7591

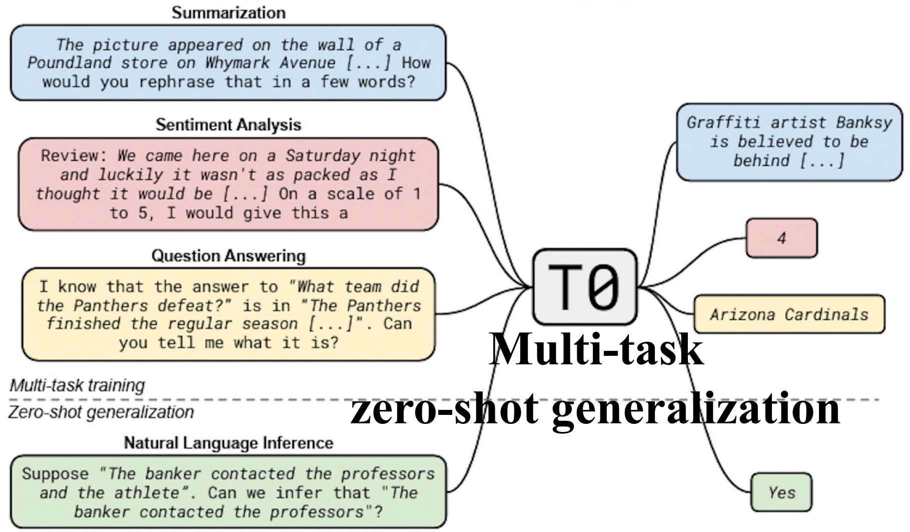
Small Models are LLM Knowledge Triggers for Medical Tabular Prediction

Jiahuan Yan[1], Jintai Chen[2,*], Chaowen Hu[1], Bo Zheng[1], Yaojun Hu[1], Jimeng Sun[3], Jian Wu[1]

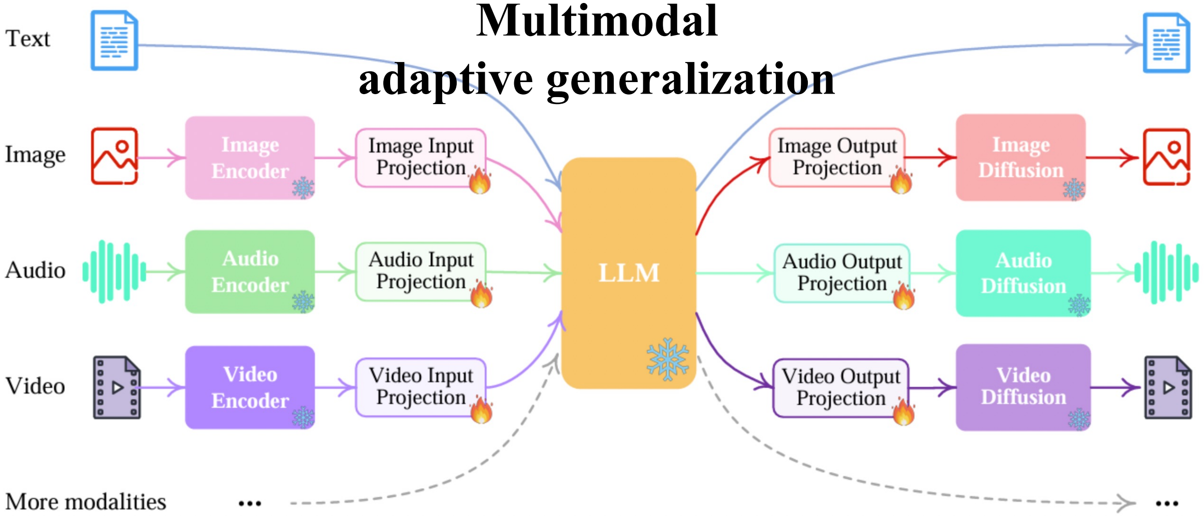
[1]Zhejiang University, Hangzhou, China
[2]Hong Kong University of Science and Technology (Guangzhou), China
[3]University of Illinois Urbana-Champaign

