Deep Representation Learning for Tabular Data

Han-Jia Ye
Nanjing University
yehj@lamda.nju.edu.cn

Jun-Peng Jiang
Nanjing University
jiangjp@lamda.nju.edu.cn

Tutorial Part 2
Tabular Data Learning with LLMs

Jun-Peng Jiang
Nanjing University
jiangjp@lamda.nju.edu.cn

Data in the “Table” Modality

- Tabular data has various modalities:
  - Textual table
    - CSV Tables
    - Markdown Tables
    - Table text
  - Table image
  - …
- Tabular data sources

Structured Table

Golfer	Country	Wins
Tiger Woods	United States	18
Geoff Ogilvy	Australia	3
Darren Clarke	Northern Ireland	2
Ernie Els	South Africa	2
Hunter Mahan	United States	2
Phil Mickelson	United States	2
Ian Poulter	England	2

Table Image

Rank	Nation	Gold	Silver	Bronze	Total
1	France	1	3	0	4
2	England	1	2	1	4
3	Ireland	1	1	0	2
–	Sweden	1	1	0	2
5	Belgium	1	0	2	3
6	Hungary	1	0	0	1
–	Netherlands	1	0	0	1
–	Spain	1	0	0	1
9	Scotland	0	1	0	1
10	Czechoslovakia	0	0	2	2
–	Italy	0	0	2	2
12	Denmark	0	0	1	1

Table Image

UCI Machine Learning Repository

OpenML

kaggle

Finance

Healthcare

Research

Opportunities and Challenges with LLMs

- The development of LLM/MMLM is advancing rapidly.
- Are LLM/MMLM truly “general-purpose”?

- Excels in content integration (QA)
- But it is not sensitive to numbers and has poor calculation ability.

Which is larger? 3.8 or 3.11?

3.11 is larger.
When comparing decimals, compare digit by digit from left to right:
Compare the first digit: 3 vs 3 → equal
Compare the second digit: 8 vs 11 → 11 > 8
Therefore, 3.11 > 3.8

How can we deal with tabular data with the help of LLMs?

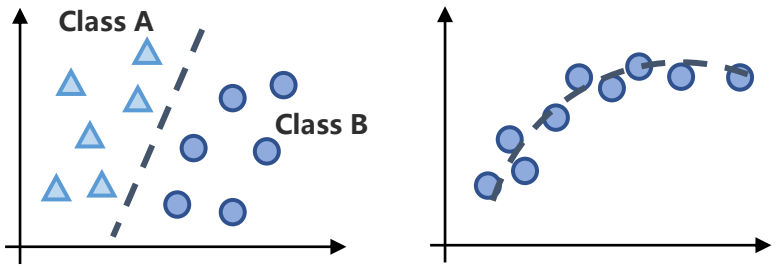
Tabular Tasks with LLMs

Traditional Prediction Task

- Given a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$,
  - $\boldsymbol{x}_i \in \mathbb{R}^d$, with ***categorical*** and ***numerical*** features/attributes
  - $y_i \in \{0,1\}$ for binary classification, $y_i \in \{1, \dots, C\}$ for C-way classification, and $y_i \in \mathbb{R}$ for regression
- The goal is to train LLMs as mapping f. Given an unseen instance $\boldsymbol{x}^*$,

$$\hat{y}^* = f(\boldsymbol{x}^*, \mathcal{D})$$

classification **regression**



Other Tasks such as QA:

- Given a table T, a corresponding query Q, and the LLM f,
- The goal is to answer the query Q according to the table T with the LLM

$$A^* = f(Q \mid T)$$

TableQA **Table Caption**

Outline

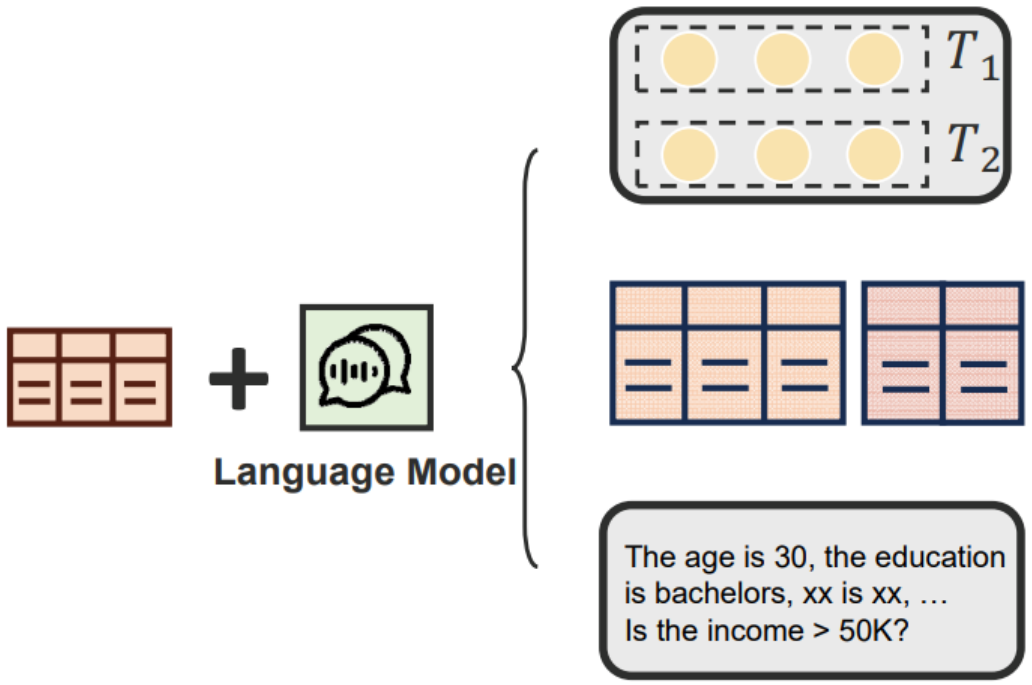
- Tabular Prediction with LLMs
- Tabular QA with LLMs
- From prediction tasks to QA tasks
- Discussions

Outline

- Tabular Prediction with LLMs
- Tabular QA with LLMs
- From prediction tasks to QA tasks
- Discussions

