Entity Matching using Large Language Models

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Ralph Peeters, Aaron Steiner and Christian Bizer

EDBT ICDT BARCELONA 2025
March 25-28, 2025, Barcelona, Spain

Entity Matching

UNIVERSITY OF MANNHEIM
Data and Web Science Group

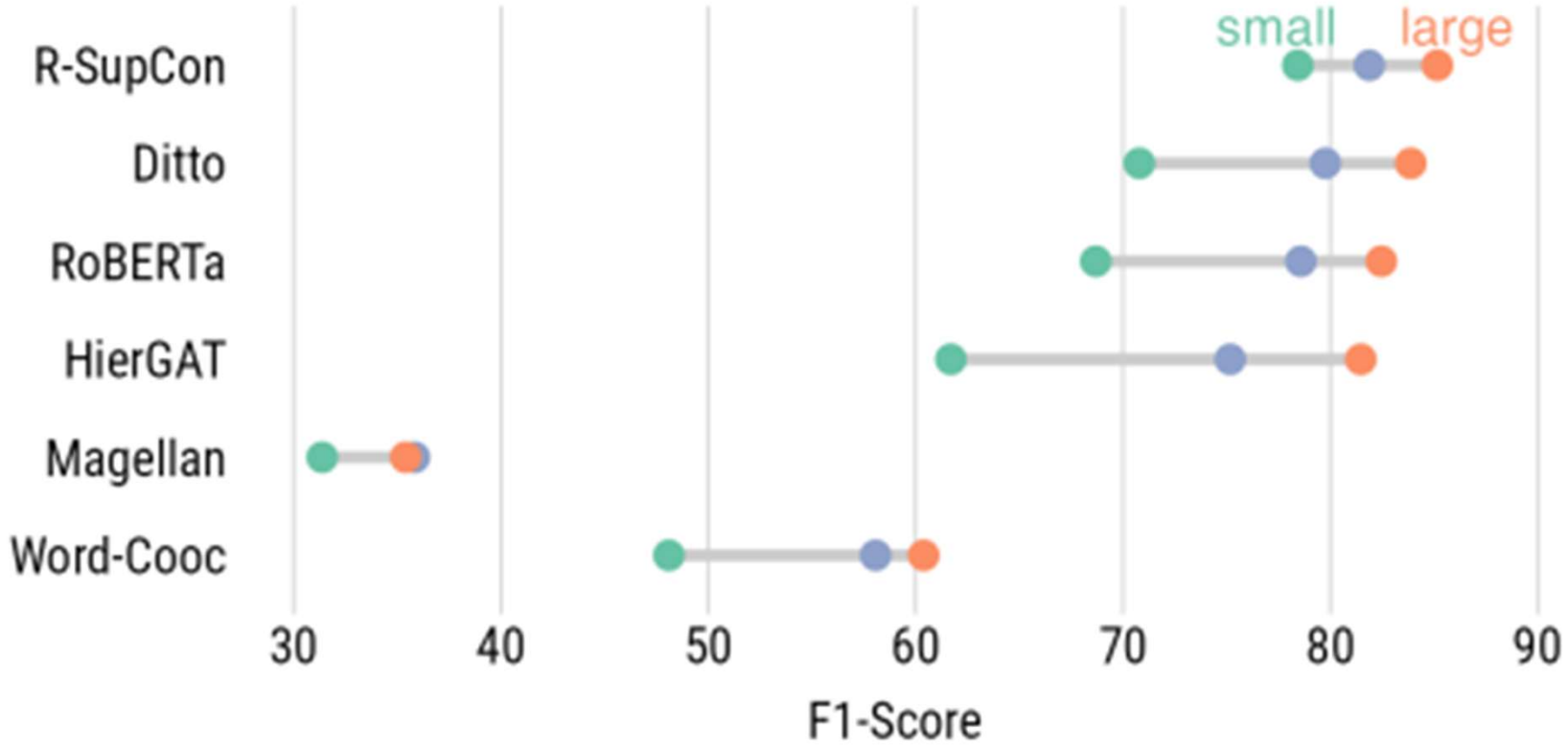
- **Goal:** Find all records that refer to the same real-world entity.
- Difficult due to **record heterogeneity**
  - different surface forms, units of measurement
  - typos, missing values, wrong values, …

