Fine-tuning LLMs for Entity Matching

Aaron Steiner, Ralph Peeters and Christian Bizer

UNIVERSITY OF MANNHEIM
Data and Web Science Group

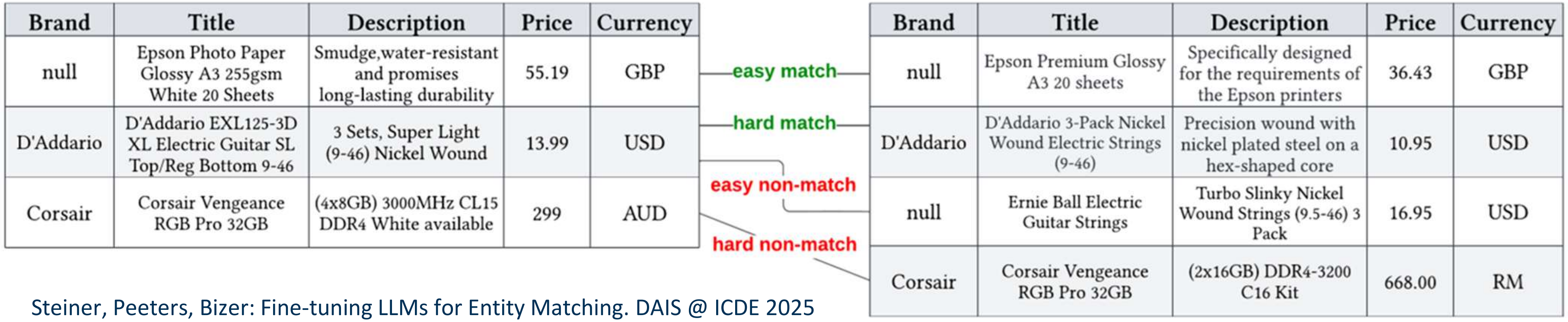
Agenda

1. Introduction to entity matching
2. LLMs for entity matching
3. Fine-tuning LLMs for entity matching
4. Example Representation
5. Example Selection and Generation

1. Entity matching - A brief introduction

- Task
  - Find entity descriptions referring to the same real-world entities in one or across multiple data sources.
- Challenges:
  - Different types of heterogeneity in entity descriptions

    → (non-)matching descriptions can look very (similar) different
  - Missing values: Some feature values may be absent in one of the descriptions



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

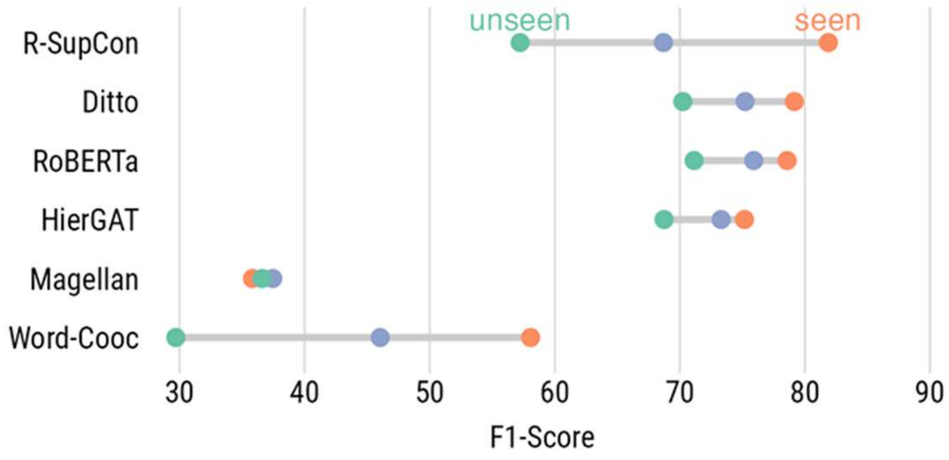
Brand	Title	Description	Price	Currency
null	Epson Premium Glossy A3 20 sheets	Specifically designed for the requirements of the Epson printers	36.43	GBP
D'Addario	D'Addario 3-Pack Nickel Wound Electric Strings (9-46)	Precision wound with nickel plated steel on a hex-shaped core	10.95	USD
null	Ernie Ball 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

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

1. Entity matching – SOTA using PLMs Limitations

- Sensitivity to Unseen Entities:
  - Poor accuracy when encountering entities or domains not represented during training
  - Challenges arise from a lack of robustness to variations in entity representation
- Requirement for large amounts Training Data:
  - Effective (~25K labels for 80% F1) performance typically demands large-scale training corpora
  - The high training data requirement limits practical applicability, especially if training data is cost intensive to acquire

  → Continuous labeling and retraining effort necessary



Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

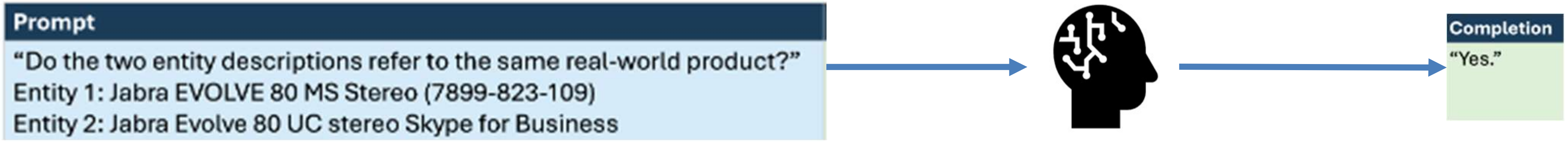
2. Motivation for Using LLMs in Entity Matching

1. Can LLMs Generalize Better to Unseen Entities?
   - Could leveraging extensive world knowledge help models adapt more effectively to new and unseen domains?
2. Can LLMs perform well with Fewer Training Examples?
3. Can LLMs Improve Interpretability and Reflect Uncertainty Better?
   - Would LLMs' richer, nuanced outputs provide better insights into model decisions compared to PLMs' overly confident softmax predictions?

