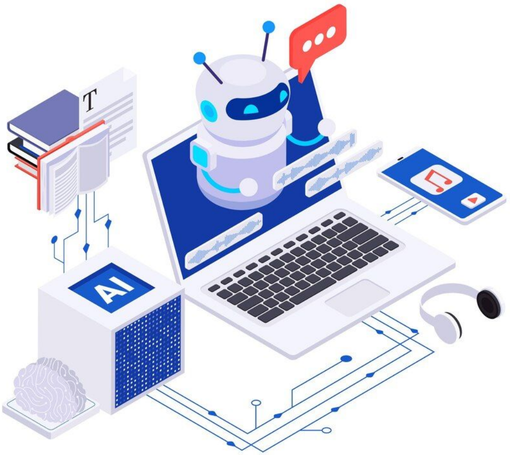
Enhancing Operational Data Synthesis and Predictive Analysis in HPC Clusters Using Large Language Models



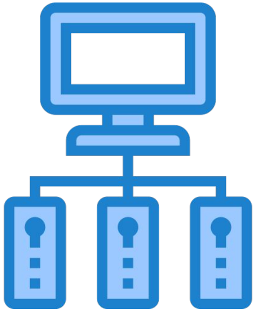
Source: freepik.com

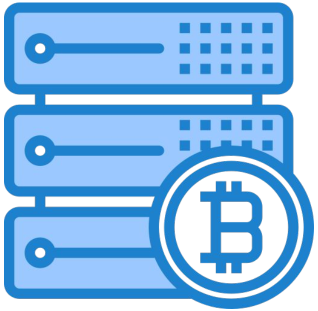
Yizhen Zang

Daily Supervisor: Xiaoyu Chu

July, 2024

HPC Clusters are Vital to Digital Society





Source: flaticon.com

HPC Clusters are Vital to Digital Society

Challenges

- ➔ Resource management
- ➔ Frequent job failures
- ➔ Lack of operational data

Source: flaticon.com

HPC Clusters are Vital to Digital Society

Impact of HPC Outages

- **Prevalence**: **80%** of users
- **Cost**: **60%** resulted in at least **$100,000** losses

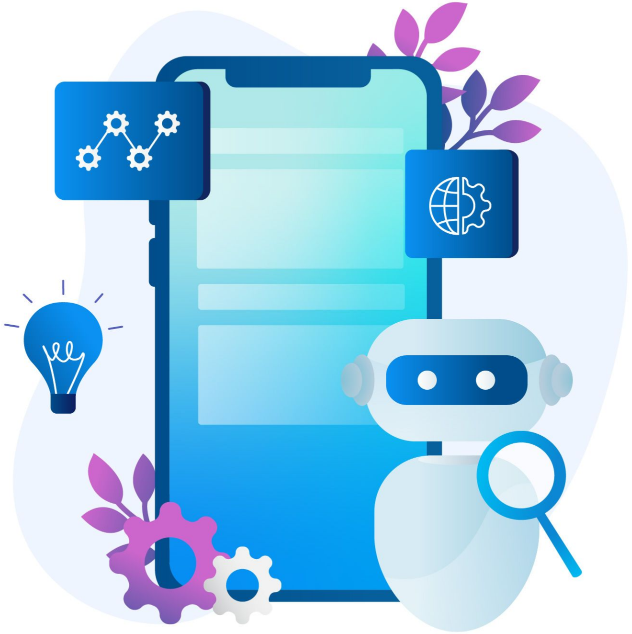
Source: Uptime Institute, Annual outages analysis 2023

Source: freepik.com

Large Language Models are Powerful

Traditional Models

- ➔ Limited Adaptability
- ➔ Simplistic Approaches
- ➔ Inflexibility

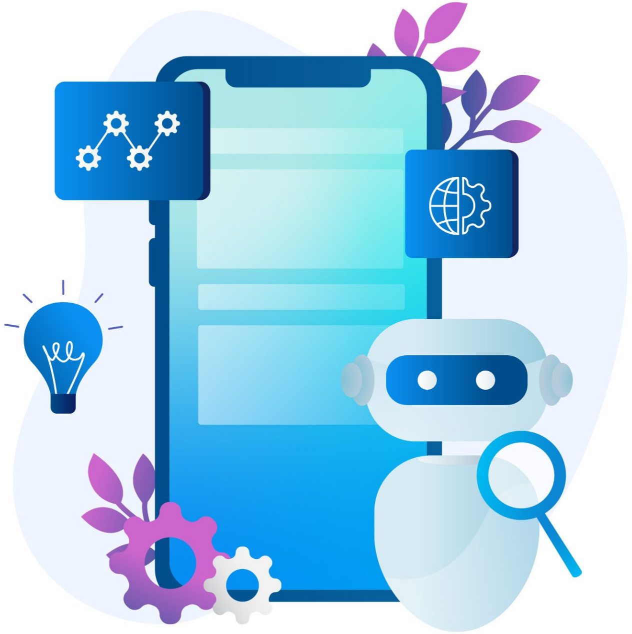


Source: freepik.com

Large Language Models are Powerful

Traditional Models
- ➔ Limited Adaptability
- ➔ Simplistic Approaches
- ➔ Inflexibility

LLMs
- ★ High Adaptability
- ★ Complex Data Handling
- ★ Flexibility and Integration



Source: freepik.com

Research Questions

The overarching RQ: How to leverage LLMs to improve operational decision-making for HPC datacenters?

Research Questions

RQ1: What are the existing applications of LLMs for HPC?

RQ1.1: How are LLMs applied in HPC for **workload synthesis**?
RQ1.2: How are LLMs applied in HPC for **workload prediction**?
RQ1.3: How are LLMs applied in **other tasks** in HPC?

Research Questions

RQ1: What are the existing applications of LLMs for HPC?

RQ2: How to design and evaluate LLM-based models for synthesizing job data in HPC clusters?

RQ2.1: How to generate synthetic job data in HPC clusters using LLMs?

RQ2.2: How are the techniques developed in RQ2.1 performing, relative to the non-LLM state-of-the-art?

Research Questions

RQ1: What are the existing applications of LLMs for HPC?

RQ2: How to design and evaluate LLM-based models for synthesizing job data in HPC clusters?

RQ3: How to design and evaluate LLM-based models for predicting job end-state in HPC clusters?

> **RQ3.1**: How to predict job failures in HPC clusters using LLMs?
> **RQ3.2**: How are the techniques developed in **RQ3.1** performing, relative to the non-LLM state-of-the-art?

Literature Review of LLMs Applications in HPC

RQ1

Reference	Applications	Models	Datasets
Shi et al.	memory workload synthesis	REaLTabFormer	SPEC2017
LM4HPC	code similarity analysis parallelism detection OpenMP Q&A	CodeBERT GraphCodeBERT gpt-3.5-turbo	POJ-104 DRB-ML OMP4Par OMPQA
MPIrigen	MPI-based parallel program	MonoCoder PolyCoder GPT-3.5	HPCorpusMPI
HPC-Coder	code completion OpenMP labeling performance prediction	GPT-2 GPT-Neo PolyCoder	self-collected
HPC-GPT	AI model management dataset management data race detection	LLaMa LLaMa 2	self-collected
LLMDB	query rewrite database diagnosis data analytics	LLMDB	self-collected

Literature Review of LLMs Applications in HPC

RQ1

Reference	Applications	Models	Datasets
Shi et al.	memory workload synthesis	REaLTabFormer	SPEC2017
LM4HPC	code similarity analysis parallelism detection OpenMP Q&A	CodeBERT GraphCodeBERT gpt-3.5-turbo	POJ-104 DRB-ML OMP4Par OMPQA
MPIrigen	MPI-based parallel program	MonoCoder PolyCoder GPT-3.5	HPCorpusMPI
HPC-Coder	code completion OpenMP labeling performance prediction	GPT-2 GPT-Neo PolyCoder	self-collected
HPC-GPT	AI model management dataset management data race detection	LLaMa LLaMa 2	self-collected
LLMDB	query rewrite database diagnosis data analytics	LLMDB	self-collected

Workload synthesis

Literature Review of LLMs Applications in HPC

RQ1

Reference	Applications	Models	Datasets
Shi et al.	memory workload synthesis	REaLTabFormer	SPEC2017
LM4HPC	code similarity analysis parallelism detection OpenMP Q&A	CodeBERT GraphCodeBERT gpt-3.5-turbo	POJ-104 DRB-ML OMP4Par OMPQA
MPIrigen	MPI-based parallel program	MonoCoder PolyCoder GPT-3.5	HPCorpusMPI
HPC-Coder	code completion OpenMP labeling performance prediction	GPT-2 GPT-Neo PolyCoder	self-collected
HPC-GPT	AI model management dataset management data race detection	LLaMa LLaMa 2	self-collected
LLMDB	query rewrite database diagnosis data analytics	LLMDB	self-collected

Workload synthesis

Code-based tasks

Literature Review of LLMs Applications in HPC

RQ1

Reference	Applications	Models	Datasets
Shi et al.	memory workload synthesis	REaLTabFormer	SPEC2017
LM4HPC	code similarity analysis parallelism detection OpenMP Q&A	CodeBERT GraphCodeBERT gpt-3.5-turbo	POJ-104 DRB-ML OMP4Par OMPQA
MPIrigen	MPI-based parallel program	MonoCoder PolyCoder GPT-3.5	HPCorpusMPI
HPC-Coder	code completion OpenMP labeling performance prediction	GPT-2 GPT-Neo PolyCoder	self-collected
HPC-GPT	AI model management dataset management data race detection	LLaMa LLaMa 2	self-collected
LLMDB	query rewrite database diagnosis data analytics	LLMDB	self-collected

Workload synthesis

Code-based tasks

Data management

Literature Review of LLMs Applications in HPC

RQ1

Reference	Applications	Models	Datasets
Shi et al.	memory workload synthesis	REaLTabFormer	SPEC2017
LM4HPC	code similarity analysis parallelism detection OpenMP Q&A	CodeBERT GraphCodeBERT gpt-3.5-turbo	POJ-104 DRB-ML OMP4Par OMPQA
MP			
HP	performance prediction	PolyCoder	
HPC-GPT	AI model management dataset management data race detection	LLaMa LLaMa 2	self-collected
LLMDB	query rewrite database diagnosis data analytics	LLMDB	self-collected

Workload synthesis

Data management

Gaps remain in workload synthesis and prediction.

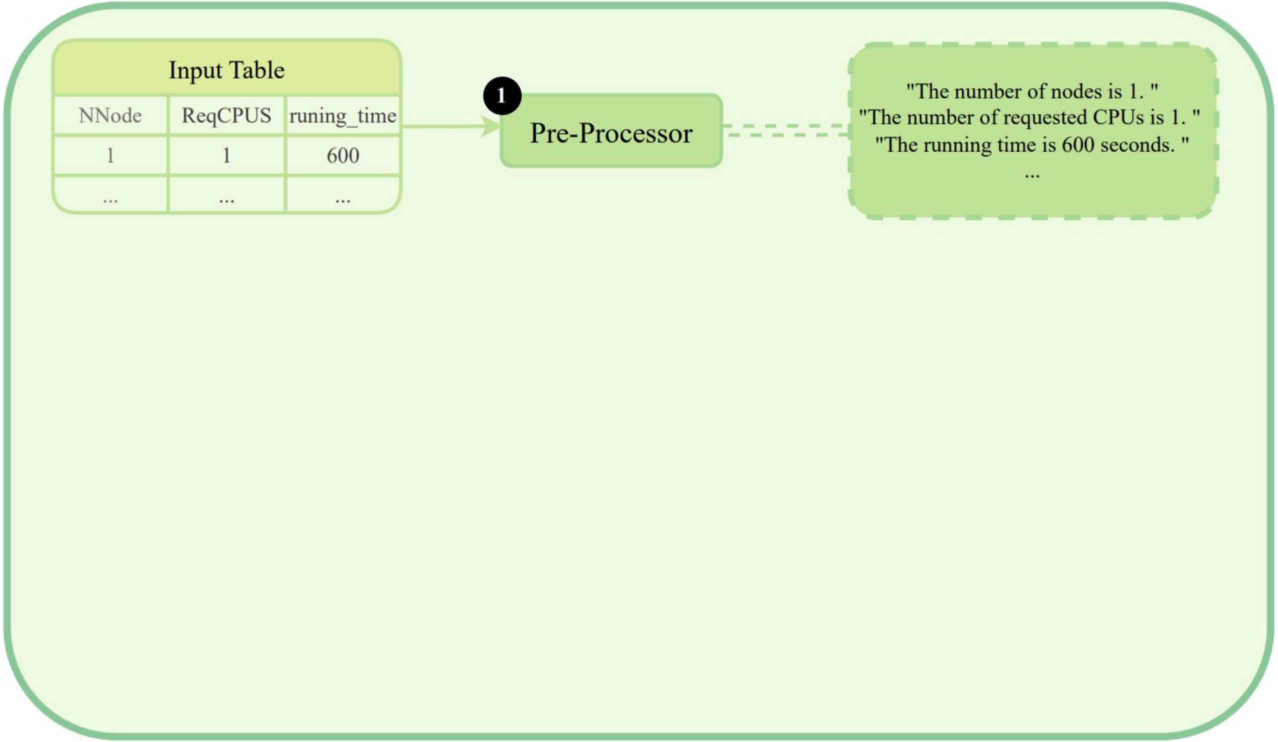
Design - How to Use LLMs for Job Data Synthesis?

RQ2.1

Input Table		
NNode	ReqCPUS	runing_time
1	1	600
...	...	...

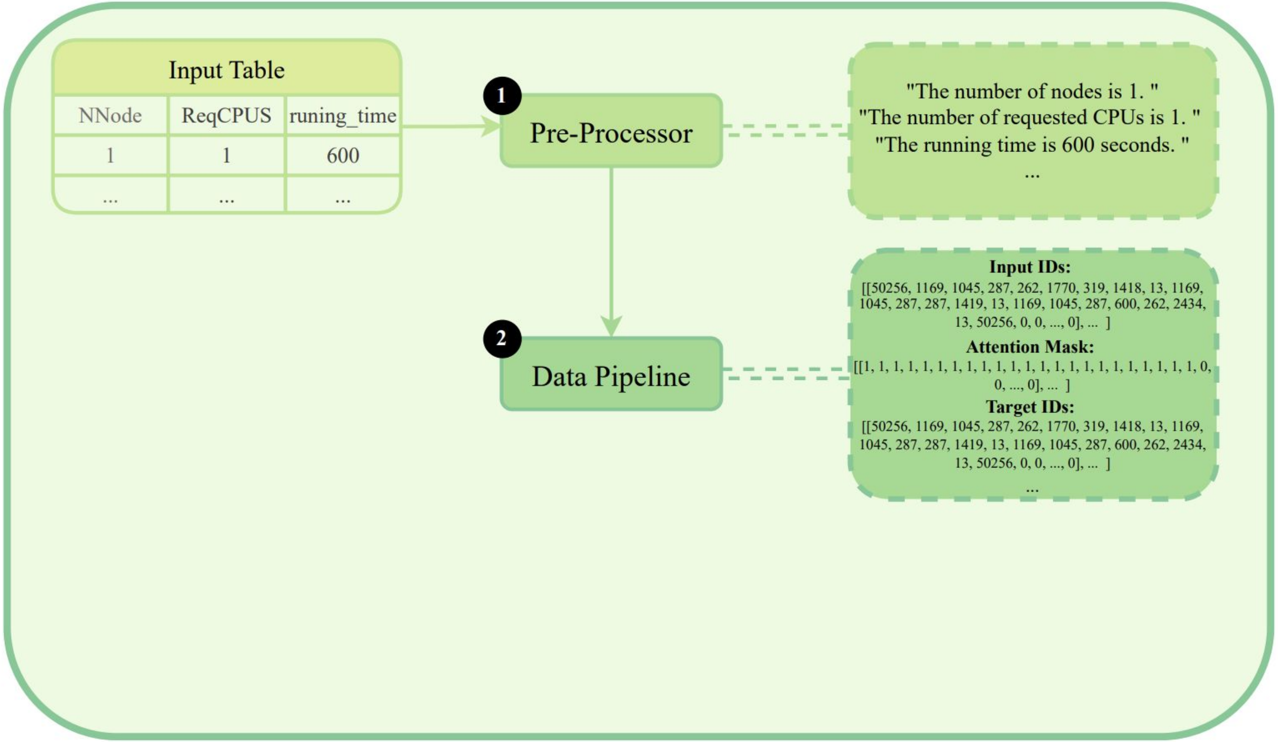
Design - How to Use LLMs for Job Data Synthesis?

