Tabular Representation, Noisy Operators, and Impacts on Table Structure Understanding Tasks in LLMs

Ananya Singha, *José Cambronero, Sumit Gulwani,*
Vu Le, Chris Parnin

Microsoft

NEURAL INFORMATION PROCESSING SYSTEMS

PROSE

Motivation

In-Context Learning

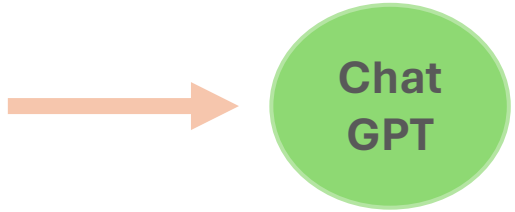
You

You are a helpful python expert. Can you please help me with the pandas program for the Query that I have about the table.

	Name	Age	City	Sex
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M
3	David	28	Boston	M
4	Emily	35	San Francisco	F
5	Frank	29	Dallas	M
6	Grace	27	Miami	F
7	Henry	32	Seattle	M

Generate python code for the query: What is the Age vs sex ration in this dataset?

Python code:



ChatGPT

Certainly! You can use the Pandas library in Python to achieve this. Here's a simple Python code snippet to find the Age vs Sex ratio in the given dataset:

python Copy code

```python
import pandas as pd

# Creating the DataFrame
data = {
    'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Emily', 'Frank', 'Grace', 'Henry'],
    'Age': [25, 30, 22, 28, 35, 29, 27, 32],
    'City': ['New York', 'Los Angeles', 'Chicago', 'Boston', 'San Francisco', 'Dallas
    'Sex': ['F', 'M', 'M', 'M', 'F', 'M', 'F', 'M']
}

df = pd.DataFrame(data)

# Calculating Age vs Sex ratio
age_sex_ratio = df.groupby('Sex')['Age'].mean()

print(age_sex_ratio)
```

This code first creates a DataFrame using the provided data and then uses the `groupby` function to group the data by the 'Sex' column. Finally, it calculates the mean age for each sex using the `mean` function and prints the resulting Age vs Sex ratio.

Motivation

Is this the best way to represent the table?

	Name	Age	City	Sex
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M
3	David	28	Boston	M
4	Emily	35	San Francisco	F
5	Frank	29	Dallas	M
6	Grace	27	Miami	F
7	Henry	32	Seattle	M

Chat GPT

Motivation

Is this the best way to represent the table?

What is the impact of noisy data?

Chat GPT

NOISE

	Name	Age	City	Sex
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M
3	David	28	Boston	M
4	Emily	35	San Francisco	F
5	Frank	29	Dallas	M
6	Grace	27	Miami	F
7	Henry	32	Seattle	M

	Name -----Age -----City	Sex
0	Alice -----25 -----New York	F
1	Bob -----30 -----Los Angeles	M
2	Charlie -----22 -----Chicago	M
3	David -----28 -----Boston	M
4	Emily -----35 -----San Francisco	F
5	Frank -----29 -----Dallas	M
6	Grace -----27 -----Miami	F
7	Henry -----32 -----Seattle	M

