UNIVERSITY OF AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM

Efficient Table Error Detection with LLMs

Thesis Defense

Guided by Dr. Hazar Harmouch

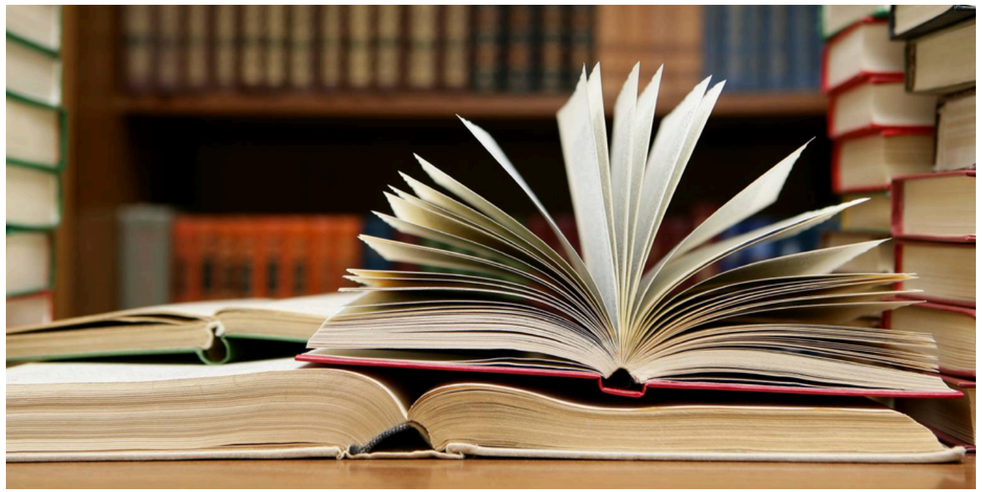
Presented by Yifeng Zhao

August 2025

UNIVERSITY OF AMSTERDAM

VU VRIJE UNIVERSITEIT AMSTERDAM

Table of Contents



1. Introduction
2. Background & Related Work
3. Methodology
4. Experiments
5. Discussion & Conclusion

Introduction

Erroreous data reduced machine learning model performance and affects enterprise decision-making.

Traditional table error detection (ED) methods use manual rules or labeled data. While Large language models (LLMs) offer an alternative, they require powerful but expensive models and more processing time.

Our work aims to explore if samll scale LLMs can detect tabular data errors without human annotation in a more efficient way.

Backgroud

1. **Error Detection in Tabular Data**
   Error types: syntactic and semantic.

   Traditional approaches
   rule, constraint, pattern, knowledge base, outlier detection.

   Model-based approaches
   feature engineering → supervised or unsupervised learning; pre-trained language models.

2. **Tokenization**
   sentences → list of tokens

   Tokenizers granularity: character or word level, subword level, and byte level.

   Examples in ED: n-gram, bag-of-characters.