Tabular Prediction with LLMs

Different roles of LLMs in tabular prediction



Feature Tokenization — TransTab, TP-BERTa, CARTE, …

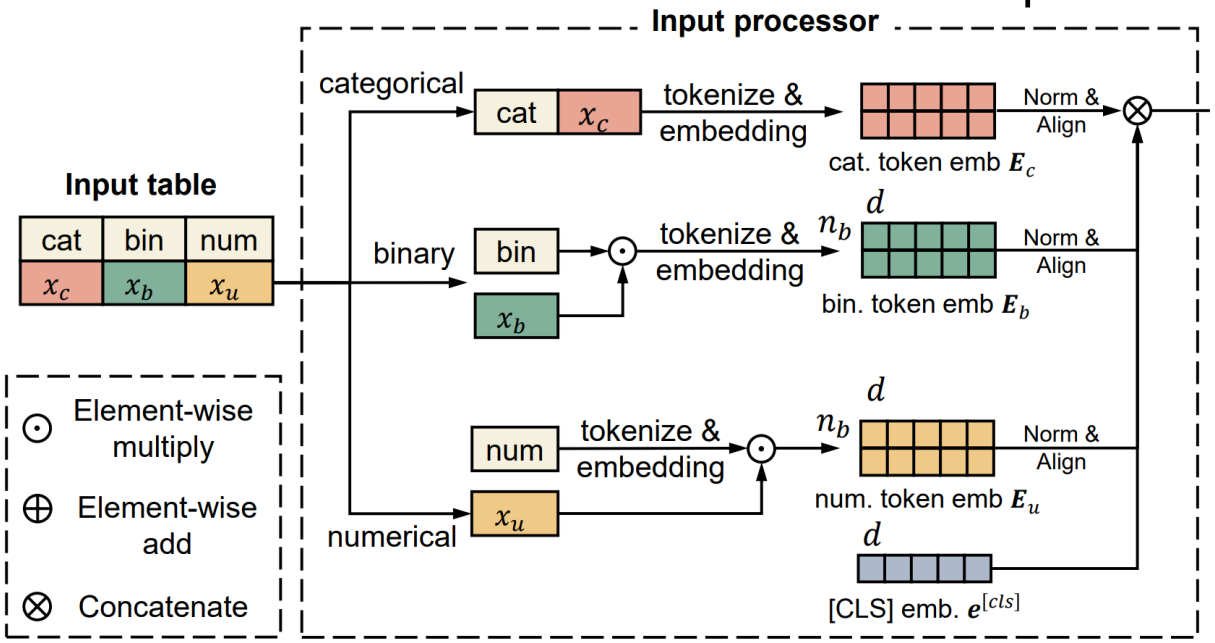
Feature Engineering — CAAFE, FeatLLM, TaPTaP, …

Textual Serialization — TabLLM,TabuLa-8B, GTL, …

TransTab

TransTab encode different kinds of features in different ways

- For categorial feature: TransTab concatenate the column name with the feature value
- For Binary feature: Be tokenized and encoded to the embeddings when the value is 1
- For numerical feature: column names and values are multiplied in embedding space



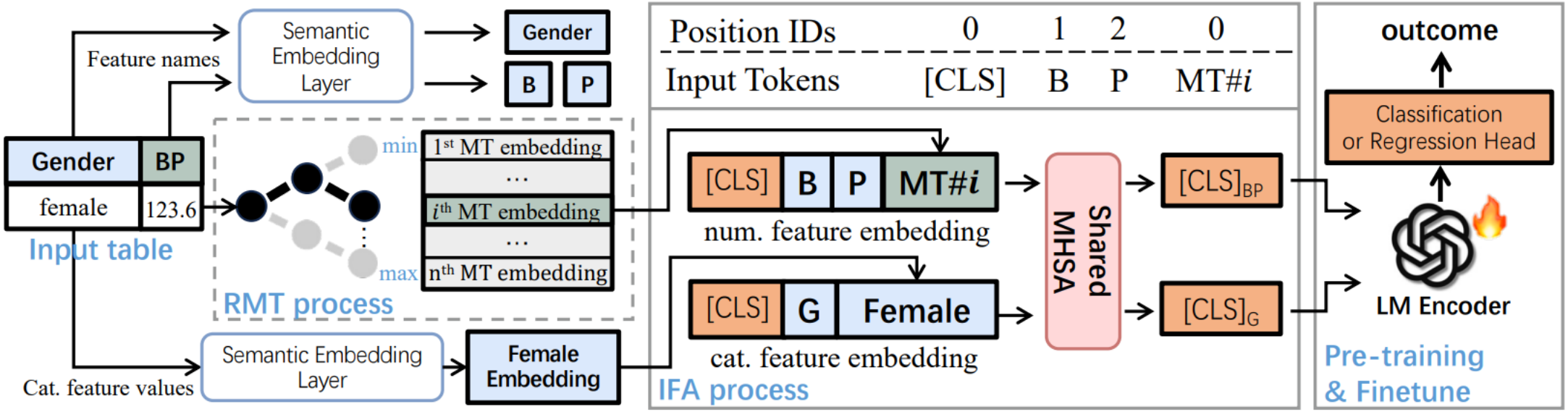
The input processor encodes the feature name and value into the token-level embedding for prediction and learning, converting each sample to a generalizable embedding vector.

Zifeng Wang, Jimeng Sun. ***TransTab: Learning Transferable Tabular Transformers Across Tables***. NeurIPS 2022.

TP-BERTa

TP-BERTa uses numerical discretization strategies and magnitude tokenization for feature encoding.

- Numerical discretization: feature binning with "C4.5 Discretization" applied to each numerical feature by recursively splitting its value range guided by its label.
- Magnitude tokenization: TP-BERTa treats each numerical bins as new words and transforms numerical values (additional tokens in vocabulary) into the language space.

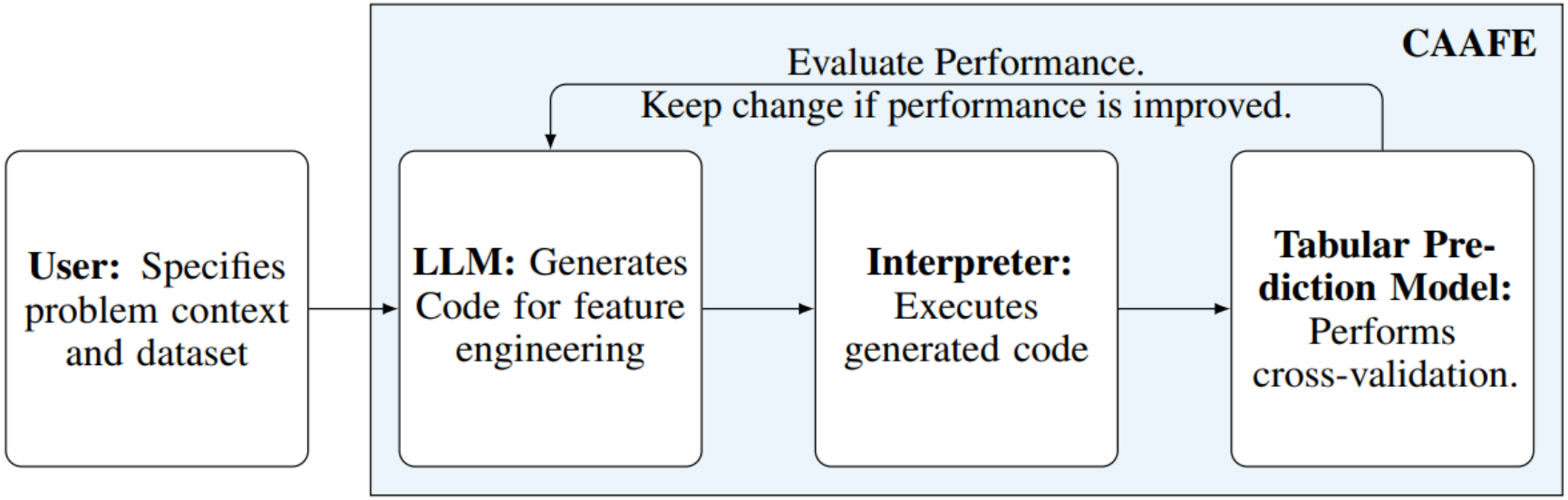


Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Z. Chen, Jimeng Sun, Jian Wu, Jintai Chen.
Making Pre-trained Language Models Great on Tabular Prediction. ICLR 2024.