Background

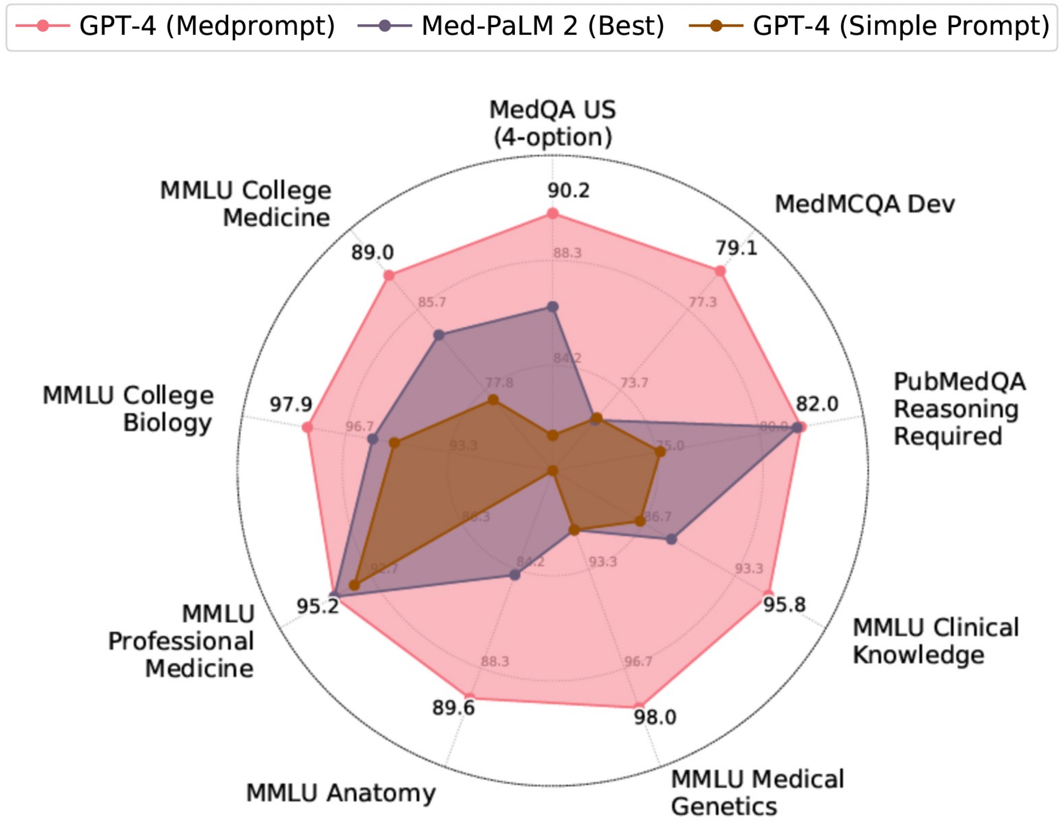
1. Universal domain proficiency of LLMs on reasoning unstructured data tasks without fine-tuning



