechanism to generate alignment data to **adapt to downstream tasks** or **strengthen LLM ability**.
- Interest: LLM for SE, Multi-modal Alignment, Agent, LLM infra for large scale training, GNN + LLM, 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

Light-weight data processing pipeline *Individual developer* *2024.07-Present*

- A lightweight Python library that supports **easily mapping the data flow graph to the actual data processing pipeline**.
- **Function coroutine as each node, connect node together as pipeline**, a pipeline could also be a node for another pipeline.
- **Using decorator to expand node** to support superstep, labelled pipeline and monitor.
- **Support both compute-intensive and IO-intensive tasks**, by using coroutine, threading pool and process pool.

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.

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

- **Design a Domain-Specific 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, decomposing and concretizing problems, concurrent optimization...

Coding

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

Tools

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

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*

北京郵電大學
Beijing University of Posts and Telecommunications

Undergraduate Transcript

Name	WANG Haoyu	Gender	Male
Student ID	2021211282	Class	2021211312
Major	Computer Science and Technology	School	School of Computer Science

Student Type	Full-time Undergraduate	Date of Enrollment	202109	Date of Graduation	202507

Course Title	Credit	Grade	Course Type	Term
Safety Education	0	Good	Compulsory	2021Fall
Practice of Innovation and Entrepreneurship	1.5	88	Elective	2021Fall
Undergraduate Psychological Health	0.5	83	Compulsory	2021Fall
The Education of Drug and AIDS prevention	2	88	Optional	2021Fall
Advanced Mathematics A (I)	5	82	Compulsory	2021Fall
Introduction to Computing and How to Program	4.5	84	Compulsory	2021Fall
Training of Thought and Morality and General Knowledge of Law	3	86	Compulsory	2021Fall
Outline of Xi Jinping's New China's Socialist Ideology	2	91	Compulsory	2021Fall
Linear Algebra	3	89	Compulsory	2021Fall
Situation and Policies I	0.4	86	Compulsory	2021Fall
Chinese Ancient Architectural Culture and Appreciation	2	84	Optional	2021Fall
Comprehensive English 3	2	90	Compulsory	2021Fall
University Physics C	4	85	Compulsory	2022Spring
Basis of Circuit Analysis and Electronic Circuit	2	87	Compulsory	2022Spring
Advanced Mathematics A (II)	5	97	Compulsory	2022Spring
Introduction to Computing and Foundation of Programming	1.5	95	Elective	2022Spring
Military Theory	2	97	Compulsory	2022Spring
Discrete Mathematics（1）	2	88	Compulsory	2022Spring
Sports Foundation	1	88	Compulsory	2022Spring
Physics Experiment A	1.5	Good	Compulsory	2022Spring
Situation and Policies II	0.4	88	Compulsory	2022Spring
The Course Introduction of Compendium of Chinese Modern History	2.5	93	Compulsory	2022Spring
The Course Introduction of Compendium of Chinese Modern History（Practice）	0.5	87	Compulsory	2022Spring
Comprehensive English 4	2	93	Compulsory	2022Spring
Probability Theory and Mathematical Statistics	4	79	Elective	2022Fall
Introduction to computer graphics and 3D game engine development	2	95	Optional	2022Fall
Introduction to Computer Systems	2	95	Compulsory	2022Fall
Course Project -- Basics of Computer Systems	0.5	94	Compulsory	2022Fall
Discrete Mathematics (2)	3	88	Compulsory	2022Fall
The Brief Introduction of Marxism	2.5	91	Compulsory	2022Fall
The Brief Introduction of Marxism（Practice）	0.5	93	Compulsory	2022Fall
Data Structures	4	90	Compulsory	2022Fall
Digital Logic and Digital System	4	90	Compulsory	2022Fall
Display technology development and game application	2	93	Optional	2022Fall
Situation and Policies III	0.4	88	Compulsory	2022Fall
English listening and speaking 2	2	86	Compulsory	2022Fall
Swimming Elective Course	1	93	Optional	2022Fall
Operations Research	2	94	Elective	2022Fall