CAAFE

CAAFE leverages LLMs to incorporate domain knowledge into the feature engineering process.

- CAAFE iteratively generate additional semantically meaningful features for tabular datasets based on the description of the dataset.
- CAAFE produces both Python code for creating new features and explanations for the utility of the generated features.
- The quality of these features is then evaluated using a general tabular model, TabPFN

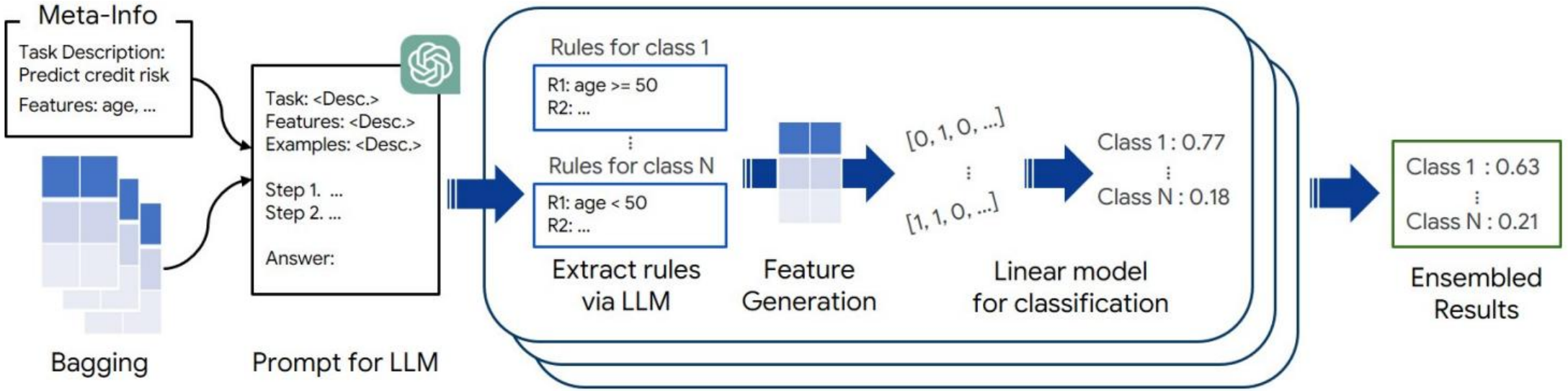


Noah Hollmann, Samuel Müller, Frank Hutter.
Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. NeurIPS 2023.

FeatLLM

FeatLLM enhances feature generation by incorporating in-context learning, enabling LLMs to create new features based on textual descriptions.

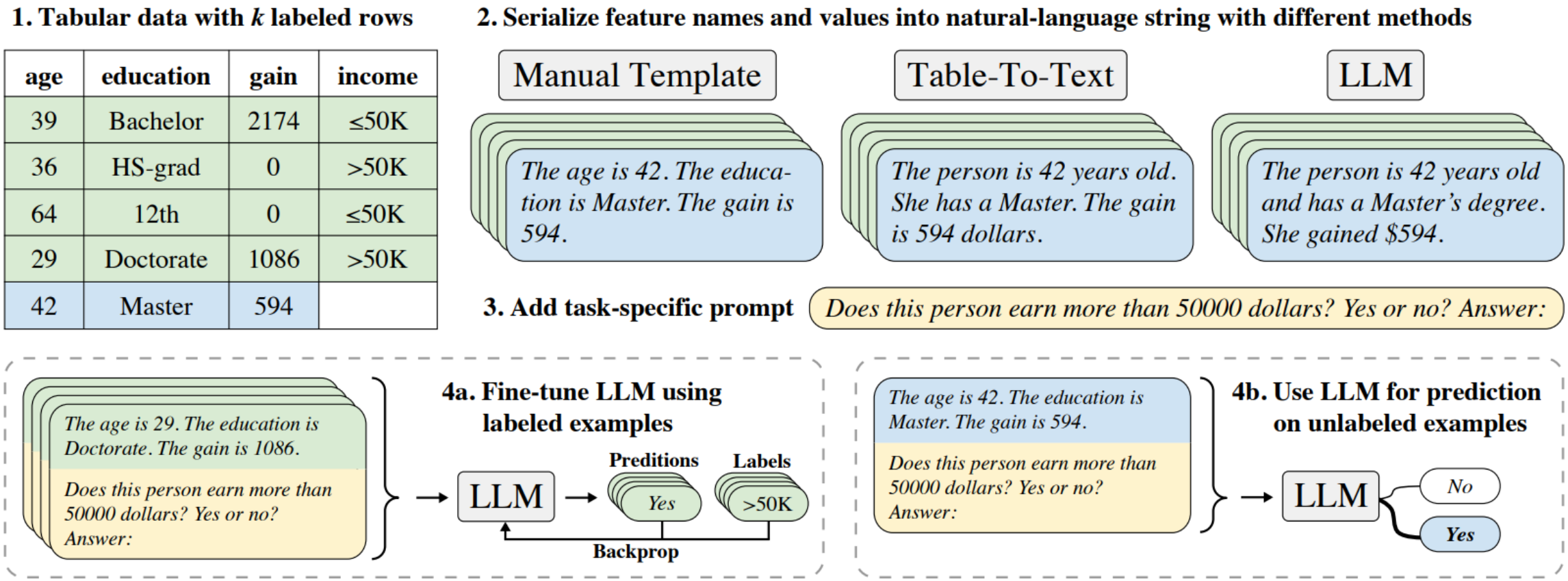
- FeatLLM only uses this simple linear model with the discovered features at inference time.
- FeatLLM repeatedly execute the entire process to create multiple inference models to make the final prediction via ensemble



Sungwon Han, Jinsung Yoon, Sercan O Arik, Tomas Pfister. ***Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning***. ICML 2024.

TabLLM

TabLLM serialize tabular data by integrating feature names into text (in the “key” is “value” format) and combining them with task descriptions, which enables LLMs to treat tabular prediction tasks as text generation problems.