Multi-task zero-shot generalization



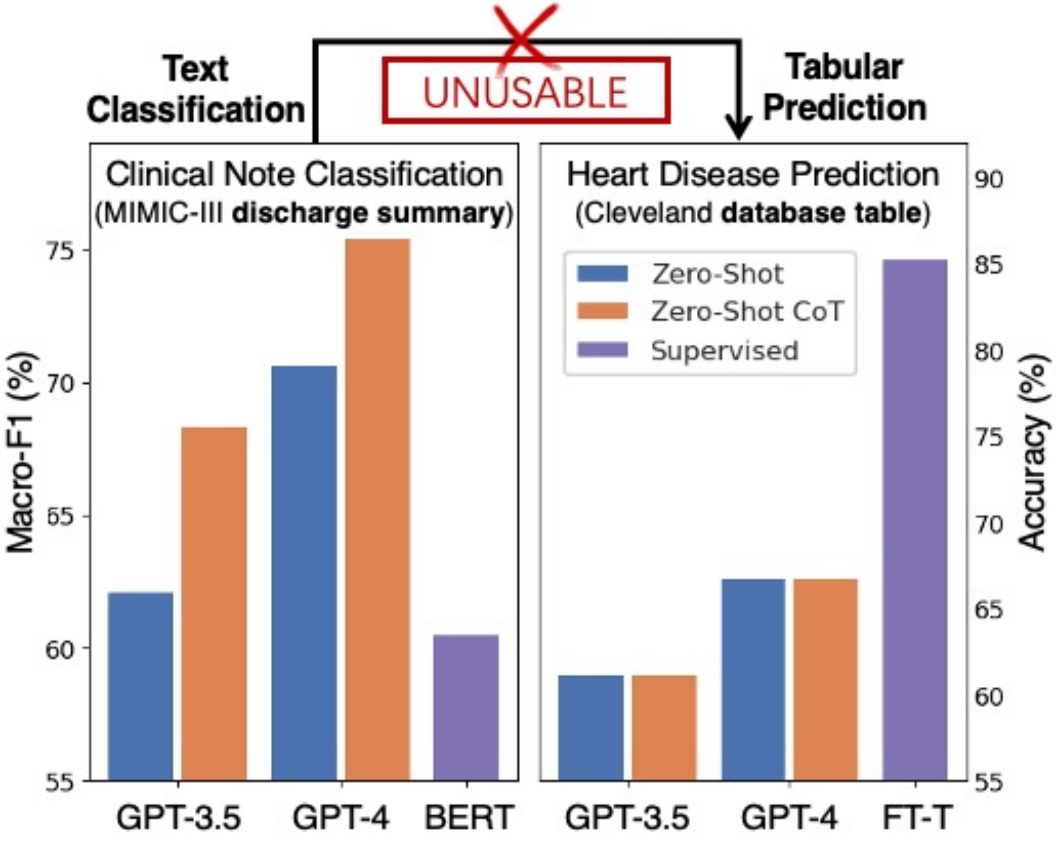
Multimodal adaptive generalization



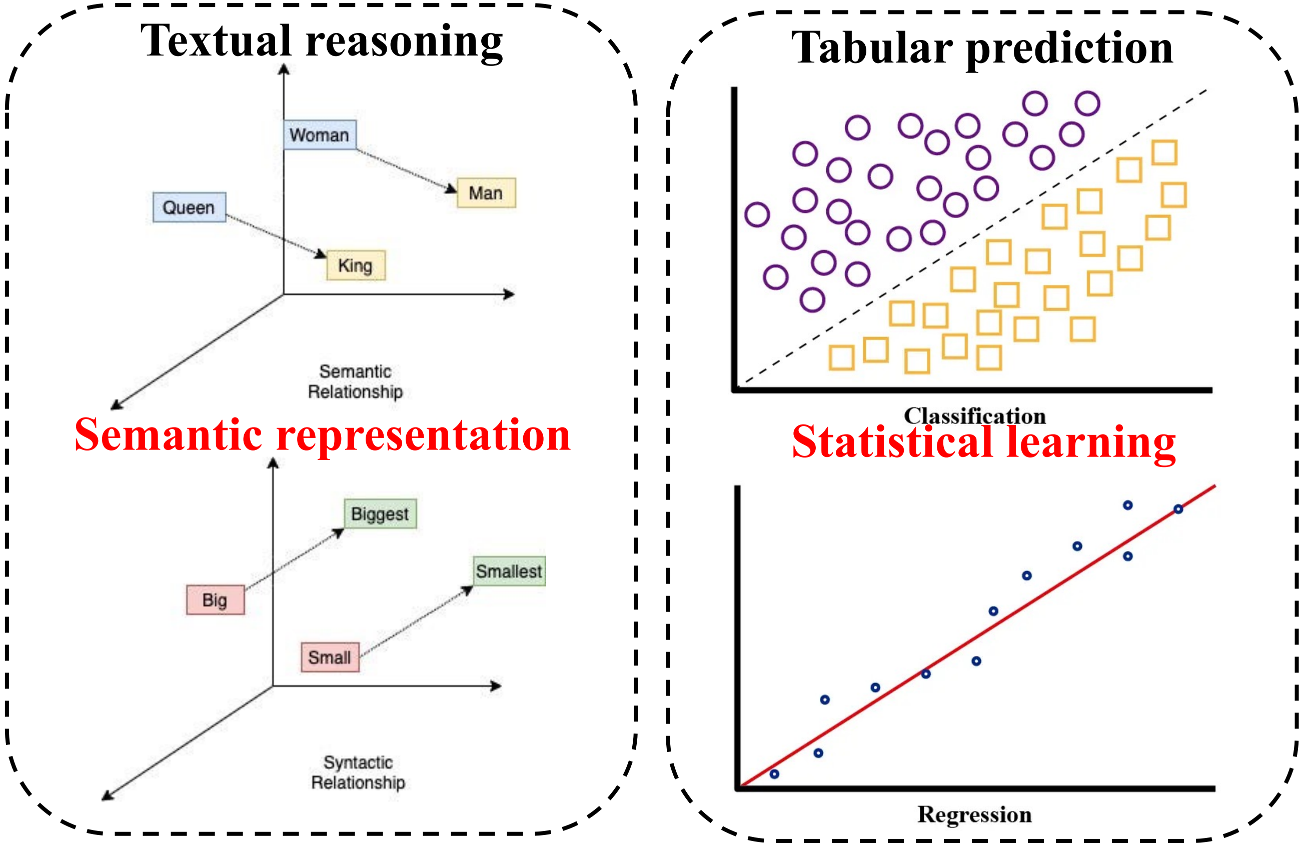
SOTA results on MedQA with meticulous prompt texts