RQ2.1



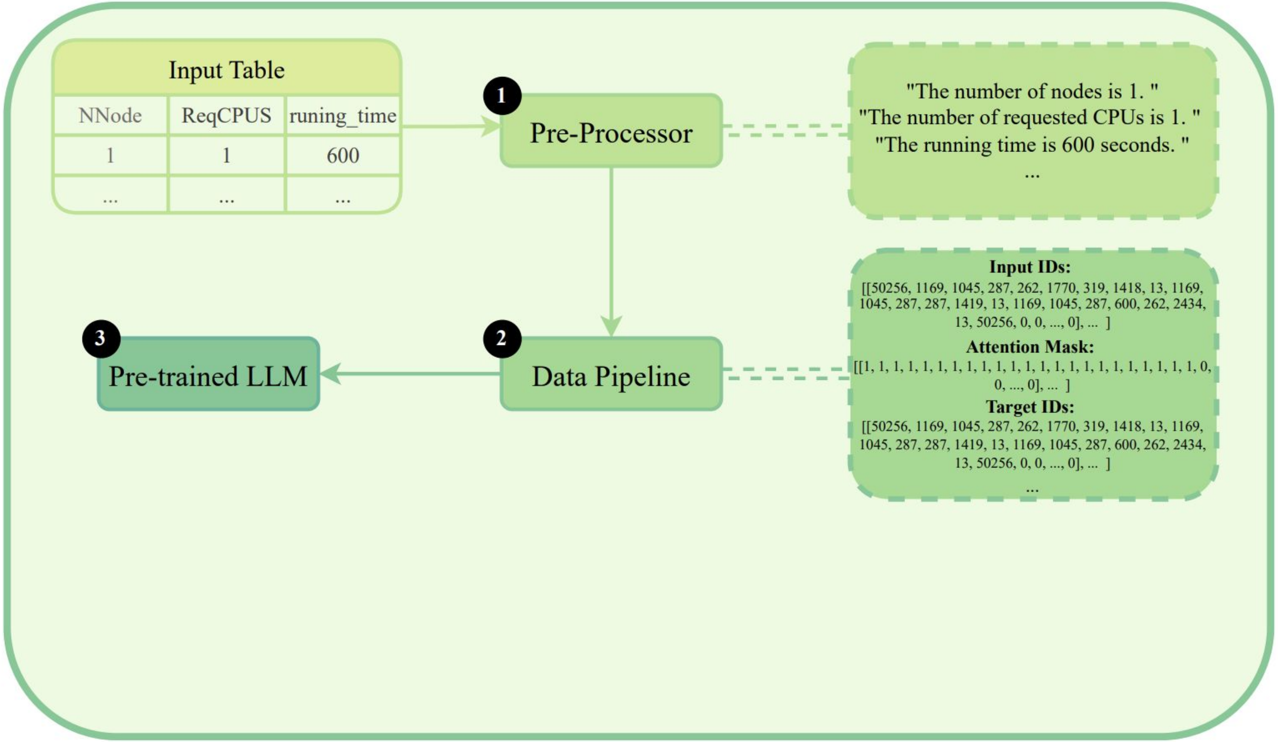
Design - How to Use LLMs for Job Data Synthesis?

RQ2.1



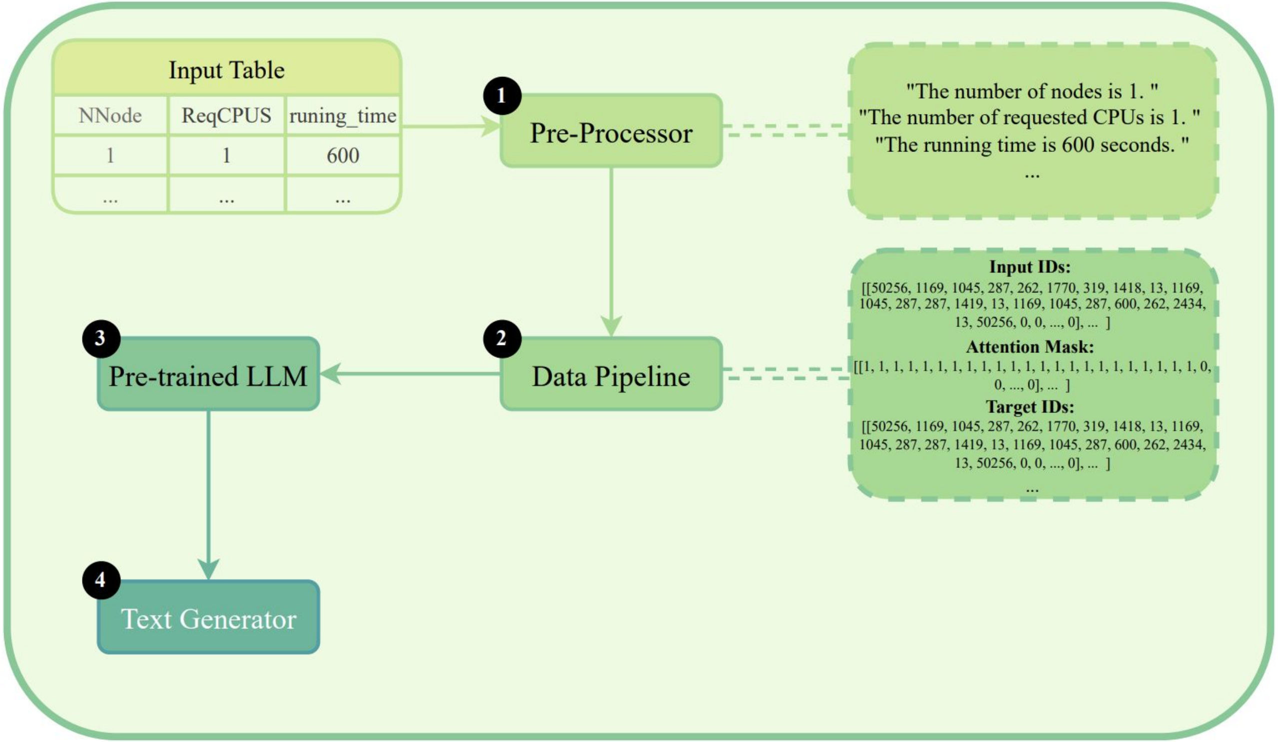
Design - How to Use LLMs for Job Data Synthesis?

RQ2.1



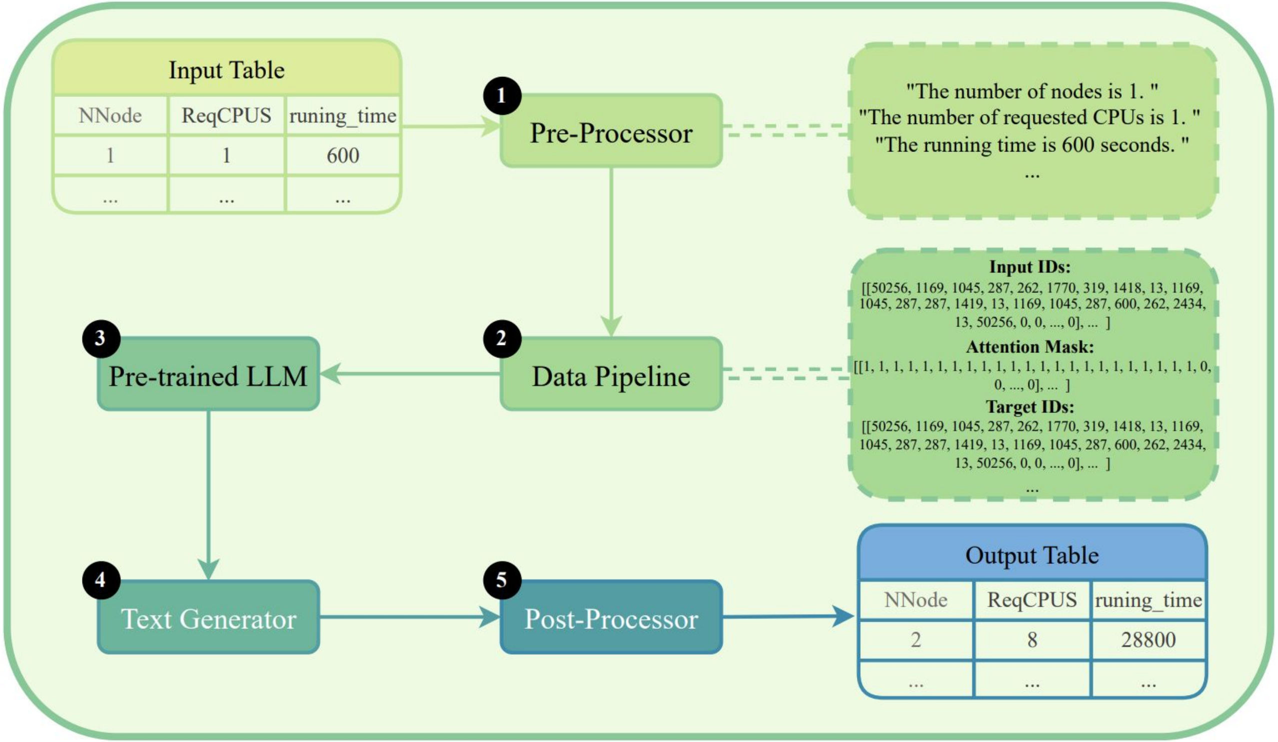
Design - How to Use LLMs for Job Data Synthesis?

RQ2.1



Design - How to Use LLMs for Job Data Synthesis?

RQ2.1



Evaluation (1/3) - How to Evaluate Synthetic Data?

RQ2.2

- **Baseline Models**: TabGAN & CTGAN

Evaluation (1/3) - How to Evaluate Synthetic Data?

RQ2.2

- **Baseline Models**: TabGAN & CTGAN

- **Distance to the Closest Record (DCR)**
  - Measures the **similarity**

Evaluation (1/3) - How to Evaluate Synthetic Data?

RQ2.2

- **Baseline Models**: TabGAN & CTGAN
- **Distance to the Closest Record (DCR)**
  - Measures the **similarity**
- **Machine Learning Efficiency (MLE)**
  - Determine the **utility** on downstream predictive tasks
    - Decision Tree (DT)
    - Random Forest (RF)

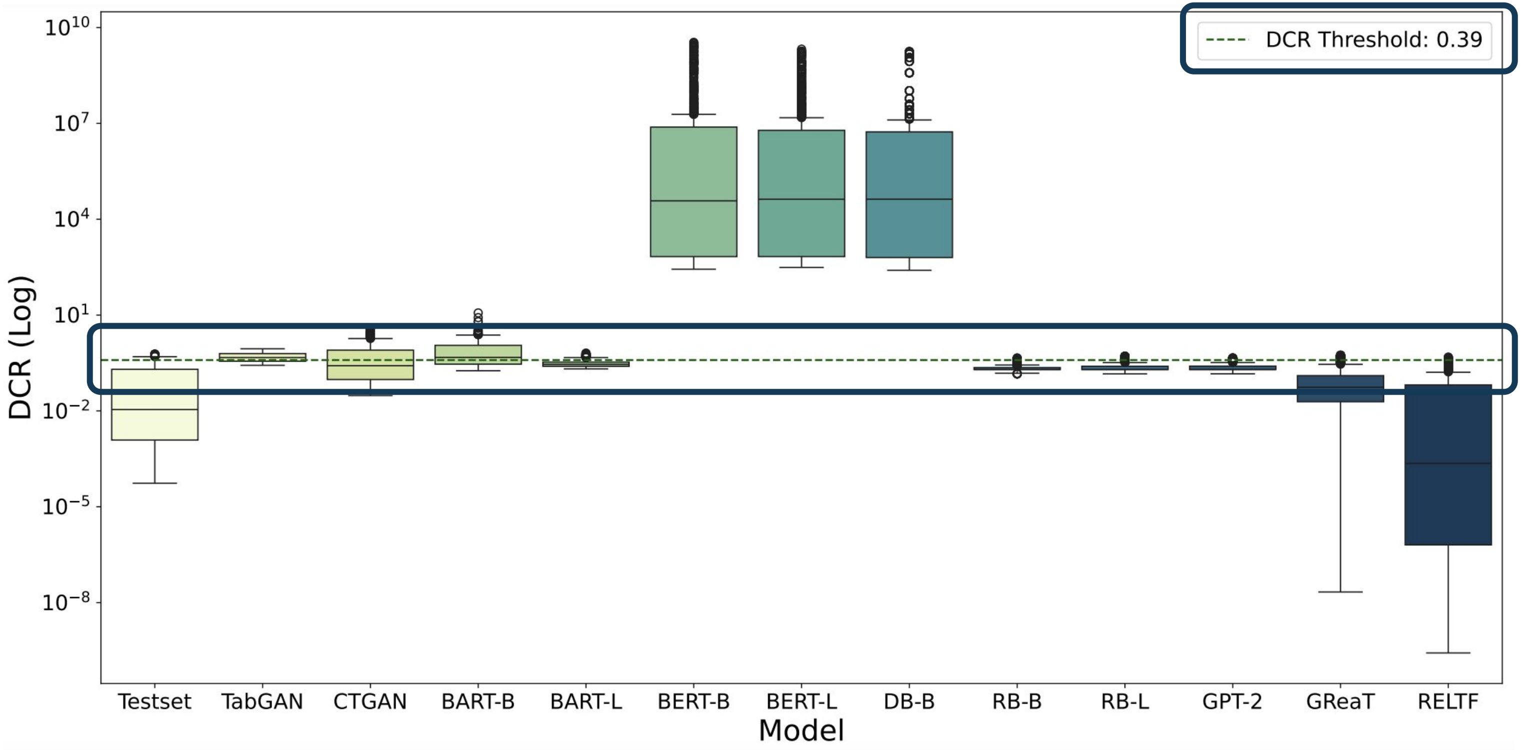
Evaluation (1/3) - How to Evaluate Synthetic Data?

