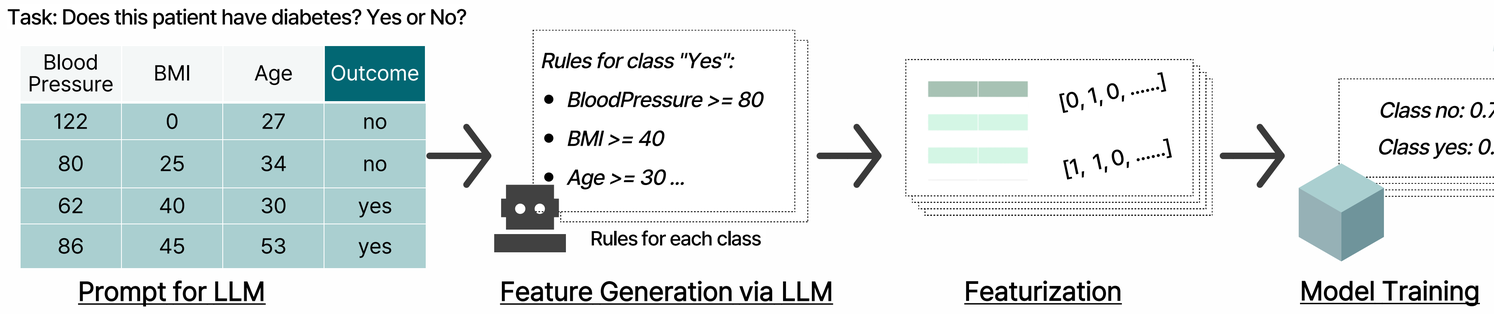
Multi-level Diagnosis and Evaluation for Robust Tabular Feature Engineering with Large Language Models

Yebin Lim, Susik Yoon

Code & Paper

Feature Engineering via Large Language Models

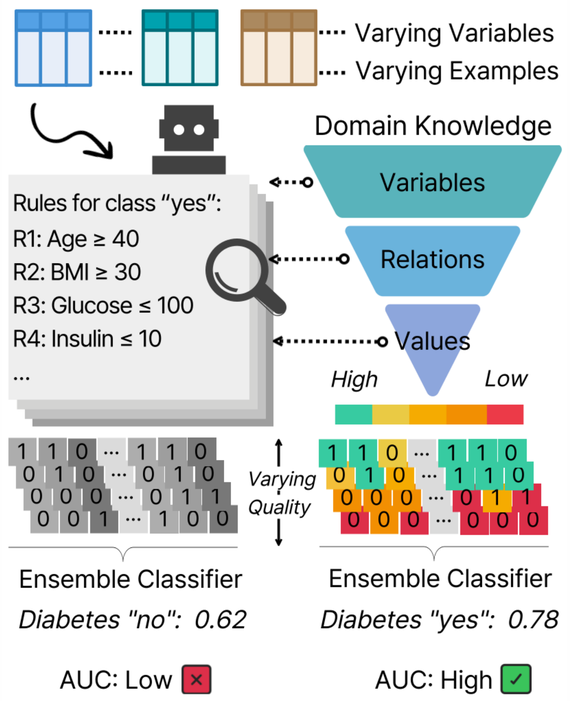
- LLMs Generate feature-wise rules related to each target class
- Outperform traditional tabular prediction methods in few-shot settings