Date:2024/7/9 18:34:03

北京郵電大學
Beijing University of Posts and Telecommunications

Course Title	Credit	Grade	Course Type	Term
Computer Networks	4	88	Compulsory	2023Spring
Curriculum Practice of Computer Networks	1.5	88	Elective	2023Spring
Computer Organization Principles	4	90	Compulsory	2023Spring
Military Skill Training	2	99	Compulsory	2023Spring
Practical Approaches to Intercultural Communication	2	92	Elective	2023Spring
Introduction to Mao Zedong Thought and the System of Theories of Socialism with Chinese Characteristics	4	93	Compulsory	2023Spring
Introduction to Mao Zedong Thought and the System of Theories of Socialism with Chinese Characteristics（Practice）	1	90	Compulsory	2023Spring
Object-Oriented Programming Design and Practice (java)	2	98	Elective	2023Spring
Ping Pong	1	88	Elective	2023Spring
Course Project -- Data Structures	1.5	93	Elective	2023Spring
Digital Logic and Digital System Curriculum Design	2	92	Elective	2023Spring
Formal Languages and Automata	2	95	Compulsory	2023Spring
Situation and Policies IV	0.4	89	Compulsory	2023Spring
Python Programming	2	96	Elective	2023Fall
Compiler Principle and Technology	3	94	Compulsory	2023Fall
Operating System	4	92	Compulsory	2023Fall
The Prictice of Programming	2	97	Elective	2023Fall
Experiments of Computer Network Technology	2	99	Elective	2023Fall
Renewable Energy and Low-Carbon Society	2	99	Optional	2023Fall
Psychology of Intimate Relationships	2	98	Optional	2023Fall
Practice of Social Innovation and Social Entrepreneurship	2	85	Optional	2023Fall
Classic Art of World Famous Museums	2	99	Optional	2023Fall
Principles of Database Systems	3	92	Compulsory	2023Fall
Design and Analysis of Algorithms	2	79	Compulsory	2023Fall
Breaststroke	1	93	Elective	2023Fall
Appreciation of Foreign Architecture	2	99	Optional	2023Fall
Network Storage Technology	2	95	Elective	2023Fall
Introduction to Western Civilizations	2	98	Optional	2023Fall
Situation and Policies V	0.4	90	Compulsory	2023Fall
Western Music in 20th Century	2	99	Optional	2024Spring
Linux Development Environment and Application	2	95	Elective	2024Spring
King of Intangible Cultural Heritage — Appreciation of Kunqu Opera	2	99	Optional	2024Spring
Cricket	1	88	Elective	2024Spring
Parallel Computation & GPU Programming	2	88	Elective	2024Spring
Operating System Course Design	1.5	83	Elective	2024Spring
The Art of Dunhuang	2	99	Optional	2024Spring
‘Internet Plus’ Thinking and Entrepreneurship practice	2	85	Optional	2024Spring
Machine Learning	2	93	Elective	2024Spring
Computer Architecture	3	96	Compulsory	2024Spring
Software Engineering	3	90	Compulsory	2024Spring
Appreciation of Shakespearian Plays	2	98	Optional	2024Spring
The Great Work——A Dream of Red Mansions	2	92	Optional	2024Spring
Modern Switching Principles	3	79	Compulsory	2024Spring
Information and Knowledge Acquisition	2	92	Elective	2024Spring
About the Forbidden City	2	98	Optional	2024Spring

NOTE:
(1) Beijing University of Posts and Telecommunications is a full-time accredited university directly under the administration of the

Ministry of Education of the People's Republic of China. It offers four-year programs for bachelor's degree. The duration for the second bachelor's degree is two years.

(2) Four grading scales are adopted in the academic transcript: 100-point scale, 5-level ordinal scale(Excellent, Good, Average, Pass, and Fail), Binary scale(Good/Fail) and Exempted. Grades that are not obtained from first-time exams are marked with *.

(3) As for the 100-point scale, credits are granted for grades that are over 60 (60 included). Grade points = 4-3×(100-X)×(100-X)÷1600 (60≤X≤100), where X is the grade obtained under the 100-point system. Grade points is 4 for 100, 1 for 60, and 0 for grades below 60. For the 5-level ordinal scale, grades between 100-90 are Excellent; 89-80 are Good; 79-70 are Average; 60-69 are Pass, and grades below 60 are Fail. For the Binary scale, grades between 100-60 are Good, and those below 60 are Fail.

(4) As for the 5-level ordinal scale, credits are granted for grades at or above Pass. One hundred points grades are assigned as: Excellent=95, Good=85, Average=75, Pass=65, and Fail=59. Grade points are assigned as: Excellent=3.95, Good=3.58, Average=2.83, Pass=1.7, and Fail=0.

(5) As for the Binary scale, credits are granted for grades at Good. One hundred points grades are assigned as: Good=80, Fail=59. Grade points are assigned as: Good=3.25, Fail=0.

(6) Students could be exempted from certain courses upon passing specific tests and granted credits accordingly. The courses will be marked as "Exempted", without specific grades on the transcript.