Wenxuan Jiang

(+86) 135-0040-2091 | ppangxuan@outlook.com | pangpang-xuan.github.io | Born in 2004.02

Education

Northeastern University, China, B.E. in **Software Engineering (AI Track)**	Sep 2022 – Present
GPA: 3.82/4.0 (89.1/100)	*Shenyang, China*

Publications

KnowCoder-X: Boosting Multilingual Information Extraction via Code — **ACL 2025 (Findings)**

Yuxin Zuo*, **Wenxuan Jiang* (co-first author)**, Wenxuan Liu*, Zixuan Li, Long Bai, Hanbin Wang, Yutao Zeng, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng.

SlideCoder: Layout-aware RAG-enhanced Hierarchical Slide Generation from Design — **EMNLP 2025 (Oral)**

Wenxin Tang*, Jingyu Xiao*, **Wenxuan Jiang**, Yuhang Wang, Xi Xiao, Xuxin Tang, Qing Li, Yuehe Ma, Junliang Liu, Shisong Tang and Michael Lyu

Research internship

Meituan (M17 Group) — Oct 2025 – Present

- We are responsible for building Data-Centric's data-training-evaluation loop, exploring **LongCat** structure, evaluation strategy, and the impact of evaluation data on model performance. As current benchmarks fail to adequately distinguish the performance differences among models, we aim to design more challenging and discriminative benchmarks.
- Based on the evaluation results, we will further improve both the training data and the model training methodology during pre-training and post training.

Tsinghua University — Apr 2025 – Jun 2025

Mentor: Xi Xiao

- We propose a new task called Reference Image-to-Slide Generation. To address this task, we introduce a framework that combines color-gradient segmentation with hierarchical retrieval-augmented generation to produce editable presentation slides from reference images.
- Additionally, we train a model on enhanced reverse-engineered data, achieving performance that approaches that of GPT-4o. Finally, we present the first benchmark for this task, featuring graded difficulty levels.

Northeastern University — Nov 2024 – Jan 2025

Mentor: Shuang Wang

- We use LLMs to generate tabular data. In the process of textualizing tabular data, we represent column elements as Python classes for pre-training. During column element ordering, we compute data correlations between elements to perform feature-based sorting of the columns.
- In the end, we instruct the model to generate the corresponding tabular data by retrieving the values associated with each class element.

Chinese Academy of Sciences (Key Laboratory of AI Safety) — Apr 2024 – Oct 2024

Mentor: Xiaolong Jin and Xueqi Cheng.

- To address the disparities in multilingual capabilities among large language models, we propose a two-phase training framework to enhance multilingual information extraction (IE) with LLMs. In the first phase, we construct schemas by defining Python classes in the target language.
- To mitigate the uneven distribution of data across languages, we develop an automatic pipeline for constructing NER parallel datasets, which are subsequently used for cross-lingual alignment in information extraction.
- In the second phase, we perform multilingual IE instruction tuning to further improve the model's cross-language generalization.

Projects

Your personal doctor bot.   Jun 2023 – Jan 2024

Mentor: Yin Zhang

- We developed a virtual medical expert capable of interacting with users through dialogue. Weintegrate voice wake-up, speech-to-text, and text-to-speech (TTS) technologies to enable users to communicate with the model through natural speech.
- We fine-tune our model using both publicly available medical data and proprietary private data. To further enhance alignment, we apply the Direct Preference Optimization (DPO) policy with LoRA, leveraging multi-GPU training via DeepSpeed. Finally, the model performs inference using vLLM in combination with retrieval-augmented generation (RAG) for improved efficiency and accuracy.

Awards & Honors

The Third Prize of Scholarship of Northeastern University   2023-2025

Skills

Coding: Proficient **Python** and **C++** , and familiar with **Linux**.

Deep Learning: Master **PyTorch**, with expertise in using and finetuning **LLM**, particularly in **NLP**.

Language: Chinese (Native), English (Fluent), with an **IELTS** score of **6.5**.