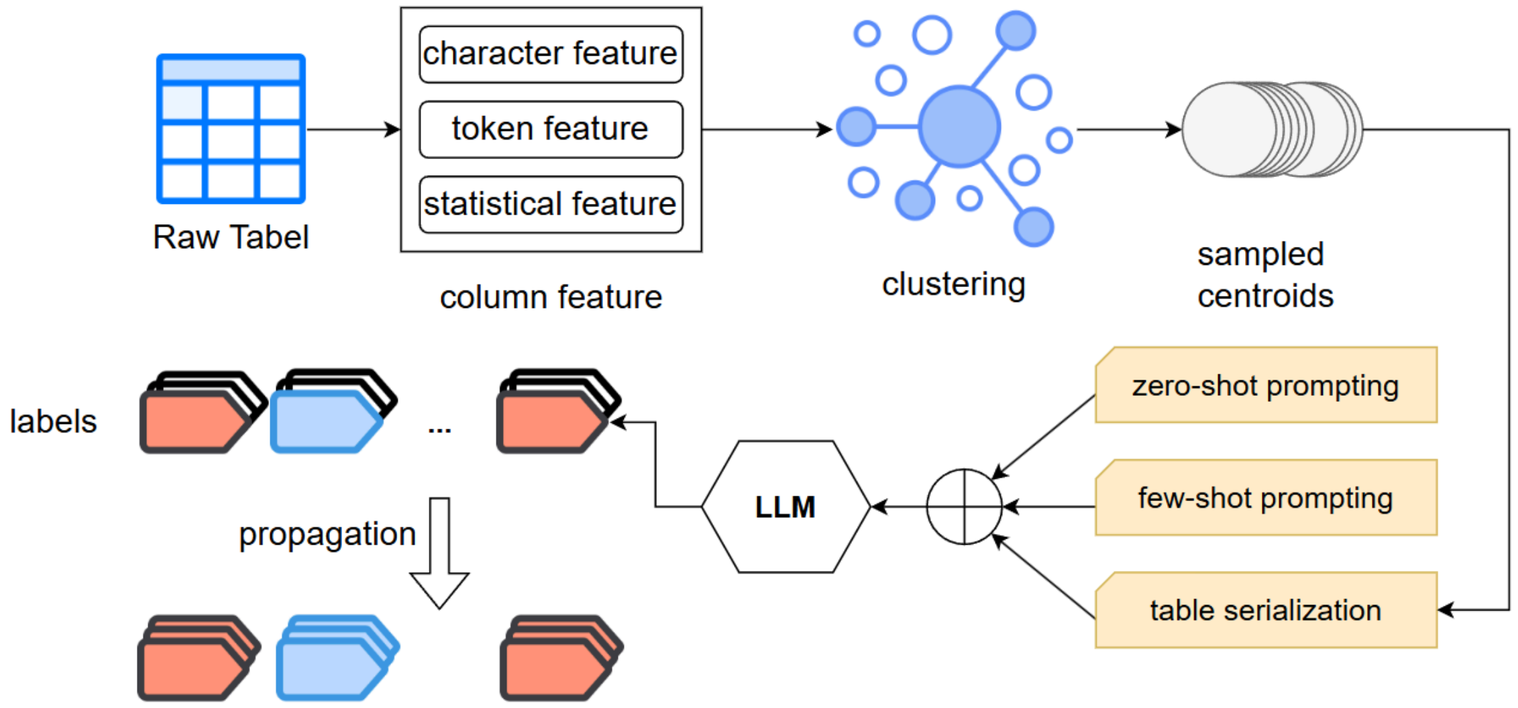
3. **Large Language Models**
   Generative Pre-trained Transformer.

Name	Designer	Players	Year	Type	Country
Ticket to Ride	Alan R. Moon	2 ~ 5	2004	Strategy	USA
Azul		2 - 4	2017	Abstract Strategy	Germany
Chess	Unknown	222	Ancient	Abstract Strategy	India
Catan	Klaus Teuber	3 - 4	1995	Resource Management	Italy
Mahjong	Unknown	4	3025	Tile-based Strategy	China
Splendorr	Marc André	2 - 4	2014	Resource Management	France
Zug um Zug	Alan R. Moon	2 - 5	2004	Strategy	USA

- Syntactic errors: typo, format / pattern violations, domain constraint violations
- Semantic errors: outliers, rule violations, missing value, duplicates

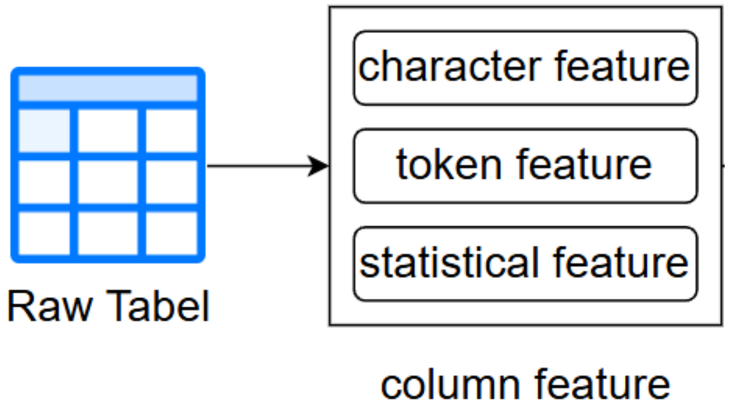
Methodology

1. Feature Extraction
2. Clustering-based Sampling
3. Error Labeling with LLMs
4. Label Propagation



Framework overview

Methodology - Feature Extraction



- **Character Feature**: character count + TF-IDF (Term Frequency-Inverse Document Frequency)
- **Token Feature**: BPE tokenizer + bin-frequency + normalization
- **Statistical Feature**: support + confidence + normalization

identify →

rare characters
rare tokens / subwords
functional dependencies

Methodology - Feature Extraction

Byte Pair Encoding (BPE): iteratively merge frequent adjacent token pairs into new tokens.

token pairs	frequency
('l', 'o')	2
('o', 'w')	2
('e', 's')	2
('s', 't')	2
('t', '</w>')	2
('w', '</w>')	1
('w', 'e')	2
...	...

