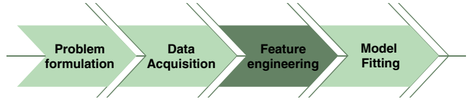
Self-Instruct CAAFE

Yanheng He, Jiahe Jin, Yuxuan Zhang
Shanghai Jiao Tong University

Background

Auto Feature Engineering: automatically create meaningful features from raw data to improve ML performance.



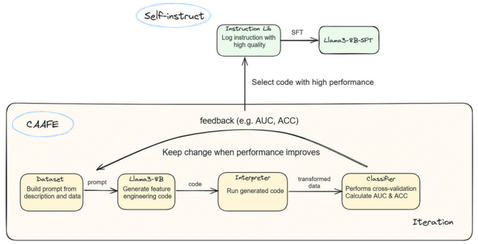
LLM Agent for Data Science: advanced AI systems that utilize LLM as central computational engine.

Llama3
General Education

Feature Engineering Agent
Professional Education

Process

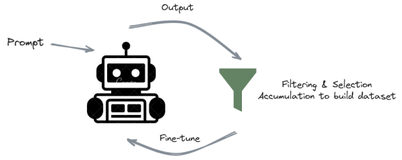
Self-Instruct CAAFE: a combination of CAAFE's basic method and the innovative Self-Instruct method



Key Contributions

Self-Instruct Method

We present an innovative experience accumulating method for building high-quality fine-tuning data automatically.



Smaller Open-Source Base Model

We replace the GPT-4 used in CAAFE with a much smaller open-source model, Llama3-8B, significantly lowering hardware barriers.

Optimized CAAFE Pipeline

We use shorter prompts and optimized feedback mechanism, achieving faster inference and better performance on smaller model.

Experiments

Setup

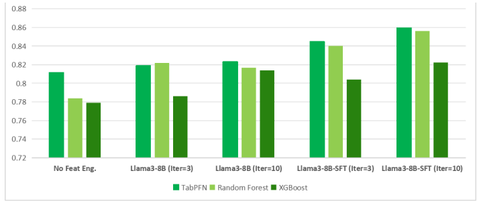
GPU: RTX 3090 (24GB VRAM)
Metrics: AUC (Area under the ROC Curve)
Base Model: Llama3-8B-Instruct
SFT Dataset: Self-Instruct dataset with 293 instructions
SFT Method: LoRA
Training Time: 3.5 hours
Downstream Classifiers: TabPFN & Random Forest & XGBoost
Evaluation Datasets: 10 OpenML datasets

Evaluation results

We adapted Llama3-8B to CAAFE and set iteration time = 10, random seed = 42. It achieved good improvements compared to no feature engineering.

Dataset	No Feat. Eng.	CAAFE with Llama3-8B
balance-scale	1.0000	1.0000
breast-w	1.0000	1.0000
cmc	0.7672	0.7671
credit-g	0.7333	0.7467
diabetes	0.8639	0.8667
tic-tac-toe	0.3810	0.4762
eucalyptus	0.9282	0.9278
pc1	0.8669	0.8649
airlines	0.7324	0.7704
jungle_chess...	0.8438	0.8579
Average	0.8117	**0.8278**

We changed the random seed and evaluated for twice, calculating the variance of the AUC. After fine-tuning, Llama3-8B achieves great improvement steadily on all of three classifiers.



Discussion

Limitations

In our preliminary exploration of experience generalization, about 1K instructions across 41 datasets were insufficient for Llama3-8B to generalize these feature engineering experiences to unseen datasets.

Hypothesis

We hypothesize that larger-scale experience accumulation from more datasets is required to meet Scaling Law requirements and achieve emergent generalization capabilities.