RQ2.2

- **Baseline Models**: TabGAN & CTGAN
- **Distance to the Closest Record (DCR)**
  - Measures the **similarity**
- **Machine Learning Efficiency (MLE)**
  - Determine the **utility** on downstream predictive tasks
    - Decision Tree (DT)
    - Random Forest (RF)
  - Accuracy, precision, recall, F1-score, AUC-ROC

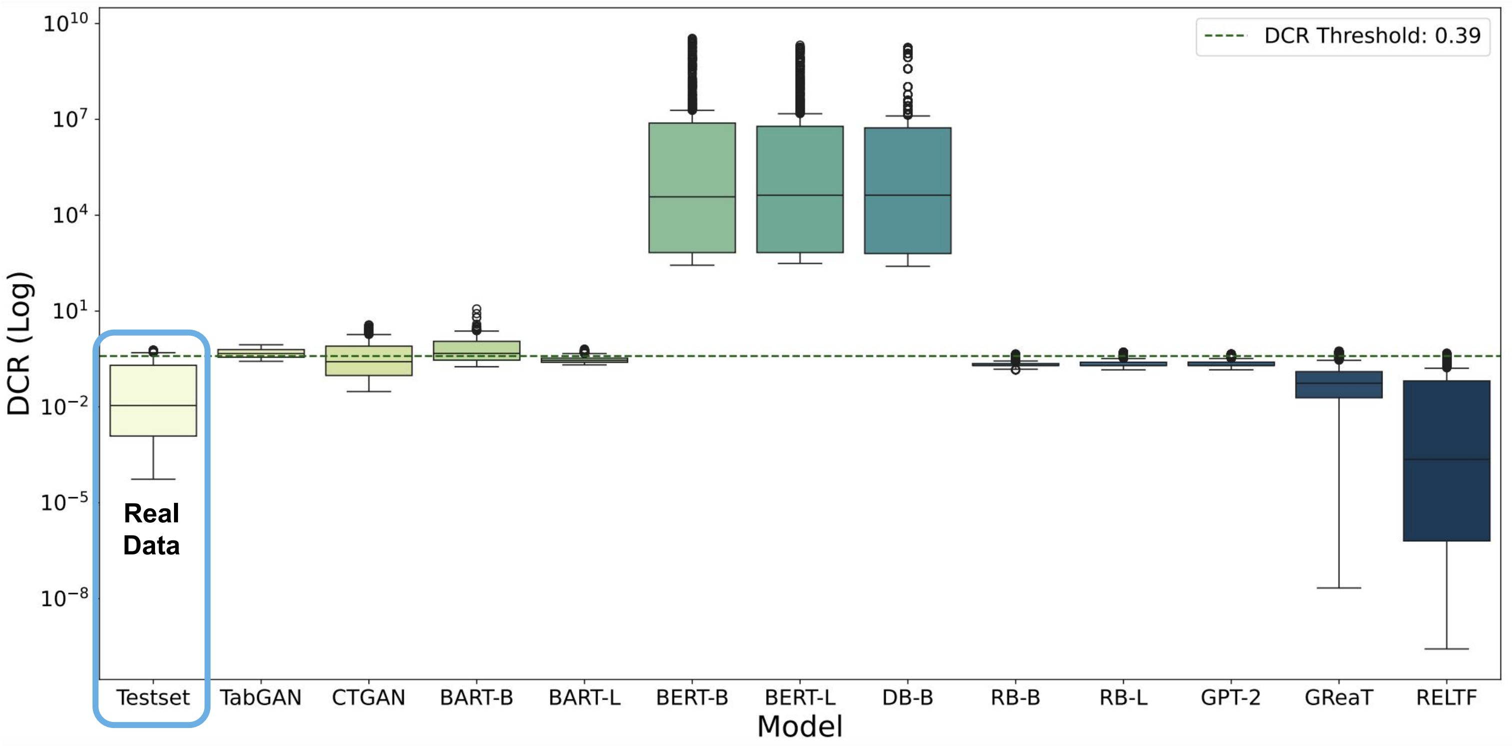
Evaluation (2/3) - Distance to the Closest Record

RQ2.2



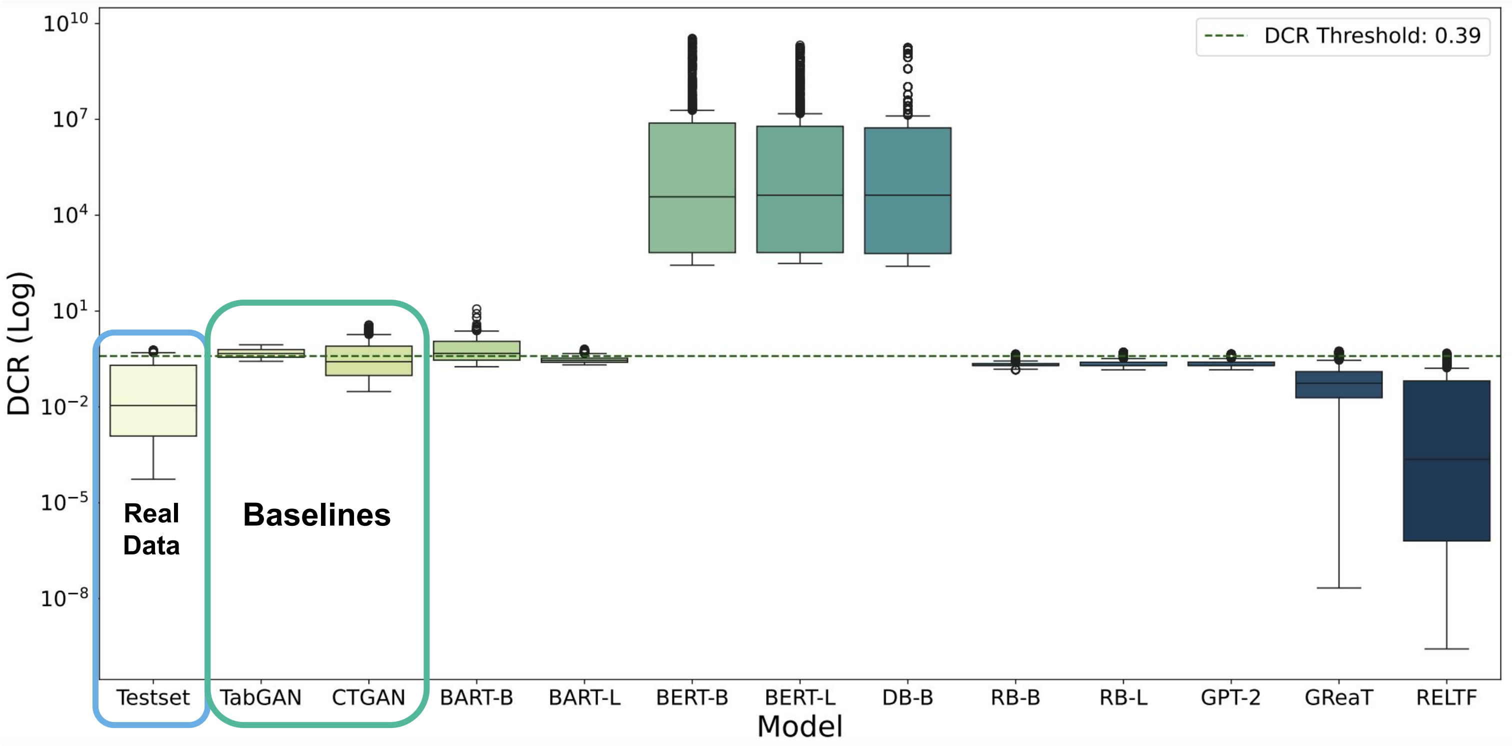
Evaluation (2/3) - Distance to the Closest Record

RQ2.2



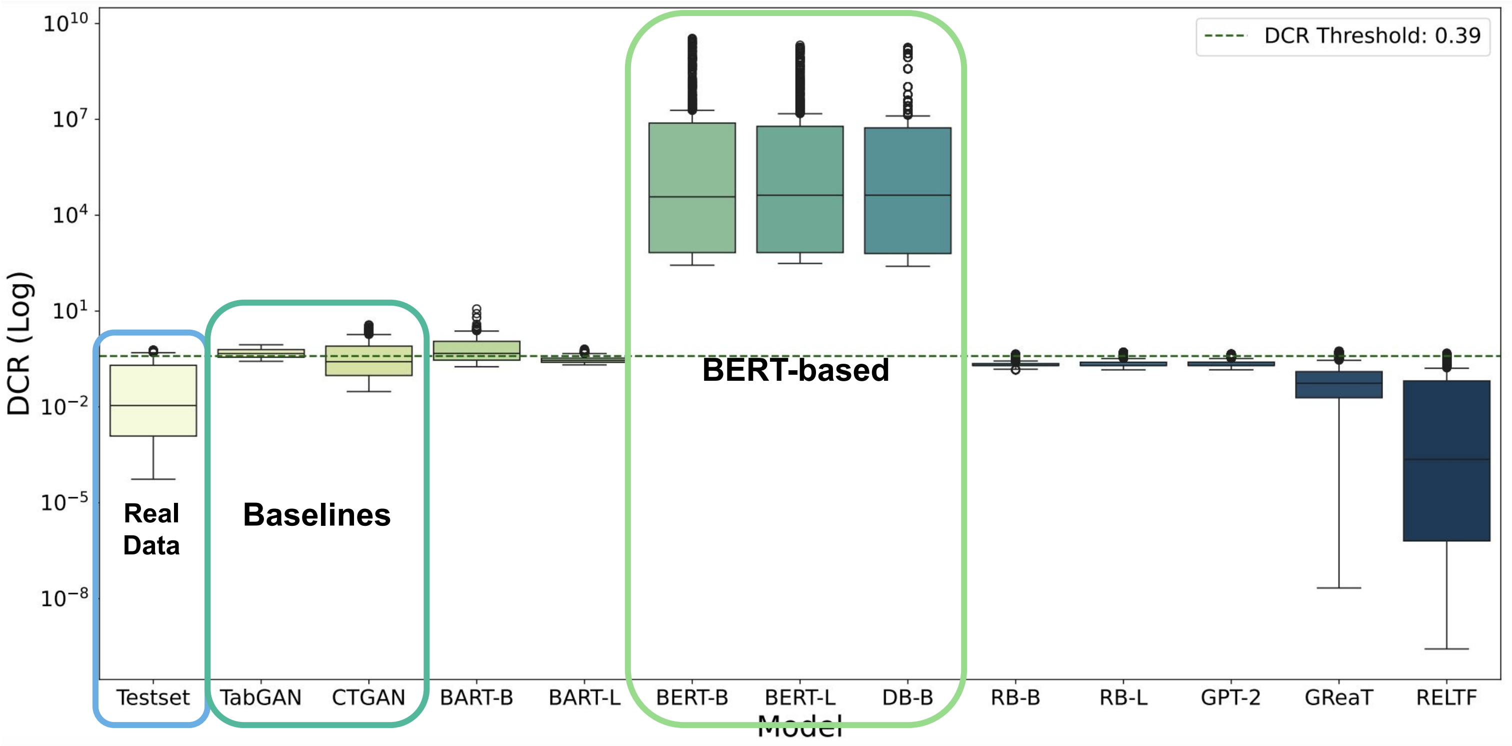
Evaluation (2/3) - Distance to the Closest Record

RQ2.2



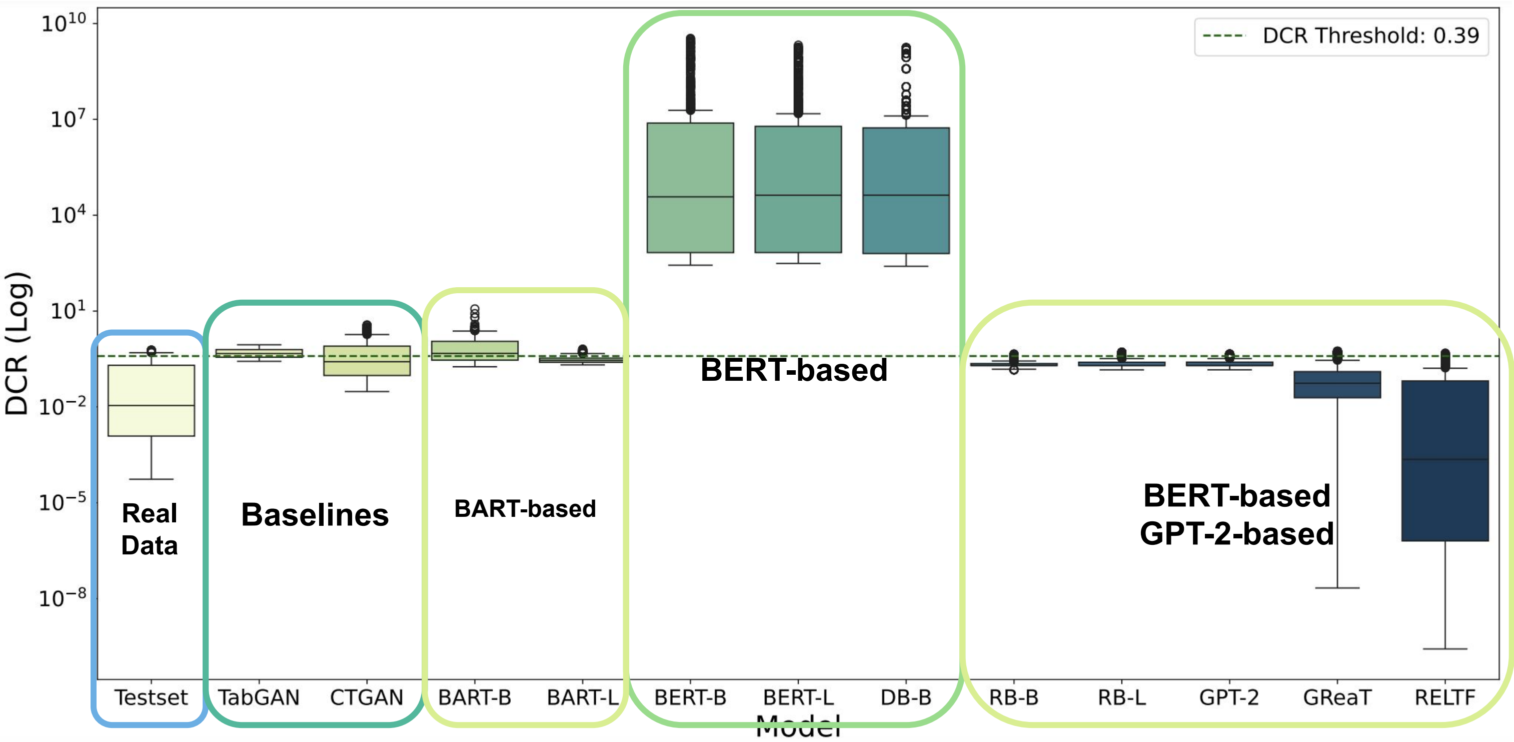
Evaluation (2/3) - Distance to the Closest Record

