



Table Learning w/ LLM

Teacher: Zheng Wang

TA: Weichen Li

Social Network Analysis (NIS8023)

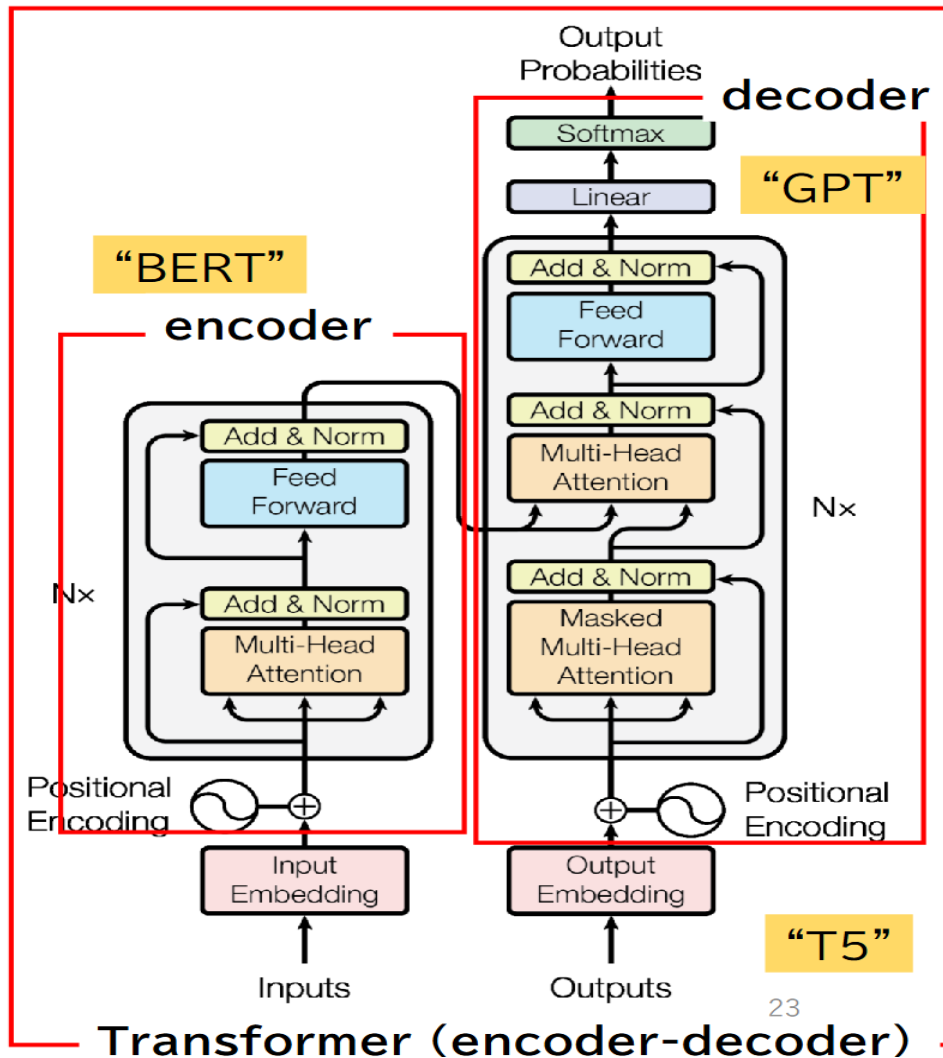
Shanghai Jiao Tong University



Outline

- **Background**
- Classical methods vs. LLM
- Table Learning w/ LLM
 - w/ Finetuned LLM
 - w/o Finetuned LLM

Transformer: The Cornerstone of LLMs



- Self-Attention Mechanism
- Encoder Generates contextualized representations for each input token.
- Decoder Generates the output one sequence token at a time
- Can be used individually or combined together.

LLM in Context Learning



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8.

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4.

(d) Zero-shot-CoT

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

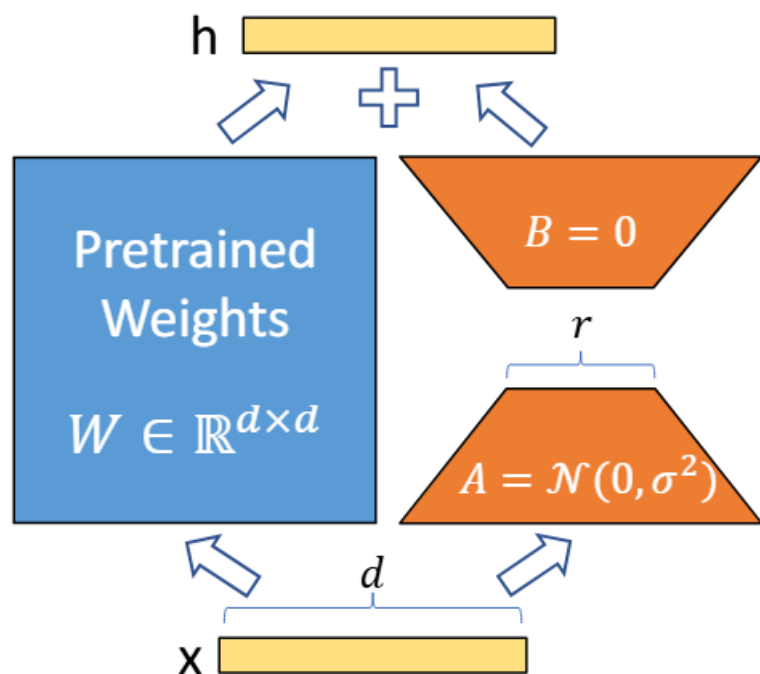
■ Frozen LLMs

■ Zero-Shot

■ Few-Shot

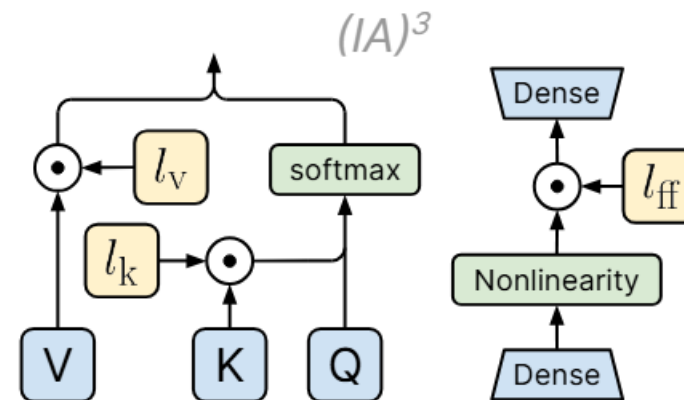
■ Chain-of-Thought

LLM Finetune: PEFT (Parameter-Efficient Fine-Tuning)



LoRA

Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.