2. Using LLMs for EM

Note:

- Various prompt formulations can be employed, affecting the model's accuracy and clarity.
- While different prompts were tested, **this presentation primarily focuses on fine-tuning strategies and example representation**, rather than prompt optimization.



Prompt

"Do the two entity descriptions refer to the same real-world product?"
Entity 1: Jabra EVOLVE 80 MS Stereo (7899-823-109)
Entity 2: Jabra Evolve 80 UC stereo Skype for Business

Completion

"Yes."

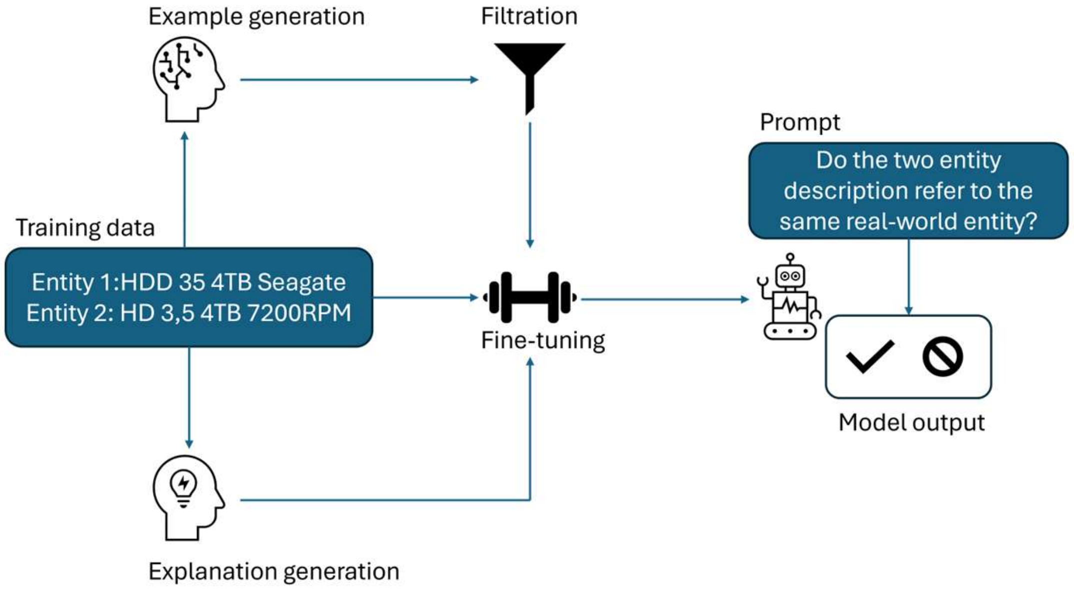
2. Fine-tuning approaches
Dimensions to Explore

We explore:

- Fine-tuning without alterations
- Fine-tuning with example augmentation
- Fine-tuning with example generation and filtration

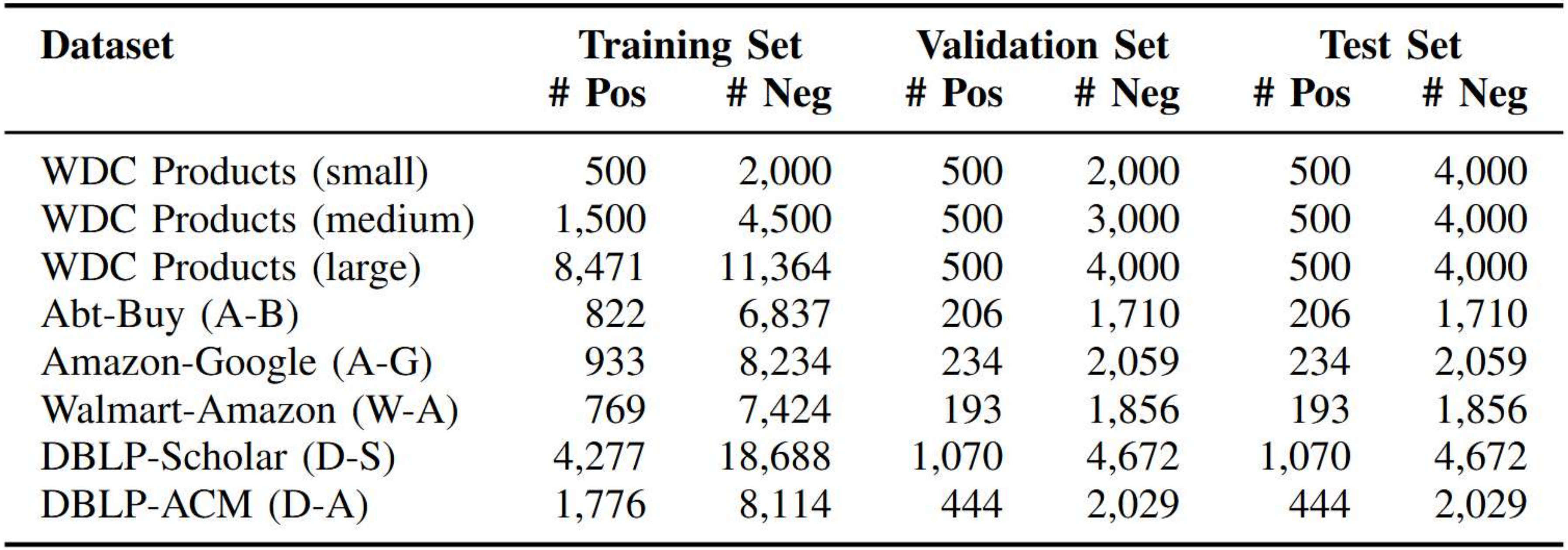
Effects are compared:

- No domain shift
- In-domain shift
- Cross-domain shift (paper only)


Example generation
Filtration
Prompt
Do the two entity description refer to the same real-world entity?
Training data
Entity 1:HDD 35 4TB Seagate
Entity 2: HD 3,5 4TB 7200RPM
Fine-tuning
Model output
Explanation generation

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

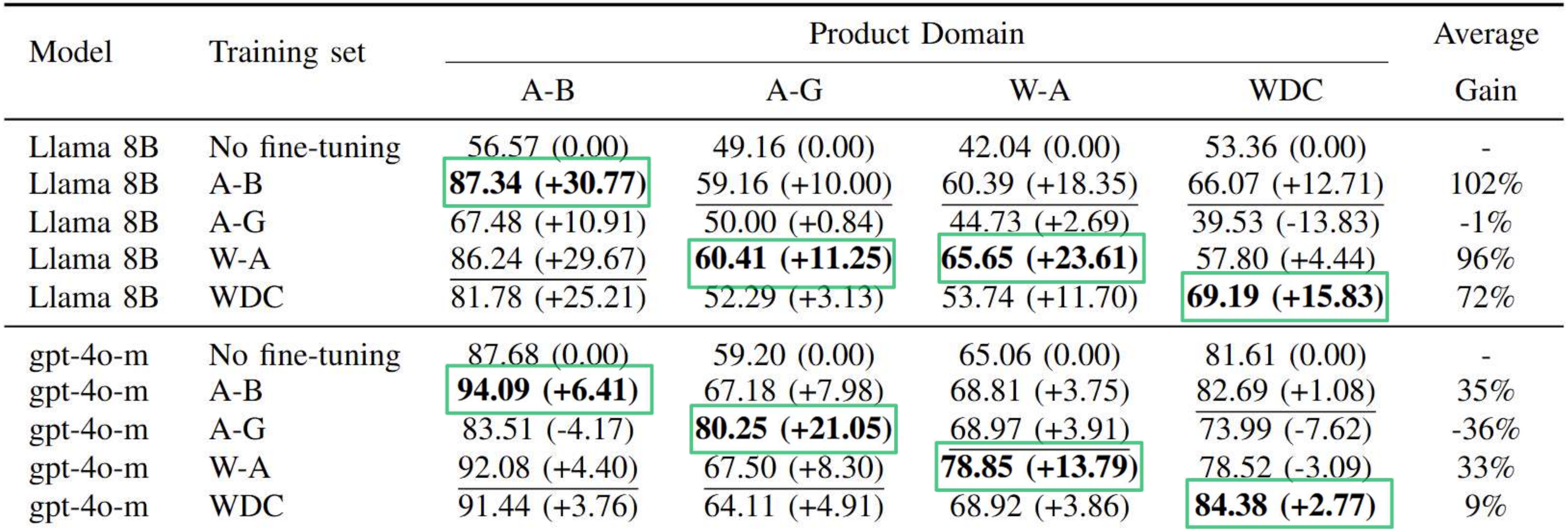
3. Evaluation Datasets used for Experiments



Dataset	Training Set		Validation Set		Test Set	
	# Pos	# Neg	# Pos	# Neg	# Pos	# Neg
WDC Products (small)	500	2,000	500	2,000	500	4,000
WDC Products (medium)	1,500	4,500	500	3,000	500	4,000
WDC Products (large)	8,471	11,364	500	4,000	500	4,000
Abt-Buy (A-B)	822	6,837	206	1,710	206	1,710
Amazon-Google (A-G)	933	8,234	234	2,059	234	2,059
Walmart-Amazon (W-A)	769	7,424	193	1,856	193	1,856
DBLP-Scholar (D-S)	4,277	18,688	1,070	4,672	1,070	4,672
DBLP-ACM (D-A)	1,776	8,114	444	2,029	444	2,029

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

3. Results of Standard Fine-tuning



Model	Training set	Product Domain				Average
		A-B	A-G	W-A	WDC	Gain
Llama 8B	No fine-tuning	56.57 (0.00)	49.16 (0.00)	42.04 (0.00)	53.36 (0.00)	-
Llama 8B	A-B	**87.34 (+30.77)**	59.16 (+10.00)	60.39 (+18.35)	66.07 (+12.71)	102%
Llama 8B	A-G	67.48 (+10.91)	50.00 (+0.84)	44.73 (+2.69)	39.53 (-13.83)	-1%
Llama 8B	W-A	86.24 (+29.67)	**60.41 (+11.25)**	**65.65 (+23.61)**	57.80 (+4.44)	96%
Llama 8B	WDC	81.78 (+25.21)	52.29 (+3.13)	53.74 (+11.70)	**69.19 (+15.83)**	72%
gpt-4o-m	No fine-tuning	87.68 (0.00)	59.20 (0.00)	65.06 (0.00)	81.61 (0.00)	-
gpt-4o-m	A-B	**94.09 (+6.41)**	67.18 (+7.98)	68.81 (+3.75)	82.69 (+1.08)	35%
gpt-4o-m	A-G	83.51 (-4.17)	**80.25 (+21.05)**	68.97 (+3.91)	73.99 (-7.62)	-36%
gpt-4o-m	W-A	92.08 (+4.40)	67.50 (+8.30)	**78.85 (+13.79)**	78.52 (-3.09)	33%
gpt-4o-m	WDC	91.44 (+3.76)	64.11 (+4.91)	68.92 (+3.86)	**84.38 (+2.77)**	9%