RQ2.2



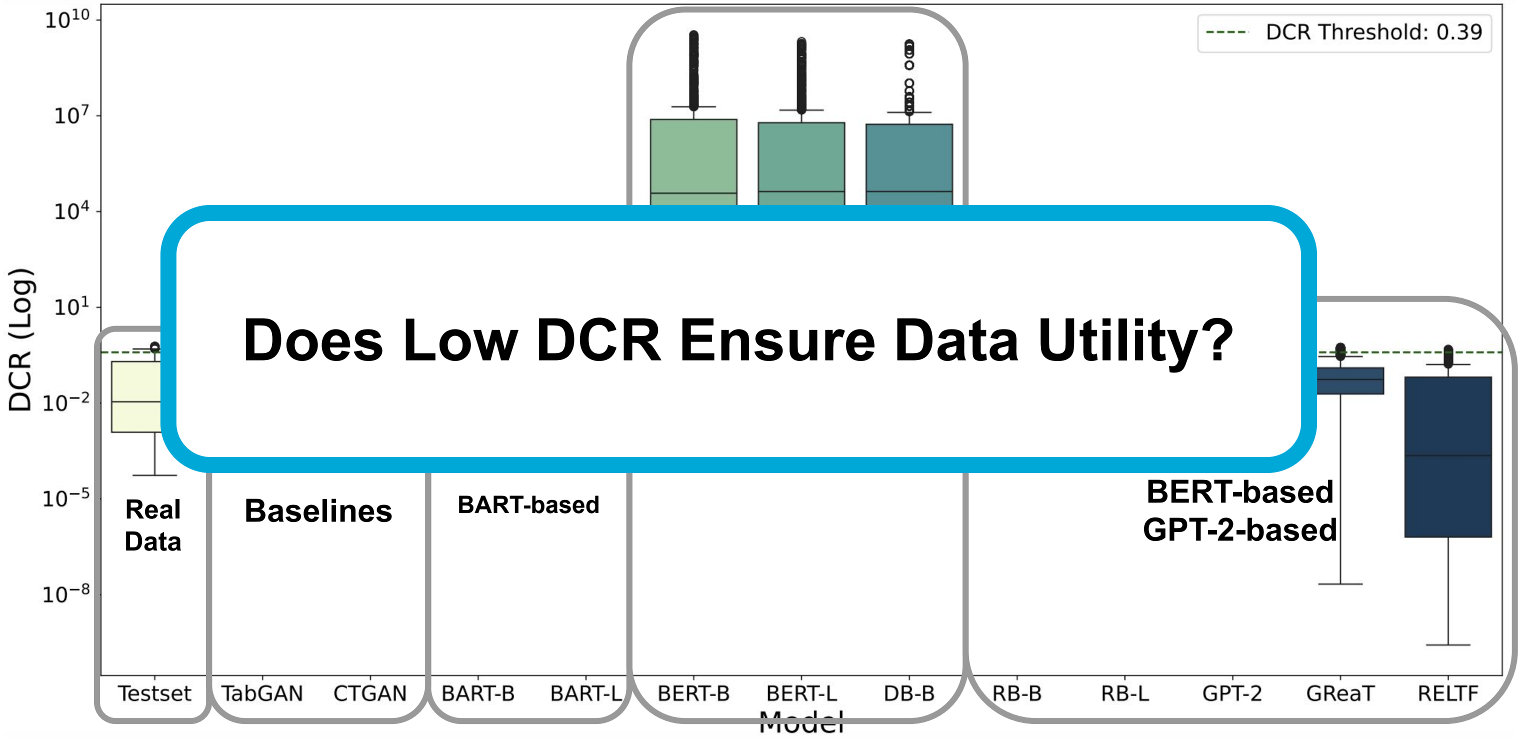
Evaluation (2/3) - Distance to the Closest Record

RQ2.2



Evaluation (2/3) - Distance to the Closest Record

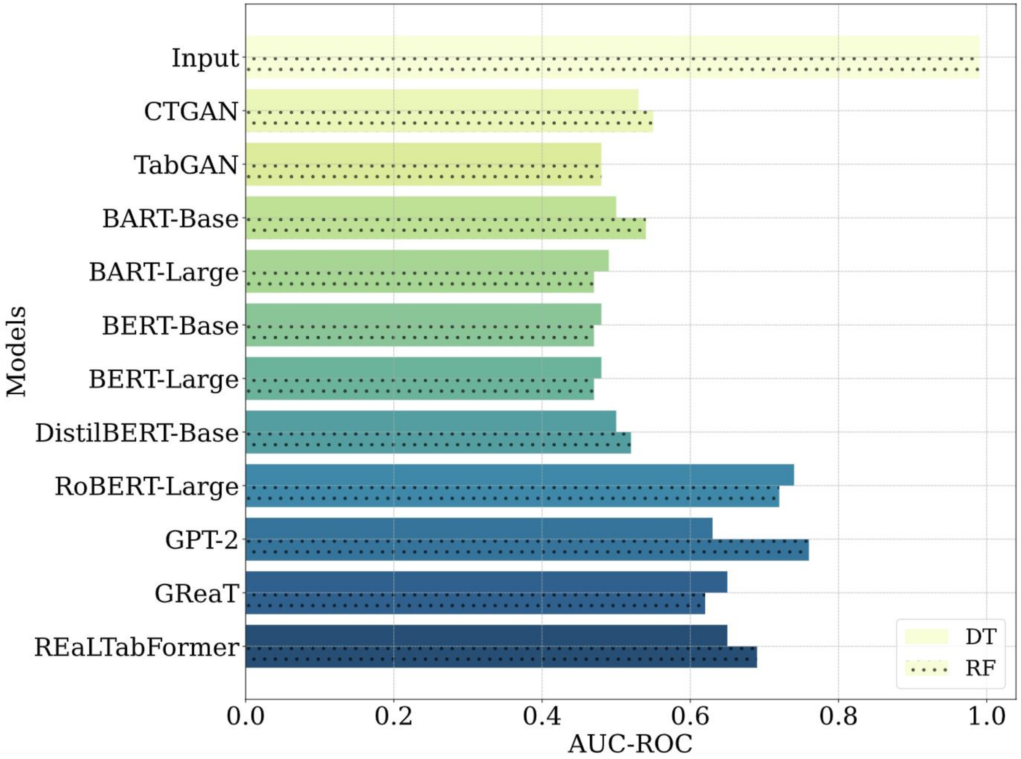
RQ2.2



Does Low DCR Ensure Data Utility?

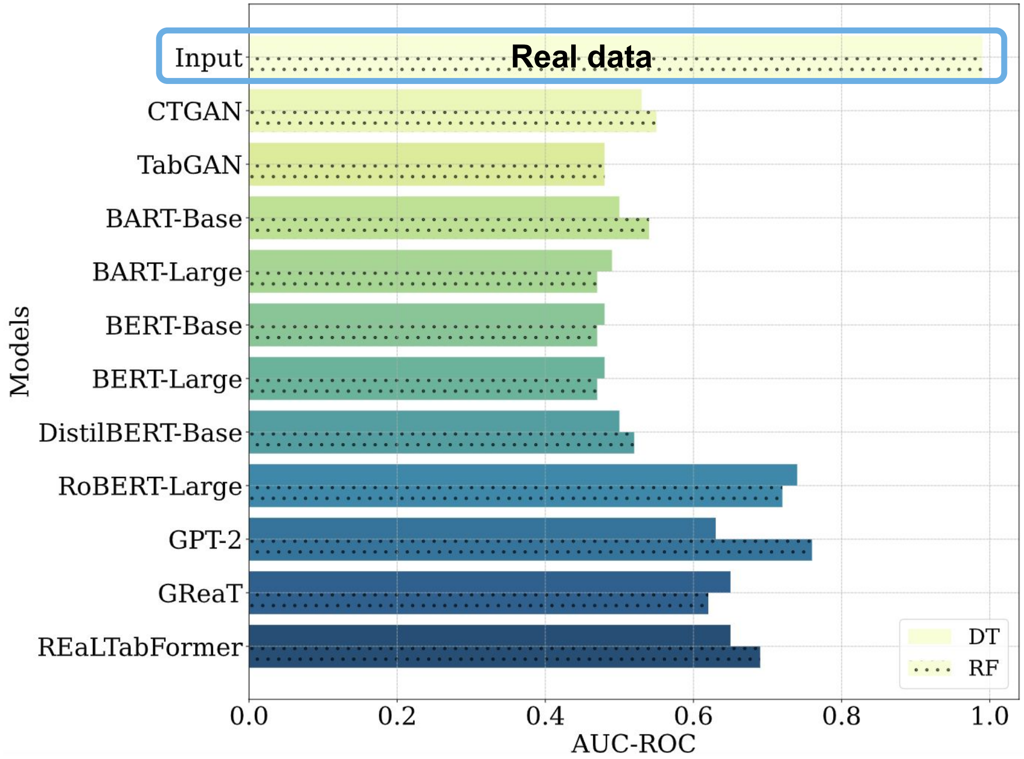
Evaluation (3/3) - Machine Learning Efficiency

RQ2.2



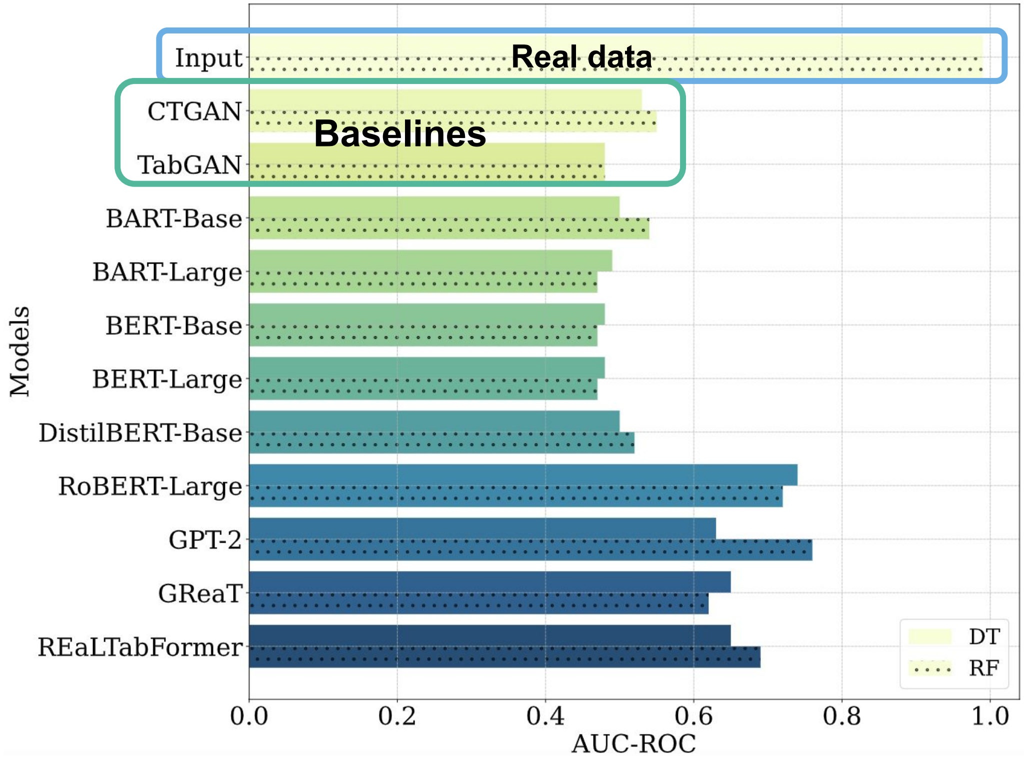
Evaluation (3/3) - Machine Learning Efficiency

RQ2.2



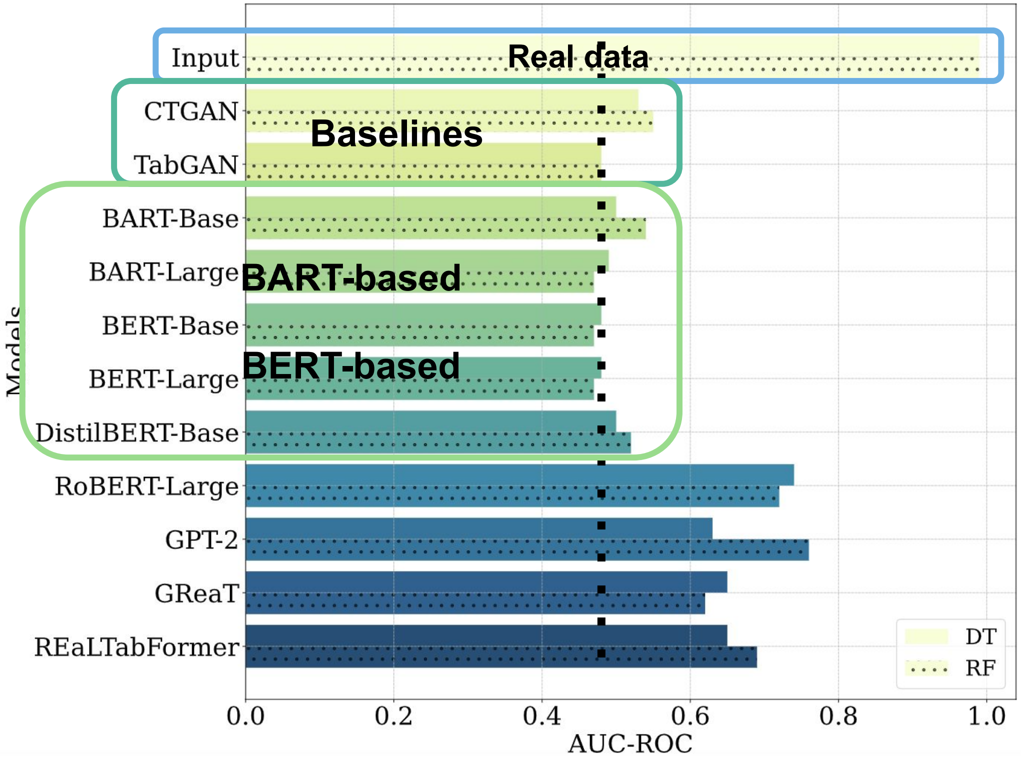
Evaluation (3/3) - Machine Learning Efficiency

RQ2.2



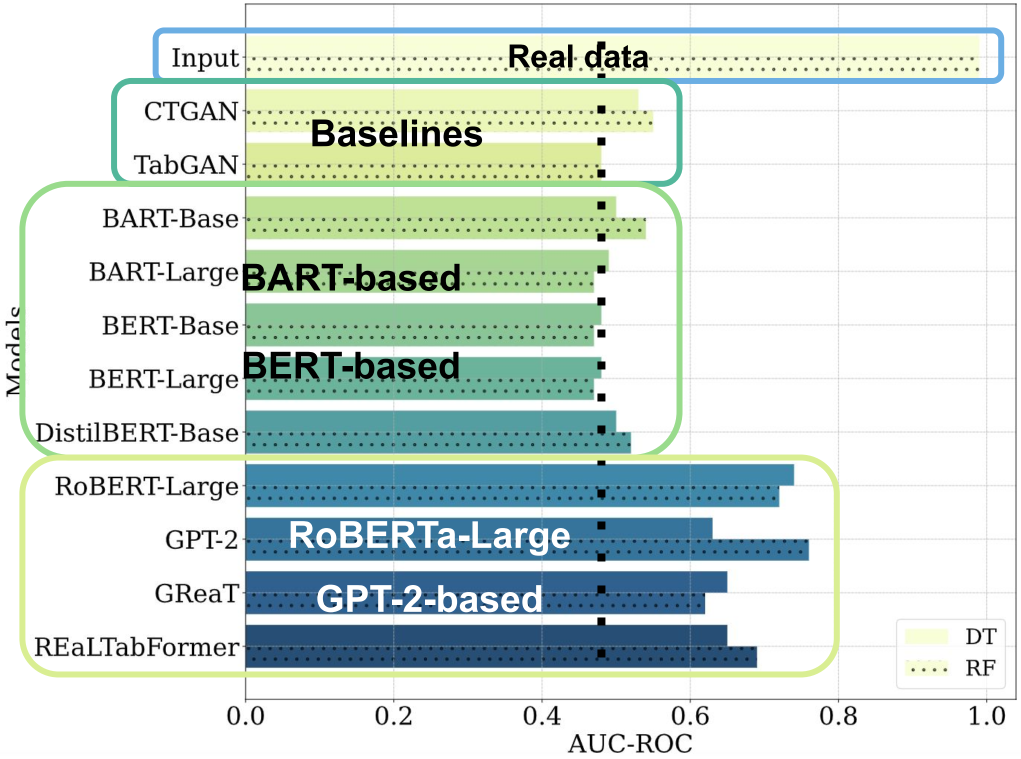
Evaluation (3/3) - Machine Learning Efficiency

RQ2.2



Evaluation (3/3) - Machine Learning Efficiency

RQ2.2



Summary of Job Data Synthesis Using LLMs

RQ2

- **F1** (DCR) - **RoBERTa-Large** and **GPT-2-based models** demonstrate the most promise with low DCR values.