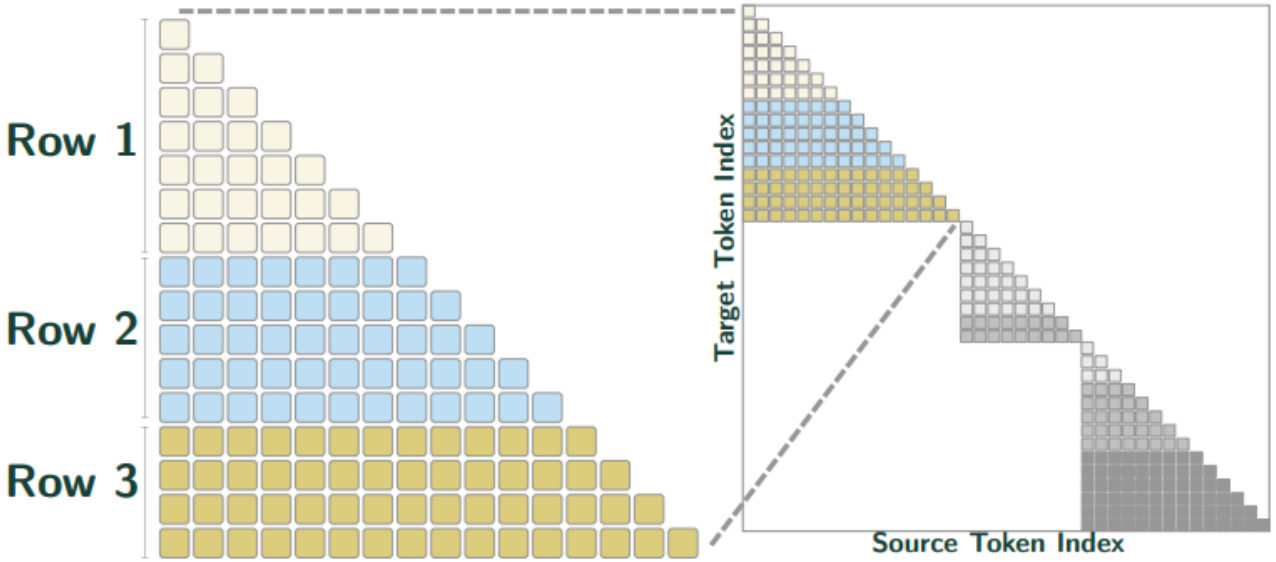


Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, David Sontag.
TabLLM: Few-shot Classification of Tabular Data with Large Language Models. AISTATS 2023.

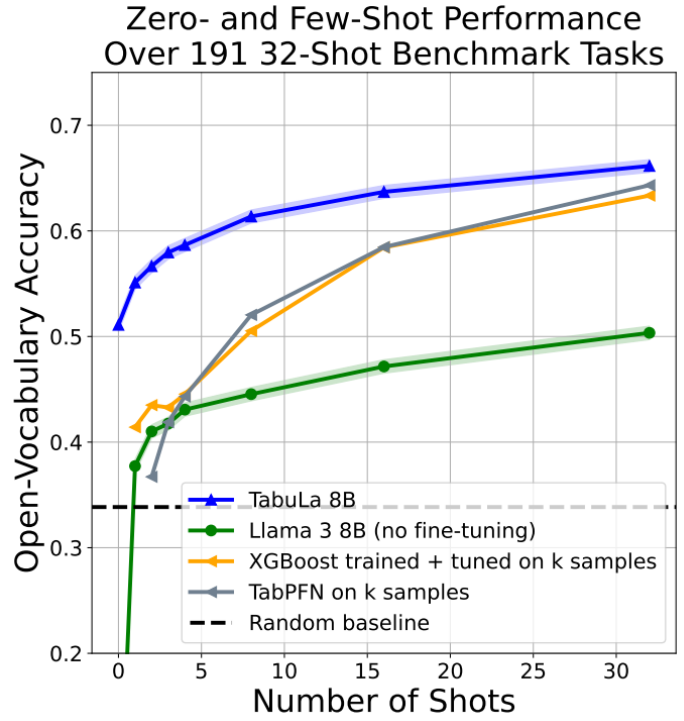
TabuLa-8B

TabuLa-8B fine-tunes a Llama 3-8B LLM for tabular data prediction (classification and binned regression) using a new packing and attention scheme for tabular prediction.

A new dataset T4, sample from TabLib, for tabular prediction



The row-causal tabular masking (RCTM), tailored to few-shot tabular prediction, allow the model to **attend to all previous samples from the same table** in a batch, but **not to samples from other tables**

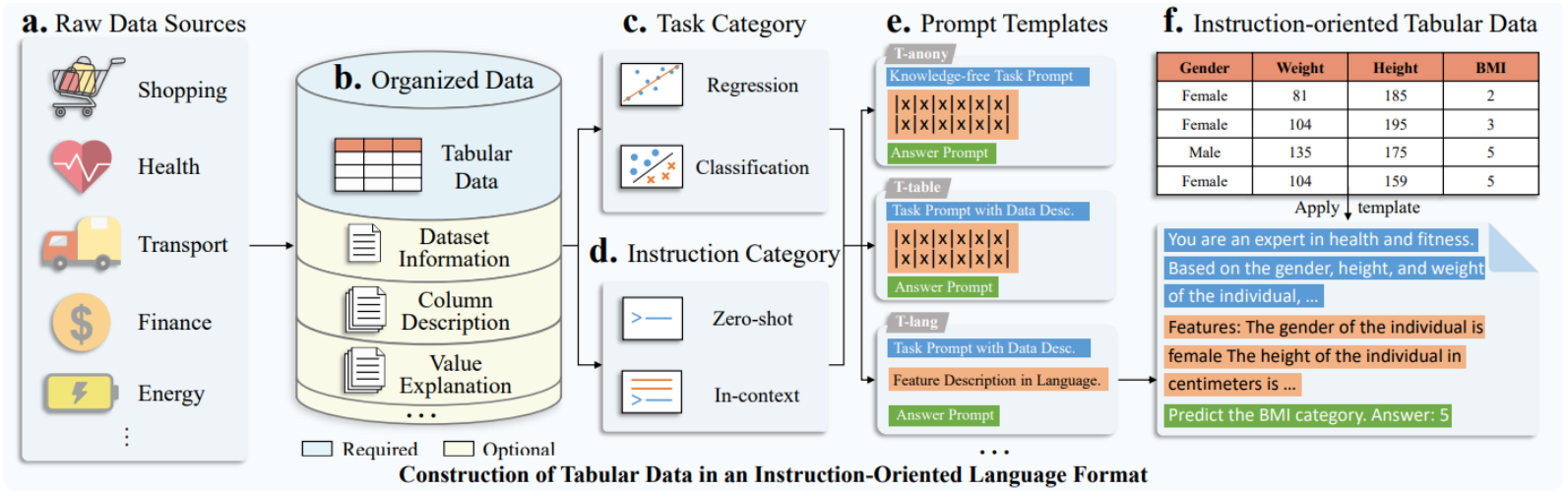


TABULA-8B outperforms baselines across 0-32-shot tasks from five tabular benchmarks

Josh Gardner, Juan C. Perdomo, Ludwig Schmidt. ***Large Scale Transfer Learning for Tabular Data via Language Modeling***. NeurIPS 2024.