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

4. Key Findings: Example Augmentation

- Main Findings
  - Adding explanations to the fine-tuning process boosts performance, especially for smaller models.
  - Explanations improve in-domain generalization.
  - The surface form of the explanation matters — structured > textual in most cases.

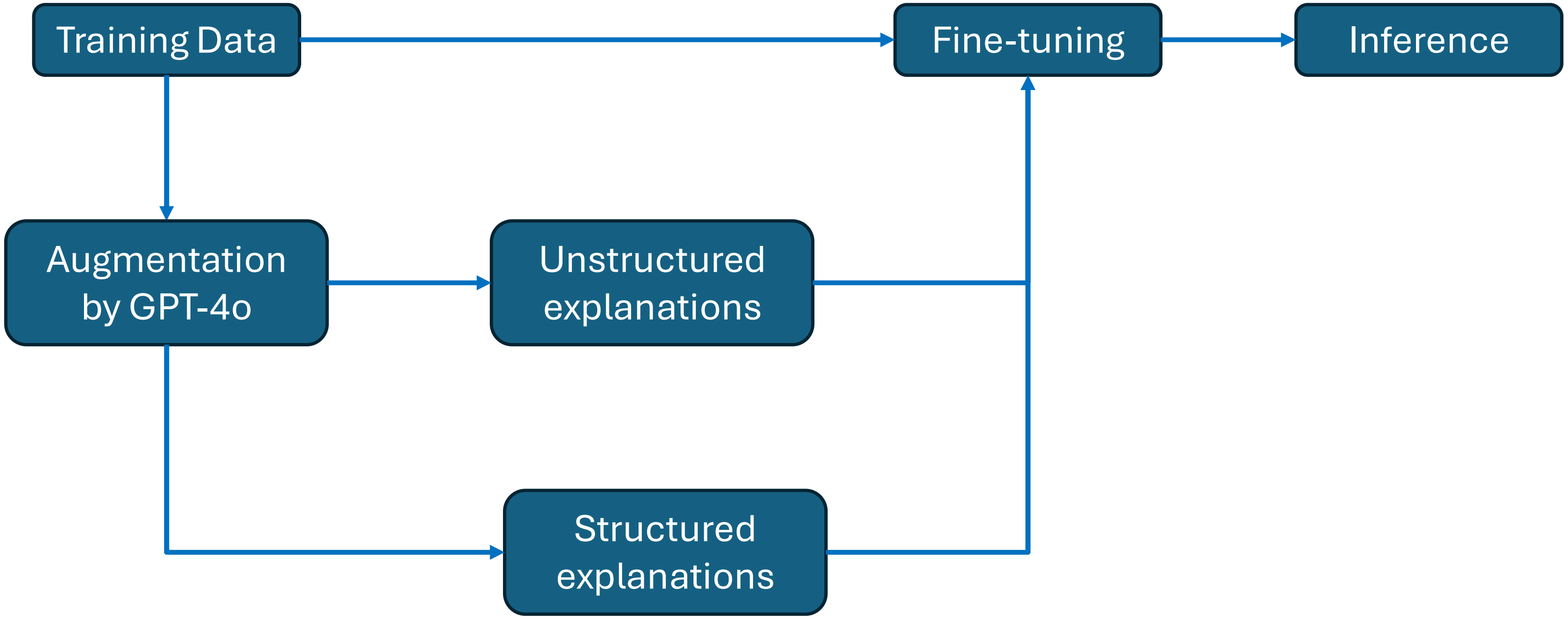
→ Should be tried due to low generation cost and no impact on inference cost.

- Future Work:
  - Explore alternative use cases for LLM-generated explanations
    - User-facing explanations