- **F2** (MLE) - Predictive models trained on synthetic datasets show **varied performance**, falling short compared to real data.

- **F3** - There is a need for **improvements** in synthetic data generation techniques.

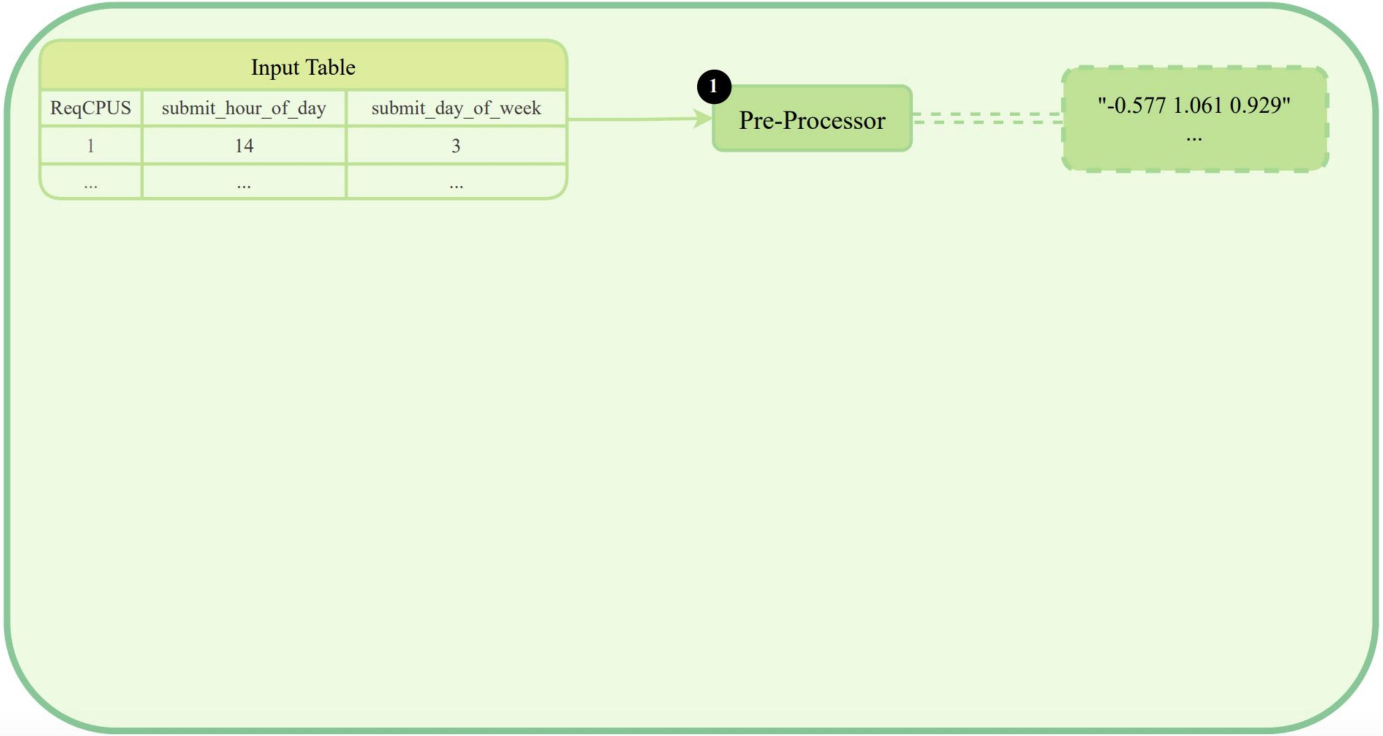
Design - How to Use LLMs to Predict Job End-State?

RQ3.1

Input Table		
ReqCPUS	submit_hour_of_day	submit_day_of_week
1	14	3
...	...	...

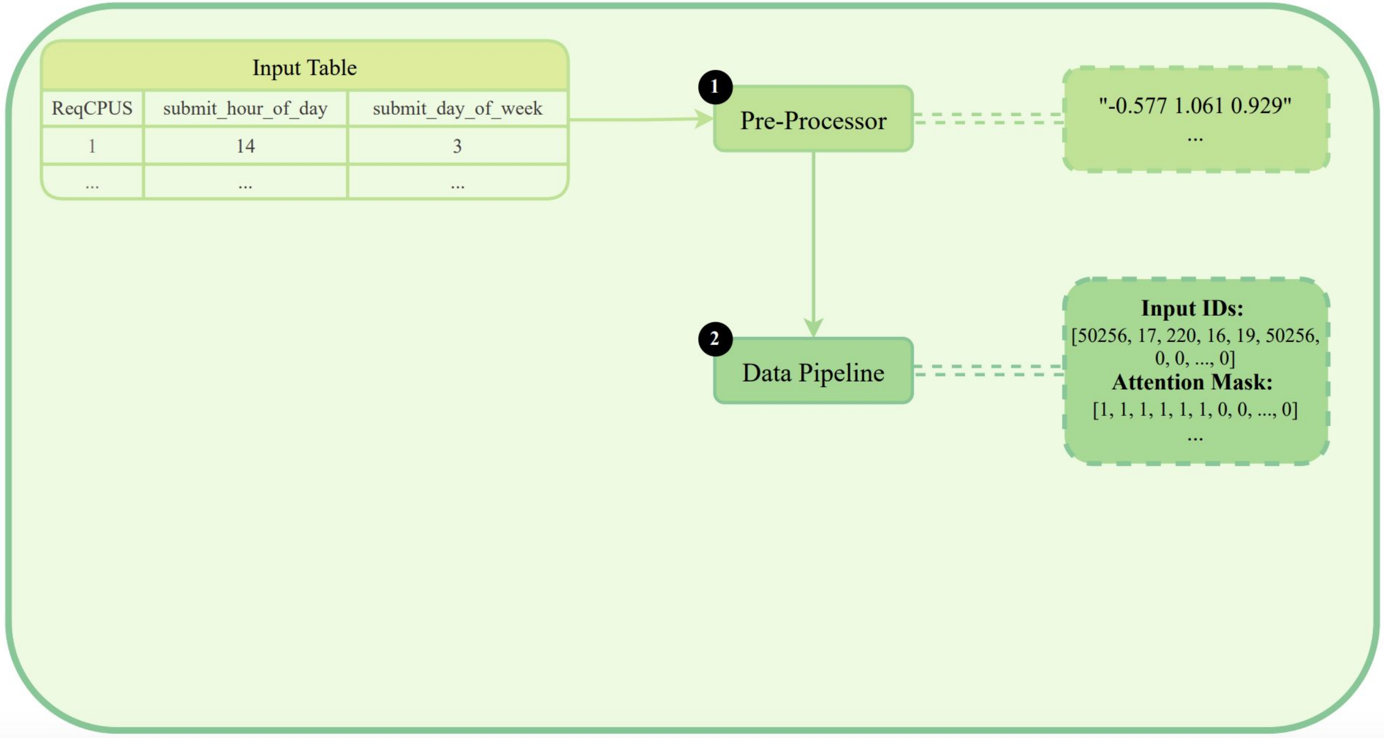
Design - How to Use LLMs to Predict Job End-State?

RQ3.1



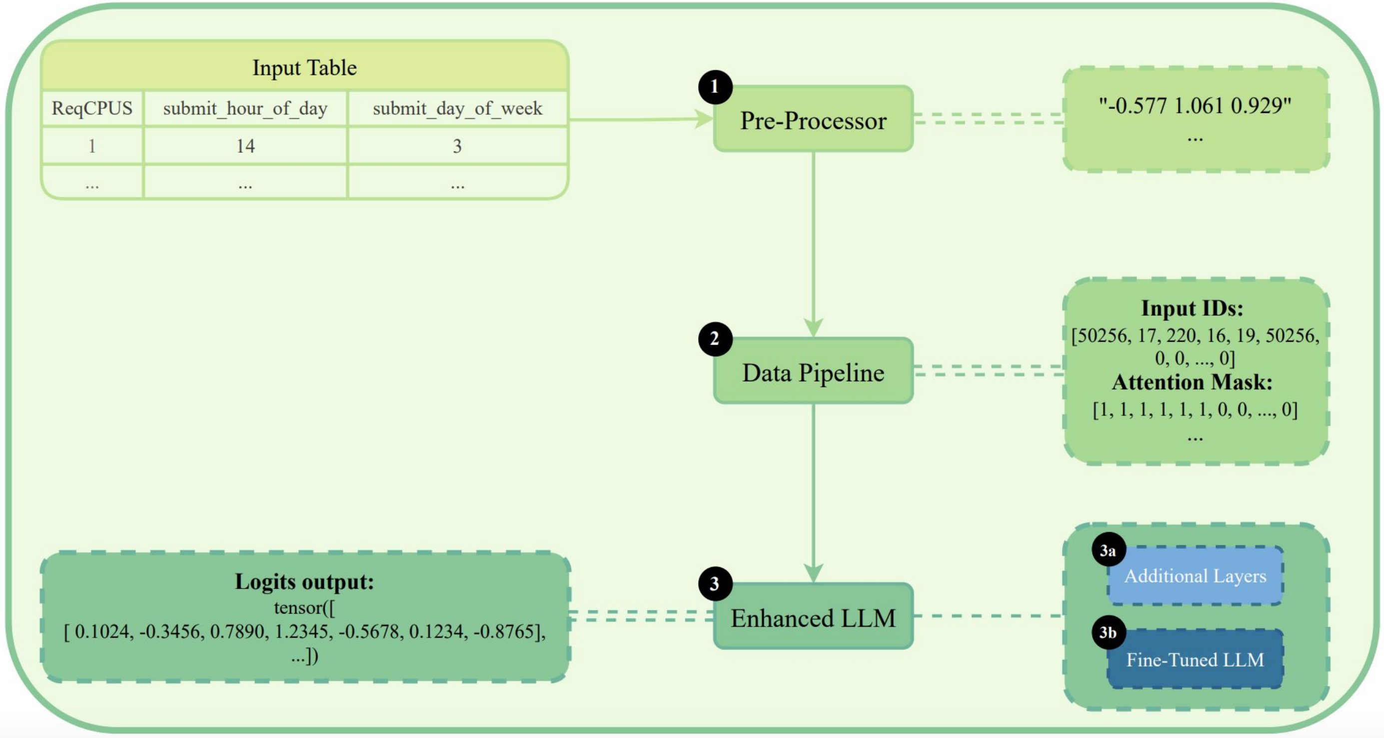
Design - How to Use LLMs to Predict Job End-State?

RQ3.1



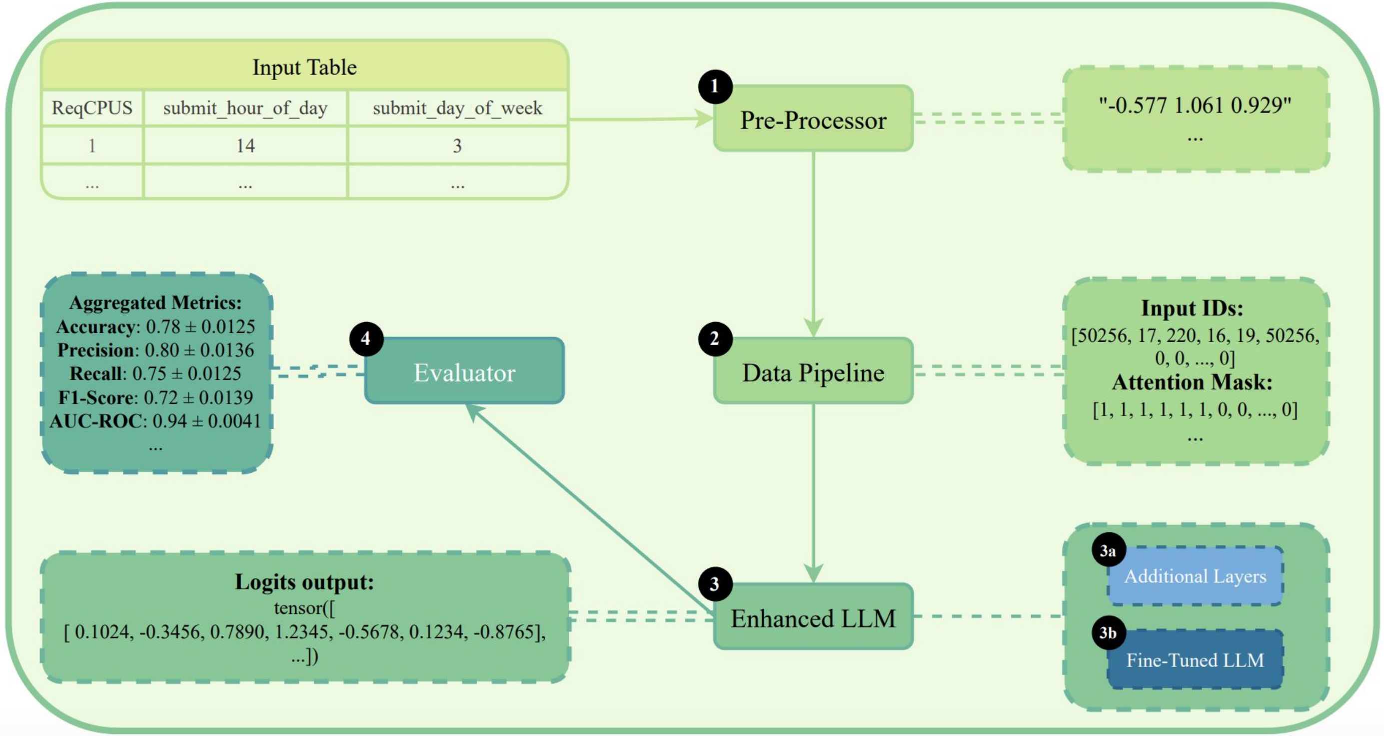
Design - How to Use LLMs to Predict Job End-State?