Brand	Title	Description	Price	Currency
null	Epson Photo Paper Glossy A3 255gsm White 20 Sheets	Smudge,water-resistant and promises long-lasting durability	55.19	GBP
D'Addario	D'Addario EXL125-3D XL Electric Guitar SL Top/Reg Bottom 9-46	3 Sets, Super Light (9-46) Nickel Wound	13.99	USD
Corsair	Corsair Vengeance RGB Pro 32GB	(4x8GB) 3000MHz CL15 DDR4 White available	299	AUD

easy match
hard match
easy non-match
hard non-match

Brand	Title	Description	Price	Currency
null	Epson Premium Glossy A3 20 sheets	Specifically designed for the requirements of the Epson printers	36.43	GBP
D'Addario	D'Addario 3-Pack Nickel Wound Electric Strings (9-46)	Precision wound with nickel plated steel on a hex-shaped core	10.95	USD
null	Ernie Ball Electric Guitar Strings	Turbo Slinky Nickel Wound Strings (9.5-46) 3 Pack	16.95	USD
Corsair	Corsair Vengeance RGB Pro 32GB	(2x16GB) DDR4-3200 C16 Kit	668.00	RM

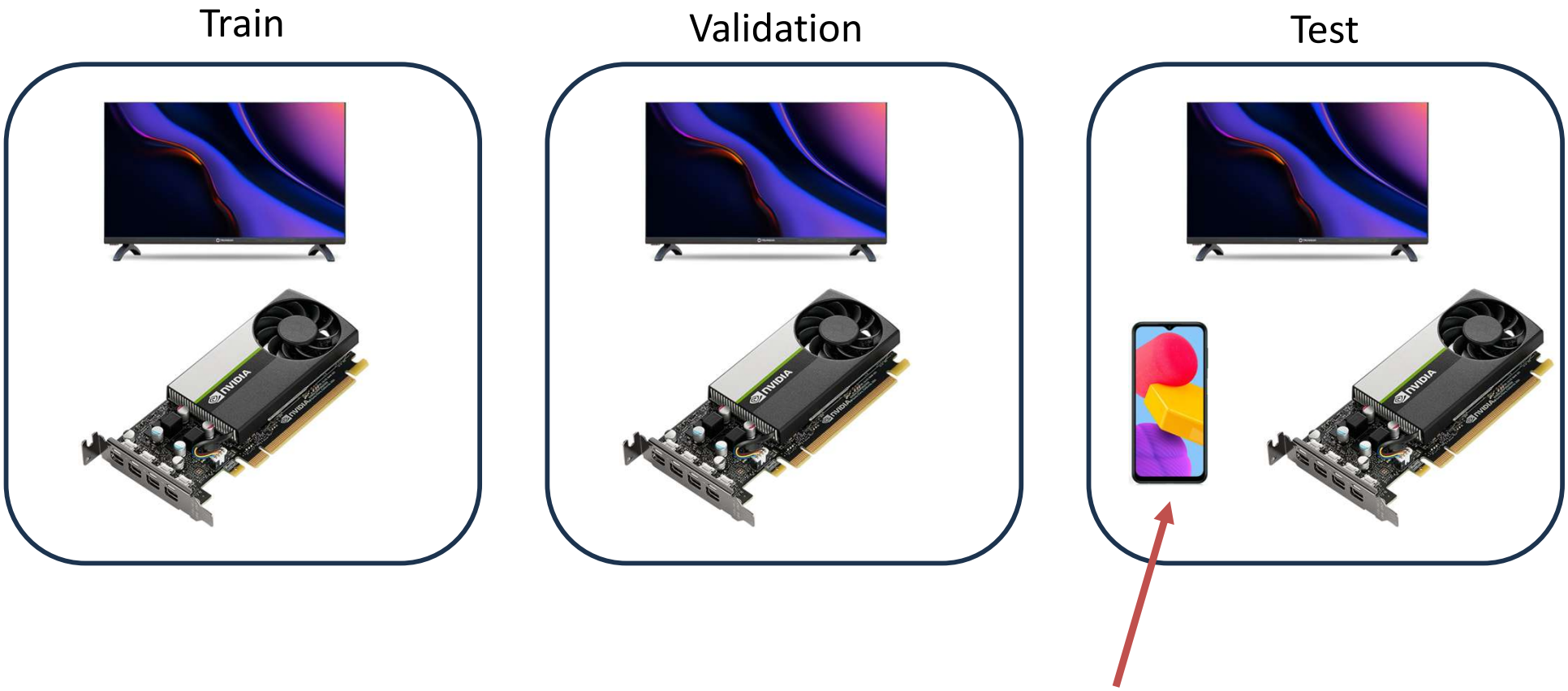
Matchers can handle the heterogeneity given enough training data