Corpus:
[l o w </w>]
[l o w e r </w>]
[n e w e s t </w>]
[w i d e s t </w>]

↓ merge 'es'

[l o w </w>]
[l o w e r </w>]
[n e w es t </w>]
[w i d es t </w>]

↓

...

↓

Token List: [low, er, new, est, wide, </w>]

newest → [new, est]
newst → [new, s, t]

condition	tokens	bin1	bin2	bin3
heart attack	['heart', 'Ġattack']	0	1	1
heart attack	['heart', 'Ġattack']	0	1	1
heart attaxk	['heart', 'Ġatt', 'ax', 'k']	3	0	1
heart attackk	['heart', 'Ġattack', 'k']	1	1	1
heart attack	['heart', 'Ġattack']	0	1	1

binned token frequency feature

token	freq	bin
heart	5	bin3
Ġattack	4	bin2
k	2	bin1
Ġatt	1	bin1
ax	1	bin1

Methodology - Feature Extraction

$$\text{Support}(x, y) = \frac{\#\{\text{records where } X = x \wedge Y = y\}}{\#\{\text{all records}\}}$$

$$\text{Confidence}(y \mid x) = \frac{\#\{\text{records where } X = x \wedge Y = y\}}{\#\{\text{records where } X = x\}}$$

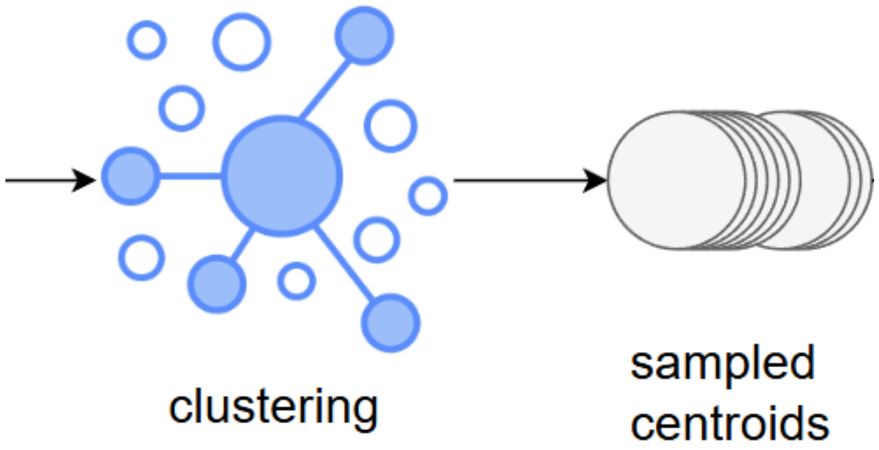
Functional Dependencies (FDs): $X \rightarrow Y$ means rows with the same X value must have the same Y value.

Use support (coverage) and confidence (accuracy) to measure the strength of a functional dependency.

city	zip	county	s(city, zip)	c(city | zip)	s(city, county)	c(city | county)
elba	36,323	coffee	4/8=0.5	4/5=0.8	3/8=0.375	3/6=0.5
elba	36,323	coffee	0.5	0.8	0.375	0.5
elba	36,323	coffee	0.5	0.8	0.375	0.5
enterprise	36,330	cxffee	2/8=0.25	2/3=0.667	1/8=0.125	1
enterprise	36,330	coffee	0.25	0.667	1/8=0.125	1/6=0.167
elba	36,323	coffxx	0.5	0.8	1/8=0.125	1
xntxrprisx	36,330	coffee	1/8=0.125	1/3=0.333	1/8=0.125	1/6=0.167
exba	36,323	coffee	1/8=0.125	1/5=0.2	1/8=0.125	1/6=0.167