RQ3.1



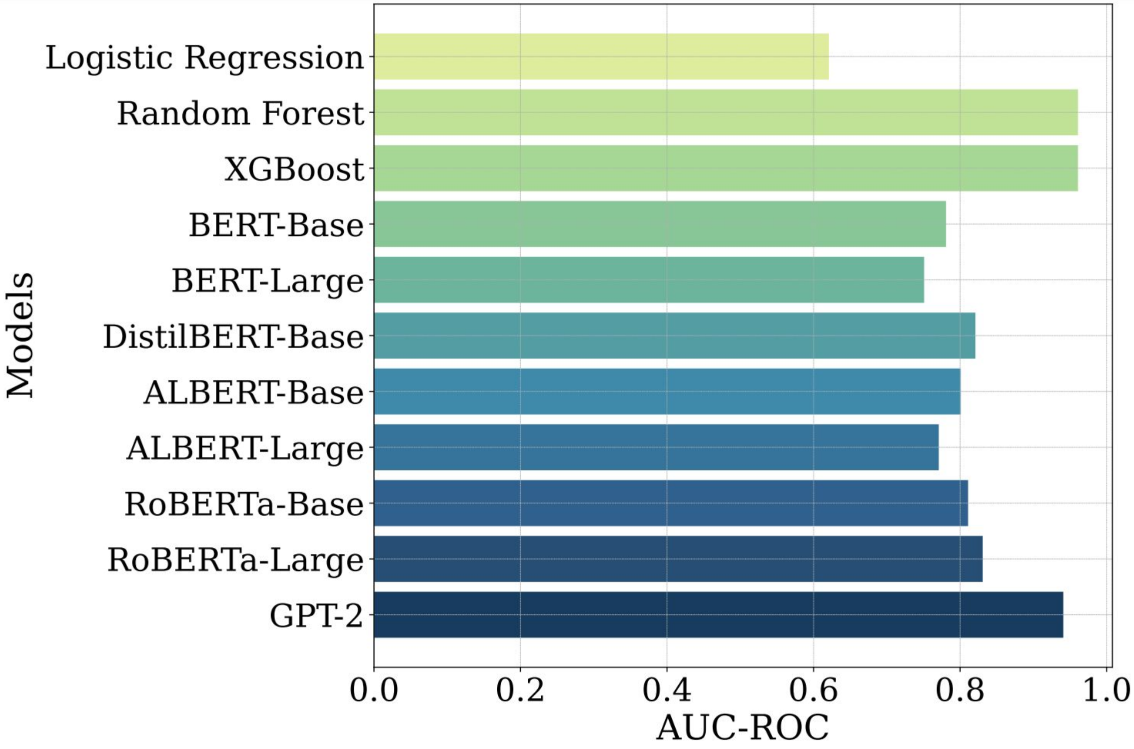
Design - How to Use LLMs to Predict Job End-State?

RQ3.1



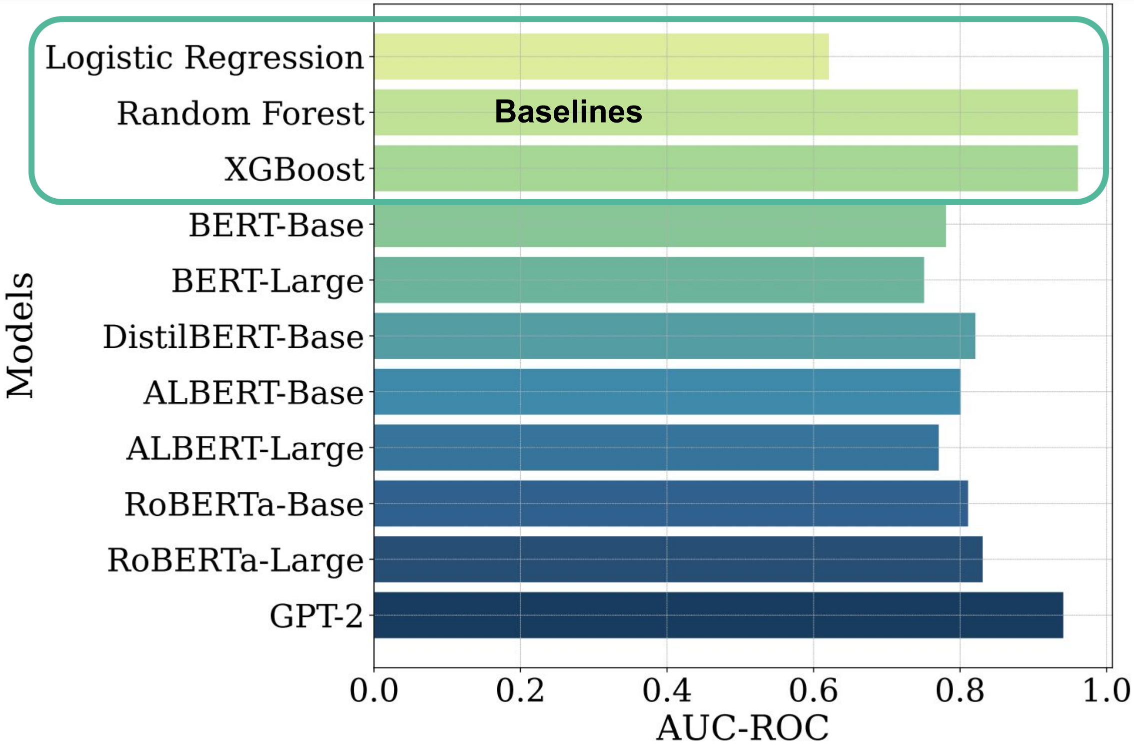
Evaluation - How Well Do LLMs Work in Prediction?

RQ3.2



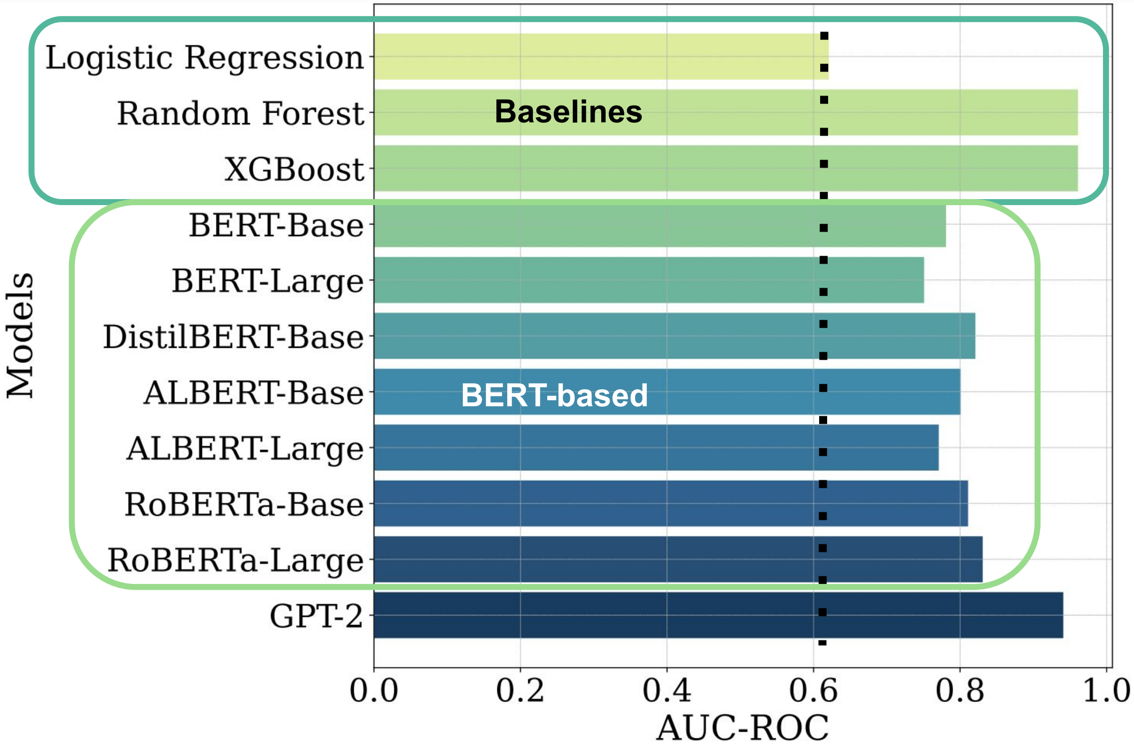
Evaluation - How Well Do LLMs Work in Prediction?

RQ3.2



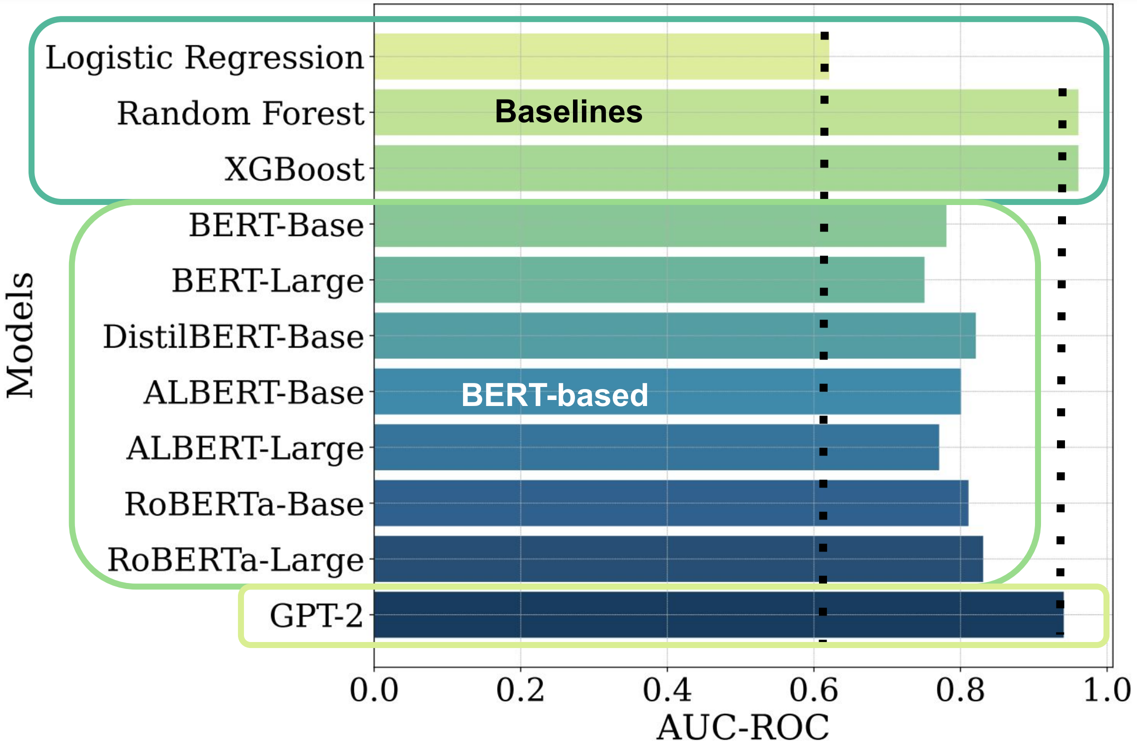
Evaluation - How Well Do LLMs Work in Prediction?

RQ3.2



Evaluation - How Well Do LLMs Work in Prediction?

RQ3.2



Summary of Job End-State Prediction Using LLMs

RQ3

- **F1** - **GPT-2** shows competitive performance against the best-performing traditional models.

- **F2** - **RoBERTa-Large** and **DistilBERT-Base** also show promising results.

- **F3** - LLMs demonstrate **significant potential** in predicting job termination states within HPC clusters.

Conclusion & Future Work

Conclusion & Future Work

RQ1

- Existing LLM applications primarily focus on **programming-related tasks**.
- Significant **potential** in workload synthesis and predictive analysis.

Conclusion & Future Work

RQ2

- Existing LLM applications primarily focus on **programming-related tasks**.
- Significant **potential** in workload synthesis and predictive analysis.

- Models such as **RoBERTa-Large** and **GPT-2** excel in data synthesis.
- LLM-generated data shows promise but doesn't match real data for training predictive models.

Conclusion & Future Work

RQ3

- Existing LLM applications primarily focus on **programming-related tasks**.
- Significant **potential** in workload synthesis and predictive analysis.

- Models such as **RoBERTa-Large** and **GPT-2** excel in data synthesis.
- LLM-generated data shows promise but doesn't match real data for training predictive models.

- **GPT-2** shows competitive performance with traditional models in predictive tasks.
- Other LLMs show varying levels of effectiveness.

Conclusion & Future Work

- Existing LLM applications primarily focus on **programming-related tasks**.
- Significant **potential** in workload synthesis and predictive analysis.

- Models suc… s.
- LLM-gener… a for training predictive r…

- **GPT-2** show… redictive tasks.
- Other LLMs show varying levels of effectiveness.

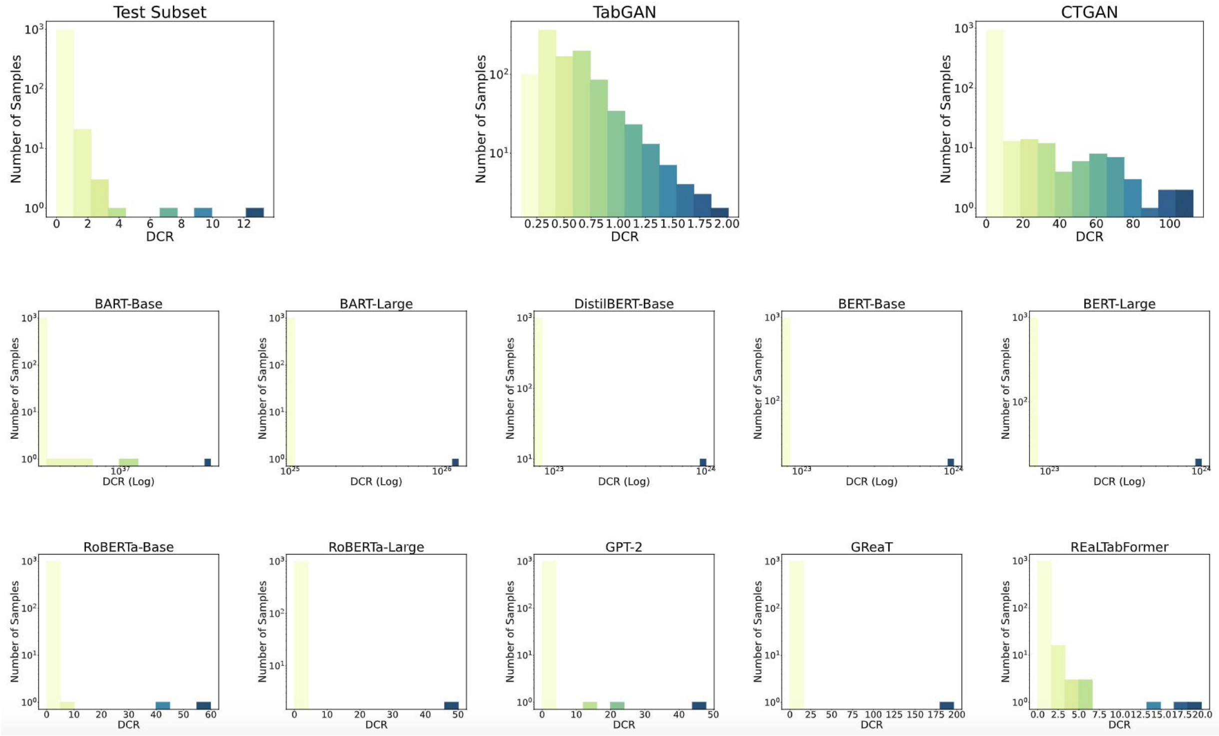
Future Work

➔ A wider range of LLM architectures.
➔ Additional metrics for synthetic data evaluation.
➔ Advanced techniques for textual encoding.