However,

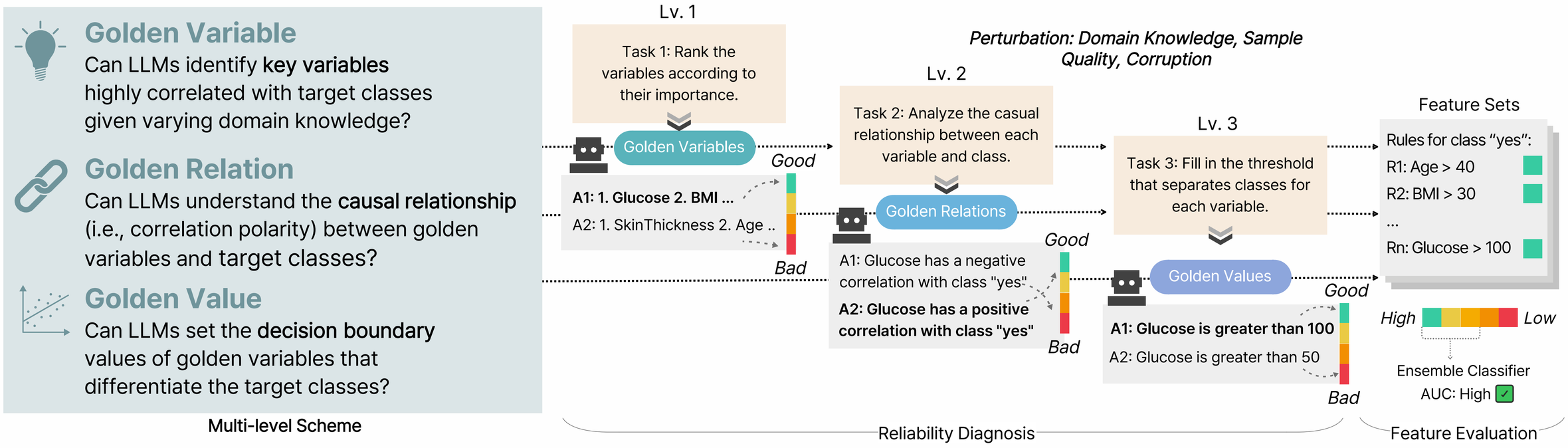
Rely on LLM with arbitrary domain knowledge and a few samples

Inconsistency and unreliability in outputs



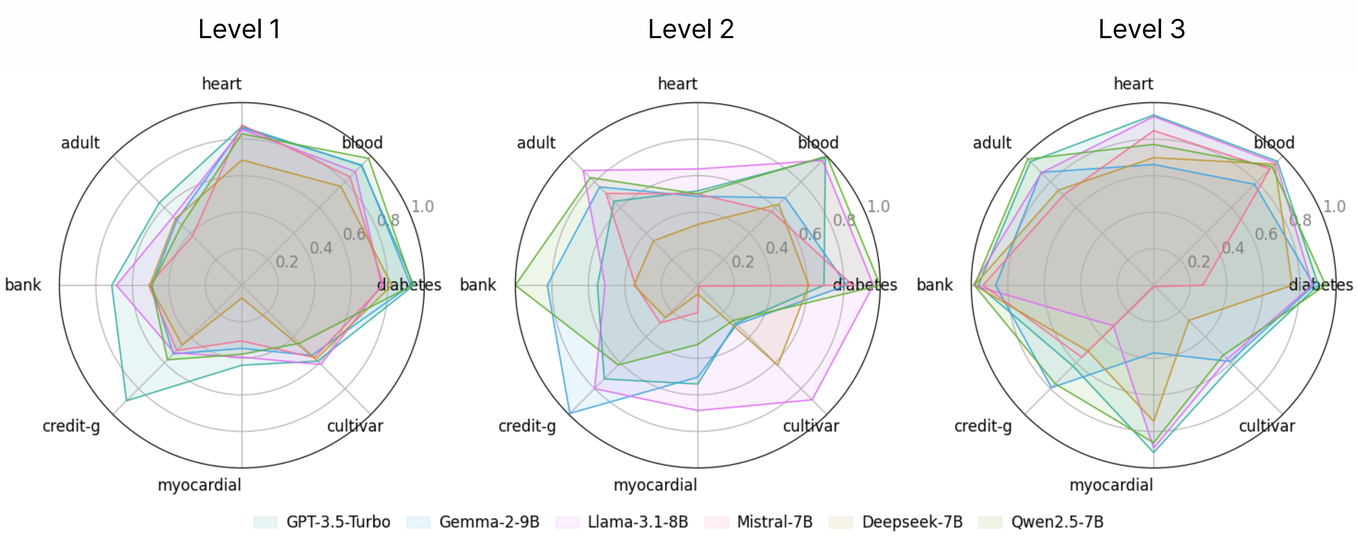
Proposed Framework for Diagnosis & Evaluation

- Drawing inspiration from real-world practices of domain experts, identify three core elements considered in feature engineering: Golden Variable, Golden Relation, Golden Value
- Based on the multi-level scheme,
  - ***Reliability Diagnosis:*** Assess the consistency in LLM respponses across varying conditions at each level
  - ***Feature Evaluation:*** Investigate how high-quality features can enhance the effectivenss of LLM-driven feature engineering

