

# TabXEval



## Why this is a Bad Table? An eXhaustive Rubric for Table Evaluation

\*Vihang Pancholi<sup>1</sup>, \*Jainit Bafna<sup>2</sup>, \*Tejas Anvekar<sup>1</sup>,  
Manish Shrivastava<sup>2</sup>, Vivek Gupta<sup>1</sup>

<sup>1</sup>Arizona State University, <sup>2</sup>IIIT Hyderabad



\* Equal Contribution

# Why Tables—and Table Generation—Matter?

The collage displays four distinct tables used in business operations:

- Employee Timesheet:** A table for tracking employee hours. It includes fields for Company Name, Employee Name, Supervisor Name, and Week of. The main table has columns for Day of Week, Regular [h:mm], Overtime [h:mm], Sick [h:mm], Vacation [h:mm], and Holiday [h:mm].
- INVOICE:** A table for billing services. It includes fields for LOGO, INVOICE #, DATE, CLIENT, and PLEASE MAKE PAYMENT TO. The main table has columns for DATE, DESCRIPTION, HOURS, and AMOUNT.
- Break-Even Analysis:** A table for financial analysis. It includes fields for Proposed Product, For the Period, Selling Price (P), Break-Even Units (X), and Break-Even Sales (S). The main table has columns for Fixed Costs, Variable Costs, and Total Fixed Costs (TFC).
- Project Management:** A table for project scheduling. It includes fields for Project Start Date, Project Manager, and Display Weeks. The main table has columns for Task Category, Lead, Start, End, Step, and Work.

- **Ubiquitous:** Spreadsheets & dashboards power every sector.
- **Decision-critical:** One row/col error can upend budgets & diagnoses.
- **LLM workflows:** AI now auto-creates tables for insight delivery.
- **High-stakes fidelity:** Tables need exact values—small slips echo widely.
- **Trust & transparency:** Strong table-eval keeps data-driven AI safe.

# Are Tables same as Text?

Film		Pre-nomination (before Jan. 14)	Post-nomination (Jan. 14 – Feb. 28)	Total	
The Martian		\$226.6 mil	\$1.8 mil	\$228.4 mil	
The Revenant					
Mad Max: Fury Road					
Bridge of Spies					
The Big Short					
Blocks	Player	Description	Due Date	2016	2015
2	Harrison	5.00%	09/20	599	599
	Barnes	4.75%	12/45	514.9	598
	Draymond	3.50%	06/24	597	597
	Green	4.60%	06/44	549	549
	Andrew	2.875%	05/26	605	545
	Bogut	8.205%	01/27	521	521
	Klay	3.125%	05/16	500	500
	Thompson	2.80%	01/21	514.9	399
	Stephen	4.00%	11/23	349	349
	Curry	6.25%	09/40	298	298
	Andre	4.76%	03/18	322	271
	Iguodala	4.45%	05/43	248	249
	Marreese	4.25%	12/42	196	196
	Speights	3.50%	09/15	599	573.3
	Shaun				

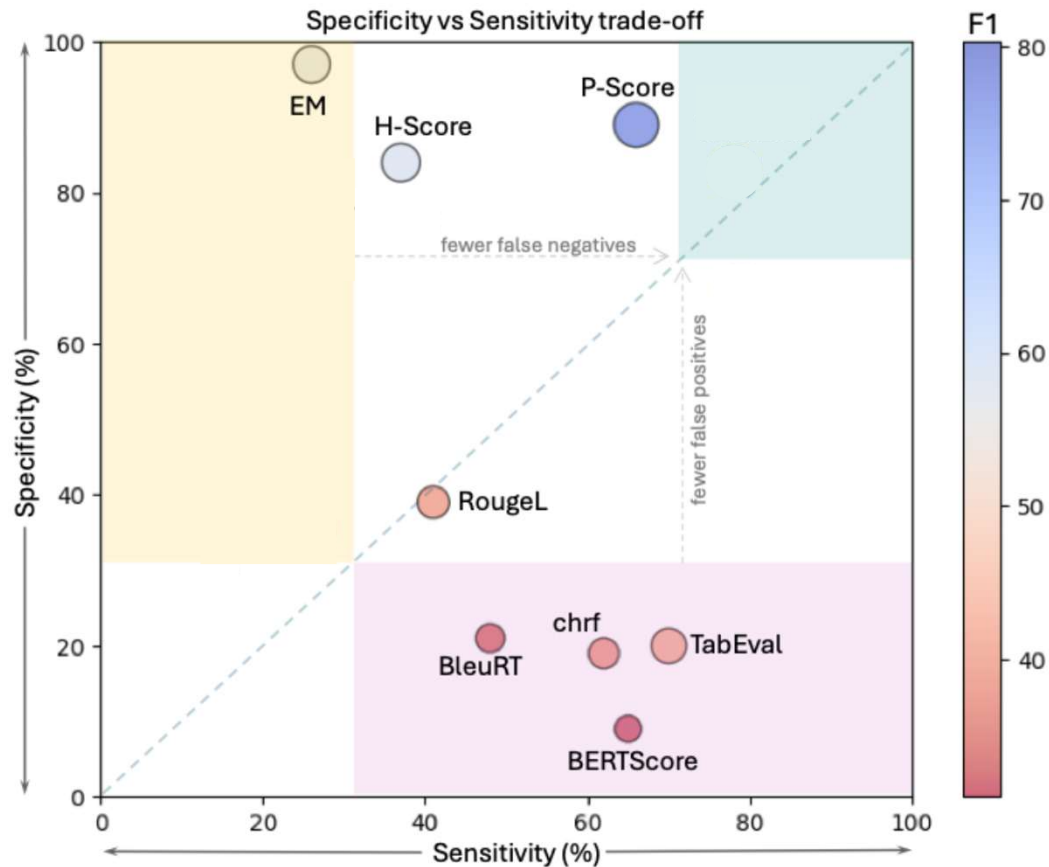
Tables contain both semantic and structural information.  
We need to consider various aspects:

- **Row / column ordering**
- **Headers and labels:** the meaning of each column / row
- **Cell values & data types** – numbers, dates, strings, units
- **Missing / extra entries** – gaps, duplicate-rows, nulls
- **Formatting & units** – 1,000 vs 1k, USD vs EUR, %, °C, etc.

“We need a metric which takes semantics and structure both into consideration”



# Gap in Current Evaluation Methods



- **Text-only view:** BLEU/ROUGE ignore table layout
- **Semantics-only:** BERTScore misses row/col swaps
- **Black-box scores:** P-/H-Score give no error trace
- **Entailment heavy:** TabEval, numeric-blind because of use of BertScore

# TabXEval

Ground Truth Table

Name	Term Start	Term End	Net Worth(est.)
Pier Ruggero Piccio	1 January 1926	6 February 1927	2.5
Armando Armani	10 February 1927	13 October 1928	1.8
Giuseppe Valle	22 February 1930	23 November 1933	3.2
Antonio Bosio	23 November 1933	22 March 1934	0.95
Ferdinando Raffaelli	10 November 1955	1 February 1958	1.5

Reference Table

Name	Term Start	Term End	Net Worth(est.)	Profession
P. Piccio	1926-01-11	6 February 1927	\$2,700,000 USD (est.)	Italian Air Force
A. Armani	1927-02-20	13 October 1928	\$1,950,000 USD (est.)	Air Force General
Giuseppe Valle	1930-02-22	23 November 1933	\$3,500,000 USD (est.)	Air Force General
Antonio Bosio	23 November 1933	22 March 1934	\$880,000 USD (est.)	Air Force Officer