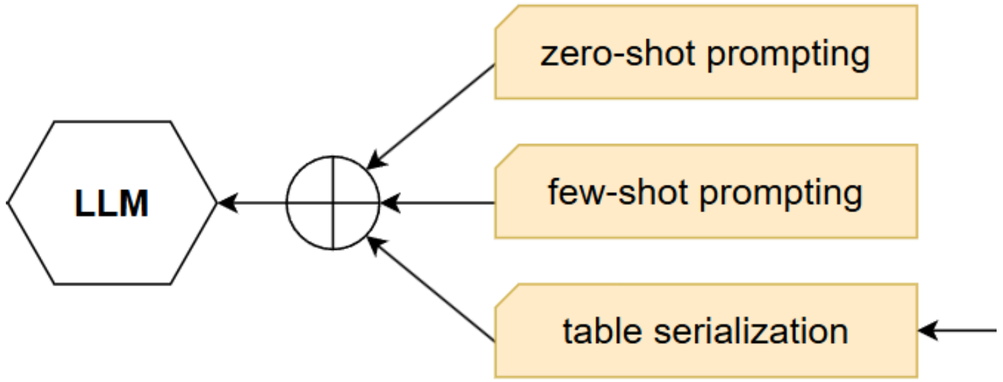
→ statistical feature

Methodology - Sampling



- **Clustering Algorithm**: use Mini-Batch K-Means, a fast approximation of k-means on large datasets, with k clusters.
- **Centroids Sampling**: for each cluster, compute the distance from all points to the centroid; select the nearest point as the cluster representative.

Methodology - Labeling with LLMs



You are an error detection assistant.
Respond ONLY in valid JSON.
[Zero-shot prompt]
[Few-shot prompt]
[Table representation]
[Output template]

prompt overview

1. Zero-shot Prompting

Use general instructions or specific rules.

Instructions:
- You are given a list of data rows extracted from a dataset, where the target column named `county`. Label each value from the target column as either an error or not an error. An error could be a typo or a formatting issue.
- Only evaluate the `county` value.
- Only label values you are confident about. If you are not sure, do not label it as an error.

```
[{
   "name": "county",
   "meaning": "County where the healthcare provider is located.",
   "data_type": "string",
   "format_rule": "no specific format",
   "null_value_rule": "not allowed"
},
...
]
```

Methodology - Labeling with LLMs

2. **Few-shot Prompting**
   Provide a small number of annotated examples.

3. **Table Representation**
   Serialize each table row into a JSON-style structure.

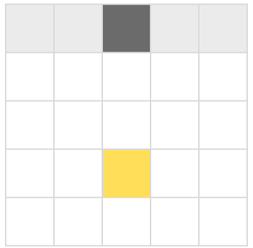
```
Here are some examples:

Example 1:
Row: [Data Instance 1]
Column: county
Original value: dx kalb
Corrected value: de kalb
```

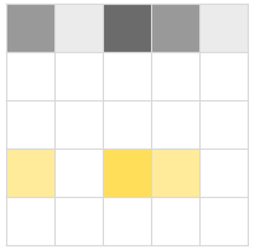
few-shot example

```
{
  "row_id": <row index>,
  "attr_1": <value_1>,
  "attr_2": <value_2>,
  ...
  "attr_n": <value_n>
}
```

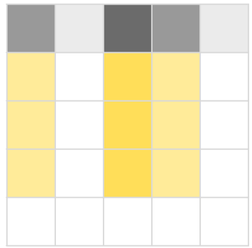
serialization example



target only



related attr.