IA3

Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965.

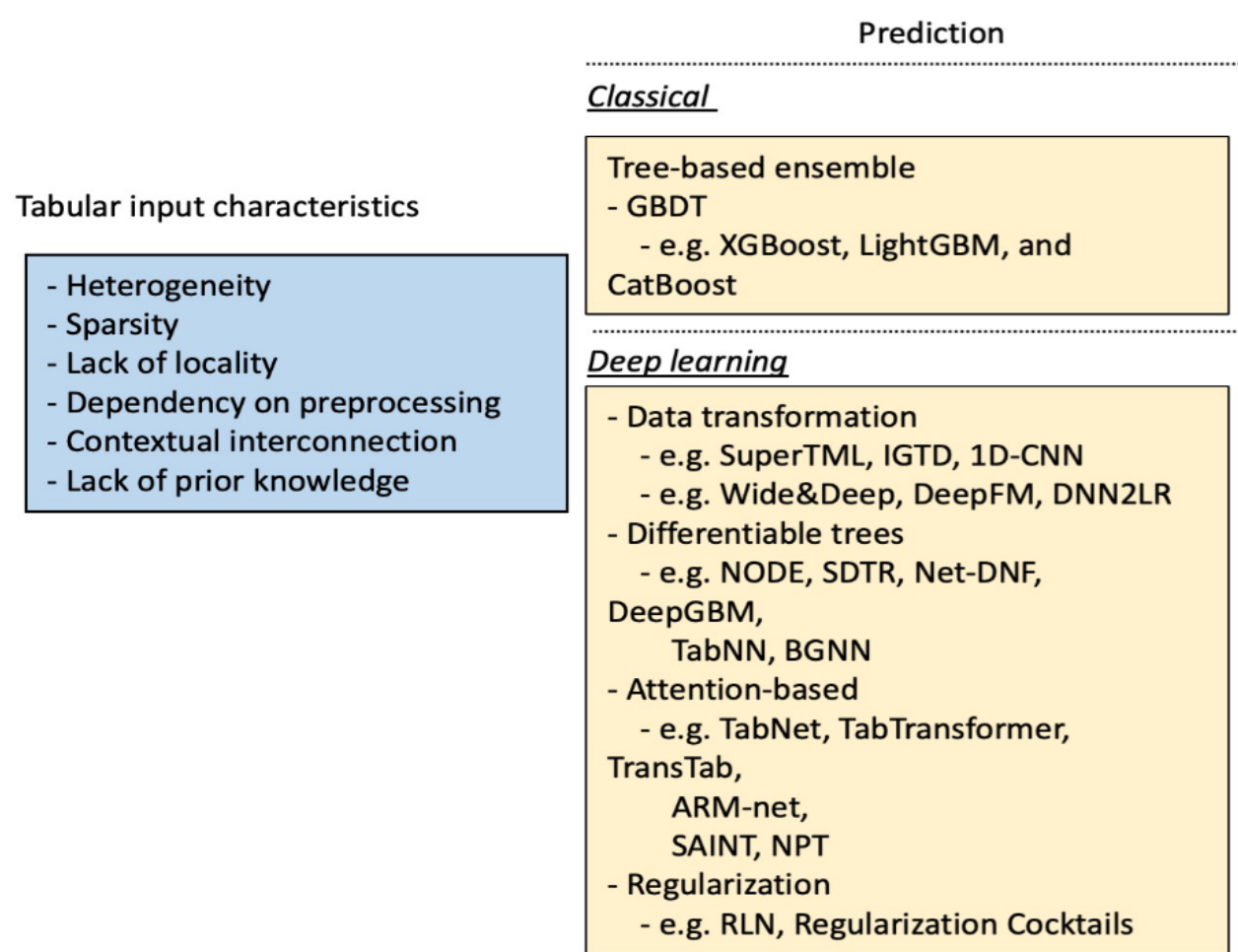
.....



Outline

- Background
- **Classical methods vs. LLM**
- Table Learning w/ LLM
 - w/ Finetuned LLM
 - w/o Finetuned LLM

