Linxi Wu

linxiw@unc.edu | (984) 261-7719 | Chapel Hill, NC

EDUCATION

University of North Carolina at Chapel Hill *Aug 2022 – May 2026, Expected*

Computer Science (B.S) & Math (B.S.) & Statistic Minor **GPA**: 3.913/4.0 (Dean List)

- Relevant Studies: Machine Learning; Deep Learning; Algorithm & Analysis; Data Structure; Computer Organization; Real Analysis; Linear Algebra; Differential Equations; Optimization; Probability

TECHNICAL SKILLS

- **Model Structure**: LLaMA, LLaVA, GPT2, Qwen, Transformer, BERT
- **Training Methods:** LoRA, QLORA, SFT, Instruct turning, PPO, DPO
- **Inference Frames:** vLLM, TensorRT vLLM, TGI, GPTQ, BNB
- **NLP Algorithm:** BPE, BBPE, SentencePiece, Greedy Search, Beam Search, Viterbi
- **Programming Skills:** Python, C++, GoLang, Java, HTML, CSS, R, JavaScript, SQL(MySQL), MongoDB

EXPERIENCE

Algorithm Developer Intern

China Unicom Artificial Intelligence Innovation Center June – Aug 2024

- Standardized and converted fuzzy time information from extensive customer conversation data into precise, accurate time formats, ensuring the consistency and reliability of data used in form filling, leading to a **20%** improvement in data accuracy and enhancing **30%** of the quality of customer interaction records.
- Utilized the **LLaMA** model to extract fuzzy time information from extensive dialogue data, then fine-tuned the model with **LoRA and QLoRA**, leading to a **30%** increase in processing efficiency and enhanced accuracy in time extraction.
- Validated time accuracy generated by OpenAI GPT-4 using **10,000** data points from an open-source customer conversation dataset from China Mobile, optimizing model generalization through strategic data splitting.
- Fine-tuned models including LLaMA3-8B-Instruct, LLaMA3-70B-Instruct-GPTQ-Int8, and LLaMA2-13B using LoRA and QLoRA on an A100-80G*4 machine. with the **LLaMA3-70B-Instruct-GPTQ-Int8** model achieving the best performance and higher accuracy.
- Enhanced the vertical-domain validation set F-score from **77.6% to 88.5%** and the general-purpose validation set F-score from **90.6% to 92.4%** using the fine-tuned LLaMA3-70B-Instruct-GPTQ-Int8 model, significantly improving performance across different domains.

Algorithm Developer Intern

Xi'an Xinfang Electronic Technology Company Dec 2023 – Jan 2024

- Implemented large language model inference service to an enterprise-oriented text extraction and summary generation project, meeting the demand for high efficiency and accuracy in text processing.
- Utilized the **vLLM** inference framework to deploy the fine-tuned Qwen2-72B-GPTQ-Int4 model, ensuring high accuracy and stability for online services while increasing processing speed by **20%**.
- Analyzed the data from previous service version, finding the ratio of prompt length processed by prefill to decode length generated was approximately **9:1**, resulting in a **40%** reduction in computing stress through prompt caching.
- Conducted multi-machine, multi-instance testing using **Nginx** with the vLLM inference framework for **load balancing**, resulting in a **70%** improvement in machine throughput compared to the original llama.cpp reasoning scheme.

Data Analyst Intern

Dunhuang Smart Tourism June – Aug 2023

- **Collaborated** with the team to identify **20** key performance indicators (KPIs), including visitor flow, length of stay, repeat rate and satisfaction, to provide a quantitative basis for business decisions, incased the company identify business growth points and increased tourist satisfaction **15%** through marketing activities.
- Developed a data cleaning process to remove duplicate records, correct formatting errors and fill in missing values, improving data quality by **40%**.
- Implemented the data standardization process to formate the data and standardized measurement, the accuracy of data analysis was improved by **20%**, and the decision-making errors were reduced.
- Analyzed more than **100,000** visitor feedback data to identify the key factors influencing visitor satisfaction using **Tableau** and created visualization of visitors' demographic information and behavior patterns using dynamic dashboards, grouped tourists into different market segments, increased bookings for specific tourism products by **30%**.