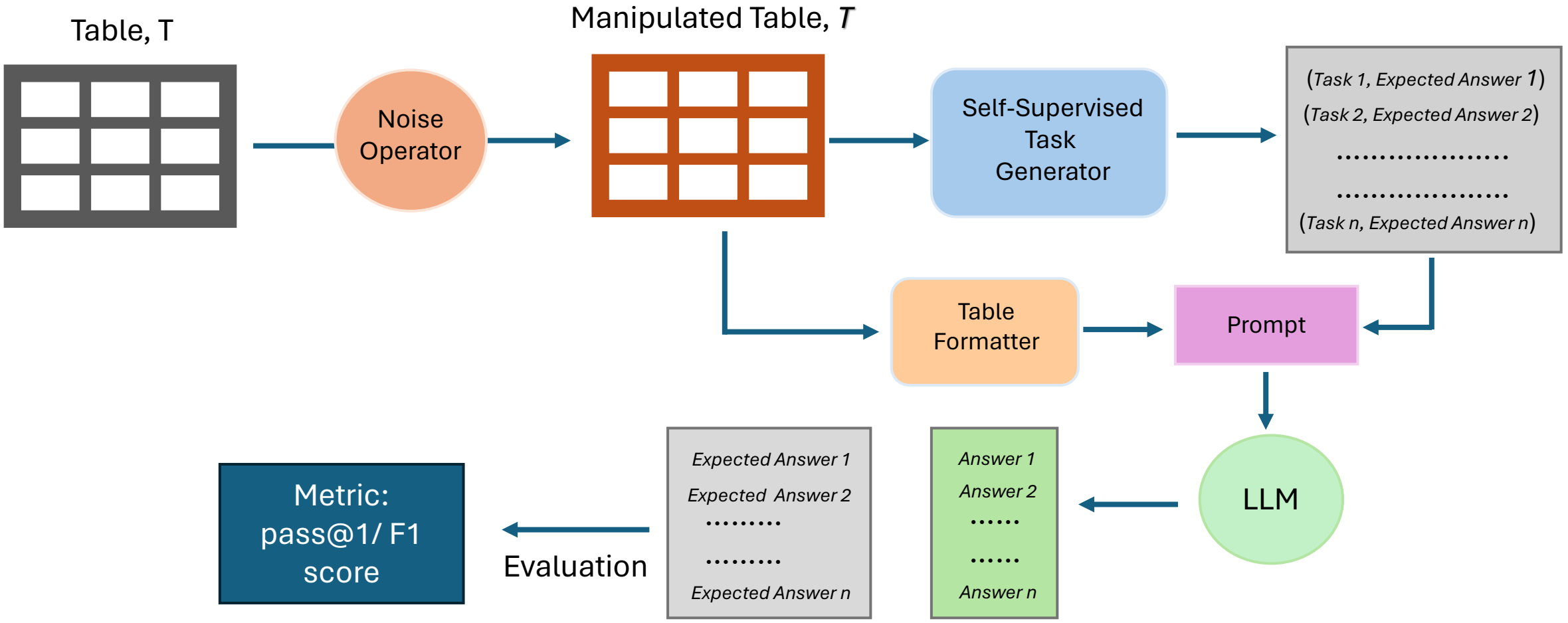
Overview

We evaluate the LLM ability to answer the self supervised structural tasks over

- ✓ Eight Tabular Formats
- ✓ Eight Noise Operation

On Seven popular data-science datasets from Kaggle.

Approach


Table, T
Noise Operator
Manipulated Table, T
Self-Supervised Task Generator
(Task 1, Expected Answer 1)
(Task 2, Expected Answer 2)
.....
.....
(Task n, Expected Answer n)
Table Formatter
Prompt
LLM
Answer 1
Answer 2
.....
.....
Answer n
Expected Answer 1
Expected Answer 2
.....
.....
Expected Answer n
Evaluation
Metric: pass@1/ F1 score

NEURAL INFORMATION PROCESSING SYSTEMS

Tabular Formats

A range of eight popular table representation format used within data-science workflows.

DFLoader Format

```
pd.DataFrame({
        Name : ['Alice', 'Bob', 'Charlie'],
        Age : [25, 30, 22],
        City : ['New York', 'Los Angeles', 'Chicago']
        },
        index=[0, 1, 2])
```

Json Format

```
{
   "0": {"Name":"Alice","Age":25,"City":"NewYork"},
   "1": {"Name":"Bob","Age":30,"City":"Los Angeles"},
   "2":{"Name":"Charlie","Age":22,"City":"Chicago"}
}
```

Data-Matrix Format

```
[
   ['', 'Name', 'Age', 'City'],
   [0, 'Alice', 25, 'New York'],
   [1, 'Bob', 30, 'Los Angeles'],
   [2, 'Charlie', 22, 'Chicago']
]
```