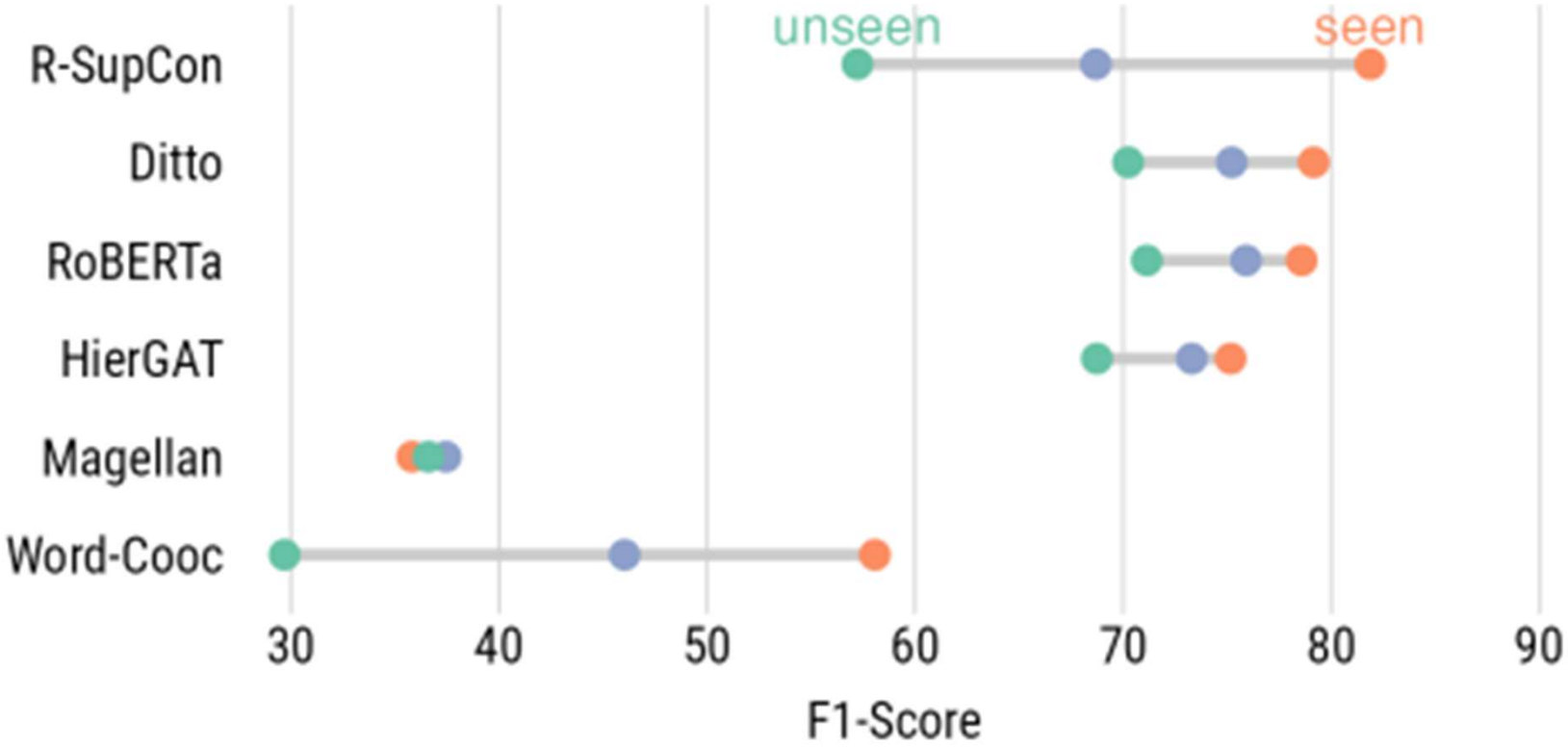
Benchmark: WDC Products, 80% corner cases