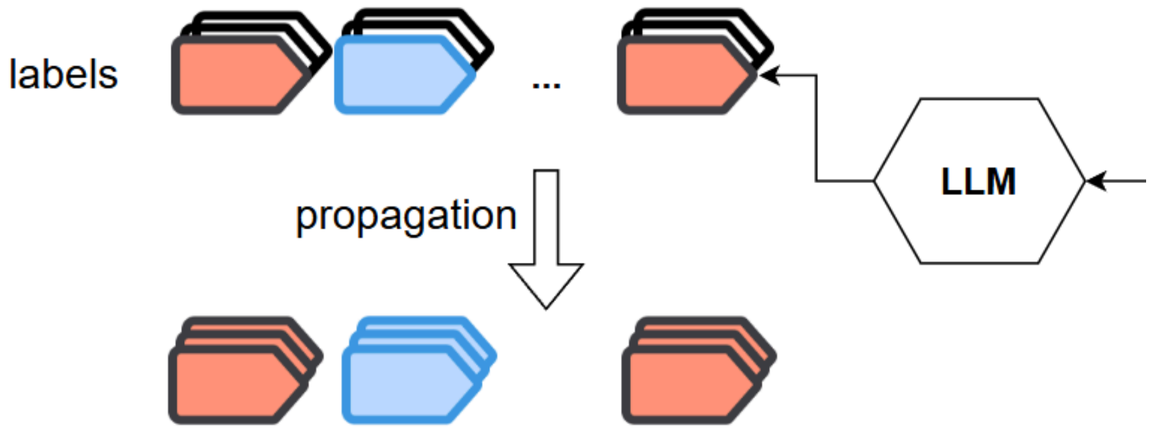


batch

- **target attribute**: target value only.
- **related attributes**: attributes with high normalized mutual information (NMI).
- **batch prompts**: multipe data instances.

Methodology - Label Propagation

- **parse LLM responses**: extract row index and predicted label from JSON.
- **cluster assumption**: instances within the same cluster are likely to share the same class label.
- **robustness handling**: If response format is invalid or incomplete → default label as normal value.



Experiments

Experimental Setup: dataset, metrics, baseline

Results: comparison study, ablation study

Experiments - Experimental Setup

Dataset	Err. Rate	Tuples	Attributes	Err. Types
Hospital	2.96%	1,000	18	T, RV
Flights	29.58%	2,376	7	RV, MV, PV
Beers	16.71%	2,410	11	RV, MV, PV
Spotify	10.00%	10,000	11	T, RV, MV, O

Summary of benchmark datasets used for evaluation. Error types consist of typo (T), rule violation (RV), missing value (MV), pattern violations (PV) and outlier (O).

Metrics

Precision, Recall and F1 score.

Baseline

Traditional method: Raha

LLM-based method: ZeroED

Datasets

Flights, *Hospital*, and *Beers* datasets are commonly used, *Spotify* is a new synthetic dataset.

Simulate errors by 3 operations: character replacement, deletion and duplication.

Experiments - Results

Approach	Hospital			Flights			Beers			Spotify		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Raha	81.9	54.3	64.1	84.6	**81.2**	**82.8**	96.2	**96.0**	96.0	58.8	43.6	49.4
Zero-ED	93.6	71.5	81.1	93.5	58.6	72.2	88.8	68.9	77.4	-	-	-
Llama-3.1-8B	23.8	**93.1**	37.9	56.9	57.9	57.4	51.8	74.5	61.1	41.9	**70.4**	52.5
Qwen3-8B	**99.8**	85.7	**92.2**	95.2	57.9	72.0	96.7	91.2	93.9	50.4	67.0	57.5
Qwen3-4B	82.1	85.0	83.5	91.6	57.2	70.4	96.0	81.0	87.9	53.3	64.5	58.3
Gemma-3-27B-IT	**99.8**	85.7	**92.2**	**100**	57.7	73.1	90.2	93.0	91.6	**79.5**	67.4	**73.0**
Gemma-3-12B-IT	86.8	87.4	87.1	96.6	57.6	72.2	**96.8**	95.7	**96.3**	70.9	62.2	66.3
Gemma-3-4B-IT	60.5	79.9	68.9	88.9	45.9	60.5	91.1	89.8	90.5	54.3	65.1	59.2
Ground Truth	99.8	93.1	96.3	77.6	89.9	79.2	96.8	95.7	96.3	75.4	70.9	73.0

Comparative evaluation of different error detection approaches, measured in percentages (%).
"–" indicates data not reported or applicable in prior works for this dataset.

Comparison Study

- Gemma-3–27B-IT → Best or near-best performance across datasets
- Our Method: Best when typos dominate
- LLM Limitations: abstract / complex values, numeric data, statistical / distributional errors