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

4. Dimension 1:
Example Augmentation

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Training Data

Fine-tuning

Inference

Augmentation by GPT-4o

Unstructured explanations

Structured explanations

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

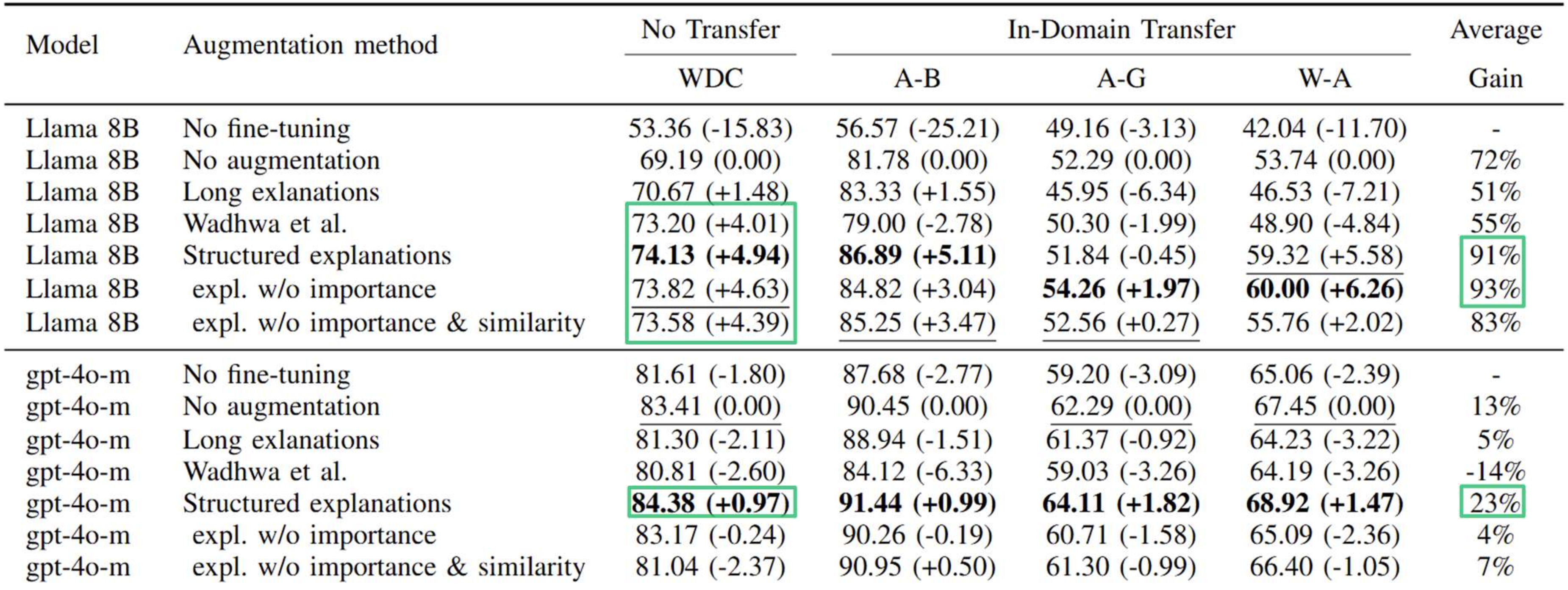
4. Dimension 1:
Unstructured Example Augmentation

Actor	Prompt contents
User	“Do the two entity descriptions refer to the same real-world product?” Entity 1: Jabra EVOLVE 80 MS Stereo (7899-823-109) Entity 2: Jabra Evolve 80 UC stereo Skype for Business
AI	“Yes. Both entities refer to the Jabra Evolve 80 headset, which is designed for professional use, specifically for communication and audio purposes. Despite the slight differences in naming conventions (MS vs. UC), they both describe the same product line and model, indicating that they are compatible with similar functionalities. Therefore, they are considered a match.”

4. Dimension 1:
Structured Example Augmentation

<table>
<tr><th>Actor</th><th>Prompt contents</th></tr>
<tr><td>User</td><td>“Do the two entity descriptions refer to the same real-world product?”
Entity 1: Jabra EVOLVE 80 MS Stereo (7899-823-109)
Entity 2: Jabra Evolve 80 UC stereo Skype for Business</td></tr>
<tr><td>AI</td><td>“Yes.
attribute=brand | importance=0.05 | values=Jabra\#\#\#Jabra | similarity=1.00
attribute=model | importance=0.95 | values=EVOLVE 80 MS Stereo\#\#\#Evolve 80 UC | similarity=0.90
attribute=type | importance=0.90 | values=Stereo \#\#\#Stereo | similarity=1.00
attribute=compatibility | importance=0.85 | values=MS (Microsoft)\#\#\#Skype for Business | similarity=0.80
attribute=SKU | importance=0.00 | values=7899-823-109 \#\#\#missing | similarity=0.00”</td></tr>
</table>