Backup Slides

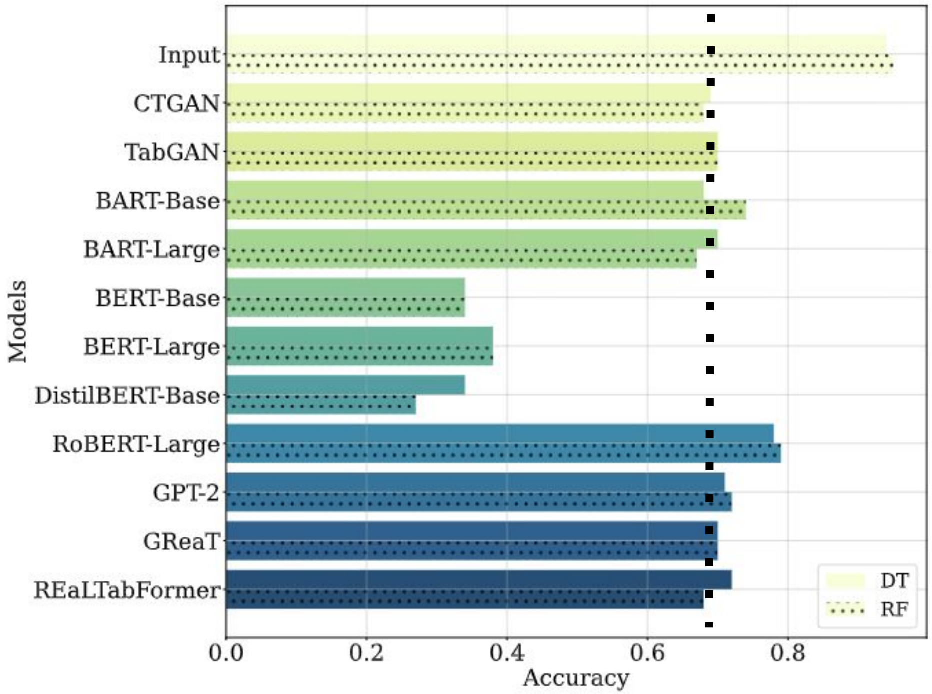
DCR Distributions

RQ2.2



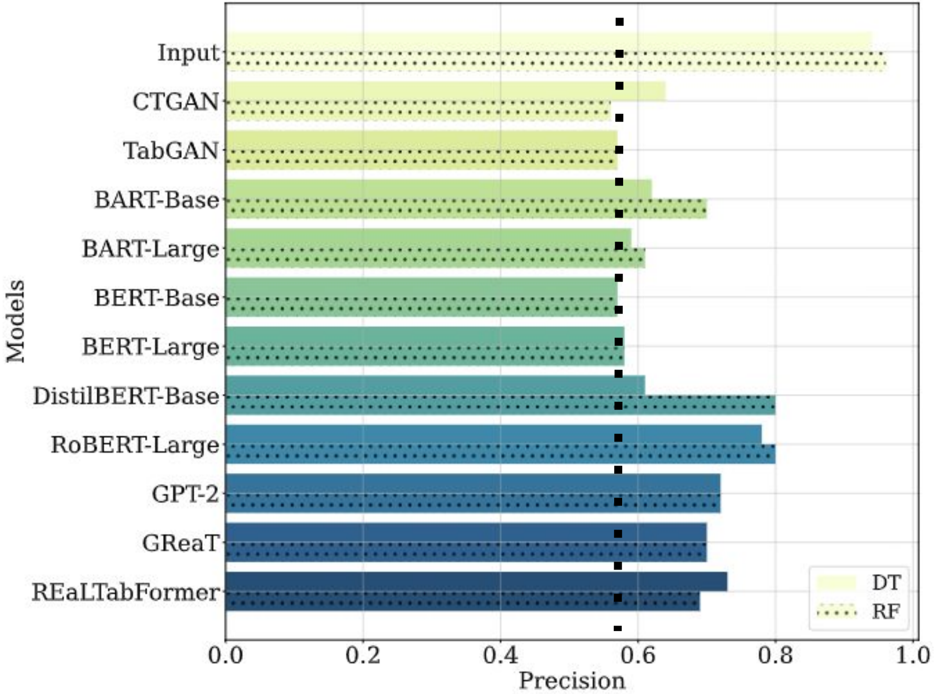
MLE Metrics (1/4) - Accuracy

RQ2.2



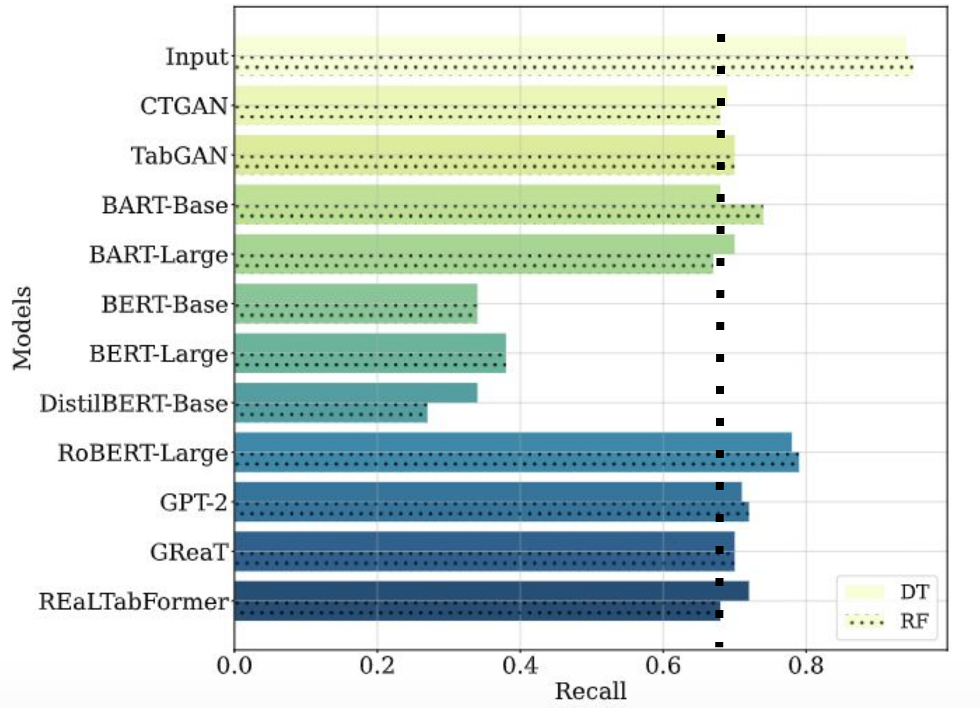
MLE Metrics (2/4) - Precision

RQ2.2



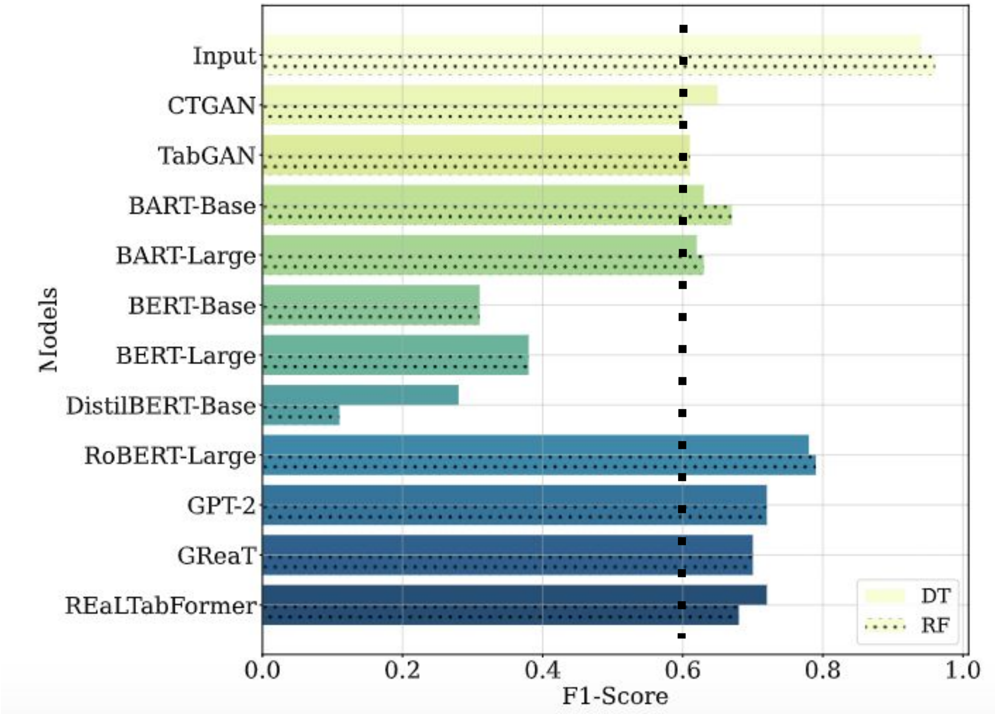
MLE Metrics (3/4) - Recall

RQ2.2



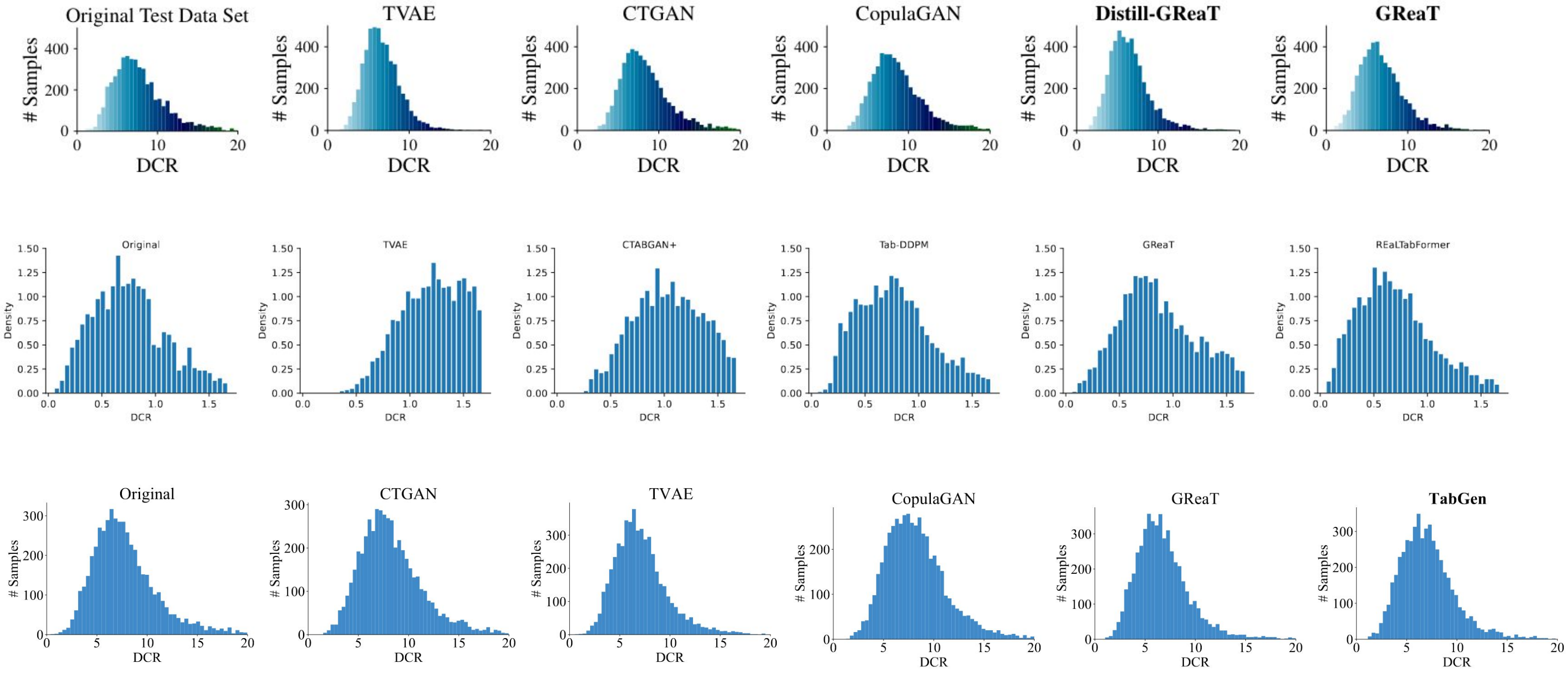
MLE Metrics (4/4) - F1 - Score

RQ2.2



DCR Distribution

RQ2.2



Source: Borisov et al., 2023;
Solatorio and Dupriez, 2023;
Zhang et al., 2023.