HTML Format

```
<table>
 <thead>
  <tr>
   <th></th>
   <th>Name</th>
   ........
   <th>Sex</th>
  </tr>
 </thead>
 <tbody>
  <tr>
   <th>0</th>
   <td>Alice</td>
    .......
   <td>F</td>
  </tr>
  ......
 </tbody>
</table>
```

Comma Separated Format

```
, Name, Age, City
0, Alice, 25, New York
1, Bob, 30, Los Angeles
2, Charlie, 22, Chicago
```

Tab Separated Format

	Name	Age	City
0	Alice	25	New York
1	Bob	30	Los Angeles
2	Charlie	22	Chicago

Markdown Format

```
|    | Name    |  Age | City          |
|---:|:--------|-----:|:--------------|
|  0 | Alice   |   25 | New York      |
|  1 | Bob     |   30 | Los Angeles   |
|  2 | Charlie |   22 | Chicago       |
```

HTML No Space Format

```
<table><thead><tr><th></th><th>Name</th><th>Age</th>
<th>City</th><th>Sex</th></tr></thead><tbody><tr><th>0
</th><td>Alice</td><td>25</td><td>New York</td><td>F<
/td></tr><tr><th>1</th><td>Bob</td><td>30</td><td>Los
Angeles</td><td>M</td></tr><tr><th>2</th><td>Charlie</
td><td>22</td><td>Chicago</td><td>M</td></tr></tbody>
</table>
```

NEURAL INFORMATION PROCESSING SYSTEMS

Noise Operation

COMBINED COLUMNS

Dirty

Description	Quantit
lotion Benzylbenzoate lotion	0Bottle
Methylated spirit 100ml	0Bottle
susp Magnessium Trisilicate 200ml	0Bottle
Susp. Amoxicillin 125mg/5ml	0Bottle
Susp. Erythromycin 125mg/5ml	0Bottle
Syp Ascorbic acid	0Bottle
Syp Multivite	0Bottle
Syr Albendazole 100mg/5ml	0Bottle
syr Cough Syrup (A) 100ml	
syr Cough Syrup (P) 100ml	
syr Ferric amonium citrate 400mg/5ml	
syr Vitamin B complex	
Syrup Paracetamol 125mg/5ml	
lotion Benzylbenzoate lotion	
Methylated spirit 100ml	
susp Magnessium Trisilicate 200ml	
susp Metronidazole 100mg/5ml	
Susp. Amoxicillin 125mg/5ml	
susp. Ampiclox 125mg/5ml	
Susp. Co-trimoxazole 240mg/5ml	
Susp. Erythromycin 125mg/5ml	

Clean

Description	Quantity	Measure
lotion Benzylbenzoate lotion	0	Bottle
Methylated spirit 100ml	0	Bottle
susp Magnessium Trisilicate 200ml	0	Bottle
Susp. Amoxicillin 125mg/5ml	0	Bottle
Susp. Erythromycin 125mg/5ml	0	Bottle
Syp Ascorbic acid	0	Bottle
Syp Multivite	0	Bottle
Syr Albendazole 100mg/5ml	0	Bottle
rup (A) 100ml	0	Bottle
rup (P) 100ml	0	Bottle
nium citrate 400mg/5ml	0	Bottle
B complex	0	Bottle
tamol 125mg/5ml	0	Bottle
benzoate lotion	0	Bottle
spirit 100ml	0	Bottle
sium Trisilicate 200ml	0	Bottle
dazole 100mg/5ml	0	Bottle
cillin 125mg/5ml	0	Bottle
ox 125mg/5ml	0	Bottle
oxazole 240mg/5ml	0	Bottle
omycin 125mg/5ml	0	Bottle