4. Results of Example Augmentation

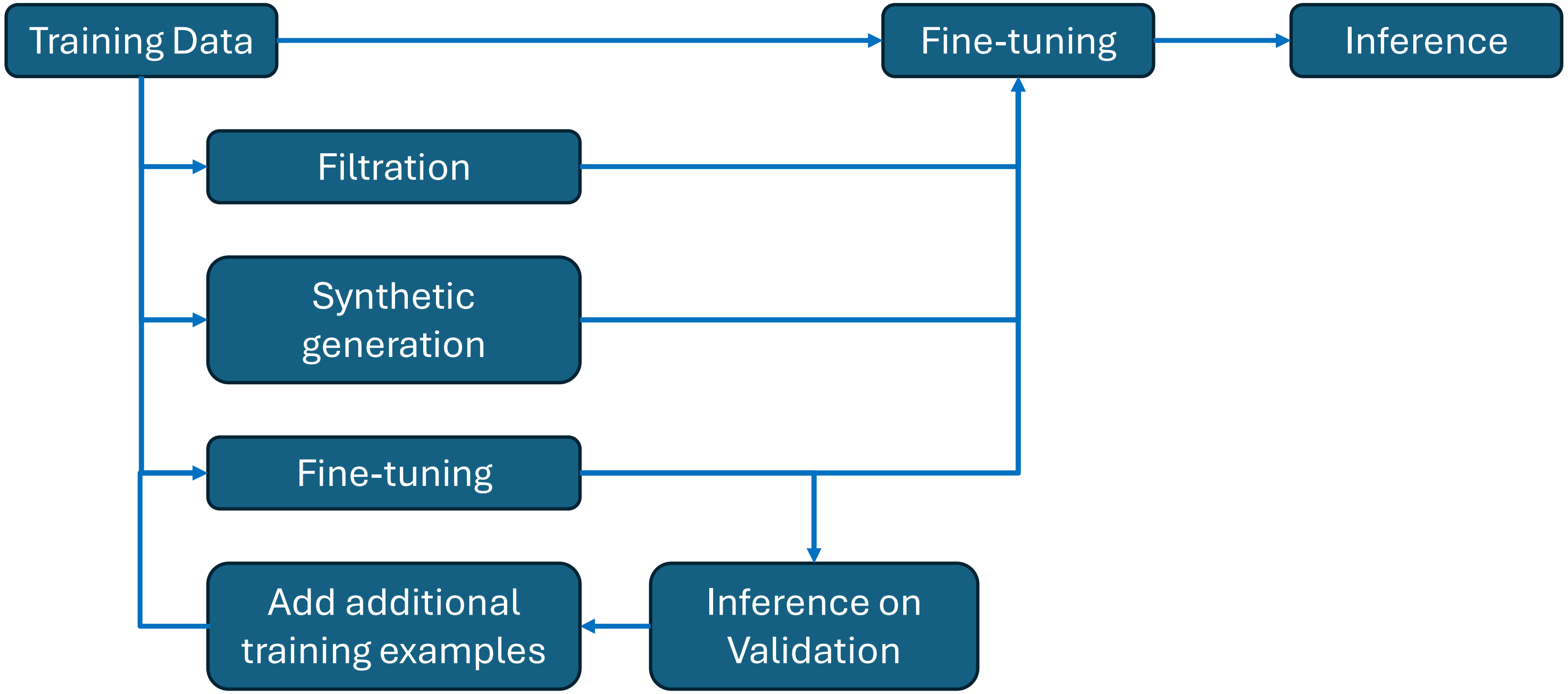


Model	Augmentation method	No Transfer	In-Domain Transfer			Average
		WDC	A-B	A-G	W-A	Gain
Llama 8B	No fine-tuning	53.36 (-15.83)	56.57 (-25.21)	49.16 (-3.13)	42.04 (-11.70)	-
Llama 8B	No augmentation	69.19 (0.00)	81.78 (0.00)	52.29 (0.00)	53.74 (0.00)	72%
Llama 8B	Long exlanations	70.67 (+1.48)	83.33 (+1.55)	45.95 (-6.34)	46.53 (-7.21)	51%
Llama 8B	Wadhwa et al.	73.20 (+4.01)	79.00 (-2.78)	50.30 (-1.99)	48.90 (-4.84)	55%
Llama 8B	Structured explanations	**74.13 (+4.94)**	**86.89 (+5.11)**	51.84 (-0.45)	59.32 (+5.58)	91%
Llama 8B	expl. w/o importance	73.82 (+4.63)	84.82 (+3.04)	**54.26 (+1.97)**	**60.00 (+6.26)**	93%
Llama 8B	expl. w/o importance & similarity	73.58 (+4.39)	85.25 (+3.47)	52.56 (+0.27)	55.76 (+2.02)	83%
gpt-4o-m	No fine-tuning	81.61 (-1.80)	87.68 (-2.77)	59.20 (-3.09)	65.06 (-2.39)	-
gpt-4o-m	No augmentation	83.41 (0.00)	90.45 (0.00)	62.29 (0.00)	67.45 (0.00)	13%
gpt-4o-m	Long exlanations	81.30 (-2.11)	88.94 (-1.51)	61.37 (-0.92)	64.23 (-3.22)	5%
gpt-4o-m	Wadhwa et al.	80.81 (-2.60)	84.12 (-6.33)	59.03 (-3.26)	64.19 (-3.26)	-14%
gpt-4o-m	Structured explanations	**84.38 (+0.97)**	**91.44 (+0.99)**	**64.11 (+1.82)**	**68.92 (+1.47)**	23%
gpt-4o-m	expl. w/o importance	83.17 (-0.24)	90.26 (-0.19)	60.71 (-1.58)	65.09 (-2.36)	4%
gpt-4o-m	expl. w/o importance & similarity	81.04 (-2.37)	90.95 (+0.50)	61.30 (-0.99)	66.40 (-1.05)	7%

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

5. Dimension 2:
Example Selection and Generation

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

4. Dimension 1:
Relevancy-filtration example

Actor	Prompt contents
User	“I am creating an entity matching benchmark and need to develop a training split that helps the model learn the intricacies of entity matching. I will provide you with two entity descriptions. Your task is to evaluate whether they form an interesting pair for training purposes. Please limit your response to ‘Yes’ or ‘No’.” **Entity 1: HDD 35 4TB Seagate IronWolf Pro NAS ST4000NE001** **Entity 2: HD 3,5 4TB 7200RPM IRONWOLF PRO 128 MB SATA3 SEAGATE**
AI	**Yes**
User	“I am creating an entity matching benchmark and need to develop a training split that helps the model learn the intricacies of entity matching. I will provide you with two entity descriptions. Your task is to evaluate whether they form an interesting pair for training purposes. Please limit your response to ‘Yes’ or ‘No’.” **Entity 1: Buy Quality Replica Omega Seamaster Planet Ocean 600M Steel Chronometer Chronograph Watch 215.30.46.51.01.001** **Entity 2: GIGABYTE Radeon RX 5500 XT OC - 4GB GDDR6 RAM - Grafikkort**
AI	**No**