Background

2. Classical textual prompting is ineffective to trigger LLMs' knowledge on tabular data prediction tasks



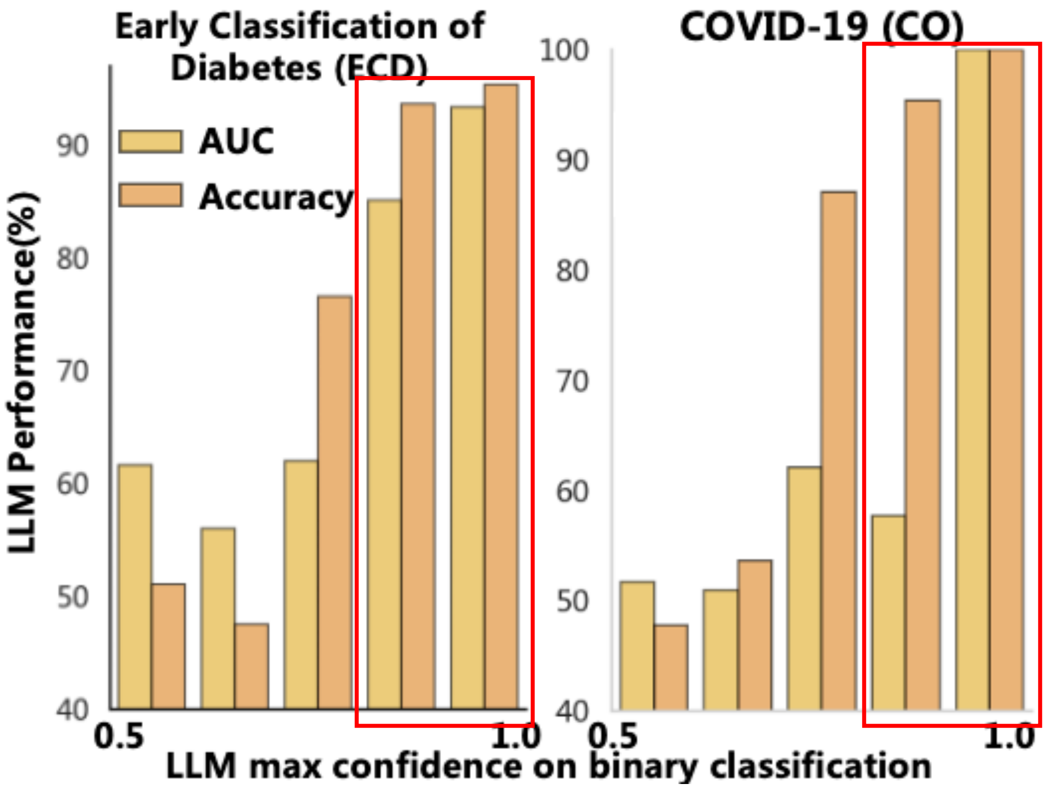
Linguistic prompting gap



Modality discrepancy

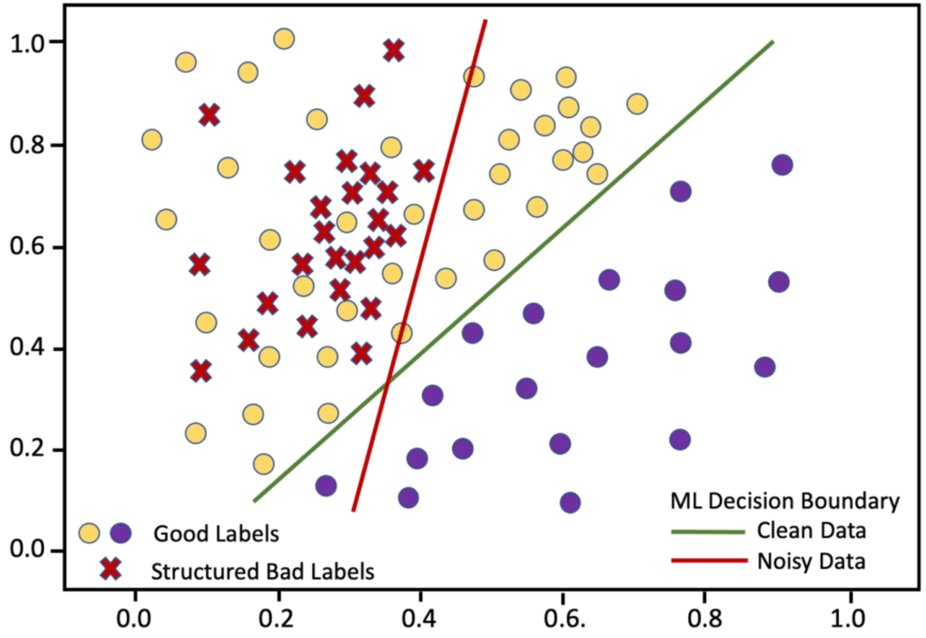
Observation & Motivation