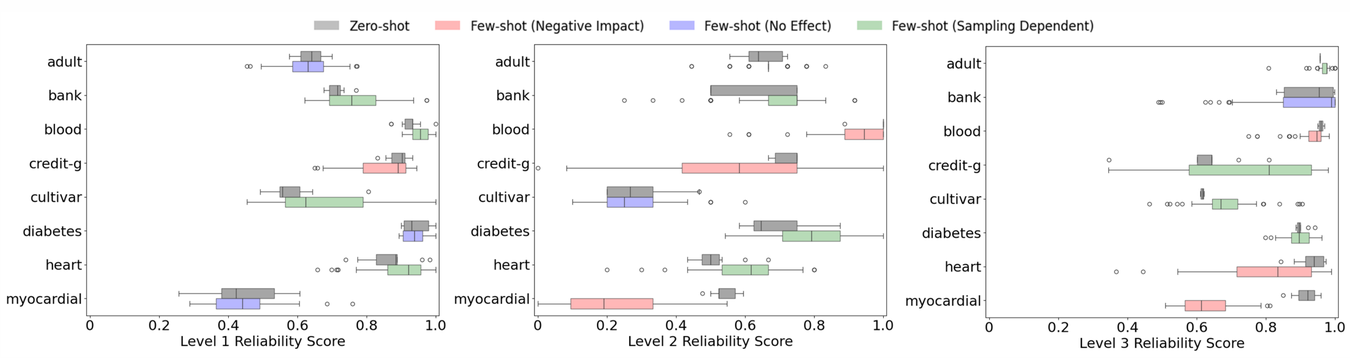


Reliability Diagnosis Results

- Variation of reliability scores of each level for different models and datasets



- Effects of varying the number of examples on reliability scores at each level for GPT-3.5-Turbo