GTL

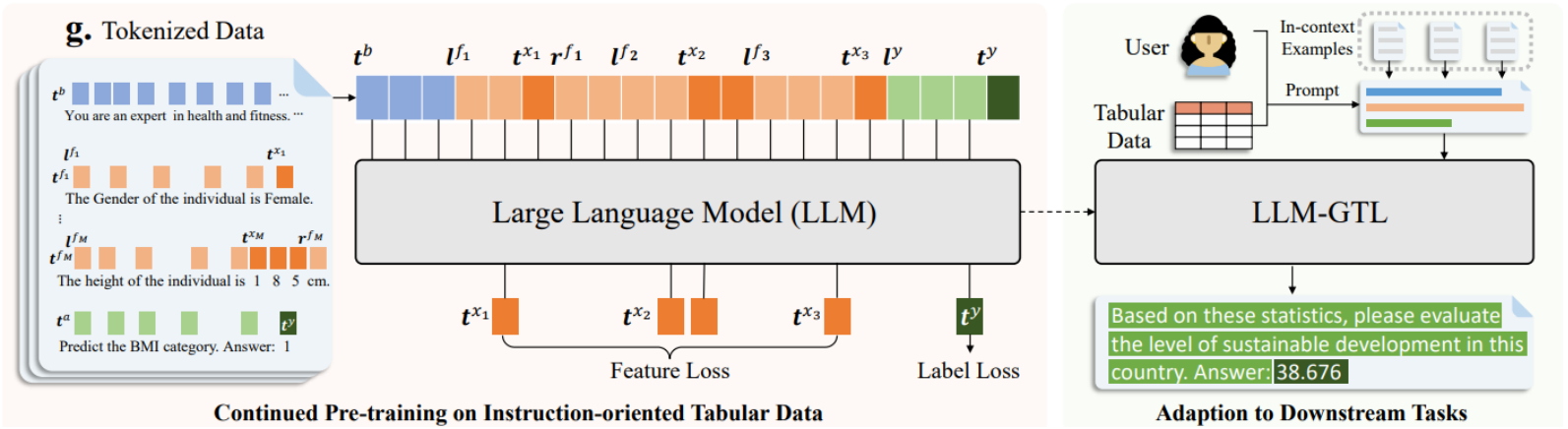
GTL transforms tabular datasets into an instruction-oriented language format, facilitating the continued pre-training of LLMs on instruction-oriented tabular data



Task Instruction

Table

Answer



Xumeng Wen, Han Zhang, Shun Zheng, Wei Xu, Jiang Bian.

From supervised to generative: A novel paradigm for tabular deep learning with large language models. KDD 2024.

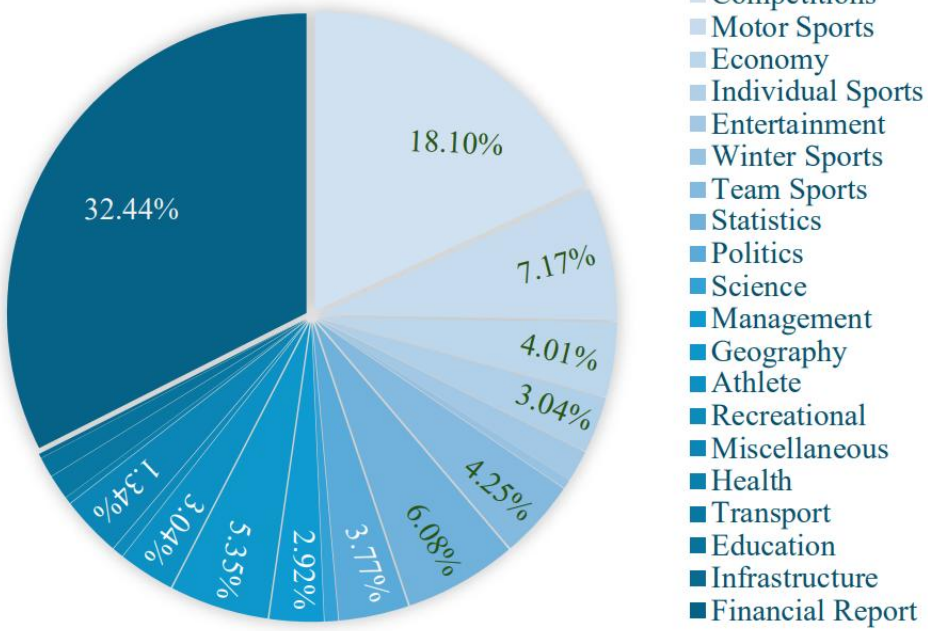
Outline

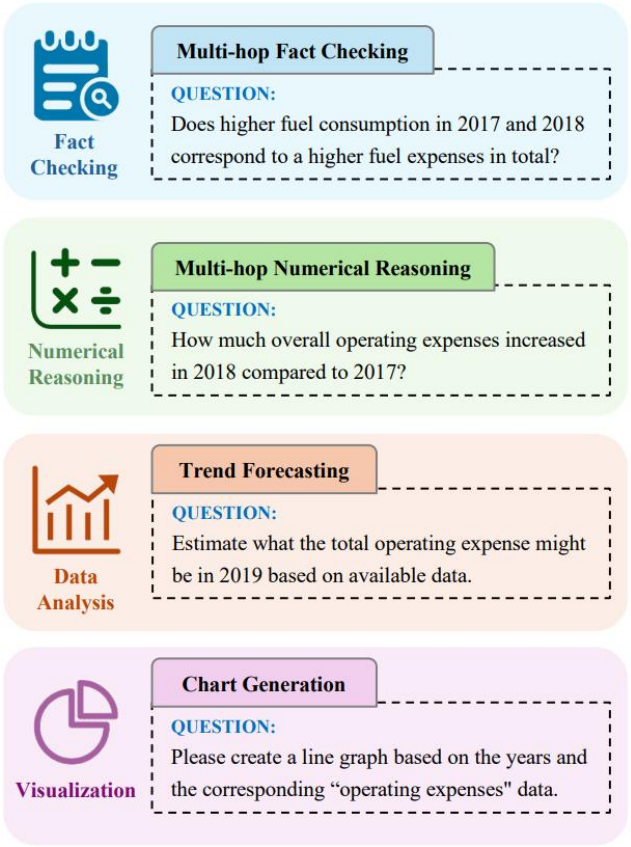
- Tabular Prediction with LLMs
- Tabular QA with LLMs
- From prediction tasks to QA tasks
- Discussions

TableBench

A detailed investigation into the application of tabular data in **industrial scenarios**

A comprehensive and complex benchmark TableBench, including 18 fields within four major categories of table question answering (TableQA) capabilities