3. Prediction on tabular samples with extreme LLM confidence is still relatively more reliable



Prediction results in high LLM confidence subsets are more reliable

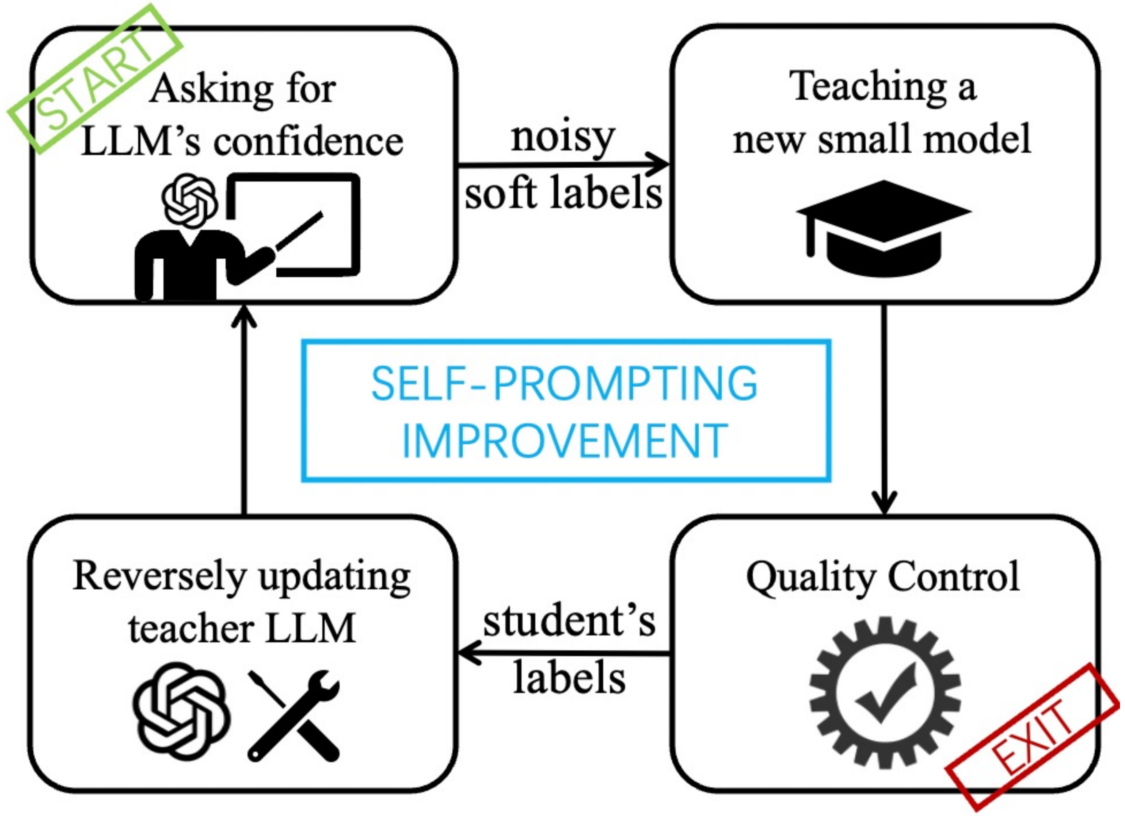
How to extract & refine LLMs' knowledge for tabular prediction?