Experiments - Results

	tu…	src	flight	sched_dep…	act_dep…	sched_arr…	act_arr…
1	28	aa	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
2	102	helloflight	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:43 p.m.
3	201	boston	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
4	389	airtravelcen…	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:43 p.m.
5	488	flightview	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
6	588	flightstats	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
7	680	panynj	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
8	765	flightexplor…	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:43 p.m.
9	862	flights	AA-1007-MIA-PHX	<null>	5:08 p.m.	<null>	7:55 p.m.
10	1121	foxbusiness	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
11	1275	myrateplan	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:43 p.m.
12	1374	orbitz	AA-1007-MIA-PHX	4:55 p.m.	4:56 p.m.	8:05 p.m.	7:55 p.m.
13	1539	flytecomm	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:43 p.m.
14	1613	mytripandmore	AA-1007-MIA-PHX	4:55 p.m.	4:56 p.m.	8:05 p.m.	7:55 p.m.
15	1700	wunderground	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	7:38 p.m.	7:43 p.m.
16	1809	flylouisville	AA-1007-MIA-PHX	<null>	5:08 p.m.	<null>	7:55 p.m.
17	1909	quicktrip	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
18	2001	allegiantair	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
19	2101	businesstrav…	AA-1007-MIA-PHX	<null>	5:08 p.m.	<null>	7:55 p.m.
20	2206	gofox	AA-1007-MIA-PHX	<null>	5:09 p.m.	<null>	7:53 p.m.
21	2294	flightaware	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	7:38 p.m.	7:43 p.m.

flights - dirty

	tı	src	flight	sched_…	act_de…	sched_…	act_ar…
1	28	aa	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
2	102	helloflight	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
3	201	boston	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
4	389	airtravelcen…	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
5	488	flightview	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
6	588	flightstats	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
7	680	panynj	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
8	765	flightexplorer	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
9	862	flights	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
10	1121	foxbusiness	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
11	1275	myrateplan	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
12	1374	orbitz	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
13	1539	flytecomm	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
14	1613	mytripandmore	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
15	1700	wunderground	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
16	1809	flylouisville	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
17	1909	quicktrip	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
18	2001	allegiantair	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
19	2101	businesstrav…	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
20	2206	gofox	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.
21	2294	flightaware	AA-1007-MIA-PHX	4:55 p.m.	5:08 p.m.	8:05 p.m.	7:55 p.m.