LLMs in Classification

RQ3.2

Dataset	Method	0	4	8	16	32	64	128	256	512	all
		Number of Shots									
Bank	XGBoost	—	$0.50_{.00}$	$0.56_{.09}$	$0.68_{.04}$	$0.76_{.03}$	**$0.83_{.02}$**	$0.85_{.03}$	$0.88_{.01}$	$0.90_{.01}$	**$0.94_{.00}$**
	TabPFN	—	$0.59_{.14}$	**$0.66_{.08}$**	**$0.69_{.02}$**	$0.76_{.03}$	$0.82_{.03}$	**$0.86_{.02}$**	**$0.89_{.00}$**	$0.90_{.00}$	$0.91_{.00}$
	TabLLM	**$0.63_{.01}$**	$0.59_{.10}$	$0.64_{.05}$	$0.65_{.05}$	$0.64_{.06}$	$0.69_{.03}$	$0.82_{.05}$	$0.87_{.01}$	$0.88_{.01}$	0.92 †
Blood	XGBoost	—	$0.50_{.00}$	$0.58_{.07}$	$0.66_{.04}$	$0.67_{.06}$	$0.68_{.05}$	$0.71_{.06}$	$0.70_{.07}$	$0.67_{.06}$	$0.71_{.04}$
	TabPFN	—	$0.52_{.08}$	$0.64_{.04}$	**$0.67_{.01}$**	**$0.70_{.04}$**	**$0.73_{.04}$**	**$0.75_{.04}$**	**$0.76_{.04}$**	**$0.76_{.03}$**	**$0.74_{.03}$**
	TabLLM	**$0.61_{.04}$**	**$0.58_{.09}$**	**$0.66_{.03}$**	$0.66_{.07}$	$0.68_{.04}$	$0.68_{.04}$	$0.68_{.06}$	$0.70_{.08}$	$0.68_{.04}$	$0.70_{.04}$
Calhousing	XGBoost	—	$0.50_{.00}$	$0.62_{.10}$	$0.74_{.03}$	$0.79_{.04}$	$0.82_{.04}$	$0.87_{.01}$	$0.90_{.01}$	$0.92_{.01}$	**$0.97_{.00}$**
	TabPFN	—	$0.63_{.13}$	**$0.63_{.11}$**	**$0.80_{.03}$**	**$0.85_{.03}$**	**$0.89_{.01}$**	**$0.91_{.01}$**	**$0.92_{.00}$**	**$0.93_{.00}$**	$0.94_{.00}$
	TabLLM	**$0.61_{.01}$**	$0.63_{.05}$	$0.60_{.07}$	$0.70_{.08}$	$0.77_{.08}$	$0.77_{.04}$	$0.81_{.02}$	$0.83_{.01}$	$0.86_{.02}$	$0.95_{.00}$
Car	XGBoost	—	$0.50_{.00}$	$0.59_{.04}$	$0.70_{.08}$	$0.82_{.03}$	$0.91_{.02}$	$0.95_{.01}$	$0.98_{.01}$	$0.99_{.01}$	$1.00_{.00}$
	TabPFN	—	$0.64_{.06}$	$0.75_{.05}$	**$0.87_{.04}$**	**$0.92_{.02}$**	**$0.97_{.00}$**	**$0.99_{.01}$**	**$1.00_{.00}$**	$1.00_{.00}$	$1.00_{.00}$
	TabLLM	**$0.82_{.02}$**	**$0.83_{.03}$**	**$0.85_{.03}$**	$0.86_{.03}$	$0.91_{.02}$	$0.96_{.02}$	$0.98_{.01}$	$0.99_{.00}$	$1.00_{.00}$	$1.00_{.00}$
Credit-g	XGBoost	—	$0.50_{.00}$	$0.51_{.07}$	$0.59_{.05}$	$0.66_{.03}$	$0.67_{.06}$	$0.68_{.02}$	$0.73_{.02}$	$0.75_{.03}$	**$0.78_{.04}$**
	TabPFN	—	$0.58_{.08}$	$0.59_{.03}$	$0.64_{.06}$	$0.69_{.07}$	$0.70_{.07}$	**$0.72_{.06}$**	**$0.75_{.04}$**	$0.75_{.02}$	$0.75_{.03}$
	TabLLM	**$0.53_{.05}$**	**$0.69_{.04}$**	**$0.66_{.04}$**	**$0.66_{.05}$**	**$0.72_{.06}$**	$0.70_{.07}$	$0.71_{.07}$	$0.72_{.03}$	$0.72_{.02}$	$0.70_{.02}$
Diabetes	XGBoost	—	$0.50_{.00}$	$0.59_{.16}$	**$0.72_{.07}$**	$0.69_{.08}$	$0.73_{.05}$	$0.78_{.05}$	$0.80_{.03}$	$0.80_{.01}$	**$0.84_{.03}$**
	TabPFN	—	$0.61_{.13}$	**$0.67_{.11}$**	$0.71_{.07}$	**$0.77_{.03}$**	**$0.82_{.03}$**	**$0.83_{.03}$**	**$0.83_{.03}$**	**$0.81_{.02}$**	$0.81_{.03}$
	TabLLM	**$0.68_{.06}$**	$0.61_{.09}$	$0.63_{.08}$	$0.69_{.07}$	$0.68_{.04}$	$0.73_{.03}$	$0.79_{.04}$	$0.78_{.02}$	$0.78_{.04}$	$0.80_{.04}$
Heart	XGBoost	—	$0.50_{.00}$	$0.55_{.14}$	$0.84_{.07}$	$0.88_{.04}$	$0.91_{.01}$	$0.91_{.01}$	$0.90_{.01}$	$0.92_{.01}$	$0.94_{.01}$
	TabPFN	—	**$0.84_{.06}$**	**$0.88_{.05}$**	$0.87_{.06}$	**$0.91_{.02}$**	**$0.92_{.02}$**	**$0.92_{.02}$**	$0.92_{.01}$	$0.92_{.02}$	$0.92_{.02}$
	TabLLM	**$0.54_{.04}$**	$0.76_{.14}$	$0.83_{.05}$	$0.87_{.04}$	$0.87_{.06}$	$0.91_{.01}$	$0.90_{.01}$	$0.92_{.01}$	$0.92_{.01}$	$0.94_{.01}$
Income	XGBoost	—	$0.50_{.00}$	$0.59_{.06}$	$0.77_{.02}$	$0.79_{.03}$	$0.82_{.02}$	$0.84_{.01}$	$0.87_{.01}$	$0.88_{.00}$	**$0.93_{.00}$**
	TabPFN	—	0.73 [illegible]	0.71 [illegible]	0.76 [illegible]	0.80 [illegible]	0.82 [illegible]	0.84 [illegible]	$0.86_{.01}$	$0.87_{.01}$	$0.89_{.00}$
	TabLLM	**$0.84_{.00}$**	**$0.84_{.01}$**	**$0.84_{.02}$**	**$0.84_{.04}$**	**$0.84_{.01}$**	**$0.84_{.02}$**	**$0.86_{.01}$**	$0.87_{.00}$	**$0.89_{.01}$**	$0.92_{.00}$
Jungle	XGBoost	—	$0.50_{.00}$	$0.58_{.07}$	**$0.72_{.05}$**	$0.78_{.03}$	$0.81_{.02}$	$0.84_{.02}$	$0.87_{.01}$	$0.91_{.01}$	$0.98_{.00}$
	TabPFN	—	**$0.65_{.08}$**	**$0.72_{.04}$**	$0.71_{.07}$	$0.78_{.02}$	$0.81_{.01}$	$0.84_{.01}$	**$0.88_{.01}$**	$0.91_{.00}$	$0.93_{.00}$
	TabLLM	**$0.60_{.00}$**	$0.64_{.01}$	$0.64_{.02}$	$0.65_{.03}$	$0.71_{.02}$	$0.78_{.02}$	$0.81_{.02}$	$0.84_{.01}$	$0.89_{.01}$	**1.00** †

#Shots	Method	Bank	Blood	C. Hous.	Car	Creditg	Diabetes	Heart	Income	Jungle
64	Logistic Reg.	$0.84_{.02}$	**$0.74_{.02}$**	$0.88_{.01}$	$0.93_{.02}$	$0.66_{.07}$	$0.80_{.02}$	$0.91_{.01}$	$0.83_{.03}$	$0.79_{.01}$
	LightGBM	$0.77_{.03}$	$0.69_{.04}$	$0.81_{.02}$	$0.85_{.06}$	$0.61_{.09}$	$0.79_{.02}$	$0.91_{.01}$	$0.78_{.03}$	$0.79_{.02}$
	XGBoost	$0.83_{.02}$	$0.68_{.05}$	$0.82_{.04}$	$0.91_{.02}$	$0.67_{.06}$	$0.73_{.05}$	$0.91_{.01}$	$0.82_{.02}$	**$0.81_{.02}$**
	SAINT	$0.81_{.03}$	$0.67_{.05}$	$0.81_{.02}$	$0.92_{.02}$	$0.66_{.06}$	$0.79_{.03}$	$0.90_{.04}$	**$0.84_{.02}$**	**$0.81_{.01}$**
	TabNet	$0.71_{.06}$	$0.63_{.06}$	$0.72_{.03}$	$0.73_{.07}$	$0.56_{.05}$	$0.71_{.04}$	$0.83_{.05}$	$0.71_{.04}$	$0.73_{.04}$
	NODE	$0.78_{.02}$	$0.71_{.05}$	$0.80_{.01}$	$0.80_{.02}$	$0.63_{.04}$	$0.77_{.04}$	$0.88_{.02}$	$0.75_{.02}$	$0.75_{.04}$
	TabPFN	$0.82_{.03}$	$0.73_{.04}$	**$0.89_{.01}$**	**$0.97_{.00}$**	**$0.70_{.07}$**	$0.82_{.03}$	**$0.92_{.02}$**	$0.82_{.04}$	**$0.81_{.01}$**
	TabLLM	$0.69_{.03}$	$0.68_{.04}$	$0.77_{.04}$	$0.96_{.02}$	**$0.70_{.07}$**	$0.73_{.03}$	$0.91_{.01}$	**$0.84_{.02}$**	$0.78_{.02}$
	LLaMA	$0.62_{.02}$	$0.66_{.03}$	$0.57_{.04}$	$0.90_{.02}$	$0.67_{.09}$	$0.78_{.05}$	$0.88_{.02}$	**$0.84_{.02}$**	$0.63_{.04}$
	LLaMA-GTL	**$0.86_{.01}$**	$0.72_{.05}$	$0.78_{.04}$	$0.96_{.01}$	**$0.70_{.09}$**	**$0.83_{.04}$**	$0.88_{.05}$	**$0.84_{.01}$**	$0.69_{.04}$
512	Logistic Reg.	$0.89_{.00}$	**$0.76_{.03}$**	$0.91_{.00}$	$0.98_{.00}$	**$0.76_{.02}$**	$0.83_{.02}$	**$0.93_{.01}$**	$0.88_{.00}$	$0.80_{.00}$
	LightGBM	$0.89_{.00}$	$0.67_{.05}$	$0.92_{.00}$	$0.99_{.01}$	$0.75_{.02}$	$0.79_{.03}$	$0.92_{.01}$	$0.88_{.00}$	**$0.91_{.00}$**
	XGBoost	**$0.90_{.01}$**	$0.67_{.06}$	$0.92_{.01}$	$0.99_{.01}$	$0.75_{.03}$	$0.80_{.01}$	$0.92_{.01}$	$0.88_{.00}$	**$0.91_{.01}$**
	SAINT	$0.88_{.01}$	$0.73_{.02}$	$0.91_{.02}$	$0.99_{.00}$	$0.73_{.03}$	$0.77_{.03}$	$0.92_{.01}$	$0.88_{.00}$	$0.90_{.00}$
	TabNet	$0.83_{.03}$	$0.72_{.02}$	$0.87_{.01}$	$0.98_{.01}$	$0.66_{.04}$	$0.74_{.07}$	$0.88_{.03}$	$0.83_{.02}$	$0.84_{.01}$
	NODE	$0.86_{.01}$	**$0.76_{.03}$**	$0.87_{.01}$	$0.96_{.01}$	$0.70_{.02}$	$0.83_{.02}$	$0.92_{.03}$	$0.83_{.01}$	$0.80_{.00}$
	TabPFN	**$0.90_{.00}$**	**$0.76_{.03}$**	**$0.93_{.00}$**	**$1.00_{.00}$**	$0.75_{.02}$	$0.81_{.02}$	$0.92_{.02}$	$0.87_{.01}$	**$0.91_{.00}$**
	TabLLM	$0.88_{.01}$	$0.68_{.04}$	$0.86_{.02}$	**$1.00_{.00}$**	$0.72_{.02}$	$0.78_{.04}$	$0.92_{.01}$	**$0.89_{.01}$**	$0.89_{.01}$
	LLaMA	$0.77_{.02}$	$0.72_{.05}$	$0.86_{.02}$	$0.99_{.00}$	$0.72_{.04}$	$0.83_{.04}$	$0.92_{.02}$	**$0.89_{.01}$**	$0.85_{.03}$
	LLaMA-GTL	**$0.90_{.00}$**	$0.75_{.04}$	$0.89_{.02}$	$0.99_{.01}$	$0.74_{.05}$	**$0.85_{.03}$**	**$0.93_{.02}$**	**$0.89_{.01}$**	$0.89_{.01}$
All	Logistic Reg.	$0.91_{.00}$	**$0.76_{.03}$**	$0.92_{.00}$	$0.98_{.00}$	**$0.79_{.03}$**	$0.83_{.02}$	$0.93_{.01}$	$0.90_{.00}$	$0.81_{.00}$
	LightGBM	**$0.94_{.00}$**	$0.74_{.04}$	**$0.97_{.00}$**	**$1.00_{.00}$**	$0.78_{.02}$	$0.83_{.03}$	**$0.94_{.01}$**	**$0.93_{.00}$**	$0.98_{.00}$
	XGBoost	**$0.94_{.00}$**	$0.71_{.04}$	**$0.97_{.00}$**	**$1.00_{.00}$**	$0.78_{.04}$	$0.84_{.03}$	**$0.94_{.01}$**	**$0.93_{.00}$**	$0.98_{.00}$
	SAINT	$0.93_{.00}$	$0.74_{.03}$	$0.95_{.00}$	**$1.00_{.00}$**	$0.77_{.04}$	$0.83_{.03}$	$0.93_{.01}$	$0.91_{.00}$	**$1.00_{.00}$**
	TabNet	$0.93_{.00}$	$0.71_{.03}$	$0.96_{.00}$	**$1.00_{.00}$**	$0.64_{.03}$	$0.81_{.03}$	$0.89_{.03}$	$0.92_{.00}$	$0.99_{.00}$
	NODE	$0.76_{.02}$	$0.74_{.03}$	$0.87_{.01}$	$0.93_{.01}$	$0.65_{.03}$	$0.83_{.03}$	$0.92_{.03}$	$0.82_{.00}$	$0.81_{.00}$
	TabPFN	$0.91_{.00}$	$0.74_{.03}$	$0.94_{.00}$	**$1.00_{.00}$**	$0.75_{.03}$	$0.81_{.03}$	$0.92_{.02}$	$0.89_{.00}$	$0.93_{.00}$
	TabLLM	$0.92_{.00}$	$0.70_{.04}$	$0.95_{.00}$	**$1.00_{.00}$**	$0.70_{.02}$	$0.80_{.04}$	**$0.94_{.01}$**	$0.92_{.00}$	**$1.00_{.00}$**
	LLaMA	**$0.94_{.00}$**	$0.72_{.04}$	**$0.97_{.00}$**	**$1.00_{.00}$**	$0.76_{.07}$	$0.84_{.03}$	$0.93_{.01}$	**$0.93_{.00}$**	**$1.00_{.00}$**
	LLaMA-GTL	**$0.94_{.00}$**	$0.75_{.05}$	$0.96_{.00}$	**$1.00_{.00}$**	$0.76_{.06}$	**$0.85_{.04}$**	$0.93_{.01}$	**$0.93_{.00}$**	**$1.00_{.00}$**

AUC performance.

Source: Hegselmann et al., 2022; Zhang et al., 2023.