TabCompare

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2	Worth (est.) (Millions USD).T1 / Net Worth (est.).T2	Profession.T1 / Profession.T2
Pier Ruggero Piccio / P. Piccio	1 January 1926 / 1926-01-11	EM	2.5 / \$2,700,000 USD (est.)	EI
Armando Armani / A. Armani	10 February 1927 / 1927-02-20	EM	1.8 / \$1,950,000 USD (est.)	EI
EM	22 February 1930 / 1930-02-22	EM	3.2 / \$3,500,000 USD (est.)	EI
EM	EM	EM	0.95 / \$880,000 USD (est.)	EI
MI	MI	MI	MI	EI

EM: Exact Match, MI: missing Information, EI: Extra Information

Forming comparison Tuple using LLM

Name.T1 / Name.T2

[String/String, Person/Person, None/None, None, abbreviated string]

Term start.T1 / Term start.T2

[Date/Date, Date/Date, None/None, None, absolute difference:10 days:]

TabAlign

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2	Worth (est.) (Millions USD).T1 / Net Worth (est.).T2	Profession.T1 / Profession.T2
Antonio Bosio / Antonio Bosio	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934		
Pier Ruggero Piccio / P. Piccio	1 January 1926 / 1926-01-11	6 February 1927 / 6 February 1927	2.5 / \$2,700,000 USD (est.)	-/ Italian Air Force
Armando Armani / A. Armani	10 February 1927 / 1927-02-20	13 October 1928 / 13 October 1928	1.8 / \$1,950,000 USD (est.)	-/Air Force General
Giuseppe Valle / Giuseppe Valle	22 February 1930 / 1930-02-22	23 November 1933 / 23 November 1933	3.2 / \$3,500,000 USD (est.)	-/Air Force General
Antonio Bosio / Antonio Bosio	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934	0.95 / \$880,000 USD (est.)	-/Air Force Officer
Ferdinando Raffaelli / -	10 November 1955 / -	1 February 1958 / -	1.5 / -	-/-

String Matching

LLM based

Structure Descriptor:

Row Missing

Extra Column

Column Descriptor:

Extra Column : Profession → Extra Column type: String

Cell level Descriptor:

Partial : 9 String : 2 Date : 3 Numeric : 4

Granular Cell level:

Formatted Date (Term Start), Abbreviate String (Name), Formatted Numerical (Net Worth)



# **TabXEval** PHASE 1: *TabAlign*

**Ground Truth Table**

Name	Term Start	Term End	Net Worth(est.)
Pier Ruggero Piccio	1 January 1926	6 February 1927	2.5
Armando Armani	10 February 1927	13 October 1928	1.8
Giuseppe Valle	22 February 1930	23 November 1933	3.2
Antonio Bosio	23 November 1933	22 March 1934	0.95
Ferdinando Raffaelli	10 November 1955	1 February 1958	1.5

**Reference Table**

Name	Term Start	Term End	Net Worth(est.)	Profession
P. Piccio	1926-01-11	6 February 1927	\$2,700,000 USD (est.)	Italian Air Force
A. Armani	1927-02-20	13 October 1928	\$1,950,000 USD (est.)	Air Force General
Giuseppe Valle	1930-02-22	23 November 1933	\$3,500,000 USD (est.)	Air Force General
Antonio Bosio	23 November 1933	22 March 1934	\$880,000 USD (est.)	Air Force Officer

**TabAlign**

Name.T1 / Name.T2		Term start.T1 / Term start.T2		Term end.T1 / Term end.T2	
Antonio Bosio / Antonio Bosio		23 November 1933 / 23 November 1933		22 March 1934 / 22 March 1934	
Pier Ruggero Piccio / P. Piccio		1 January 1926 / 1926-01-11		6 February 1927 / 6 February 1927	
Armando Armani / A. Armani		10 February 1927 / 1927-02-20		13 October 1928 / 13 October 1928	
Giuseppe Valle / Giuseppe Valle		22 February 1930 / 1930-02-22		23 November 1933 / 23 November 1933	
Antonio Bosio / Antonio Bosio		23 November 1933 / 23 November 1933		22 March 1934 / 22 March 1934	
Ferdinando Raffaelli / -		10 November 1955 / -		1 February 1958 / -	

- **Hybrid matcher:** exact look-ups augmented with LLM mapping
- **Robust alignment:** resolves synonyms, merged columns, and transposed layouts
- **Gap-aware output:** returns a fully aligned table with “-” placeholders for missing cells

# **TabXEval** PHASE 2: *TabCompare*

- **Cell-wise tuples:** create a detailed comparison record for every cell
- **Row / column checks:** spot missing, extra, and reordered rows or columns
- **Delta detection:** catch unit swaps, format changes, and numeric shifts

## TabCompare

Name.T1 / Name.T2	Term start.T1 / Term start.T2	Term end.T1 / Term end.T2	Worth (est.) (Millions USD).T1 / Net Worth (est.).T2	Profession.T1 / Profession.T2
Pier Ruggero Piccio / P. Piccio	1 January 1926 / 1926-01-11	EM	2.5 / \$2,700,000 USD (est.)	EI
Armando Armani / A. Armani	10 February 1927 / 1927-02-20	EM	1.8 / \$1,950,000 USD (est.)	EI
EM	22 February 1930 / 1930-02-22	EM	3.2 / \$3,500,000 USD (est.)	EI
EM	EM	EM	0.95 / \$880,000 USD (est.)	EI
MI	MI	MI	MI	EI

EM: Exact Match, MI: missing Information, EI: Extra Information

Forming comparison Tuple using LLM

Name.T1 / Name.T2	Term start.T1 / Term start.T2
[String/String, Person/Person, None/None, None, abbreviated string]	[Date/Date, Date/Date, None/None, None, absolute difference:10 days:]

## Structure Descriptor:

Row Missing

Extra Column

## Column Descriptor:

Extra Column : Profession → Extra Column type: String

## Cell level Descriptor:

Partial : 9 String : 2 Date : 3 Numeric : 4

## Granular Cell level:

Formatted Date (Term Start), Abbreviate String (Name), Formatted Numerical (Net Worth)

# TabXEval Scoring

$$\text{TabXEval} = \sum_{I \in \{Missing, Extra, Partial\}} \beta_I \times \left( \sum_{E \in \{row, column, cell\}} \alpha_E \frac{f_E}{N_E} \right) \gamma_p$$

$\beta_I$  – weight for each error class (Missing, Extra, Partial)

$\alpha_E$  – weight for each entity type (Row, Column, Cell)