Classical Methods for Table Learning

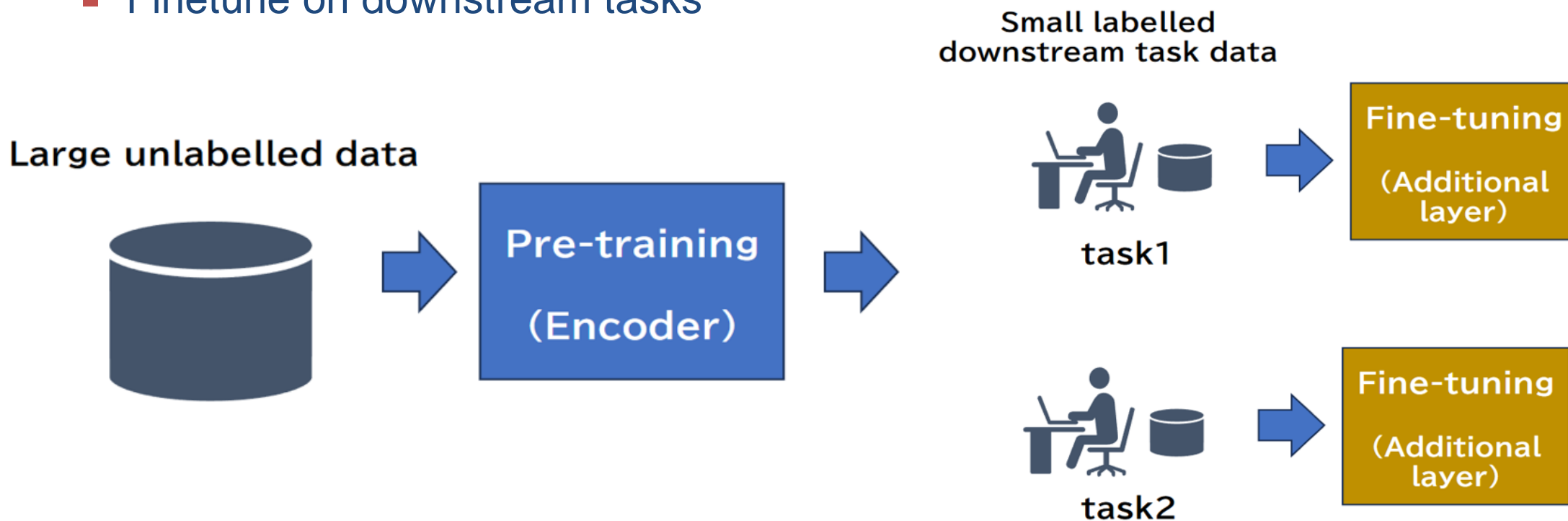


- Only focus on Prediction for simple.
- Data preprocessing requires manual feature engineering.
- Semantic understanding is insufficient.

Encoder For tables



- Pretrain-and-Finetune (“BERT-like”)
 - Learning good table representation (embedding) with table pretraining tasks
 - Finetune on downstream tasks





Why LLM is Promising for Table Learning

Evidence

The Itzulia Basque Country cycling event, held annually in the picturesque and rugged terrain of the ...

Rank	Cyclist	Team	Time
1	Davide Rebellin (ITA)	Gerolsteiner	3'42"
2	David Moncoutié (FRA)	Cofidis	3'56"
...

Question

TableQA

Which country had the most cyclists?

Table Fact Verification

The Spain had the most cyclists finish.

Table-to-Text

Describe the cyclist with the 1st rank.

Text-to-SQL

Show the team of the cyclist whose rank is 1.

...



Table Reasoning

Answer

TableQA

Italy

Table Fact Verification

False

Table-to-Text

Davide Rebellin is a Italy cyclist and...

Text-to-SQL

SELECT Team FROM table WHERE Rank = 1

...

- Tabular data can be easily serialized as text for LLM processing.
- LLM excelling in zero-shot or few-shot scenarios.
- LLM offer strong interpretability and flexibility.



Outline

- Background
- Classical methods vs. LLM
- **Table Learning w/ LLM**
 - w/ Finetuned LLM
 - w/o Finetuned LLM

Overview



- How to serialize table data into text data?
- How to construct an effective prompt?
- Which PEFT method should be selected?

Index	feature1	feature2
1	xxx	xxx
2	xxx	xxx

Serialize

Corpora

Prompt



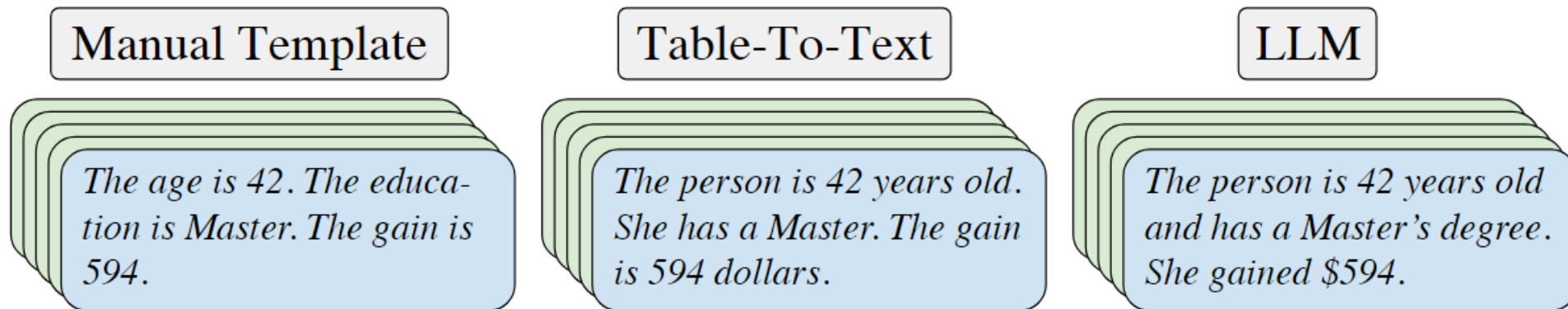
Finetune LLM





TabLLM: Serialize Table Data

- How to serialize table data into text data?
 - Text Template: An textual enumeration of all features as "The column name is value."
 - Table-To-Text: Use a specialized table-to-text generation model.
 - Text LLM: Use an LLM for table-to-text generation.
 - Json format, LaTeX format, Markdown format.....



TabLLM: Construct Prompt



- How to construct an effective prompt?
 - Use dataset-relevant prompt

Bank Dataset:

```
answer_choices: 'No ||| Yes'
jinja: '{{serialization}}

Does this client subscribe to a term
deposit? Yes or no?
Answer:
|||
{{ answer_choices[label] }}'
```

Heart Dataset:

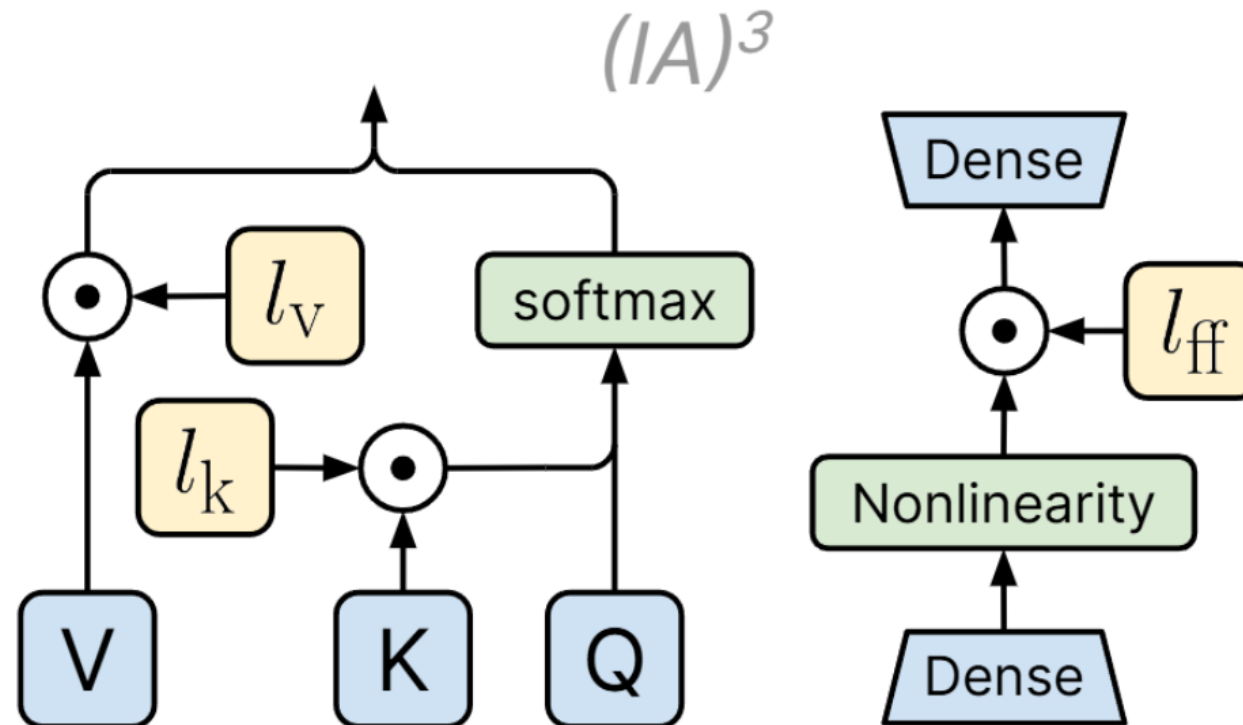
```
answer_choices: 'No ||| Yes'
jinja: '{{serialization}}

Does the coronary angiography of this
patient show a heart disease? Yes or
no?
Answer:
|||
{{ answer_choices[label] }}'
```




TabLLM: Choose a Proper PEFT Method

- Which PEFT method should be selected?
 - Here TabLLM has chosen IA3 for finetuning LLM.



TabLLM Pipeline



1. Tabular data with k labeled rows

age	education	gain	income
39	Bachelor	2174	≤50K
36	HS-grad	0	>50K
64	12th	0	≤50K
29	Doctorate	1086	>50K
42	Master	594	

2. Serialize feature names and values into natural-language string with different methods

Manual Template

The age is 42. The education is Master. The gain is 594.

Table-To-Text

The person is 42 years old. She has a Master. The gain is 594 dollars.

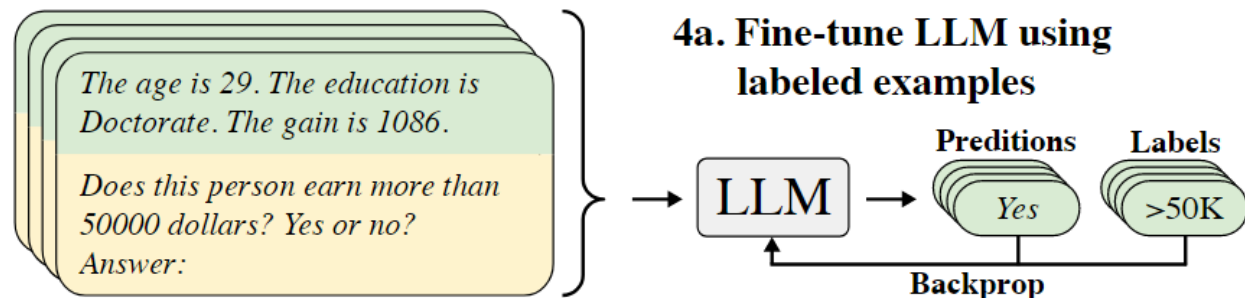
LLM

The person is 42 years old and has a Master's degree. She gained \$594.

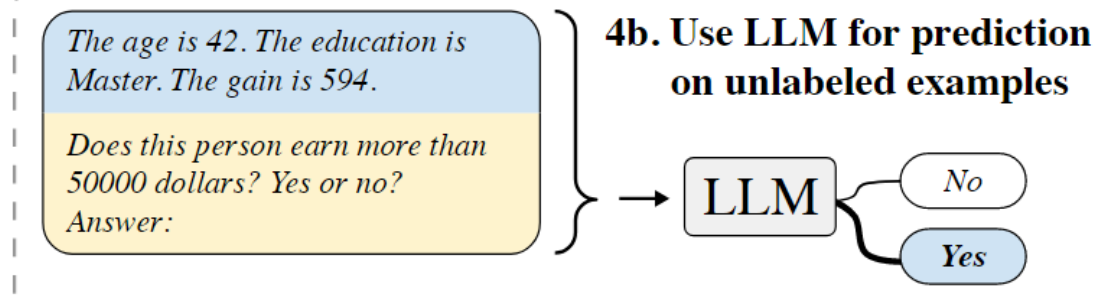
3. Add task-specific prompt

Does this person earn more than 50000 dollars? Yes or no? Answer:

4a. Fine-tune LLM using labeled examples



4b. Use LLM for prediction on unlabeled examples





TableLlama: Serialization & Prompt

(a) Column Type Annotation

1958 Nippon Professional Baseball season

Central League			
Stat	Player	Team	Total
Wins	Masaichi Kaneda	Kokutetsu Swallows	31
Losses	Noboru Akiyama	Taiyo Whales	23
Earned run average	Masaichi Kaneda	Kokutetsu Swallows	1.3
Strikeouts	Masaichi Kaneda	Kokutetsu Swallows	311
Innings pitched	Motoshi Fujita Noboru Akiyama	Yomiuri Giants Taiyo Whales	359

Instruction:

This is a **column type annotation** task. The goal for this task is to choose the correct types for one selected column of the table from the given candidates. The Wikipedia page, ... provide important information for choosing the correct column types.