Small: 5K examples, medium: 10K examples, large: 25K examples

Challenge: Unseen Entities



Matchers to not generalize well to unseen entities



Motivation of this work

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Generative LLMs have shown **strong zero-shot performance** on a wide range of tasks in natural language processing

- Potentials of generative LLMs for entity matching
  1. require less task-specific training data
  2. ability to match unseen entities

Contribution: Broad Experimental Study on Using LLMs for Entity Matching

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Test various prompting techniques
  - Scenario 1: No training data
  - Scenario 2: With training data
- Investigate prompt sensitivity
- Compare LLMs and PLMs
- Fine-tune LLMs for entity matching
- LLMs for explanations and automated error analysis

Evaluated LLMs and Baselines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Large Language Models:

- GPT-mini (gpt-4o-mini-2024-07-18) – Hosted
- GPT-4 (gpt4-0613) – Hosted
- GPT-4o (gpt-4o-2024-08-06) – Hosted
- Llama2 (Llama-2-70B-chat-hf) – Open-source
- Llama3.1 (Meta-Llama-3.1-70B-Instruct) – Open-source
- Mixtral (Mixtral-8x7B-Instruct-v0.1) – Open-source

Baselines:

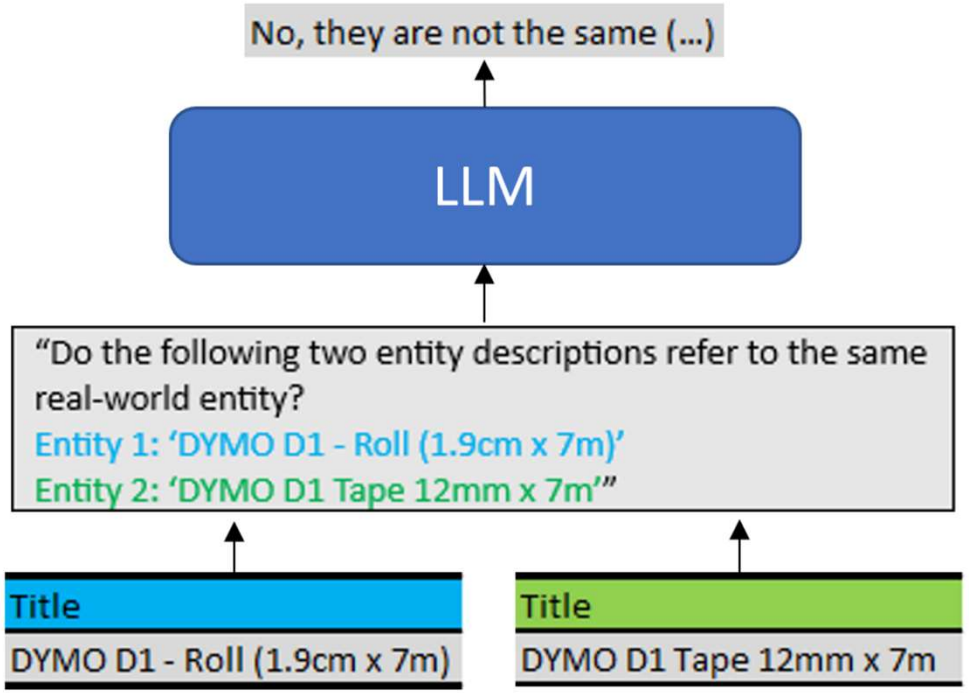
- RoBERTa: PLM fine-tuned for entity matching
- Ditto: PLM-based matching system using data augmentation