$f_E$  – # of correctly matched entities

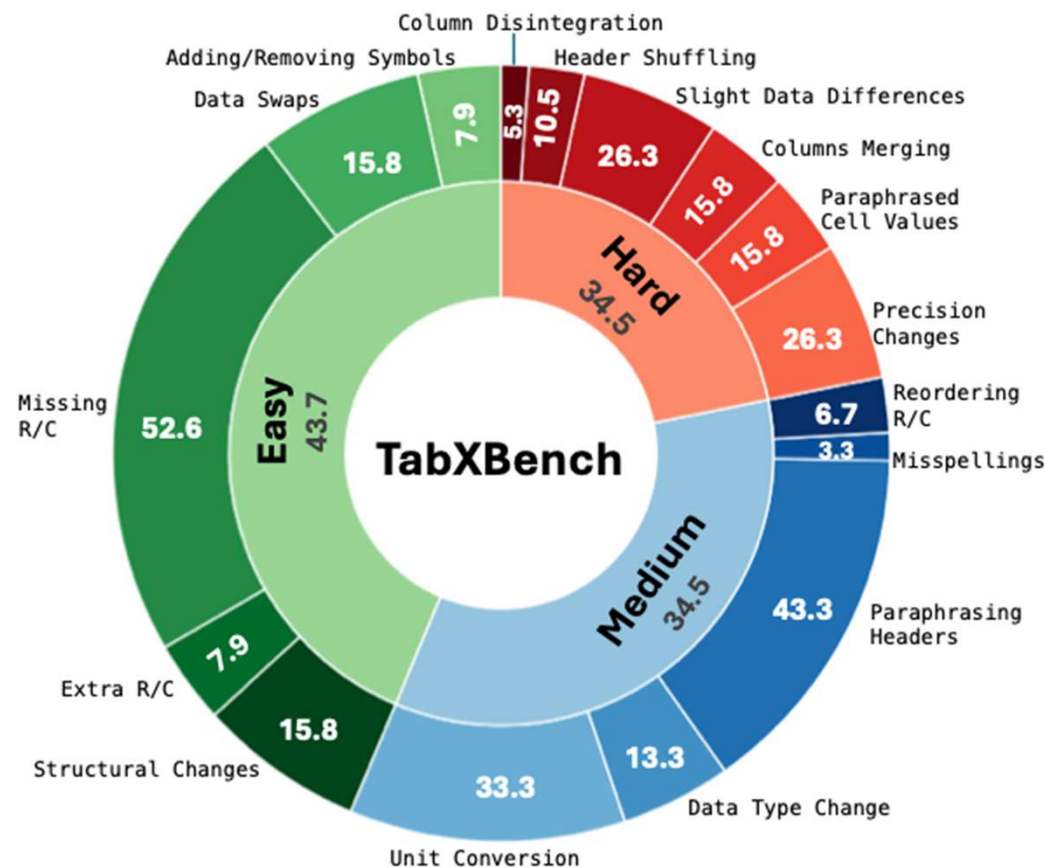
$N_E$  – total entities in the ground-truth table

For partial matches at the cell level, the modifier  $\gamma_p$  is defined as:

$$\gamma_p = \begin{cases} 1, & \text{if no partial cell,} \\ \omega_p \left| \frac{GT-Ref}{Ref} \right|, & \text{if partial cell detected.} \end{cases}$$

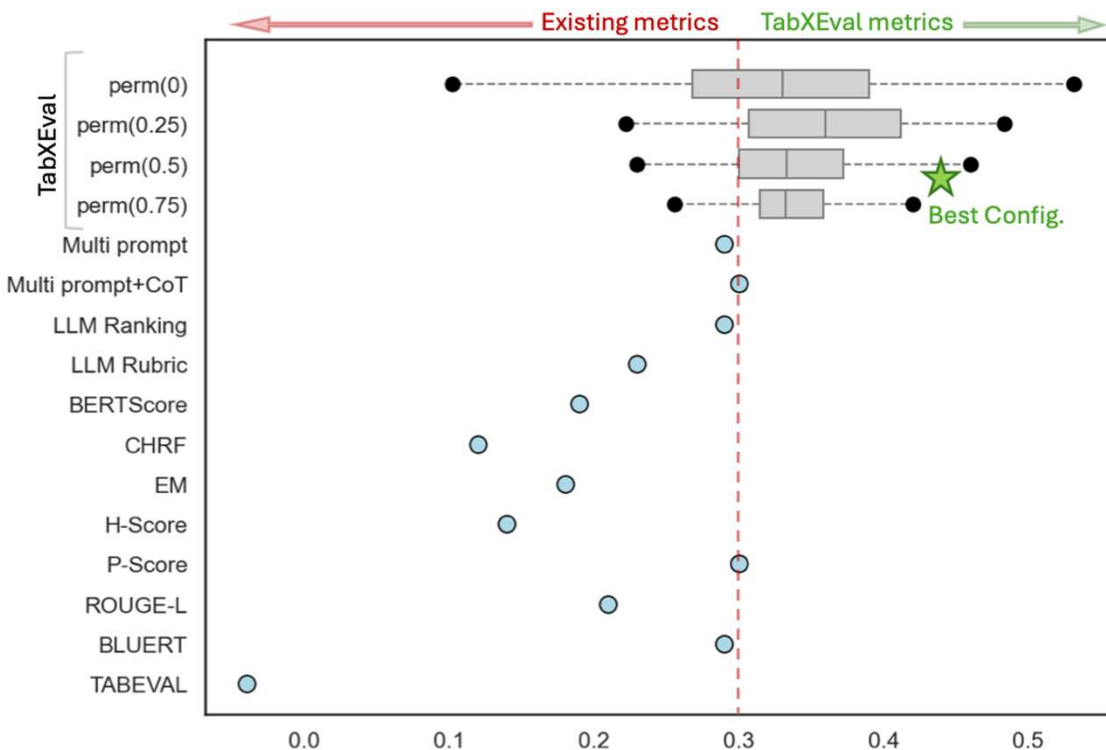


# TabXBench



- 250 test cases
- 6 diverse sources
- 16 distinct error types
- 3 difficulty tiers
- Balanced stress test

# How our Metric is Domain Agnostic



- **Domain-agnostic core** – outperforms all baselines on every corpus without tuning
- **Domain-adaptive knobs** – re-weight structure vs semantics (row/col vs cell) for task focus.
- **Balanced ★ config** – highest human-correlation, lowest variance → fidelity + flexibility.

# Comparison of TabXEval and Other Metrics on TabXBench

Method	Avg %	Spearman's	Kendal's $\tau$	RBO
<b>TabXEval</b>	<b>39.3%</b>	<b>44%</b>	<b>40%</b>	<b>34%</b>
P-Score	29.3%	30%	27%	31%
MP + CoT	28.0%	30%	25%	29%
LLM Rank	27.6%	29%	24%	30%
MP	27.6%	29%	24%	30%
BLUERT	27.0%	29%	25%	27%
LLM Rubric	22.3%	23%	16%	28%
ROUGE-L	22.0%	21%	18%	27%
EM	20.0%	18%	16%	26%
BERTScore	19.6%	19%	15%	25%
H-Score	17.6%	14%	11%	28%
CHRF	16.0%	12%	11%	25%
<b>TabEval</b>	<b>5.0%</b>	<b>-4%</b>	<b>-4%</b>	<b>23%</b>

- **TabXEval +15–20 pp** over LLM-prompted metrics → best structure + semantics
- **P-Score / BLEURT**: Plateau despite improvements over lexical baselines.
- **TabEval lowest** – many false positives on numeric / unit errors

# Explainable Error Forensics for Text-to-Table Tasks

	LLaMA-3.3 70B			GPT-4o			Gemini-2.0-flash		
	MI	EI	EM	MI	EI	EM	MI	EI	EM
WikiTable									
Row	8.69	15.82	25.59	23.44	11.51	27.11	20.81	10.67	26.03
Col	0.37	0.92	1.67	4.47	0.07	1.02	2.97	0.37	1.55
WikiBio									
Row	25.16	29.33	16.17	26.09	27.63	19.39	30.08	24.38	16.89
Col	0.10	0.0	0.05	0.05	0.025	0.0	0.12	0.0	0.0
TANQ									
Row	7.6	5.83	10.97	8.27	2.80	13.00	8.51	4.01	13.01
Col	2.69	1.82	23.89	2.24	0.19	22.42	2.82	0.63	21.78
RotoWire									
Row	3.48	39.22	17.62	1.32	28.65	38.41	3.10	22.82	13.80
Col	10.87	16.21	52.70	17.57	5.71	60.24	16.35	10.52	48.51

Table 4: **Table-Level Performance Analysis:** Row/Column MI, EI, and EM rates on WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o leads with highest Row EM (27.11) and lowest Col EI (0.07) on WikiTables.