Input:

[TLE] The Wikipedia page is about 1958 Nippon Professional Baseball season. The Wikipedia section is about Central League. The table caption is Pitching leaders. [TAB] col: | stat | player | ... [SEP] row 1: | Wins | Masaichi Kaneda | ... [SEP] row 2: | Losses | ...

Question:

The column 'player' contains the following entities: <Masaichi Kaneda>, <Noboru Akiyama>, ... The column type candidates are: **tv.tv_producer, astronomy.star_system_body, ...** What are the correct column types for this column (column name: player; entities: <Masaichi Kaneda>, ... , etc)?

Response: sports.pro_athlete, baseball.baseball_player, people.person.

■ Table interpretation

■ Table augmentation

■ Question answering

■ Fact verification

■

Zhang T, Yue X, Li Y, et al. Tablellama: Towards open large generalist models for tables[J]. arXiv preprint arXiv:2311.09206, 2023.

(b) Row Population

NBA Conference Finals

Eastern Conference Finals				
Year	Champion	Coach	Result	Runner-up
1971	Baltimore Bullets	Gene Shue	4-3	New York Knicks



Instruction:

This is a table **row population** task. The goal of this task is to populate the possible entities of the selected column for a table, given the Wikipedia page title, ... You will be given a list of entity candidates. Please rank them so that the most likely entities come first.

Input:

[TLE] The Wikipedia page is about NBA conference finals. The Wikipedia section is about eastern conference finals. The table headers are: | year | champion | ... You need to populate the column: year. [SEED] The seed entity is <1971_NBA_playoffs>.

Question:

The entity candidates are: <2003_NBA_playoffs>, <1982-83_Washington_Bullets_season>, <2004_NBA_playoffs>, <Philadelphia_76ers>, <1983-84_Washington_Bullets_season>, <1952_NBA_playoffs>, ...

Response: <1972_NBA_playoffs>, <1973_NBA_playoffs>, <1974_NBA_playoffs>, <1975_NBA_playoffs>, <1976_NBA_playoffs>, ...

(c) Hierarchical Table QA

Table: Department of defense obligations for research, development, test, and evaluation, by agency: 2015-18

agency	2015	2016	2017	2018
department of defense				
rdt&e	61513.5	69306.1	70866.1	83725
total research	6691.5	7152	7178	7652.7
basic research	2133.4	2238.7	2110.1	2389.9
defense advanced research projects agency				
rdt&e	2815.6	2933.4	2894.5	3018.2
total research	1485	1535.9	1509.4	1680
basic research	359.8	378.1	391.2	458.4

Instruction:

This is a **hierarchical table question answering** task. The goal for this task is to answer the given question based on the given table. The table might be hierarchical.

Input:

[TLE] The table caption is department of defense obligations for research, development, test, and evaluation, by agency: 2015-18. [TAB] | agency | 2015 | 2016 | ... [SEP] | department of defense | department of defense | ... [SEP] | rdt&e | 61513.5 | ... [SEP] | total research | 6691.5 | ... [SEP] | basic research | 2133.4 | ... [SEP] | defense advanced research projects agency | ...

Question:

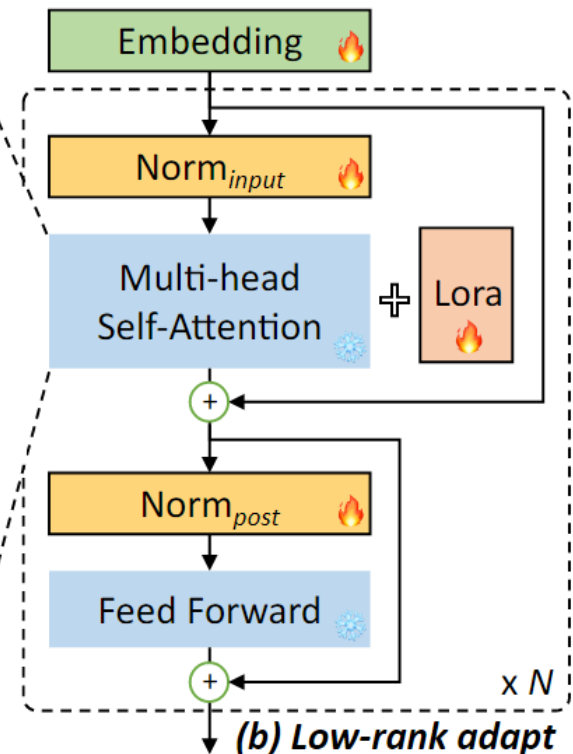
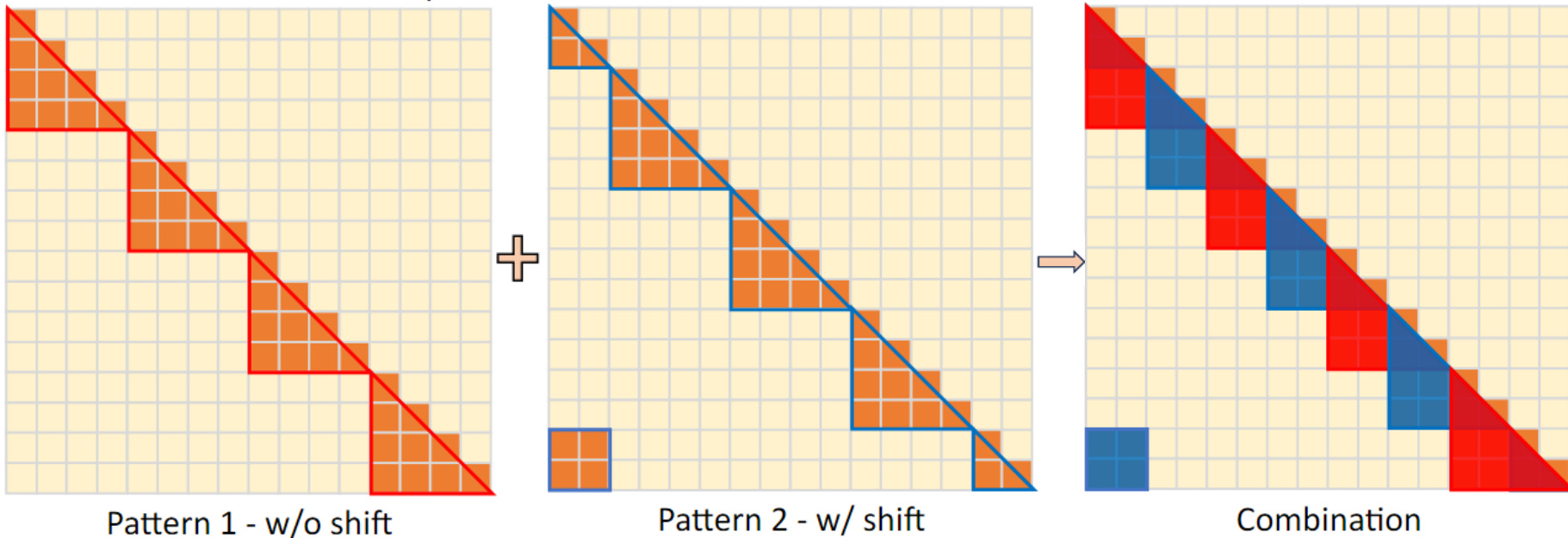
How many dollars are the difference for basic research of defense advanced research projects agency increase between 2016 and 2018?