SERIALIZED ROW

Dirty

Name Hussein Hakeem Address Number 22 Fioye Crescent Surulere Lagos Age 17 Gender Male
Name Arojoye Samuel Address 11 Omolade Close Omole Estate Lagos Age 16 Gender Male
Name Alex Ezurum Address 1 Adamu Lane, Abuja Age 14 Gender Male
Name Susan Nwaimo Address Number 58 Yaba Street, Kaduna State Age 16 Gender Female
Name Ajao Opeyemi Address No12 Olubunmi Street, Abeokuta Age 18 Gender Female
Name Banjoko Adebusola Address 34 Ngige Street, Ugheli, Delta Age 14 Gender Female
Name Muhammed Olabisi Address 13, ICAN road, Enugu Age 12 Gender Female
Name Oluwagbemi Mojisola Address ACCA Lane, Onitsha Age 13 Gender Female

Clean

Name	Address	Age	Gende
Hussein Hakeem	Number 22 Fioye Crescent Surulere Lagos	17	Male
Arojoye Samuel	11 Omolade Close Omole Estate Lagos	16	Male
Alex Ezurum	1 Adamu Lane, Abuja	14	Male
Susan Nwaimo	Number 58 Yaba Street, Kaduna State	16	Femal
Ajao Opeyemi	No12 Olubunmi Street, Abeokuta	18	Femal
Banjoko Adebusola	34 Ngige Street, Ugheli, Delta	14	Femal
Muhammed Olabisi	13, ICAN road, Enugu	12	Femal
Oluwagbemi Mojisola	ACCA Lane, Onitsha	13	Femal

MISSING HEADER

	A	B	C	D
	23	John	Doe	35
	42	Jane	Smith	28
	18	Bob	Johnson	40
	30	Alice	Williams	32
	55	Michael	Brown	45
6	37	Emily	Anderson	29
7	28	David	Miller	38
8	48	Sophia	Thomas	41
9	33	Christophe	White	36
10	25	Olivia	Martin	31
11				
12				

NEURAL INFORMATION PROCESSING SYSTEMS

Spatial Invariance

ShuffleRows

	Name	Age	City	Sex
1	Bob	30	Los Angeles	M
0	Alice	25	New York	F
2	Charlie	22	Chicago	M

ShuffleColumns

	Age	Sex	Name	City
0	25	F	Alice	New York
1	30	M	Bob	Los Angeles
2	22	M	Charlie	Chicago

TransposeTable

	0	1	2
Name	Alice	Bob	Charlie
Age	25	30	22
City	New York	Los Angeles	Chicago
Sex	F	M	M

Header Manipulation

SequentialColumnNames

	col_0	col_1	col_2	col_3
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M

ShuffleColumnNames

	Age	Sex	Name	City
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M

ArbitraryColumnNames

	0MSTtV	ViJ7d2Em	mXjqq0	xqFNEY1YnB
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M