Stat	LLaMA-3.3 70B							GPT-4o							Gemini-2.0-flash						
	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others
WikiTables																					
EI	0.05	4.33	0.00	0.17	0.00	0.00	0.13	0.03	1.17	0.00	0.02	0.00	0.00	0.13	0.02	2.33	0.00	0.11	0.00	0.00	0.07
MI	0.01	0.80	0.00	0.03	0.00	0.00	0.00	0.01	0.93	0.00	0.00	0.00	0.00	0.01	0.00	0.73	0.00	0.00	0.00	0.00	0.00
Partial	0.22	25.00	0.00	0.35	0.00	0.01	0.09	0.32	20.34	0.00	0.55	0.00	0.02	0.10	0.30	22.50	0.00	0.48	0.00	0.02	0.07
WikiBio																					
EI	0.04	2.84	0.00	0.09	0.00	0.00	0.12	0.04	2.02	0.00	0.03	0.00	0.00	0.09	0.04	2.30	0.00	0.07	0.00	0.00	0.06
MI	0.02	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.01	0.00	0.00	0.00	0.00	0.37	0.00	0.01	0.00	0.00	0.00
Partial	0.16	14.38	0.00	2.60	0.00	0.00	0.03	0.16	15.80	0.00	0.86	0.00	0.00	0.03	0.15	13.59	0.00	2.97	0.00	0.00	0.04
TANQ																					
EI	0.05	0.84	0.00	0.16	0.09	0.00	0.00	0.02	0.18	0.00	0.06	0.03	0.01	0.00	0.00	0.29	0.00	0.07	0.02	0.00	0.00
MI	0.01	0.24	0.00	0.11	0.01	0.04	0.00	0.01	0.08	0.00	0.05	0.02	0.01	0.00	0.02	0.21	0.00	0.05	0.00	0.02	0.00
Partial	2.73	20.50	0.00	4.72	4.82	2.19	0.07	1.28	11.92	0.00	3.48	3.46	1.32	0.01	1.22	9.35	0.00	2.02	2.64	1.12	0.02
RotoWire																					
EI	0.84	0.50	0.00	0.00	0.00	0.00	0.01	0.68	0.23	0.00	0.00	0.00	0.00	0.00	1.31	0.54	0.00	0.00	0.00	0.00	0.09
MI	0.97	0.32	0.00	0.00	0.00	0.00	0.04	0.56	0.08	0.00	0.00	0.00	0.00	0.00	1.06	0.28	0.00	0.00	0.00	0.00	0.02
Partial	0.66	0.92	0.00	0.00	0.00	0.00	0.00	0.31	0.69	0.00	0.00	0.00	0.00	0.00	2.72	3.87	0.00	0.00	0.00	0.00	0.03

Table 3: **Cell-Level Performance Analysis** of Extra (EI), Missing (MI), and Partial mismatches across data types numerical, string, boolean, date, list, time, and other for WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o shows fewer string EI in WikiTables and lower partial errors in numerical and string cells in TANQ and RotoWire.

Our rubric surfaces GPT-4o’s tightest alignment (27% row EM, 0.07% col EI) while pinpointing LLaMA’s noisy extras and Gemini’s numeric-heavy collapses, turning latent row/column and cell-type faults into actionable debugging insights.



# Closing Notes



# Closing Notes

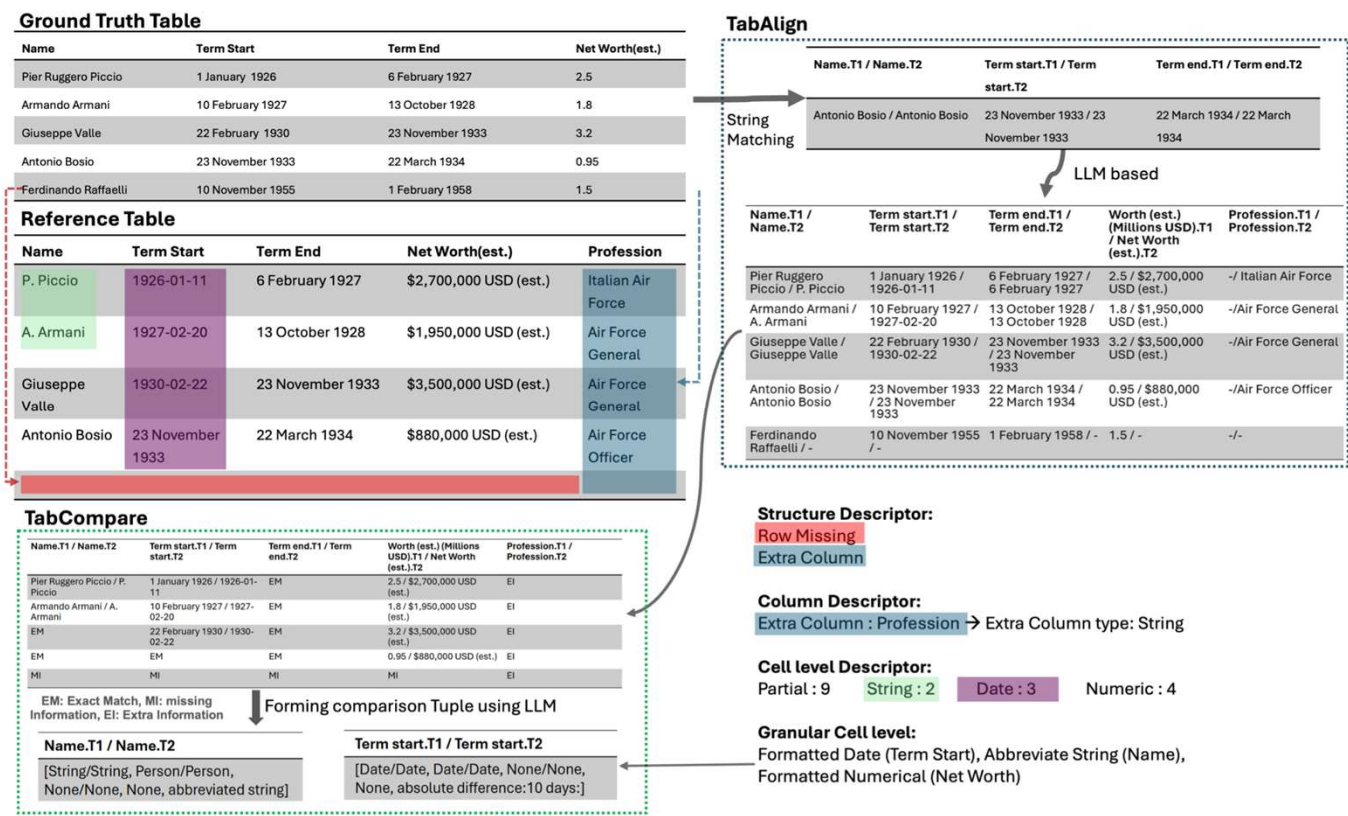
- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation





# Closing Notes

- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors



# Closing Notes

- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors

**Ground Truth Table**

Name	Term Start	Term End	Net Worth(USD)
For Reginald Pico	3 January 1927	6 February 1927	3.5
Antonio Borda	10 February 1927	10 October 1928	1.8
Guisepe Vella	22 February 1928	23 November 1933	3.2
Antonio Borda	23 November 1933	22 March 1934	0.85
Antonio Borda	10 November 1933	7 February 1934	1.0