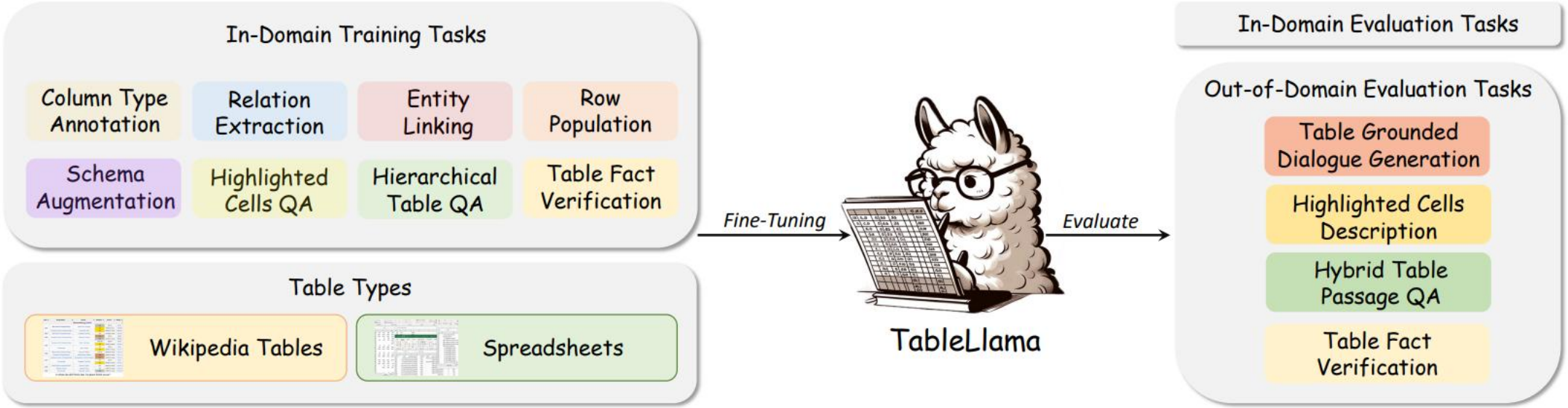
both open-source and proprietary LLMs still have significant room for improvement to meet real-world demands

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, Zhoujun Li. ***TableBench: A Comprehensive and Complex Benchmark for Table Question Answering***. AAAI 2025.

TableLlama

TableInstruct, a new dataset with a variety of realistic tables and tasks, for instruction tuning and evaluating LLMs

TableLlama: fine-tuning Llama 2 (7B) with LongLoRA to address the long context challenge for tables.



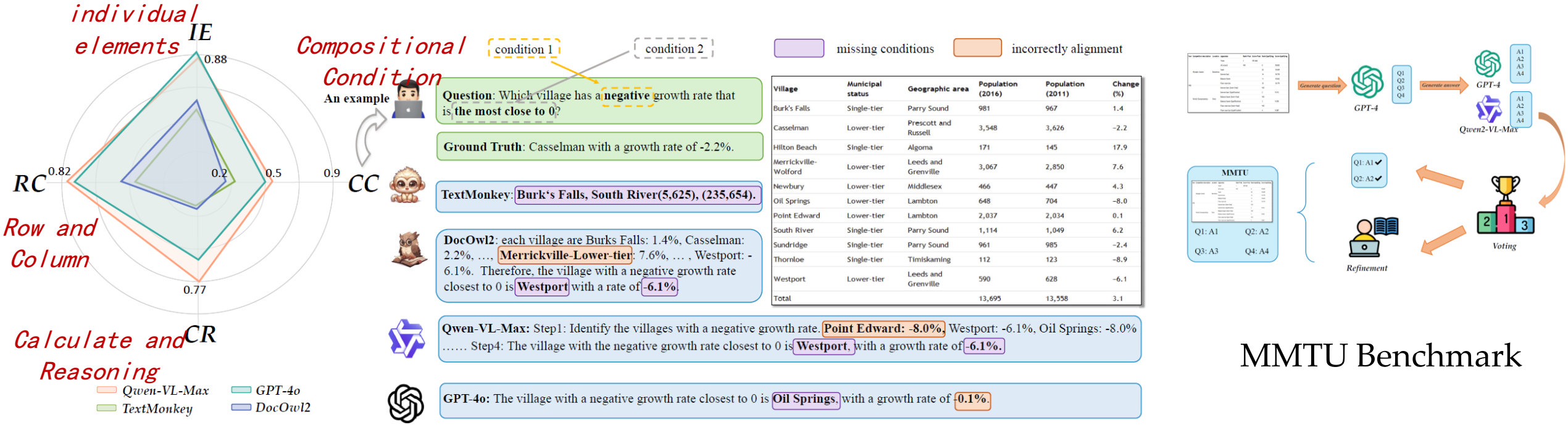
Tianshu Zhang, Xiang Yue, Yifei Li, Huan Sun. ***TableLlama: Towards Open Large Generalist Models for Tables***. CoRR 2023.

MMTU

Massive Multimodal Tabular Understanding (MMTU) benchmark comprehensively assesses the full capabilities of MLLMs in tabular understanding.

Evaluate four key aspects: element, row/column, compositional condition understanding, and basic calculations/reasoning

MLLMs still limited in certain simple scenarios, particularly when handling compositional conditions