Semi-Structured Content

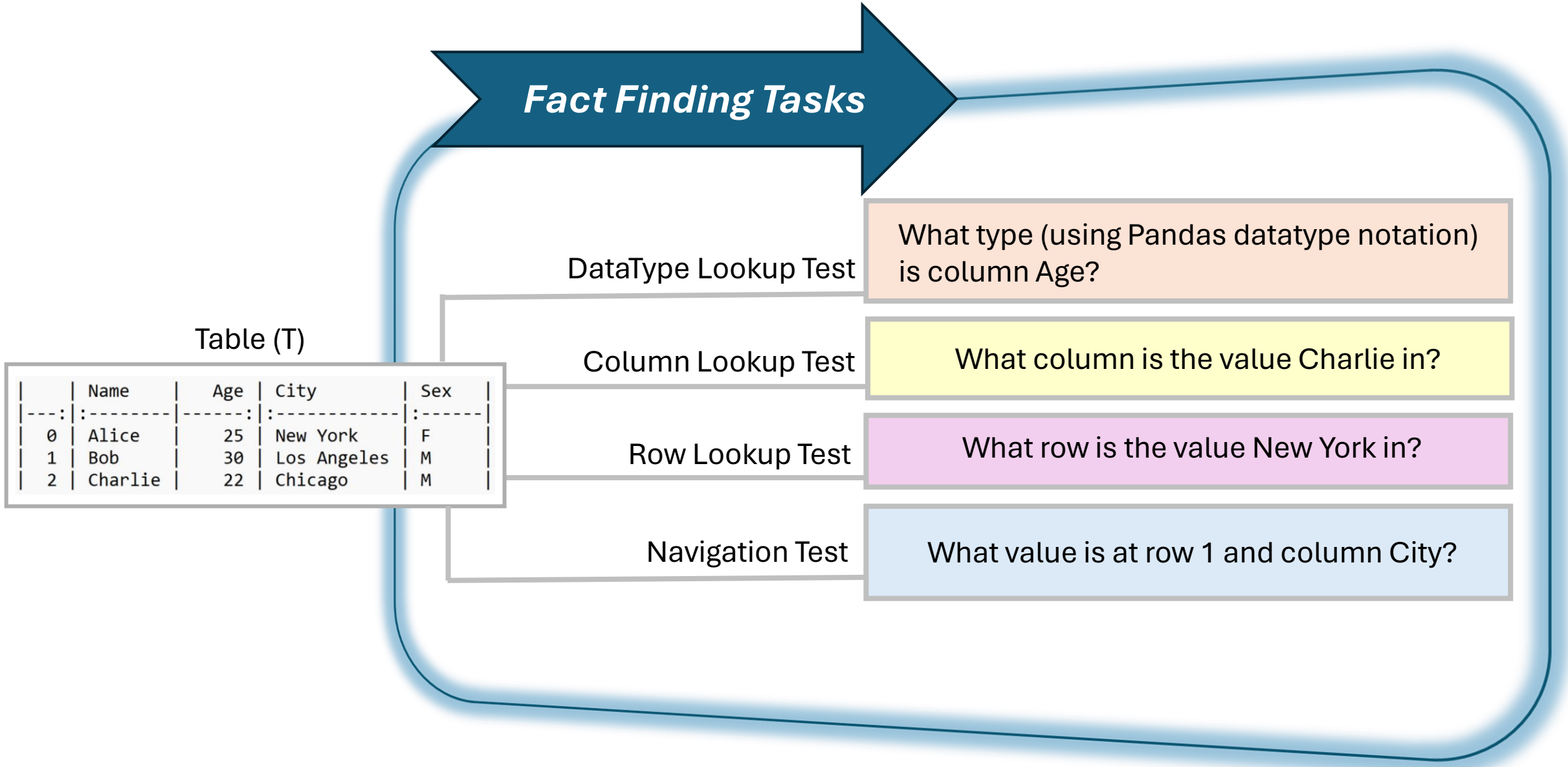
SerializeTable

```
Name:Alice,Age:25,City:New York,Sex:F
Name:Bob,Age:30,City:Los Angeles,Sex:M
Name:Charlie,Age:22,City:Chicago,Sex:M
```

ColumnMerger

	Name-----Age-----City	Sex
0	Alice-----25-----New York	F
1	Bob-----30-----Los Angeles	M
2	Charlie-----22-----Chicago	M

Self-Supervised Structural Tasks



Self-Supervised Structural Tasks

Table Transformation Tasks

Table (T)

	Name	Age	City	Sex
0	Alice	25	New York	F
1	Bob	30	Los Angeles	M
2	Charlie	22	Chicago	M

Table Transpose Test

Can you transpose the table?

Table Reconstruction Test

Can you reconstruct the table by deserializing the table above?

Table Column Reorder Test

Can you reorder the table such that the column are in this new order ['Sex', 'Name', 'Age', 'City']?

Results: Impact of Formats

Table 1: Average Pass@1 for fact-finding tasks. DFLoader provides overall high pass@1 performance.