**Reference Table**

Name	Term Start	Term End	Net Worth(USD)	Profession
P. Pico	1926-01-01	6 February 1927	\$2,700,000 USD (est.)	Soldier Air Force
A. Borda	1927-02-10	10 October 1928	\$1,800,000 USD (est.)	Air Force General
Guisepe Vella	1928-02-22	23 November 1933	\$3,500,000 USD (est.)	Air Force General
Antonio Borda	23 November 1933	22 March 1934	\$880,000 USD (est.)	Air Force Officer

**TabAlign**

Name T1 / Name T2	Term start T1 / Term start T2	Term end T1 / Term end T2	Net Worth T1 / Net Worth T2	Profession T1 / Profession T2
For Reginald Pico / Antonio Borda	3 January 1927 / 10 February 1927	6 February 1927 / 10 October 1928	3.5 / 1.8	Soldier Air Force / Air Force General
Antonio Borda / Antonio Borda	10 February 1927 / 10 October 1928	10 October 1928 / 23 November 1933	1.8 / 3.2	Air Force General / Air Force General
Guisepe Vella / Antonio Borda	22 February 1928 / 23 November 1933	23 November 1933 / 22 March 1934	3.2 / 0.85	Air Force General / Air Force Officer
Antonio Borda / Antonio Borda	23 November 1933 / 10 November 1933	22 March 1934 / 7 February 1934	0.85 / 1.0	Air Force Officer / Air Force Officer

**TabCompare**

Name T1 / Name T2	Term start T1 / Term start T2	Term end T1 / Term end T2	Net Worth T1 / Net Worth T2	Profession T1 / Profession T2
For Reginald Pico / Antonio Borda	3 January 1927 / 10 February 1927	6 February 1927 / 10 October 1928	3.5 / 1.8	Soldier Air Force / Air Force General
Antonio Borda / Antonio Borda	10 February 1927 / 10 October 1928	10 October 1928 / 23 November 1933	1.8 / 3.2	Air Force General / Air Force General
Guisepe Vella / Antonio Borda	22 February 1928 / 23 November 1933	23 November 1933 / 22 March 1934	3.2 / 0.85	Air Force General / Air Force Officer
Antonio Borda / Antonio Borda	23 November 1933 / 10 November 1933	22 March 1934 / 7 February 1934	0.85 / 1.0	Air Force Officer / Air Force Officer

**Structure Descriptor:**  
Row Missing: Extra Column

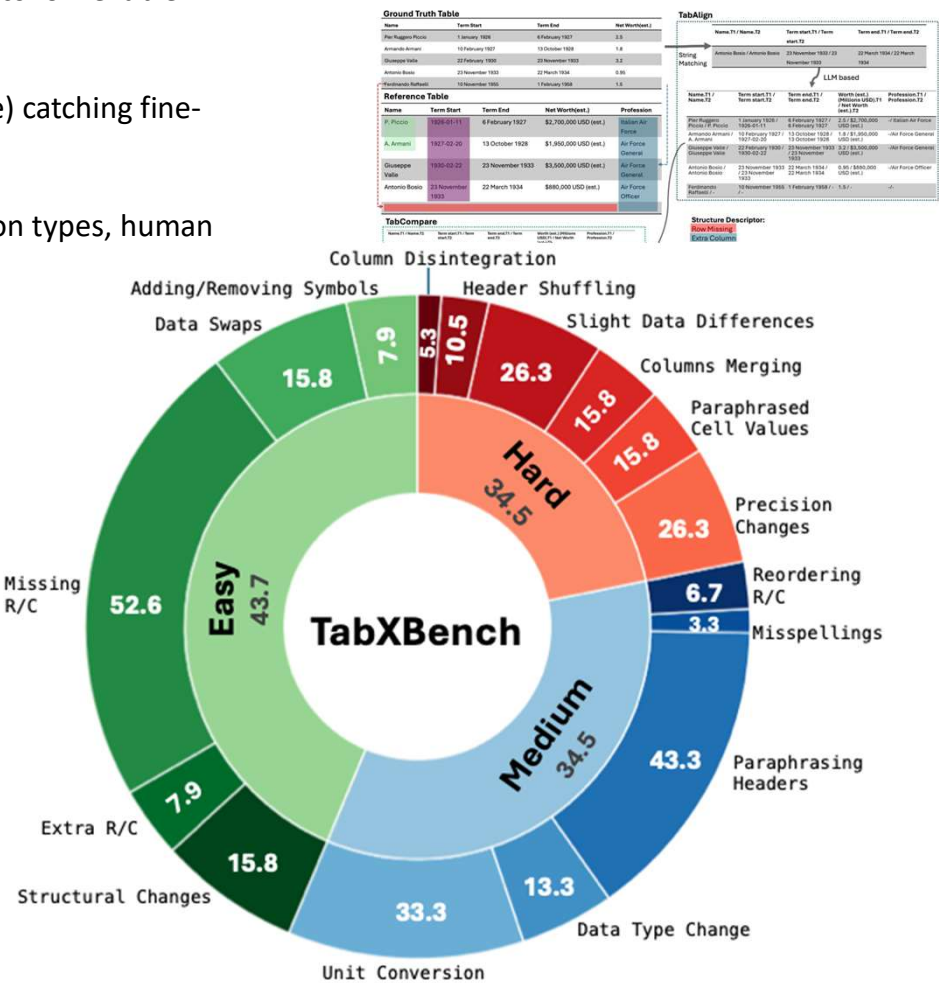
**Column Descriptor:**  
Extra Column: Profession → Extra Column type: String

**Cell level Descriptor:**  
Partial: 9 String: 2 Date: 3 Numeric: 4

**Granular Cell level:**  
Formatted Date (Term Start), Abbreviate String (Name), Formatted Numerical (Net Worth)

# Closing Notes

- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- **TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels



# Closing Notes

- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- **TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels

**Ground Truth Table**

Name	Term Start	Term End	Net Worth(k\$)
For Reginald Pico	3 January 1928	6 February 1927	3.5
Antonio Basso	10 February 1927	10 October 1928	1.8
Guiseppi Vella	22 February 1928	23 November 1933	3.2
Antonio Basso	23 November 1933	22 March 1934	0.85
Antonio Basso	10 November 1933	7 February 1934	1.0

**Reference Table**

Name	Term Start	Term End	Net Worth(k\$)	Profession
P. Pico	1928-01-01	6 February 1927	\$2,700,000 USD (net.)	Soldier Air Force
A. Basso	1927-02-10	10 October 1928	\$1,800,000 USD (net.)	Air Force General
Guiseppi Vella	1928-02-22	23 November 1933	\$3,600,000 USD (net.)	Air Force General
Antonio Basso	23 November 1933	22 March 1934	\$880,000 USD (net.)	Air Force Officer

**TabAlign**

Name T1 / Name T2	Term start T1 / Term start T2	Term end T1 / Term end T2	Net Worth T1 / Net Worth T2
Antonio Basso / Antonio Basso	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934	0.85 / 0.85

**TabCompare**

Name T1 / Name T2	Term start T1 / Term start T2	Term end T1 / Term end T2	Net Worth T1 / Net Worth T2
For Reginald Pico / For Reginald Pico	3 January 1928 / 3 January 1928	6 February 1927 / 6 February 1927	3.5 / 3.5
Antonio Basso / Antonio Basso	10 February 1927 / 10 February 1927	10 October 1928 / 10 October 1928	1.8 / 1.8
Guiseppi Vella / Guiseppi Vella	22 February 1928 / 22 February 1928	23 November 1933 / 23 November 1933	3.2 / 3.2
Antonio Basso / Antonio Basso	23 November 1933 / 23 November 1933	22 March 1934 / 22 March 1934	0.85 / 0.85