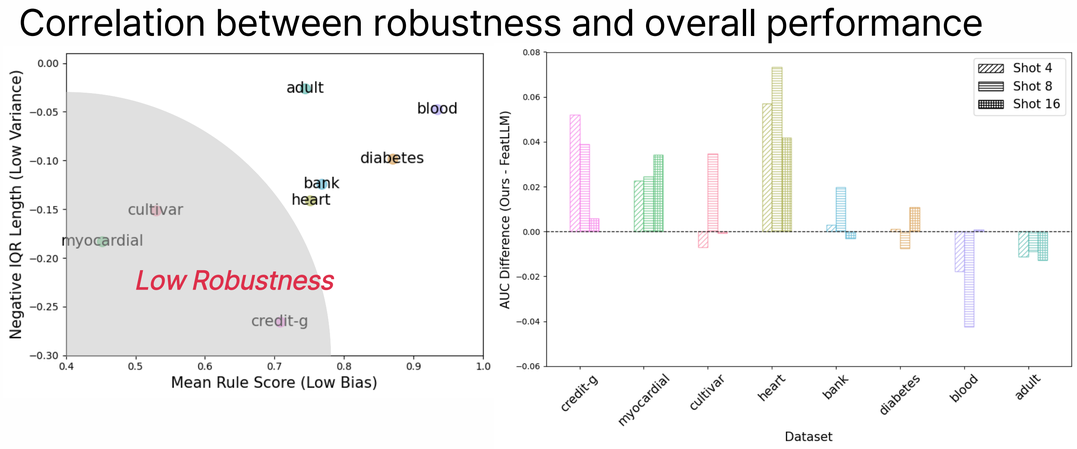


Feature Evaluation Results

- Few-shot classification performance (AUC) evaluation results

Data	Shot	LogReg	RandomForest	XGBoost	FeatLLM	Ours (Top 3)	Ours (w/o Bottom 3)	Improvement (%)
Credit-g	4	56.77 ± 11.93	51.35 ± 8.5	50.0 ± 0.0	52.27 ± 8.38	**57.77 ± 5.37**	55.65 ± 7.08	▲10.52
	8	49.7 ± 12.84	57.06 ± 8.59	49.84 ± 6.37	58.87 ± 4.69	**62.89 ± 5.72**	61.49 ± 7.65	▲6.83
	16	**64.48 ± 9.71**	64.27 ± 11.32	59.49 ± 10.36	56.47 ± 4.51	57.51 ± 1.87	58.74 ± 8.04	▲4.02
Myocardial	4	54.28 ± 5.09	**57.93 ± 2.64**	50.0 ± 0.0	54.08 ± 3.28	56.35 ± 12.34	55.46 ± 4.77	▲4.20
	8	54.25 ± 8.33	52.78 ± 2.67	**55.44 ± 5.34**	51.6 ± 7.06	54.04 ± 6.16	52.26 ± 7.69	▲4.73
	16	56.39 ± 5.57	50.96 ± 5.98	55.21 ± 5.96	58.54 ± 1.84	**61.96 ± 3.64**	60.92 ± 2.09	▲5.84
Cultivar	4	41.93 ± 9.19	44.14 ± 4.23	50.0 ± 0.0	**55.84 ± 4.99**	55.14 ± 6.45	55.63 ± 8.79	▼0.38
	8	48.67 ± 7.27	49.2 ± 4.68	48.44 ± 1.56	56.95 ± 3.52	**60.43 ± 6.79**	57.45 ± 5.37	▲6.11
	16	53.86 ± 8.89	50.28 ± 5.77	57.08 ± 5.59	57.57 ± 2.67	57.49 ± 3.22	**58.3 ± 2.46**	▲1.27
Bank	4	67.65 ± 16.53	64.28 ± 5.0	50.0 ± 0.0	74.34 ± 1.71	75.17 ± 1.6	**76.07 ± 2.87**	▲2.33
	8	75.05 ± 1.57	63.36 ± 7.13	58.52 ± 10.73	76.09 ± 2.57	77.87 ± 0.38	**78.03 ± 1.66**	▲2.55
	16	77.6 ± 2.18	77.69 ± 2.51	68.75 ± 10.87	79.57 ± 1.01	**79.59 ± 2.72**	79.5 ± 2.96	▲0.03
Heart	4	52.19 ± 1.59	**79.92 ± 7.71**	50.0 ± 0.0	73.82 ± 6.06	77.69 ± 2.7	77.18 ± 3.53	▲5.24
	8	60.86 ± 8.74	**81.84 ± 2.88**	53.76 ± 11.81	70.88 ± 13.15	76.9 ± 7.8	70.99 ± 10.31	▲8.49
	16	65.45 ± 13.36	**85.5 ± 2.39**	82.99 ± 1.69	80.31 ± 7.69	83.57 ± 9.29	81.08 ± 5.33	▲4.06
Diabetes	4	47.04 ± 12.37	56.67 ± 11.65	50.0 ± 0.0	79.55 ± 0.35	79.65 ± 0.97	**79.74 ± 0.5**	▲0.24
	8	52.73 ± 5.8	64.19 ± 6.21	39.2 ± 14.42	**80.48 ± 0.21**	79.71 ± 0.24	80.41 ± 0.76	▼0.09
	16	64.78 ± 14.34	67.3 ± 6.02	72.69 ± 2.33	79.85 ± 0.83	**80.94 ± 2.11**	80.25 ± 1.52	▲1.37
Blood	4	42.75 ± 16.56	48.66 ± 12.56	50.0 ± 0.0	**56.34 ± 6.66**	54.57 ± 10.59	55.89 ± 6.51	▼0.80
	8	60.27 ± 8.9	57.67 ± 8.98	55.87 ± 5.1	66.63 ± 0.69	62.28 ± 7.24	**66.71 ± 0.84**	▲0.12
	16	**68.59 ± 3.81**	51.9 ± 8.84	63.43 ± 8.09	67.61 ± 1.9	67.98 ± 0.31	67.08 ± 1.71	▲0.55
Adult	4	58.3 ± 7.89	70.28 ± 5.32	50.0 ± 0.0	**87.58 ± 0.29**	86.48 ± 1.21	87.55 ± 0.83	▼0.03
	8	58.97 ± 8.93	57.27 ± 21.03	59.19 ± 7.96	**87.29 ± 0.31**	86.35 ± 0.3	86.95 ± 0.15	▼0.39
	16	67.61 ± 10.76	77.93 ± 2.79	68.17 ± 9.31	87.59 ± 0.9	85.53 ± 1.74	**87.61 ± 0.97**	▲0.02

- Correlation between robustness and overall performance



More questions or relevant work? Contact Yebin Lim via yebinuni@korea.ac.kr