Table Formats	ColumnLookupTests	DataTypeLookupTests	NavigationTests	RowLookupTests	Overall
CommaSeparated	64.43	95.00	65.57	78.14	75.78
DFLoader	72.71	95.29	68.29	82.86	**79.79**
DataMatrix	62.57	84.00	56.57	**87.43**	72.64
Json	65.00	**96.43**	**71.43**	78.86	77.93
Markdown	61.43	85.86	48.71	73.29	67.32
TabSeparated	67.00	94.00	64.43	78.14	75.8
HTML	**79.83**	94.67	58.83	52.33	71.4
HTMLNoSpace	73.00	93.50	62.00	59.50	72.00

Results: Impact of Formats

Table 1: Average Pass@1 for fact-finding tasks. DFLoader provides overall high pass@1

Table Formats	ColumnLookupTests	DataTypeLookupTests	NavigationTests	RowLookupTests	Ov
[illegible]	[illegible]	95.00	65.57	78.14	75.78
[illegible]	[illegible]	95.29	68.29	82.86	**79.79**
DataMatrix	62.57	84.00	56.57	**87.43**	72.64
Json	65.00	**96.43**	**71.43**	78.86	77.93
Markdown	61.43	85.86	48.71	73.29	67.32
TabSeparated	67.00	94.00	64.43	78.14	75.8
HTML	**79.83**	94.67	58.83	52.33	71.4
HTMLNoSpace	73.00	93.50	62.00	59.50	72.00

Markdown (-12.47%) < DF Loader

Results: Impact of Formats

Markdown (-12.47%) < DF Loader

NavigationTests

Column NavigationTests

Row 3

Table Formats	ColumnLookupTests	DataTypeLookupTests	NavigationTests	RowLookupTests	Overall
CommaSeparated	64.43	95.00	65.57	78.14	75.78
DFLoader	72.71	95.29	68.29	82.86	**79.79**
DataMatrix	62.57	84.00	56.57	**87.43**	72.64
Json	65.00	**96.43**	**71.43**	78.86	77.93
Markdown	61.43	85.86	48.71	73.29	67.32
TabSeparated	67.00	94.00	64.43	78.14	75.8
HTML	**79.83**	94.67	58.83	52.33	71.4
HTMLNoSpace	73.00	93.50	62.00	59.50	72.00

Table 1: Average Pass@1 for fact-finding tasks. DFLoader provides overall high pass@1

Table Formats	ColumnLookupTests	DataTypeLookupTests	NavigationTests	RowLookupTests	Ov
[illegible]	[illegible]	95.00	65.57	78.14	75.78
DFFormat	72.71	95.29	68.29	82.86	**79.79**
DataMatrix	62.57	84.00	56.57	**87.43**	72.64
Json	65.00	**96.43**	**71.43**	78.86	77.93
Markdown	61.43	85.86	48.71	73.29	67.32
TabSeparated	67.00	94.00	64.43	78.14	75.8
HTML	**79.83**	94.67	58.83	52.33	71.4
HTMLNoSpace	73.00	93.50	62.00	59.50	72.00

Results: Impact of Formats

Overall

Markdown (-12.47%) < DF Loader

Navigation Tests

Json (+5.86%) > CommaSeparated

Results: Impact of Formats

Overall

Markdown (-12.47%) < DF Loader

Table 2: F1 scores for transformation tasks. DFLoader and JSON format, with structural element isolation and repetition, enable high performance on average across transformation tasks.

Table	TableColumnReorderTests	TableReconstructionTests	TableTransposeTests	Overall
CommaSeparated	95.33	74.33	99.00	89.55
DFLoader	99.33	**98.00**	98.33	**98.55**
DataMatrix	92.67	90.67	0.00	61.11
Json	**99.67**	85.00	**100.00**	94.89
Markdown	50.00	24.33	34.00	36.11
TabSeparated	93.33	92.33	50.00	78.55
HTML	50.00	86.00	83.33	73.11
HTMLNoSpace	83.33	84.00	83.33	83.55