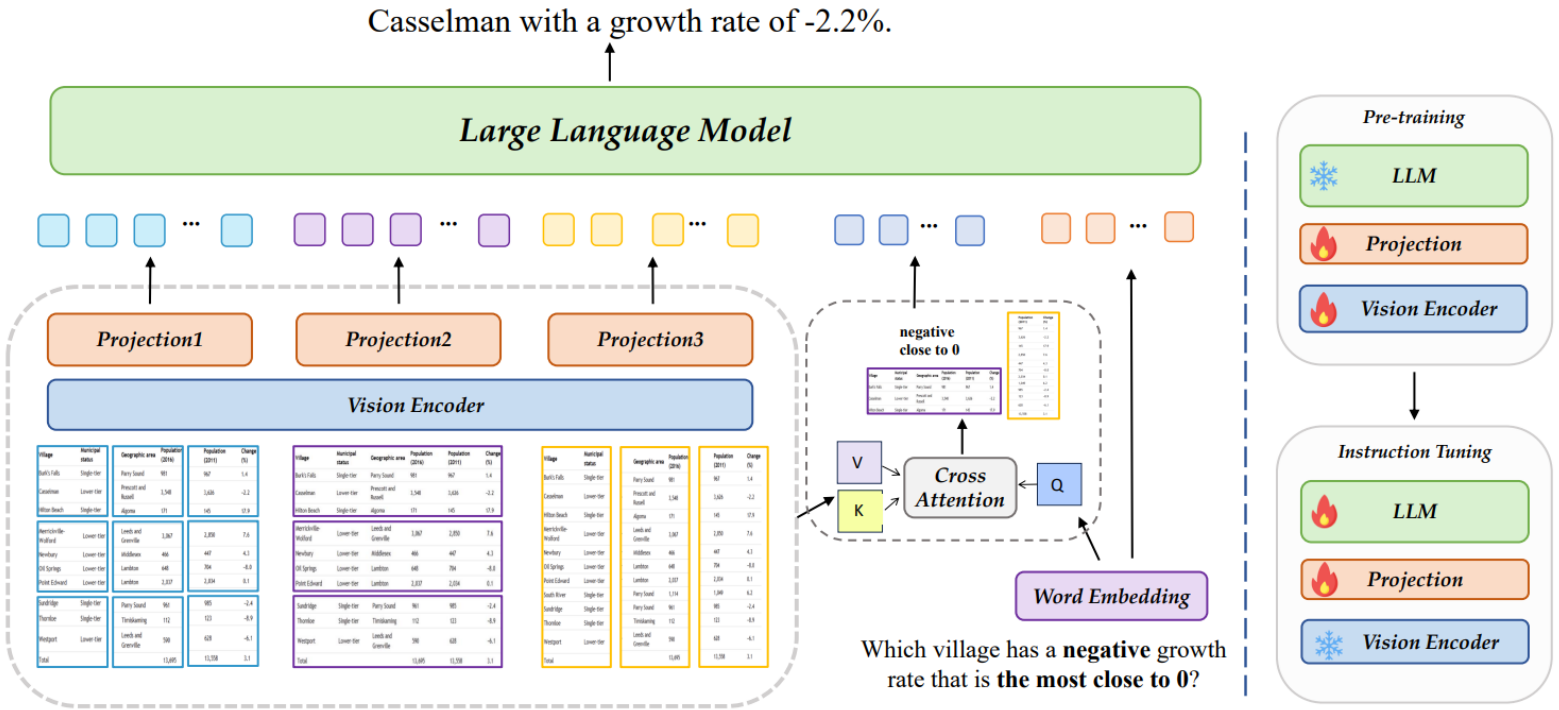


MMTU Benchmark

Jun-Peng Jiang, Tao Zhou, De-Chuan Zhan, Han-Jia Ye. ***Compositional Condition Question Answering in Tabular Understanding***. ICML 2025.

CoCoTab

By introducing **row and column patches** to extract contextual information and combining the **cross-attention mechanism** to align the table structure with the problem semantics, the challenges caused by different patch alignment errors and loss of problem information have been alleviated, and the shortcomings of MLLMs in compositional conditional tasks have been filled.



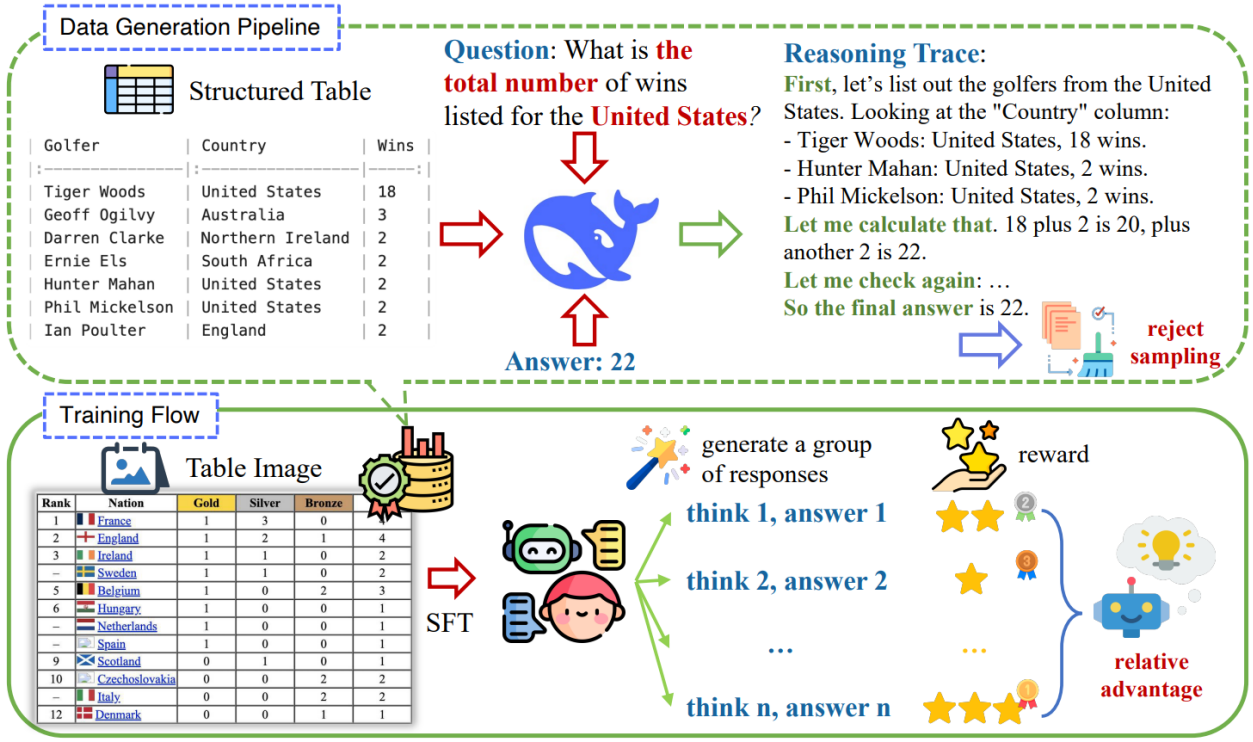
	MMTU			
	IE	RC	CC	CR
LLaVA-1.6-7B (Liu et al., 2024a)	0.50	0.32	0.12	0.06
LLaVA-1.6-13B (Liu et al., 2024a)	0.59	0.38	0.13	0.08
Monkey (Li et al., 2024)	0.39	0.24	0.28	0.06
TextMonkey (Liu et al., 2024c)	0.62	0.36	0.28	0.06
mPlug-Owl (Ye et al., 2023a)	0.11	0.08	0.15	0.06
Docowl (Ye et al., 2023b)	0.65	0.45	0.26	0.07
shareGPT4V (Chen et al., 2025)	0.17	0.09	0.16	0.04
VisCPM (Hu et al., 2023)	0.04	0.03	0.27	0.04
InstructBLIP (Dai et al., 2023)	0.06	0.04	0.08	0.04
Donut (Kim et al., 2021)	0.62	0.14	0.03	0.02
CoCoTab	**0.68**	**0.50**	**0.43**	**0.38**

Jun-Peng Jiang, Tao Zhou, De-Chuan Zhan, Han-Jia Ye. ***Compositional Condition Question Answering in Tabular Understanding***. ICML 2025.

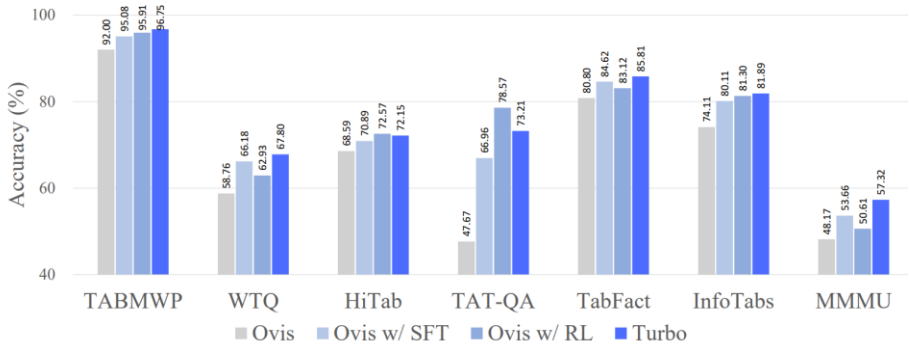
Turbo

Turbo highlight the **practical value** of using structured tables as **privileged information** during training to enhance reasoning over table images, as such structured tables are often unavailable at inference time in real-world scenarios.