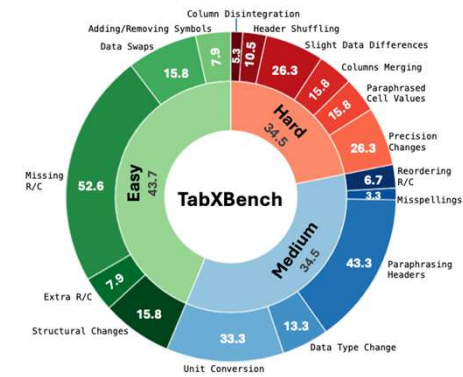
**TabXBench**

Structure Descriptor: Row Missing, Extra Column

Column Descriptor: Extra Column: Profession → Extra Column type: String

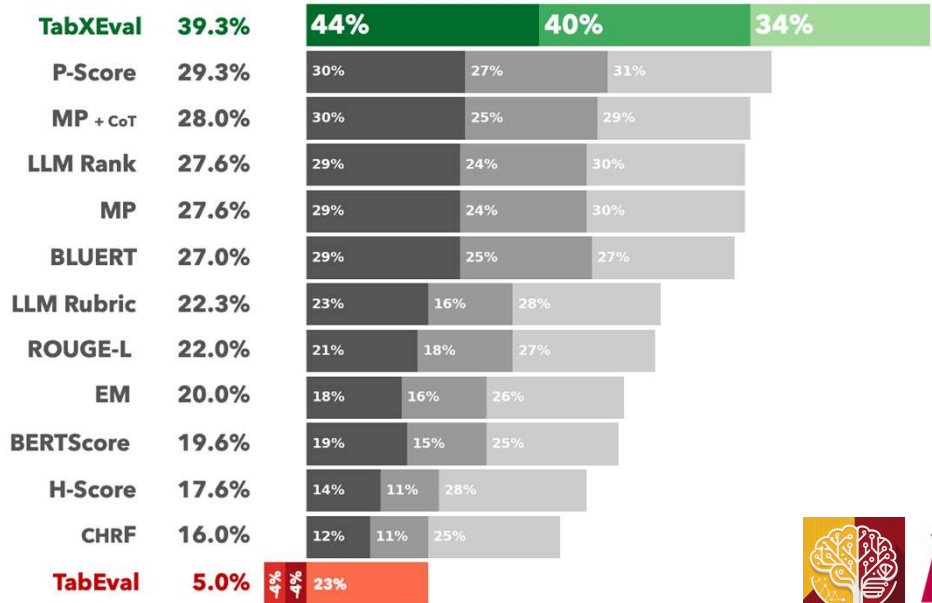
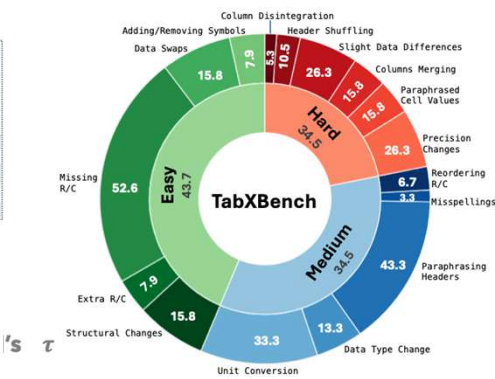
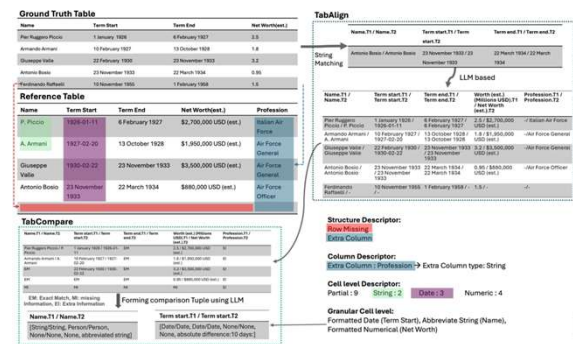
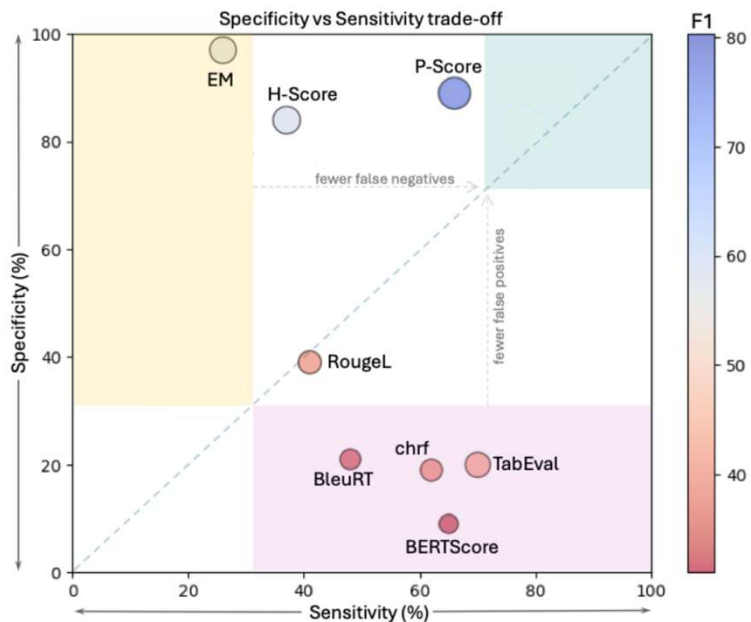
Cell level Descriptor: Partial: 9, String: 2, Date: 3, Numeric: 4

Granular Cell level: Formatted Date (Term Start), Abbreviate String (Name), Formatted Numerical (Net Worth)



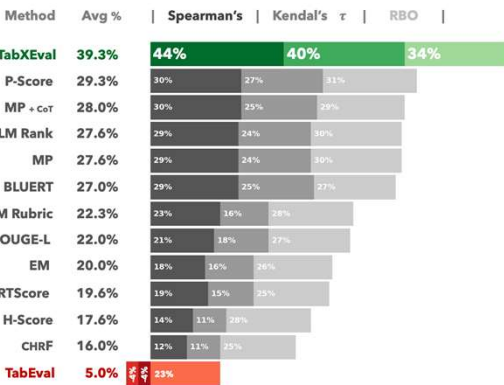
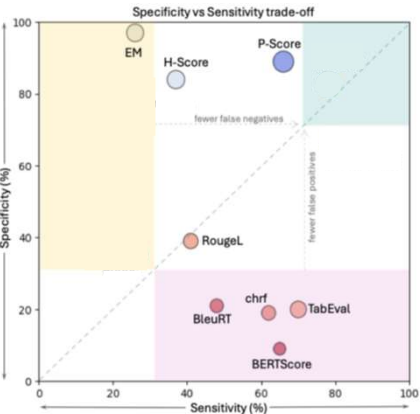
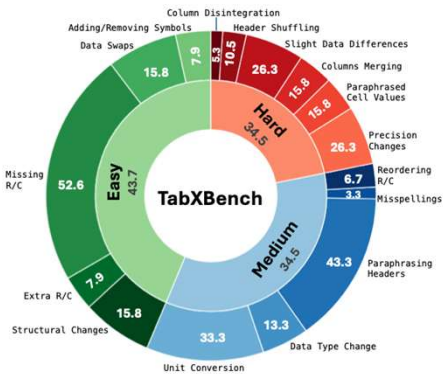
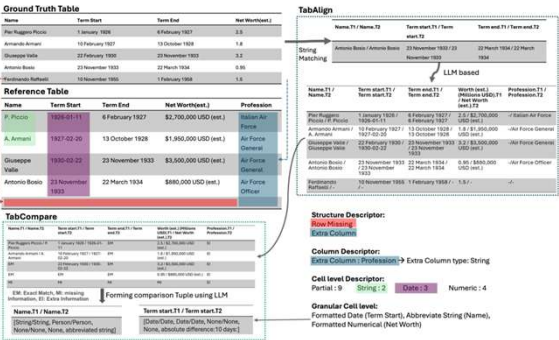
# Closing Notes

- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- **TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels
- **Metric audit** → Sensitivity vs Specificity exposes limits of current metrics



# Closing Notes

- Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels
- Metric audit** → Sensitivity vs Specificity exposes limits of current metrics





# Closing Notes

- Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels
- Metric audit** → Sensitivity vs Specificity exposes limits of current metrics
- Validated** → TabXEval beats baselines with clear, explainable feedback

	LLaMA-3.3 70B			GPT-4o			Gemini-2.0-flash		
	MI	EI	EM	MI	EI	EM	MI	EI	EM
<b>WikiTable</b>									
Row	8.69	15.82	25.59	23.44	11.51	27.11	20.81	10.67	26.03
Col	0.37	0.92	1.67	4.47	0.07	1.02	2.97	0.37	1.55
<b>WikiBio</b>									
Row	25.16	29.33	16.17	26.09	27.63	19.39	30.08	24.38	16.89
Col	0.10	0.0	0.05	0.05	0.025	0.0	0.12	0.0	0.0
<b>TANQ</b>									
Row	7.6	5.83	10.97	8.27	2.80	13.00	8.51	4.01	13.01
Col	2.69	1.82	23.89	2.24	0.19	22.42	2.82	0.63	21.78
<b>RotoWire</b>									
Row	3.48	39.22	17.62	1.32	28.65	38.41	3.10	22.82	13.80
Col	10.87	16.21	52.70	17.57	5.71	60.24	16.35	10.52	48.51

Table 4: **Table-Level Performance Analysis:** Row/Column MI, EI, and EM rates on WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o leads with highest Row EM (27.11) and lowest Col EI (0.07) on WikiTables.

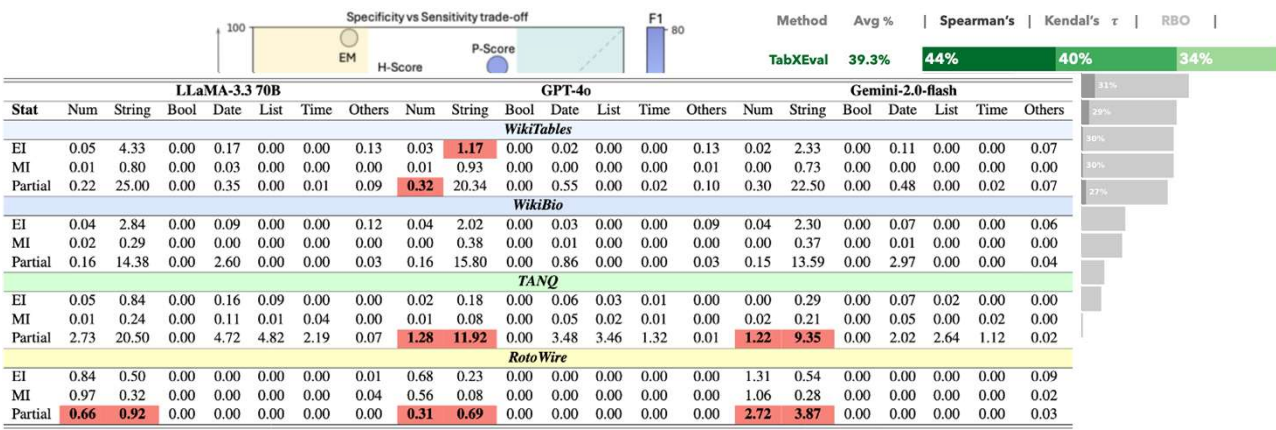
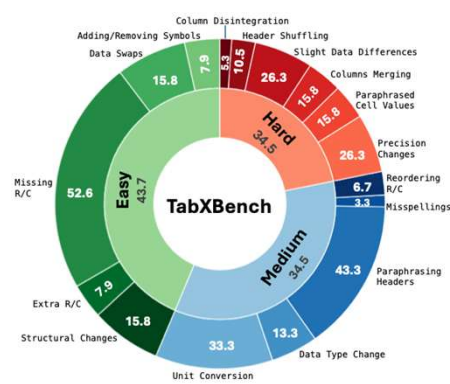
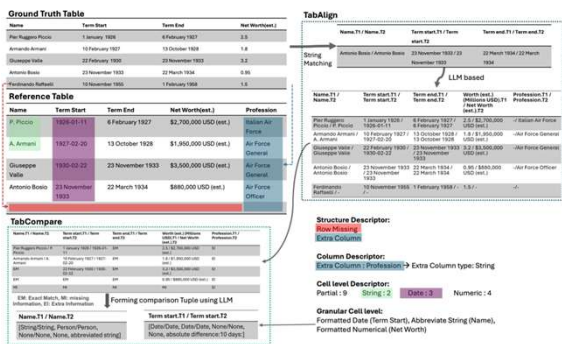


Table 3: **Cell-Level Performance Analysis** of Extra (EI), Missing (MI), and Partial mismatches across data types numerical, string, boolean, date, list, time, and other for WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o shows fewer string EI in WikiTables and lower partial errors in numerical and string cells in TANQ and RotoWire.



# Closing Notes

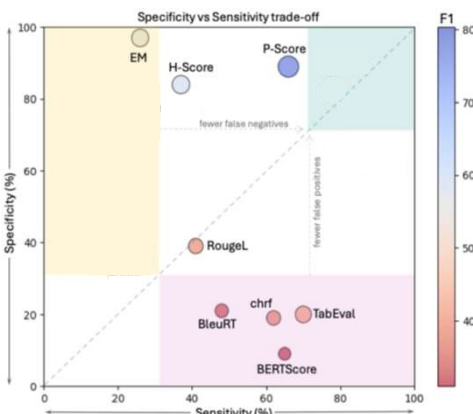
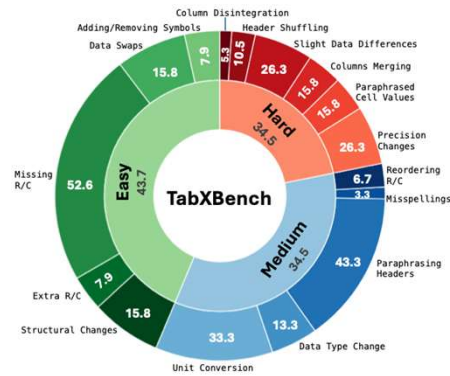
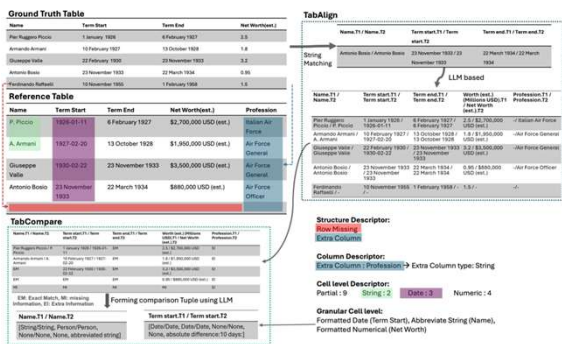
- **Rubric-driven scoring** → blends structure + semantics for reliable table evaluation
- **TabXEval** → 2-stage LLM pipeline (Align → Compare) catching fine-grained errors
- **TabXBench** → 250 tables, 6 domains, 22 perturbation types, human gold labels
- **Metric audit** → Sensitivity vs Specificity exposes limits of current metrics
- **Validated** → TabXEval beats baselines with clear, explainable feedback