Results: Impact of Formats

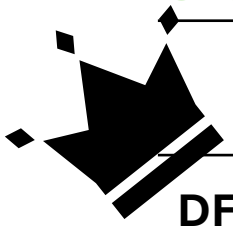
Overall

Markdown (-12.47%) < DF Loader

Navigation Tests

Json (+5.86%) > CommaSeparated

Overall



	Fact-Finding Tasks	Table-Transformation Tasks
DF Loader	79.79%	98.55%

Results: Impact of Formats

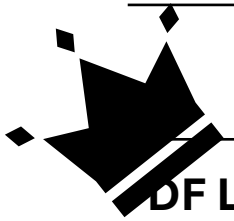
Overall

Markdown (-12.47%) < DF Loader

Navigation Tests

Json (+5.86%) > CommaSeparated

Overall



	Fact-Finding Tasks	Table-Transformation Tasks
DF Loader	79.79%	98.55%
Json	77.93%	94.89%

Results: Impact of Formats

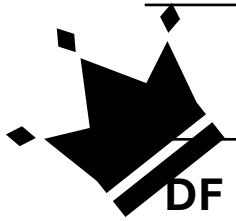
Overall

Markdown (-12.47%) < DF Loader

Navigation Tests

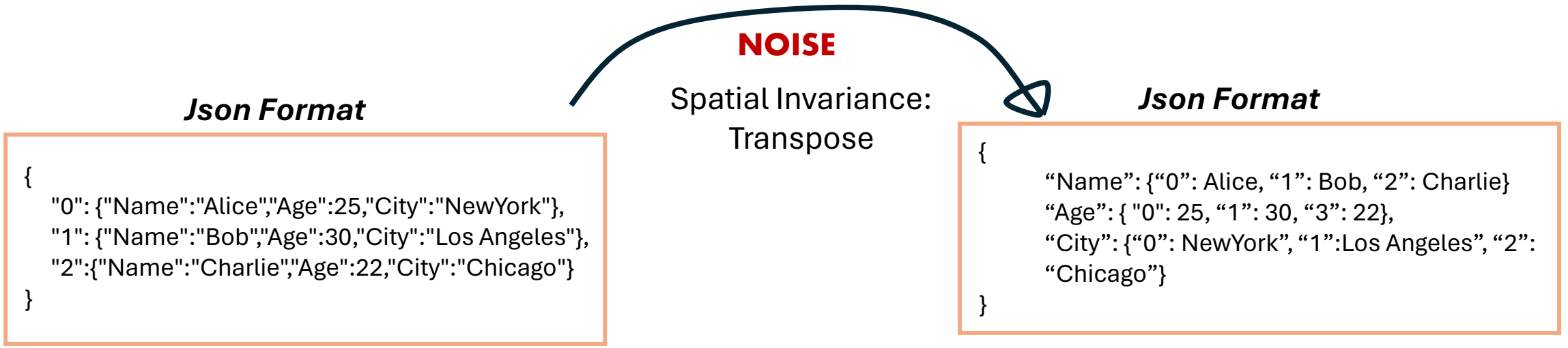
Json (+5.86%) > CommaSeparated

Overall



	Fact-Finding Tasks	Table-Transformation Tasks
DF Loader	79.79%	98.55%
Json	77.93%	94.89%