Reformulation: A noisy label learning problem

SERSAL: Synergy learning with small models

4. SERSAL is an unsupervised non-textual self-prompting method to enhance LLM tabular prediction (considering binary classification)



#1 Soft LLM pseudo labeling.
Query LLM probability confidence for each sample using simple zero-shot prompt templates.

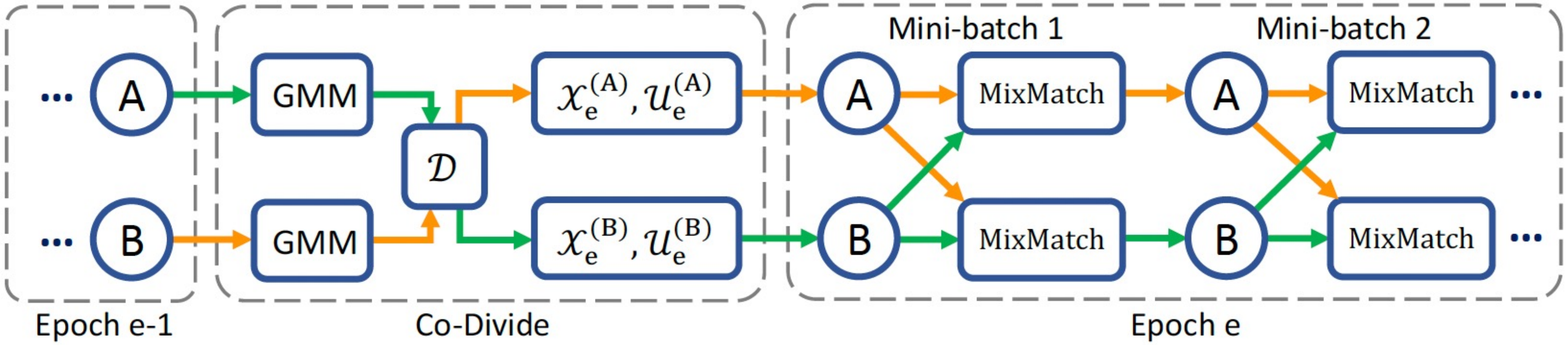
#2 Teaching a small model with LLM noisy labels.
Reformulate the extraction of LLM knowledge as learning with noisy labels (LNL) using a small tabular model.

#3 Quality control for prompting loop termination.
We use high-confidence samples for early stopping.