	LLaMA-3.3 70B			GPT-4o			Gemini-2.0-flash		
	MI	EI	EM	MI	EI	EM	MI	EI	EM
<b>WikiTable</b>									
Row	8.69	15.82	25.59	23.44	11.51	27.11	20.81	10.67	26.03
Col	0.37	0.92	1.67	4.47	6.97	1.02	2.97	0.37	1.55
<b>WikiBio</b>									
Row	25.16	29.33	16.17	26.09	27.63	19.39	30.08	24.38	16.89
Col	0.10	0.0	0.05	0.05	0.025	0.0	0.12	0.0	0.0
<b>TANQ</b>									
Row	7.6	5.83	10.97	8.27	2.80	13.00	8.51	4.01	13.01
Col	2.69	1.82	23.89	2.24	0.19	22.42	2.82	0.63	21.78
<b>RotoWire</b>									
Row	3.48	39.22	17.62	1.32	28.65	38.41	3.10	22.82	13.80
Col	10.87	16.21	52.70	17.57	5.71	60.24	16.35	10.52	48.51

Table 4: **Table-Level Performance Analysis:** Row/Column MI, EI, and EM rates on WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o leads with highest Row EM (27.11) and lowest Col EI (0.07) on WikiTables.

	LLaMA-3.3 70B										GPT-4o										Gemini-2.0-flash																					
	Stat	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others	Num	String	Bool	Date	List	Time	Others	Stat	Num	String	Bool	Date	List	Time	Others												
WikiTable																																										
EI	0.05	4.33	0.00	0.17	0.00	0.00	0.13	0.03	1.17	0.00	0.02	0.00	0.00	0.13	0.02	2.33	0.00	0.11	0.00	0.00	0.07	MI	0.01	0.80	0.00	0.03	0.00	0.00	0.01	0.53	0.00	0.00	0.00	0.01	0.73	0.00	0.00	0.00	0.00			
MI	0.01	0.80	0.00	0.03	0.00	0.00	0.00	0.01	0.53	0.00	0.00	0.00	0.00	0.01	0.00	0.73	0.00	0.00	0.00	0.00	0.00	Partial	0.22	25.00	0.00	0.35	0.00	0.01	0.09	10.34	0.00	0.55	0.00	0.02	0.10	0.30	22.50	0.00	0.48	0.00	0.02	
WikiBio																																										
EI	0.04	2.84	0.00	0.06	0.00	0.00	0.12	0.04	2.02	0.00	0.03	0.00	0.00	0.09	0.04	2.36	0.00	0.07	0.00	0.00	0.06	MI	0.02	0.29	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.01	0.00	0.00	0.60	0.37	0.00	0.01	0.00	0.00		
MI	0.02	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.01	0.00	0.00	0.60	0.37	0.00	0.00	0.00	0.00	0.00	0.00	Partial	0.16	14.38	0.00	0.26	0.00	0.00	0.03	0.16	15.80	0.00	0.86	0.00	0.00	0.03	0.15	13.59	0.00	2.97	0.00	0.00
TANQ																																										
EI	0.05	0.84	0.00	0.16	0.09	0.00	0.00	0.02	0.18	0.00	0.06	0.03	0.01	0.00	0.00	0.29	0.00	0.07	0.02	0.00	0.00	MI	0.01	0.24	0.00	0.11	0.01	0.04	0.00	0.01	0.08	0.00	0.02	0.21	0.00	0.05	0.00	0.02	0.00	0.00		
MI	0.01	0.24	0.00	0.11	0.01	0.04	0.00	0.01	0.08	0.00	0.02	0.01	0.00	0.02	0.00	0.21	0.00	0.05	0.00	0.02	0.00	Partial	2.73	20.59	0.00	4.72	4.82	2.19	0.07	1.58	19.92	0.00	3.48	3.46	1.32	0.01	1.35	0.00	2.02	2.64	1.12	0.02
RotoWire																																										
EI	0.84	0.50	0.00	0.00	0.00	0.00	0.01	0.58	0.23	0.00	0.00	0.00	0.00	1.31	0.54	0.00	0.00	0.00	0.00	0.00	0.00	MI	0.97	0.32	0.00	0.00	0.00	0.00	0.04	0.56	0.08	0.00	0.00	0.00	0.00	1.06	0.28	0.00	0.00	0.00	0.02	
MI	0.97	0.32	0.00	0.00	0.00	0.00	0.04	0.56	0.08	0.00	0.00	0.00	0.00	1.06	0.28	0.00	0.00	0.00	0.00	0.00	0.02	Partial	0.66	0.92	0.00	0.00	0.00	0.00	0.00	0.31	0.69	0.00	0.00	0.00	0.00	0.272	2.87	0.00	0.00	0.00	0.03	

Table 3: **Cell-Level Performance Analysis** of Extra (EI), Missing (MI), and Partial mismatches across data types numerical, string, boolean, date, list, time, and other for WikiTables, WikiBio, TANQ, and RotoWire. **Highlights:** GPT-4o shows fewer string EI in WikiTables and lower partial errors in numerical and string cells in TANQ and RotoWire.

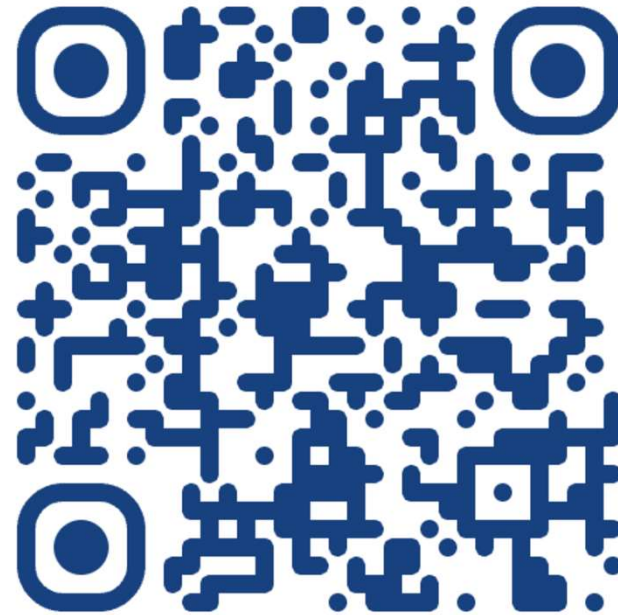


Method	Avg %	Spearman's	Kendal's $\tau$	RBO
TabXEval	39.3%	44%	40%	34%
P-Score	29.3%	30%	27%	35%
MP + coT	28.0%	30%	25%	29%
LLM Rank	27.6%	29%	24%	30%
MP	27.6%	29%	24%	30%
BLUERT	27.0%	29%	25%	27%
LLM Rubric	22.3%	23%	16%	28%
ROUGE-L	22.0%	21%	18%	27%
EM	20.0%	18%	16%	26%
BERTScore	19.6%	19%	15%	25%
H-Score	17.6%	14%	11%	28%
CHRF	16.0%	12%	11%	25%
TabEval	5.0%	4%	3%	23%



# Try TabXEval Yourself!

<https://coral-lab-asu.github.io/tabxeval/>





# Thank You



**ASU** Arizona State  
University

  
INTERNATIONAL INSTITUTE OF  
INFORMATION TECHNOLOGY  
HYDERABAD