flights - clean

Experiments - Results

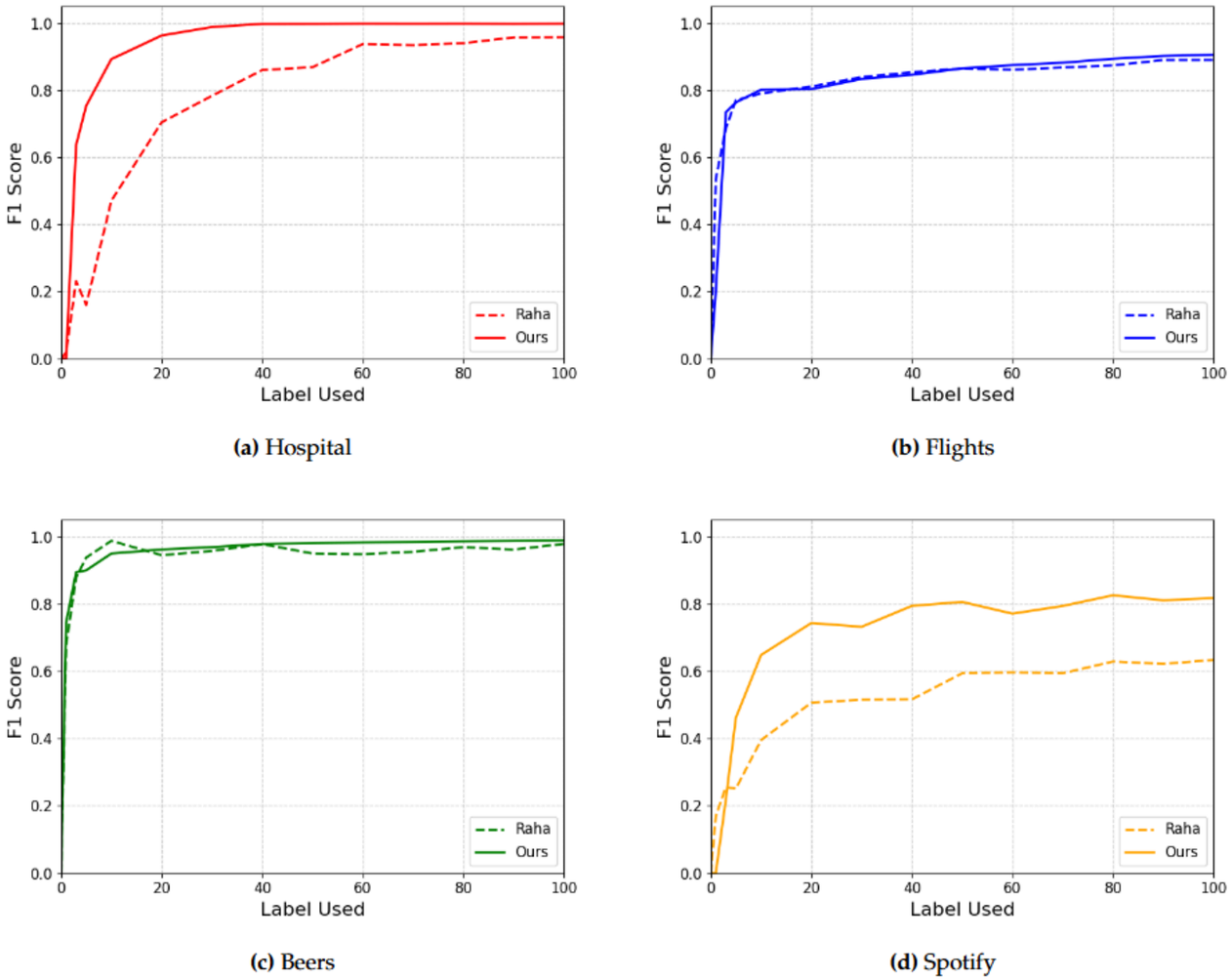


Figure 4.1: Error detection performance between Raha and our approach

Comparison Study

compare with raha when using ground truth labels.

- Better performance and efficiency in detecting typos;
- Performance of FDs violation detection is similar to Raha.

Experiments - Results

Strategy	Hospital			Flights			Beers		
	P	R	F_1	P	R	F_1	P	R	F_1
ZS-I+V	62.45	83.30	71.38	**100**	56.67	72.26	88.65	17.01	28.54
ZS-R+V	95.41	85.74	90.32	**100**	57.66	**73.15**	96.05	75.72	84.68
ZS-I+FS+V	60.71	**93.06**	73.48	**100**	41.34	58.50	95.85	38.68	55.11
ZS-R+FS+V	84.07	90.06	88.34	**100**	57.66	**73.15**	95.59	67.52	79.14
ZS-I+A	68.62	83.30	75.25	96.78	47.62	63.83	85.26	11.14	19.71
ZS-R+A	**99.78**	85.74	**92.23**	95.19	**57.93**	72.02	96.70	**91.24**	**93.89**
ZS-I+FS+A	65.50	87.62	74.96	96.68	46.10	62.43	94.48	28.66	43.98
ZS-R+FS+A	96.01	85.74	90.59	97.32	57.66	72.42	**98.86**	65.43	78.74

ZS-I: zero-shot prompting with general instructions; **ZS-R**: zero-shot prompting with specific rules; **FS**: few-shot prompting; **V**: use value from the target attribute only; **A**: use values from related attributes.

Ablation study of different prompting strategies across datasets using Qwen3-8B, measured in percentages (%). Batch size is 10.

Ablation Study

compare different prompting strategies.

ZS-R+A outperforms others overall.

FS is effective for ZS-I, but not for ZS-R.

compare different batch size.

larger batch size improve both performance and efficiency.

Batch Size	F_1 (%)	Token Count	Time (s)	Rate (token/s)
1	74.19	104,874	527	198.97
2	79.71	66,474	354	187.81
4	94.99	47,247	263	179.63
10	92.23	35,691	211	169.15
20	94.79	32,047	193	165.99

Effect of different batch sizes on performance and efficiency. Measured on the Hospital datatset, using Gemma-3-27B-IT, zero-shot prompting with specific rules and values from related attributes.

Discussion & Conclusion

1. Limitations:
   - LLM Hallucination
   - Data Leakage
   - Numerical and Temporal Data
   - Small-Scale LLMs
2. Future Work:
   - Ensemble
   - Self-Improvement
   - Code Generation
   - Data Wrangling

SUMMARY

1. BPE tokenizer in feature engineering;
2. Clustering-based sampling and small-scale LLMs → improve efficiency and reduce cost;
3. zero-shot with rules combined with related attributes → enhance performance, minimal human involvement;
4. batch prompting → improves model inference efficiency.

Thank You