ZHENGYANG YU

(+86) 13316856459 | yuzhengyang_2024@126.com | https://zhengyang-2002.github.io

EDUCATION

Lanzhou University **09/2020 – 07/2024**

B.E. in Computer Science and Technology (Data Science Track) **GPA:***86.92/100* *Lanzhou, China*

Relevant Courses: Data Science Programming (97), Data Structures (93), Data Mining (93), Machine Learning (94), Information Retrieval (90), Computing and Informatics Design (93), Calculus (95), Business Statistic (91)

INTERNSHIPS EXPERIENCE

OPPO **07/2023 – 11/2023**

Large Language Model Algorithm Intern *Shenzhen, China*

Task 1: Pre-training Pipeline Development and Optimization

- Developed a multi-process program to crawl 2TB of code-centric pre-training corpora from multiple sources and designed a cleaning pipeline tailored for code data, including metadata collection, privacy anonymization, and large-scale deduplication.
- Implemented an efficient Minhash+LSH method for deduplication and introduced a novel approach by transforming the LSH table into a connected graph, applying community detection to significantly enhance deduplication efficiency.
- Trained a tokenizer on the cleaned corpus and conducted incremental pretraining of the Starcoder-15.5B model using Deepspeed for parallel training across A100 clusters.
- Monitored metrics closely and applied targeted strategies, including rolling back checkpoints to prevent catastrophic forgetting, and removing noisy data and adjusting the learning rate to improve training stability and effectiveness.

Task 2: Fine-tuning for Real-world Integration

- Designed a code fine-tuning instruction auto-generation system based on Self-instruction and WizardLM, generating over 100K code fine-tuning data entries.
- Implemented and compared various semantic deduplication methods, identifying embedding-based clustering as the most effective approach. This method improved data quality and enabled the code model to achieve an 8% to 10% performance improvement on Human-Eval.
- Designed specific fine-tuning formats to enable the code model to use external tools during dialogue, facilitating the development and release of cod e completion and testing plugins for VSCode.

RELEVANT EXPERIENCE

Research on Psychological Counseling Based on Large Language Models **02/2024 – 06/2024**

Thesis Project, Supervised by Prof. Mingqiang Yang, Lanzhou University *Lanzhou University*

- Implemented a dual-model based dialogue data augmentation system for psychological counseling, transforming open-source, non-professional dialogues into multi-turn dialogue datasets that conform to professional counseling paradigms.
- Designed fine-tuning instructions. Applied the LoRA method for model fine-tuning, and aligned the model with human values using the Direct Preference Optimization (DPO) method to ensure safe usage.
- Developed an empathy assessment model based on bidirectional encoders and trained with Rank loss to measure the level of empathy of the generated dialogues.
- Developed an Elo-based scoring system to evaluate the model across dimensions such as professionalism, practicality, and Socratic dialogue capability; determined that Mental-GLM out- performed other evaluation models after hundreds of rounds of testing.

GoDaddy - Microbusiness Density Forecasting **11/2022 – 06/2023**

Kaggle Competition *Remote*

- Conducted in-depth analysis of the microbusiness dataset, ensuring the accuracy and reliability of the data. Detected outliers, visualized the key features, and improved model prediction accuracy by filtering out abnormal data.
- Applied regularization, differencing, and exponential weighted averaging to the time-series data to enhance the stability of subsequent model training.
- Used K-means clustering to group counties and designated counties in smaller clusters as part of a training blacklist to reduce the impact of outliers.
- Built a stacking model with XGBoost and SVM as first-level prediction models and linear regression as the meta-model. Trained the model and conducted expanding window cross-validation, fine-tuning hyperparameters to improve the model's generalization ability.
- Achieved an SMAPE of 4.09 and a top 7% ranking in the Kaggle competition, earning a bronze medal.

SKILLS

Programming: Python, C, C++, R, Java, HTML/CSS/JavaScript, PHP, Dart, SQL

Frameworks: PyTorch, TensorFlow, Deepspeed, Sklearn, NumPy, Flask, Spark, Hadoop

Tools: Git, Vim, PostgreSQL, Flutter, Docker, Kubernetes, OpenRefine, Weka