Turbo utilizes this to bridge the modality gap and enhance tabular reasoning capabilities of MLLMs.



Method	Question Answering				Fact Verification		MMMU	Average
	TABMWP	WTQ	HiTab	TAT-QA	TabFact	InfoTabs		
InternVL-2.5 [10]	90.88	43.19	45.94	34.97	66.46	55.50	41.46	54.06
Qwen2.5-VL [3]	92.48	65.85	67.09	70.54	83.01	77.91	42.07	71.28
MiniCPM-V-2.6 [75]	83.68	47.97	56.53	51.55	78.48	73.03	31.10	60.33
HIPPO [37]	87.34	55.71	63.13	61.40	82.29	75.70	-	-
HIPPO w/o ST	85.83	49.10	57.23	56.22	80.20	72.74	35.98	62.47
Table-LLaVA [86]	53.20	16.62	7.87	10.49	57.62	66.78	17.68	32.89
TabPedia [84]	10.66	23.53	6.54	13.08	35.49	2.43	2.44	13.45
Ovis2 [39]	92.00	58.76	68.59	47.67	80.80	74.11	48.17	67.16
Ovis2-CoT	92.12	60.80	66.43	48.70	81.61	72.46	50.61	67.53
TURBO	**96.75**	**67.80**	**72.15**	**73.21**	**85.81**	**81.89**	**57.32**	**76.42**



Jun-Peng Jiang, Yu Xia, Hai-Long Sun, Shiyin Lu, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, Han-Jia Ye.
Multimodal Tabular Reasoning with Privileged Structured Information. NeurIPS 2025.

Outline

- Tabular Prediction with LLMs
- Tabular QA with LLMs
- From prediction tasks to QA tasks
- Discussions

Unified Tabular Models

From prediction tasks to QA tasks

I am a **19-year-old female** living in the **southwestern** part of the United States. **I don't smoke** and my Body Mass Index (BMI) is 29.8. I don't have any dependents in my health insurance plan. Currently, my insurance cost is **$1,744.47**. I am considering moving to the **northwest** to change my job. By the time I start the new job, I will be **24 years old**. The cafeteria in the new office is quite famous. I'm worried that my diet might become unhealthy, causing my BMI to rise to around **33.3**. In this case, will my medical expenses decrease?

History Table

age	sex	bmi	children	smoker	region	charges
42	female	29	1	no	southwest	7050.642
37	male	34.1	4	yes	southwest	40182.246
.....						
29	female	25.9	0	no	southwest	3353.284
24	female	33.3	0	no	northwest	?

QA Task

↓

Prediction Task

Answer：The actual medical expenses is estimated to be approximately 2855.43, and it is expected to increase.

Current model: Only conducts **historical table retrieval** and **estimates** based on common sense **logic**, with most outputs falling within a certain range.

Let's **search in the data**... Since there are no matching results, we use a **rough** estimate... Perhaps it will double from $1744 to **around** $3000 or more.

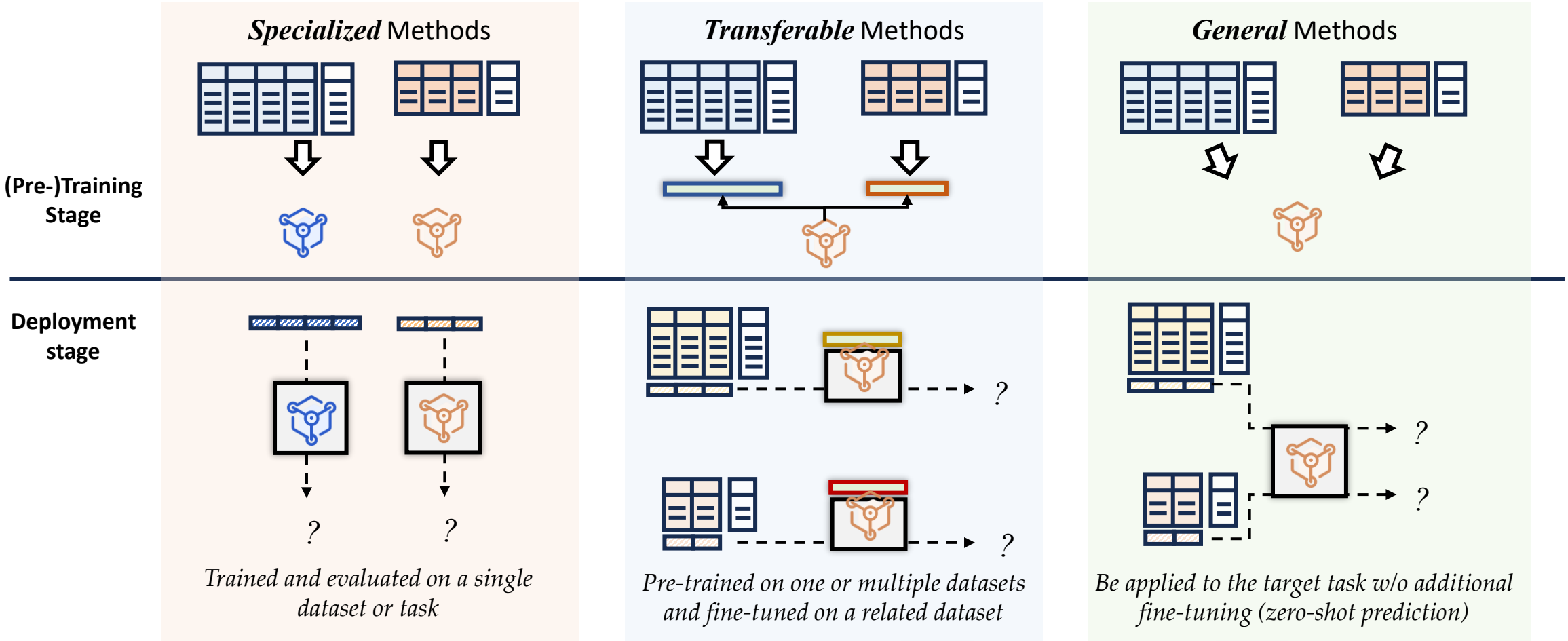
error: 400 InternalError.Algo.InvalidParameter: Range of **input length should be [1, 129024]**...

insensitive to values

have difficulty handling long contexts

An-Yang Ji, Jun-Peng Jiang, De-Chuan Zhan, Han-Jia Ye. ***A Unified Tabular Benchmark for Tabular QA with prediction tasks.*** To Appear 2026.

Tabular Data Survey



https://github.com/LAMDA-Tabular/Tabular-Survey

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, Han-Jia Ye. ***Representation Learning for Tabular Data: A Comprehensive Survey***. TPAMI 2026

Thank you

Q&A

For more discussions, please contact jiangjp@lamda.nju.edu.cn

Tutorial Slides

Tabular Toolbox

Tabular Benchmark

Tabular Survey