Benchmarks

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Dataset	Training set		Validation set		Test set	
	# Pos	# Neg	# Pos	# Neg	# Pos	# Neg
(WDC) - WDC Products	500	2,000	500	2,000	259	989
(A-B) - Abt-Buy	616	5,127	206	1,710	206	1,000
(W-A) - Walmart-Amazon	576	5,568	193	1,856	193	1,000
(A-G) - Amazon-Google	699	6,175	234	2,059	234	1,000
(D-S) - DBLP-Scholar	3,207	14,016	1,070	4,672	250	1,000
(D-A) - DBLP-ACM	1,332	6,085	444	2,029	250	1,000

Scenario 1: Zero-Shot Prompting



We experiment with different prompt variations

- Domain-specific vs. general wording
- Simple vs. complex language
- Forcing the model to respond with Yes/No vs. allowing long answers

Scenario 1: Zero-Shot - Results

Prompt	All Datasets (Average F1)					
	GPT-mini	GPT-4	GPT-4o	LLama2	Llama3.1	Mixtral
domain-complex-force	85.29	88.91	**87.00**	66.60	**84.87**	**68.29**
domain-complex-free	**85.40**	**89.46**	80.31	**69.69**	72.06	62.13
domain-simple-force	50.41	86.10	82.72	57.24	60.52	41.65
domain-simple-free	33.65	87.92	63.53	50.40	36.94	43.20
general-complex-force	83.50	87.94	85.02	63.89	83.26	59.51
general-complex-free	83.13	87.85	55.81	62.73	80.54	61.50
general-simple-force	52.88	81.12	83.65	62.31	72.65	33.59
general-simple-free	45.49	85.07	64.67	52.77	63.38	36.12
Narayan-complex	56.13	86.70	50.65	56.34	36.16	32.04
Narayan-simple	75.15	86.92	45.64	67.81	37.32	30.94
Mean	65.10	86.80	69.90	60.98	62.77	46.90
Standard deviation	18.45	2.26	14.86	6.18	18.54	13.68

- Most models are highly sensitive to prompt formulation
- Prompt engineering still matters!

Comparison to PLM Baselines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

	WDC	A-B	W-A	A-G	D-S	D-A
GPT-mini	81.15	91.93	86.58	72.18	86.11	97.60
GPT-4	**89.61**	**95.78**	**89.67**	76.38	89.82	98.41
GPT-4o	87.64	93.95	86.65	73.56	89.76	97.06
Llama2	69.09	82.03	63.91	57.93	85.46	97.62
Llama3.1	83.67	89.84	84.85	73.99	86.32	98.81
Mixtral	53.37	82.20	70.44	40.98	77.75	90.32
RoBERTa	77.53	91.21	87.02	79.27	93.88	**99.14**
Ditto	84.90	91.31	86.39	**80.07**	**94.31**	99.00
Δ best LLM/PLM	4.71	4.47	2.65	-3.69	-4.49	-0.33

– **Zero-shot LLMs** outperform **fine-tuned PLMs** on 50% of the datasets

Scenario 2: With Training Data
Few-Shot Prompting

Task Desc.	Do the following two product descriptions match?
Demonstr-ations	Product 1: 'Title: DYMO D1 19 mm x 7 m' Product 2: 'Title: Dymo D1 (19mm x 7m – BoW)'
Answer	Yes.
Task Desc.	Do the following two product descriptions match?
Demonstr-ations	Product 1: 'Title: DYMO D1 Tape 24mm' Product 2: 'Title: Dymo D1 19mm x 7m'
Answer	No.
Task Desc.	Do the following two product descriptions match?
Task Input	Product 1: 'Title: DYMO D1 – Roll (1.9cm x 7m)' Product 2: 'Title: DYMO D1 Tape 12mm x 7m'