Response: 80.3.

TableLlama : Choose a Proper PEFT Method

- Which PEFT method should be chosen?
- Here TableLlama has chosen LongLoRA for finetuning LLM.

Each pattern in half heads



TableLlama Pipeline



- In-Domain and Out-of-Domain Evaluation
 - In-Domain: train the generalist table model.
 - Out-of-Domain: test generalization ability.

In-Domain Training Tasks

Column Type
Annotation

Relation
Extraction

Entity
Linking

Row
Population

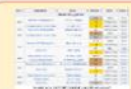
Schema
Augmentation

Highlighted
Cells QA

Hierarchical
Table QA

Table Fact
Verification

Table Types



Wikipedia Tables



Spreadsheets

Fine-Tuning



TableLlama

Evaluate

In-Domain Evaluation Tasks

Out-of-Domain Evaluation Tasks

Table Grounded
Dialogue Generation

Highlighted Cells
Description

Hybrid Table
Passage QA

Table Fact
Verification



Outline

- Background
- Classical methods vs. LLM
- **Table Learning w/ LLM**
 - w/ Finetuned LLM
 - **w/o Finetuned LLM**

Why Non-Finetune?



- High resource consumption: even using fine-tuning techniques, the entire model must be loaded.
- High costs: Fine-tuning models on large-scale tabular data is very costly.
- State-of-the-art LLMs often do not support fine-tuning.



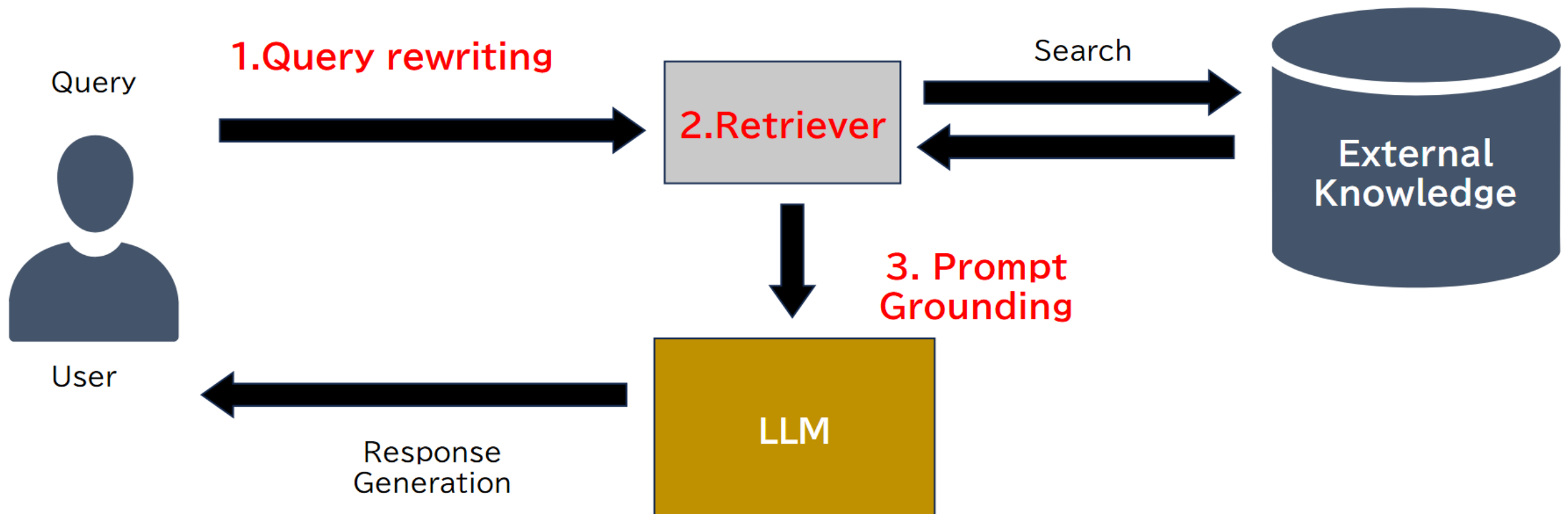
Vs



Retrieval-based Method



- What is RAG (Retrieval-Augmented Generation)?
 - RAG is an efficient way to on-demand get external knowledge