Result: Impact of Noise



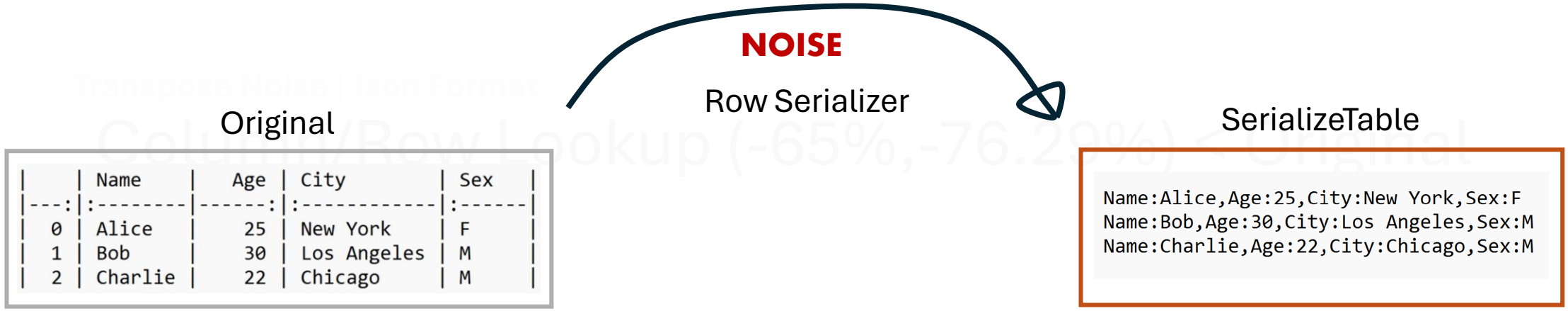
Column/Row Lookup (-65%,-76.29%)< Original

Result: Impact of Noise

Transpose Noise | Json Format

Column/Row Lookup (-65%,-76.29%) < Original

Result: Impact of Noise



DataType Lookup (-12.8%) < Original

Result: Impact of Noise

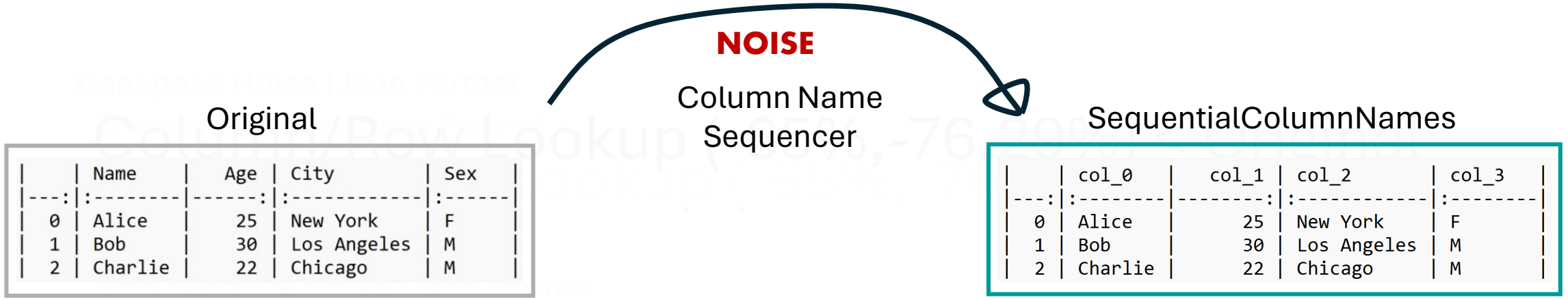
Transpose Noise | Json Format

Column/Row Lookup (-65%,-76.29%) < Original

Row Serialize Noise | Json Format

DataType Lookup (-12.8%) < Original

Result: Impact of Noise



Column Reorder (-67.33%) < Original

Result: Impact of Noise

Transpose Noise | Json Format

Column/Row Lookup (-65%,-76.29%) < Original

Row Serialize Noise | Json Format

DataType Lookup (-12.8%) < Original

Column Name Sequencing Noise | Comma Separated Format

Column Reorder (-67.33%) < Original

Future Directions

Impact of Format and Noise's on End-to-end Tasks

Cross-LLM performance

Reach Out For Questions

Ananya Singha
Microsoft
ananyasingha2000@gmail.com
[GitHub/LinkedIn/Twitter]

Full Paper: Tabular Representation, Noisy Operators, and Impacts on Table Structure Understanding Tasks in LLMs

Code: prose/misc/TRL-neurips-2023 at main · microsoft/prose (github.com)

PROSE

PROSE - Microsoft Research

NEURAL INFORMATION PROCESSING SYSTEMS

Microsoft