– Demonstration selection strategies: Random, string similarity, handpicked

Scenario 2: With Training Data Handwritten and Learned Rules

Task Desc.	The following rules regarding product features need to be observed:
Hand-Written Rules	1. The brand of matching products must be the same if available. 2. Model names of matching products must be the same if available. 3. Model numbers of matching products must be the same if available. 4. Additional features of matching products must be the same if available. 5. Matching attributes may not have the exact same surface form due to different case, typos, value formats. 6. If an attribute is missing for one description, it is likely still a match if the existing attributes match.
Task Desc.	Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
Task Input	Product 1: 'Title: DYMO D1 – Roll (1.9cm x 7m)' Product 2: 'Title: DYMO D1 Tape 12mm x 7m'

Results with Training Data

Prompt	Shots	All Datasets (Mean F1)					
		GPT4-mini	GPT4	GPT4o	LLama2	Llama3.1	Mixtral
Fewshot-related	6	73.76	90.24	90.41	65.44	82.12	50.51
	10	76.56	**90.80**	**91.21**	62.69	85.85	53.25
Fewshot-random	6	77.86	89.44	89.77	63.99	85.95	57.37
	10	80.51	89.05	89.85	65.62	**88.06**	53.94
Fewshot-handpicked	6	72.81	88.61	89.44	70.52	84.87	57.76
	10	73.93	88.76	89.52	69.91	87.60	51.03
Hand-written rules	0	81.49	87.65	86.36	51.22	85.57	**79.03**
Learned rules	0	84.14	86.64	84.96	44.23	84.11	74.53
Mean	-	77.63	88.90	88.94	61.70	85.51	59.68
Standard deviation	-	3.85	1.25	2.00	8.63	1.77	10.23
Best zero-shot	0	**85.51**	89.95	88.10	**76.01**	86.25	69.18
Δ Few-shot/zero-shot	-	-5.00	0.85	3.10	-5.49	1.81	-11.42
Δ Rules/zero-shot	-	-1.37	-2.29	-1.74	-24.78	-0.68	9.86

- Usefulness of few-shot prompting depends on model
- Few-shot generally more helpful than rules

Fine-tuning LLMs for Entity Matching

UNIVERSITY OF MANNHEIM
Data and Web Science Group

		WDC	A-B	W-A	A-G	D-S	D-A
Δ best zero-shot	Llama2	-2.28	+10.12	+26.66	+18.26	+7.34	+1.58
	Llama3.1	-5.80	+3.76	+6.16	+4.68	+6.05	+0.79
	GPT-mini	+7.74	+3.01	+6.41	+14.93	+7.84	+1.80
Δ best GPT4	Llama2	-22.8	-3.63	+0.90	-0.19	+2.98	+0.79
	Llama3.1	-11.74	-2.18	+1.34	+2.29	+2.55	+1.19
	GPT-mini	-0.72	-0.84	+3.32	+10.73	+4.13	+0.99
Best GPT4	-	**89.61**	**95.78**	89.67	76.38	89.82	98.41

- Fine-tuned GPT-mini reaches similar performance as zero-shot GPT-4
- Fine-tuned LLMs still generalize well to unseen data sets

Explaining Matching Decisions

Actor	Prompt (Walmart-Amazon)
User	"Do the two entity descriptions refer to the same real-world entity? Answer with 'Yes' if they do and 'No' if they do not. **Entity 1:** night owl night owl zeus-810 16-channel h. 264 dvr kit with 1tb hard drive 773.82 **Entity 2:** night owl optics security zeus-810 16-channel 8-camera h. 264 dvr surveillance kit with d1 recording and hdmi output zeus-810 676.63"
AI	Yes.
User	Now **explain your decision in a structured format**, listing the attributes that you compared for reaching your decision. Each attribute should be accompanied by the attribute values and a score between -1 and 1 that shows the **importance of the attribute** for the decision. If the attribute influenced the decision towards non-match the importance score should be negative. If the attribute pointed towards a match, the importance score should be positive. Also provide a **similarity score for the attribute values**. If an attribute only occurs in one item, specify the value of that attribute for the other item as "missing". An example output is the following: attribute=brand - importance=0.05 - values=Logitech###Logitech - similarity=1.00 attribute=model - importance=-0.95 - values=MX G500###MX Master 3S - similarity=0.20 attribute=color - importance=0.00 - values=missing###Graphite - similarity=0.00'