Retrieval-based Method



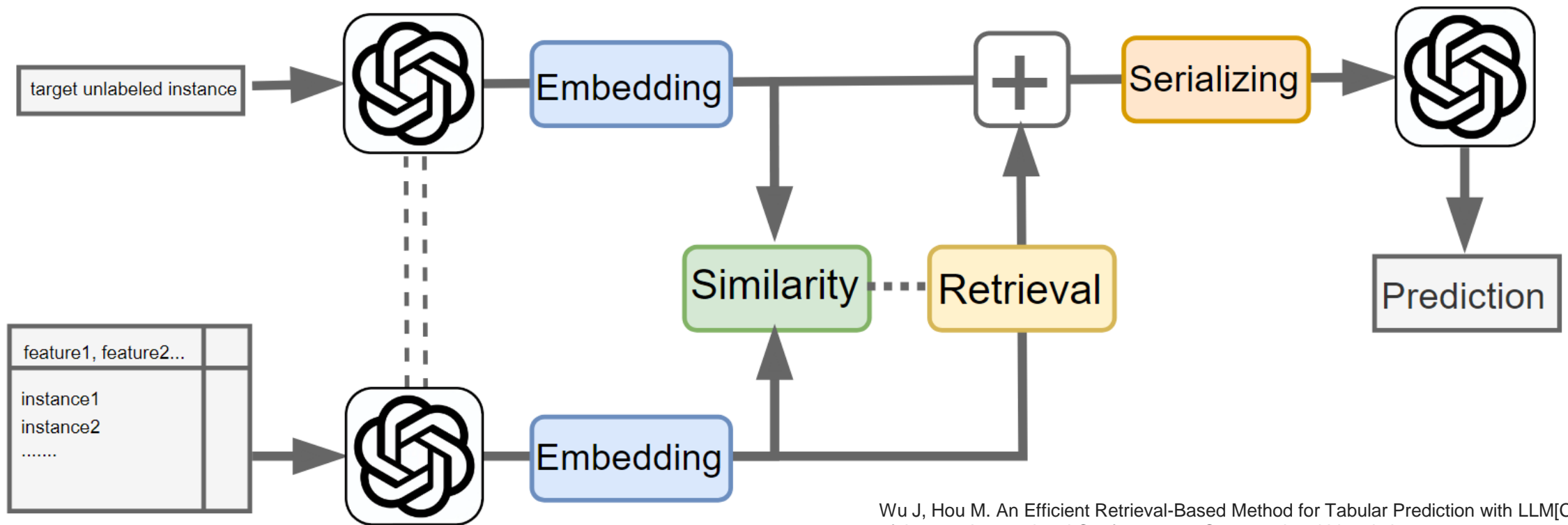
- Intuition: There is some association between certain data within the same table.
- **Knowledge is also in tables!**
- In context learning (ICL) is very important!

Income table				
id	Name	Education	Age	Gain
1	Tom	High school	30	27000\$
2	John	Master	25	89000\$
3	Lily	Master	27	75000\$

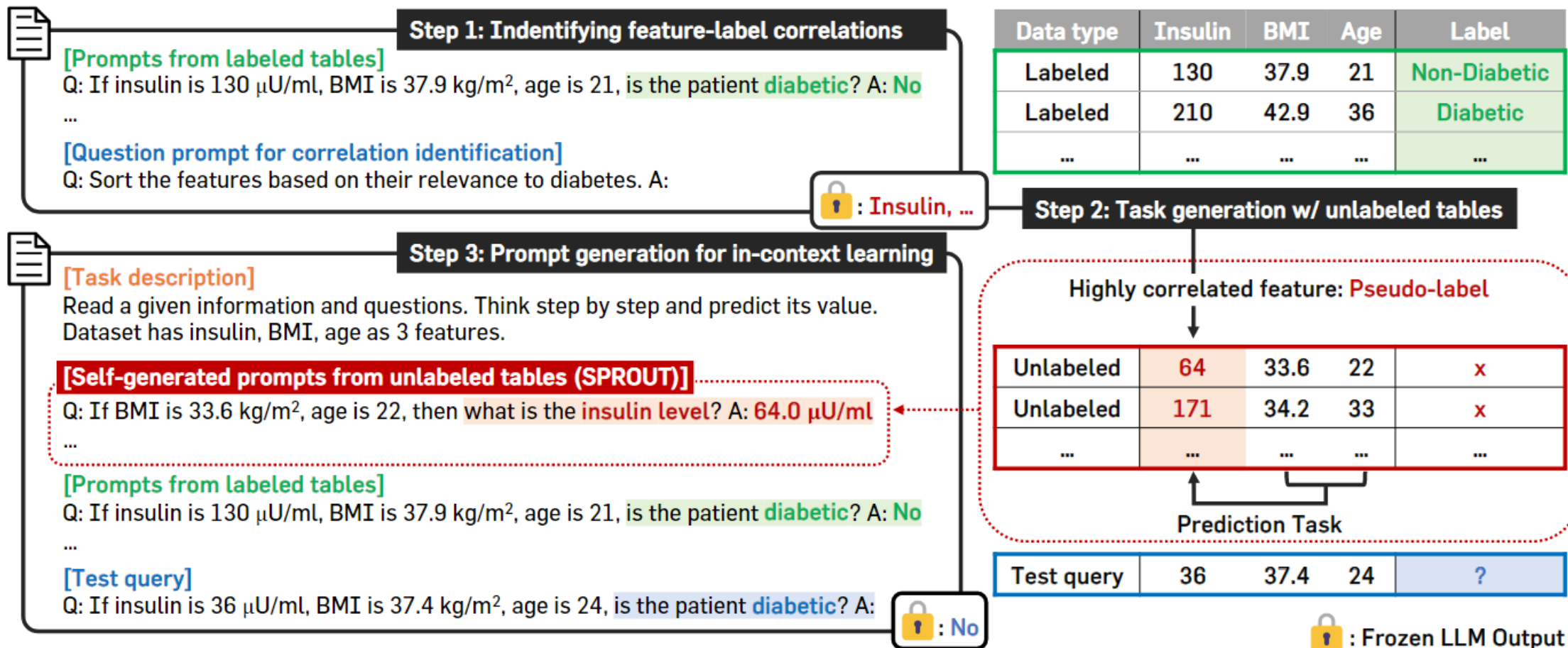
People with the same educational background tend to have similar incomes.