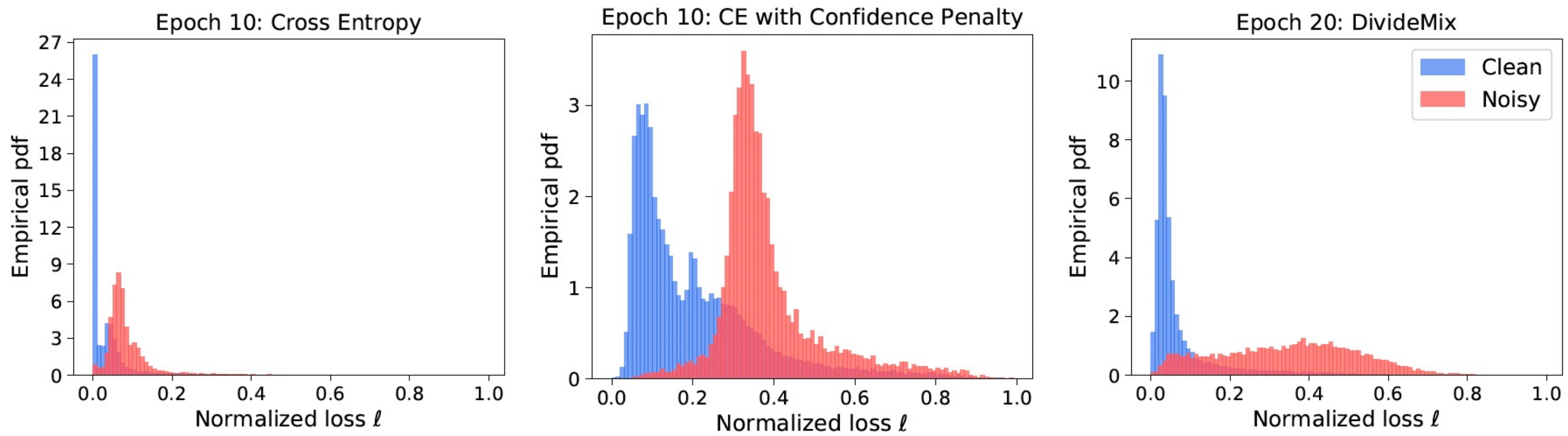
#4 Reversely teaching the LLM using the well-taught small model.
Refine LLM knowledge in a co-teaching manner.

Adapted LNL methods: DivideMix

5. We adapt classical LNL method DivideMix to teaching a small tabular model (FT-Transformer) from noisy LLM outputs



DivideMix workflow (A & B are small tabular models)



DivideMix mechanism

neural networks tend to learn simple patterns first before fitting label noise

Main experiment: medical diagnosis

	HF	LC	ECD	LI	HE	PID	FH	ST	CO	AN
Random guessing	37.22	40.18	46.25	50.28	62.73	63.24	50.39	41.76	71.55	51.28
FSSM*(supervised FT-T)	88.19	86.61	99.60	78.94	100.00	84.72	66.25	82.98	99.91	99.92
0-shot (GPT-3.5)	71.88	78.87	85.71	76.81	68.51	73.12	60.32	63.01	82.60	90.43
8-shot* (GPT-3.5)	73.65	78.87	**87.68**	76.81	68.51	73.12	58.27	60.85	77.63	87.19
CoT (GPT-3.5)	71.88	78.87	82.36	76.81	68.51	70.83	60.32	63.01	82.60	90.43
TabLLM (GPT-3.5)	76.37	78.87	87.06	78.24	74.39	75.69	61.78	68.48	85.78	89.11
LIFT (GPT-3.5)	78.23	80.69	83.92	73.60	72.57	73.12	60.32	70.92	87.93	90.43
SERSAL (GPT-3.5)	**91.39**	**85.42**	86.40	**79.39**	**85.14**	**78.97**	**63.97**	**76.36**	**96.85**	**98.37**
TabLLM+SERSAL (GPT-3.5)	93.82	85.42	88.39	80.71	89.27	82.54	65.02	81.74	97.51	98.16
SERSAL (GPT-4)	94.18	86.93	92.68	82.51	92.76	82.39	67.14	81.23	97.96	98.82

AUC score comparison of ChatGPT on 10 binary medical diagnosis datasets using different prompting schemes.