Prompt LLM to explain decision in a structured format, listing:

- relevant attributes
- attribute importance
- similarity scores

Example Explanation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Product 1: ZEBRA 800033-801 BLACK RIBBON (1000 PRINTS) GBP 9.5
Product 2: Zebra Black Monochrome Ribbon 800011-101 - 1000 Prints GBP £12.99

ASSISTANT:
attribute=brand; importance=0.20; values=ZEBRA, Zebra; similarity=1.00

attribute=model; importance=-0.40; values=800033-801, 800011-101; similarity=0.00

attribute=color; importance=0.10; values=Black, Black; similarity=1.00

attribute=prints; importance=0.20; values=1000, 1000; similarity=1.00

attribute=price; importance=0.10; values=GBP 9.5, GBP £12.99; similarity=0.00

Aggregating Structured Explanations

UNIVERSITY OF MANNHEIM
Data and Web Science Group

	Matches			Non-Matches		
Attribute	Freq.	Mean Import.	St.Dev.	Freq.	Mean Import.	St.Dev.
DBLP-Scholar						
title	0.96	**0.59**	0.40	0.95	-0.40	0.38
authors	0.78	**0.65**	0.40	0.68	**-0.66**	0.34
conference	0.50	0.35	0.37	0.29	-0.11	0.29
year	0.46	0.26	0.37	0.43	-0.16	0.25
journal	0.14	0.40	0.43	0.05	-0.15	0.25
Walmart-Amazon						
brand	0.98	**0.78**	0.34	0.99	-0.04	0.34
price	0.92	-0.03	0.27	0.86	-0.16	0.25
model	0.81	**0.63**	0.51	0.82	**-0.77**	0.37
color	0.24	0.23	0.31	0.35	-0.06	0.23
product type	0.12	**0.64**	0.48	0.11	-0.42	0.50

– Possible to derive global insights into model behaviour

Automating Error Analysis using Structured Explanations

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Actor	Prompt
User	The following list contains false positive and false negative product pairs from the output of a product matching classification system. **Given the product pairs and the associated explanations, come up with a set of error classes**, separately for both false positives and false negatives, **that explain why the classification systems fails on these examples.** **False Negatives:** **False Positives:** FN1 FP1 {Entity 1} {Entity 1} {Entity 2} {Entity 2} {Explanation} {Explanation} {additional FNs} {additional FPs}

Created Error Classes – DBLP-Scholar

False Negatives (26 overall)	# errors
1. Year Discrepancy: Differences in publication years lead to false negatives, even when other attributes match closely.	8
2. Venue Variability: Variations in how the publication venue is listed (e.g., abbreviations, full names) cause mismatches.	14
3. Author Name Variations: Differences in author names, including initials, order of names, or inclusion of middle names, lead to false negatives.	9
4. Title Variations: Minor differences in titles, such as missing words or different word order, can cause false negatives.	11
5. Author List Incompleteness: Differences in the completeness of the author list, where one entry has more authors listed than the other.	11
False Positives (26 overall)	**# errors**
1. Overemphasis on Title Similarity: High similarity in titles leading to false positives, despite differences in other critical attributes.	15
2. Author Name Similarity Overreach: False positives due to high similarity in author names, ignoring discrepancies in other attributes.	16
3. Year and Venue Ignored: Cases where the year and venue match or are close, but other discrepancies are overlooked.	5
4. Partial Information Match: Matching based on partial information, such as incomplete author lists or titles, leading to false positives.	19
5. Misinterpretation of Publication Types: Confusing different types of publications (e.g., conference vs. journal) when other attributes match.	9

Conclusion

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. Experiments have shown that LLMs require less/partly no domain-specific training data compared to PLMs
2. LLM have demonstrated the ability to match unseen entities
3. LLMs can explain matching decisions
4. LLMs can analyze matching errors

Thank you for your attention!

Ready for your questions!

Link to Paper

Link to Code