5. Dimension 2:
Training Set Filtering

- Error-based Filtering
  - Remove training examples mislabeled by the LLM
  - Result: 20% smaller training set (2,000 vs. 2,500)
- Relevancy-based Filtering
  - Keep only “interesting” entity pairs (e.g., corner cases)
  - Filtered by GPT-4o using its own judgment
  - Result: Very compact dataset (608 examples)

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

5. Dimension 2:
Synthetic Example Generation

- All approaches start from seed examples in WDC-small
- Brief Prompting
  - Short task prompt + seed → 3 non-matches + 1 match
- Detailed Prompting
  - Longer prompt with explanation of corner cases and categories
- Demonstration-based Prompting
  - Adds 6 diverse example pairs to guide generation

- Combined with relevancy filtering to remove low-quality outputs
- Manual inspection revealed low correctness and diversity issues in generations

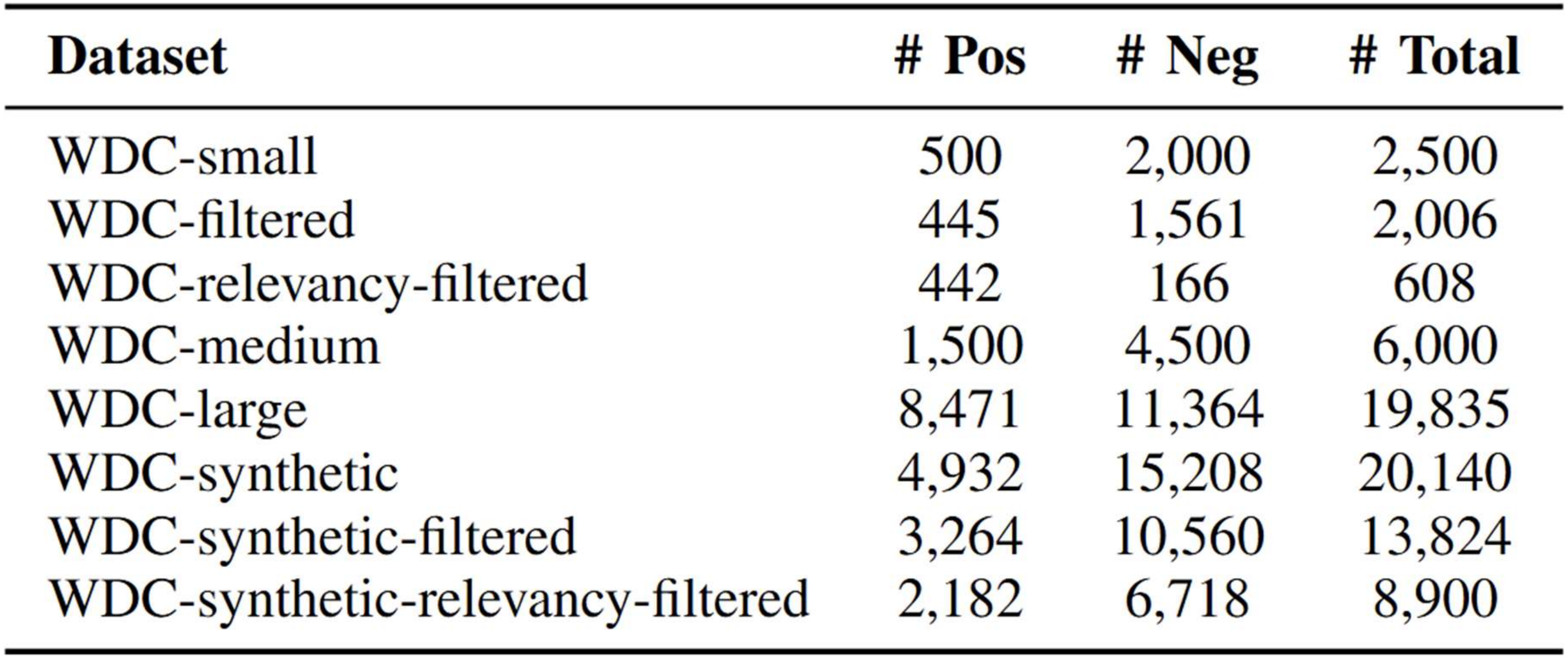
Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

5. Dimension 2:
Error-based Example Selection

- Identify incorrect predictions on the validation set after initial fine-tuning.
- Use embedding-based similarity to find examples similar to the errors.
- Iteratively expand the training set with these challenging examples.
- 5 iterations: Start with 2,500 base examples → add 2,500 error-similar examples → retrain each time.
- Select the best-performing model based on validation F1 score.