Other experiment

Ablation on key designs

	HF	LC	ECD	LI	HE	PID	FH	ST	CO	AN
SERSAL	91.39	85.42	86.40	79.39	85.14	78.97	63.97	76.36	96.85	98.37
w/o soft pseudo	84.58	76.58	87.24	78.25	75.79	75.93	62.58	75.05	93.97	97.53
w/o ES	84.03	74.11	75.92	59.39	47.41	68.43	57.08	74.70	90.57	97.57

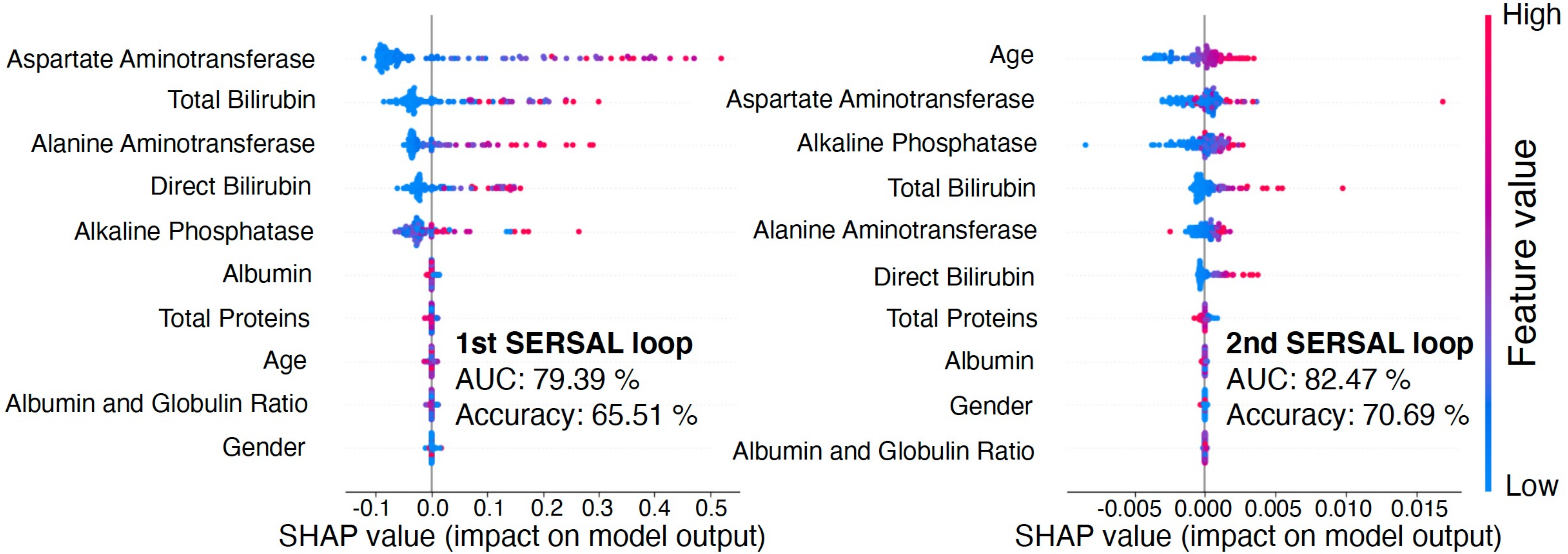
#1 Substitute soft probability LLM pseudo labels with hard ones
#2 Remove heuristic early stopping

Effectiveness of multi-loop SERSAL

# Loop	ECD		LI	
	SERSAL	LLM 0-shot	SERSAL	LLM 0-shot
1	86.40	85.71	79.39	76.81
2	87.00	86.42	82.47	80.26
3	89.00	87.81	84.07	82.91

AUC variation of ChatGPT & small tabular model as SERSAL iteration time increases

Interpretability



Interpretability from feature importance perspective
the variation of the Shapely Values (using small model outputs as targets)
on Indian Liver Patient Records dataset

Conclusion

- **Firstly bring the common challenge** of existing general-purpose LLMs on statistical-learning tabular data prediction tasks to the spotlight.

- **A novel unsupervised self-prompting method** to adapt LLM's capability to tabular data prediction by synergy learning with small models to extract correct patterns from LLM intrinsic knowledge.

- Experiments on 10 widely recognized medical diagnosis binary tabular datasets reveal the **consistent effectiveness of SERSAL** compared to common textual prompting methods.

Project repo https://github.com/jyansir/sersal
Personal homepage https://jyansir.github.io