Simple Retrieval Method Example

- Retrieve similar instances.
- Few-shot prompt for LLM prediction.

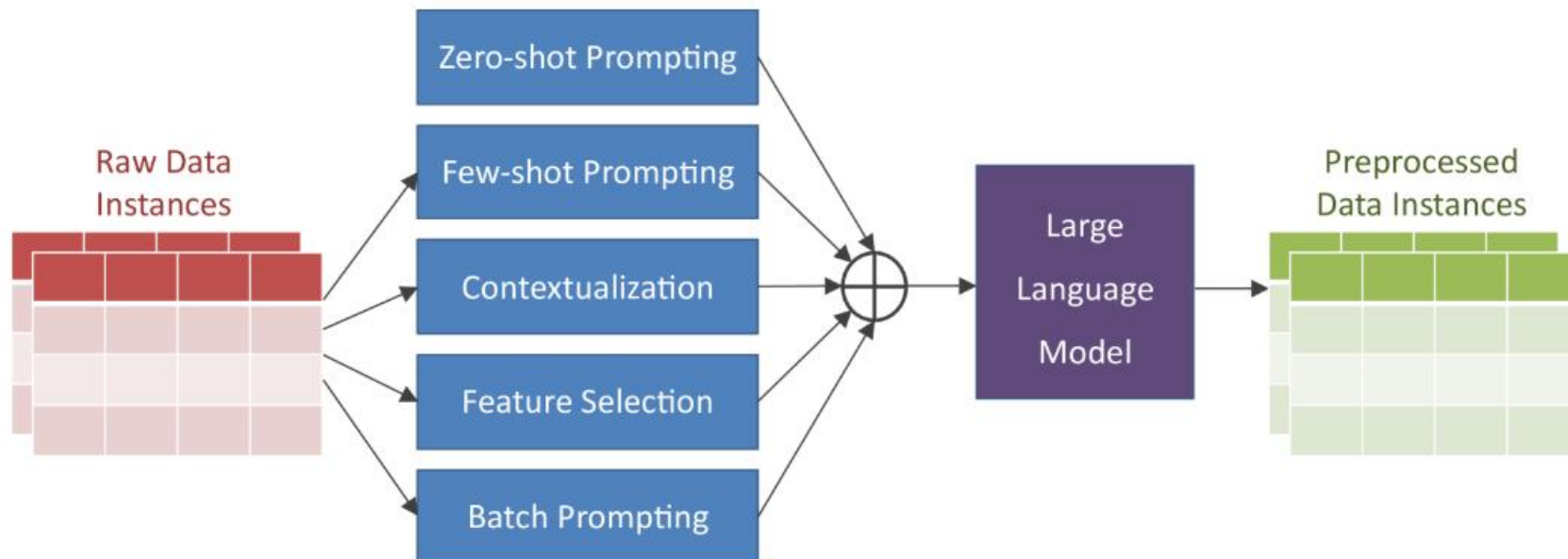


Retrieval Method Example in Limited Labeled Samples



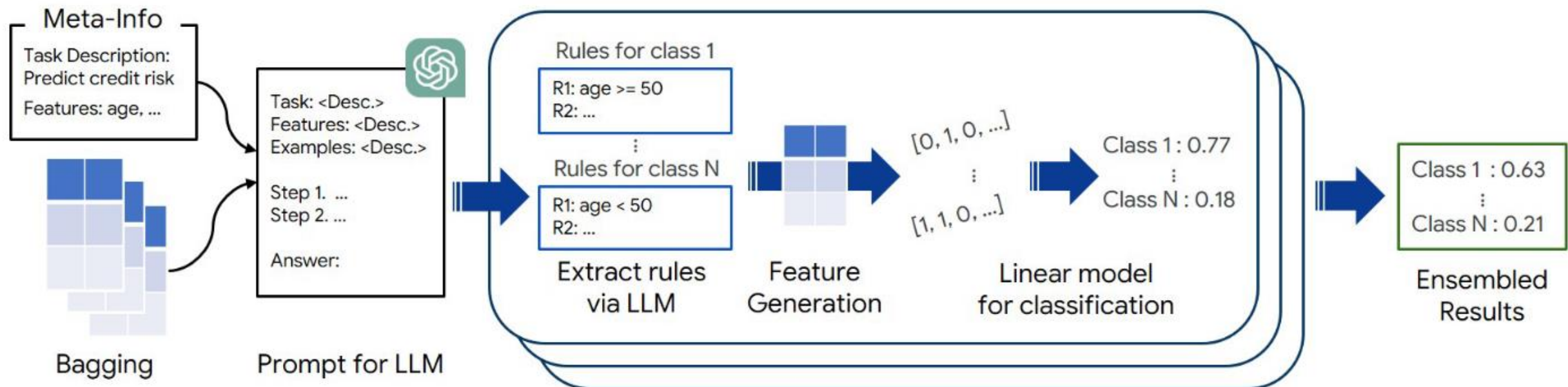
Data Augmentation/Filtering Method

- Intuition: LLM can automatically capture latent patterns and relationships within the data.
- High-quality prompts are required.
- LLM may not be the backbone of the pipeline.



FeatLLM: LLM as Feature Engineer

- LLMs can identify and extract the most relevant features for classification/regression tasks.
- Simple MLP can be used as classifiers and regressors.



Conclusion



- Strengths
 - Extensive knowledge coverage
 - Effective in-context learning and zero-shot capabilities
 - Strong performance in interactive text generation tasks
- Weakness
 - Relatively slow response times
 - High operational cost
 - Limited effectiveness with:
 - Mathematical task
 - Large tables

Challenges



- Can LLMs be integrated with traditional tree-based methods and deep learning approaches?
- Analyzing tabular data row by row with a large model incurs enormous cost, how can this cost be reduced?
- How can relational (multi-table) data be analyzed using LLMs?

Thanks for your time.
QA.