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

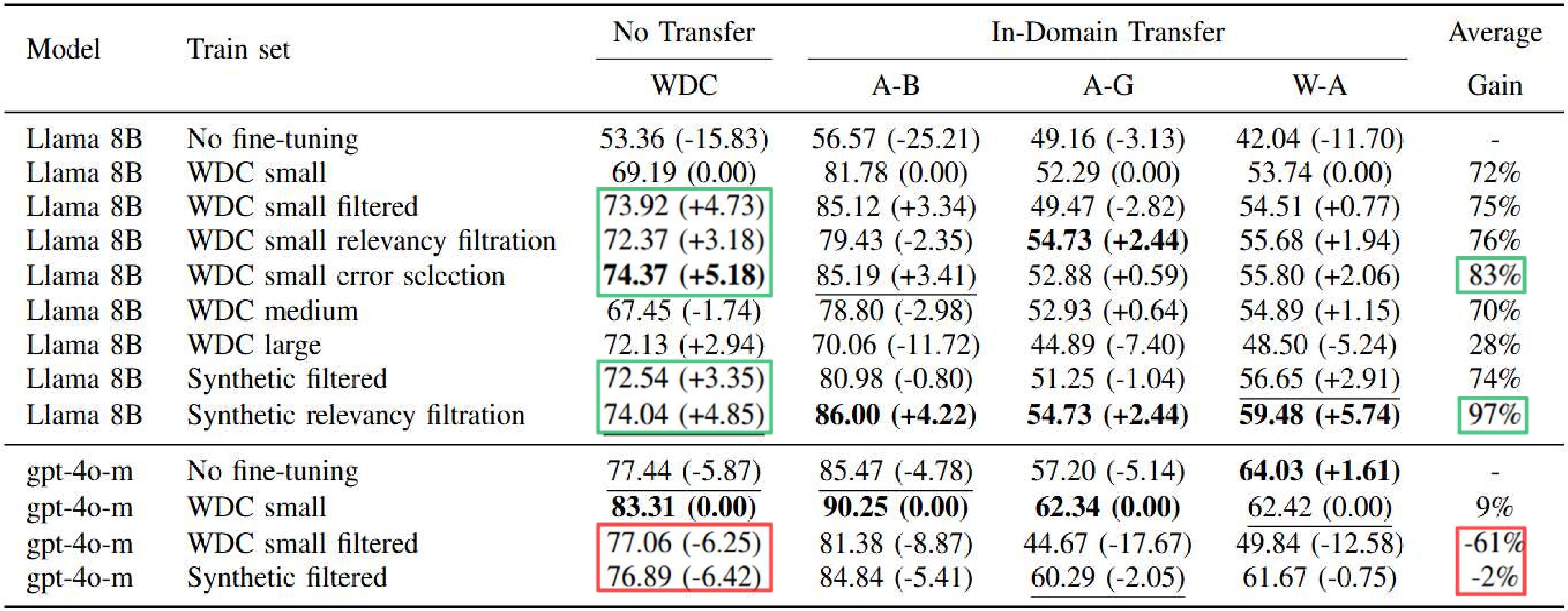
5. Dimension 2:
Example Selection and Generation



Dataset	# Pos	# Neg	# Total
WDC-small	500	2,000	2,500
WDC-filtered	445	1,561	2,006
WDC-relevancy-filtered	442	166	608
WDC-medium	1,500	4,500	6,000
WDC-large	8,471	11,364	19,835
WDC-synthetic	4,932	15,208	20,140
WDC-synthetic-filtered	3,264	10,560	13,824
WDC-synthetic-relevancy-filtered	2,182	6,718	8,900

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

5. Results of Example Selection and Generation



Model	Train set	No Transfer	In-Domain Transfer			Average
		WDC	A-B	A-G	W-A	Gain
Llama 8B	No fine-tuning	53.36 (-15.83)	56.57 (-25.21)	49.16 (-3.13)	42.04 (-11.70)	-
Llama 8B	WDC small	69.19 (0.00)	81.78 (0.00)	52.29 (0.00)	53.74 (0.00)	72%
Llama 8B	WDC small filtered	73.92 (+4.73)	85.12 (+3.34)	49.47 (-2.82)	54.51 (+0.77)	75%
Llama 8B	WDC small relevancy filtration	72.37 (+3.18)	79.43 (-2.35)	**54.73 (+2.44)**	55.68 (+1.94)	76%
Llama 8B	WDC small error selection	**74.37 (+5.18)**	85.19 (+3.41)	52.88 (+0.59)	55.80 (+2.06)	83%
Llama 8B	WDC medium	67.45 (-1.74)	78.80 (-2.98)	52.93 (+0.64)	54.89 (+1.15)	70%
Llama 8B	WDC large	72.13 (+2.94)	70.06 (-11.72)	44.89 (-7.40)	48.50 (-5.24)	28%
Llama 8B	Synthetic filtered	72.54 (+3.35)	80.98 (-0.80)	51.25 (-1.04)	56.65 (+2.91)	74%
Llama 8B	Synthetic relevancy filtration	74.04 (+4.85)	**86.00 (+4.22)**	**54.73 (+2.44)**	**59.48 (+5.74)**	97%
gpt-4o-m	No fine-tuning	77.44 (-5.87)	85.47 (-4.78)	57.20 (-5.14)	**64.03 (+1.61)**	-
gpt-4o-m	WDC small	**83.31 (0.00)**	**90.25 (0.00)**	**62.34 (0.00)**	62.42 (0.00)	9%
gpt-4o-m	WDC small filtered	77.06 (-6.25)	81.38 (-8.87)	44.67 (-17.67)	49.84 (-12.58)	-61%
gpt-4o-m	Synthetic filtered	76.89 (-6.42)	84.84 (-5.41)	60.29 (-2.05)	61.67 (-0.75)	-2%

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

Key findings

- Fine-tuning enhances performance, especially for smaller models
- Incorporating structured explanations into the fine-tuning process boosts performance—particularly for smaller models—without increasing inference cost, due to ordering (Not Chain-of-Thought reasoning)
- Example selection strategies (e.g., error-based filtering, relevance-based filtering) improve performance in both non-transfer settings and in-domain generalization. However, their effectiveness currently appears limited to the LLaMA model series.
- Cross-domain generalization remains low

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025

Thank you for listening

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Paper

Github

LinkedIn

Steiner, Peeters, Bizer: Fine-tuning LLMs for Entity Matching. DAIS @ ICDE 2025