

# Entity Matching using Large Language Models

Ralph Peeters  
Data and Web Science Group  
University of Mannheim  
Mannheim, Germany  
ralph.peeters@uni-mannheim.de

Aaron Steiner  
Data and Web Science Group  
University of Mannheim  
Mannheim, Germany  
aasteine@mail.uni-mannheim.de

Christian Bizer  
Data and Web Science Group  
University of Mannheim  
Mannheim, Germany  
christian.bizer@uni-mannheim.de

## ABSTRACT

Entity matching is the task of deciding whether two entity descriptions refer to the same real-world entity. Entity matching is a central step in most data integration pipelines. Many state-of-the-art entity matching methods rely on pre-trained language models (PLMs) such as BERT or RoBERTa. Two major drawbacks of these models for entity matching are that (i) the models require significant amounts of task-specific training data and (ii) the fine-tuned models are not robust concerning out-of-distribution entities. This paper investigates using generative large language models (LLMs) as a less task-specific training data-dependent and more robust alternative to PLM-based matchers. The study covers hosted and open-source LLMs which can be run locally. We evaluate these models in a zero-shot scenario and a scenario where task-specific training data is available. We compare different prompt designs and the prompt sensitivity of the models. We show that there is no single best prompt but that the prompt needs to be tuned for each model/dataset combination. We further investigate (i) the selection of in-context demonstrations, (ii) the generation of matching rules, as well as (iii) fine-tuning LLMs using the same pool of training data. Our experiments show that the best LLMs require no or only a few training examples to perform comparably to PLMs that were fine-tuned using thousands of examples. LLM-based matchers further exhibit higher robustness to unseen entities. We show that GPT4 can generate structured explanations for matching decisions and can automatically identify potential causes of matching errors by analyzing explanations of wrong decisions. We demonstrate that the model can generate meaningful textual descriptions of the identified error classes, which can help data engineers to improve entity matching pipelines.

## 1 INTRODUCTION

Entity matching [3, 8, 15] is the task of discovering entity descriptions in different data sources that refer to the same real-world entity. Entity matching is a central step in data integration pipelines [9] and forms the foundation of interlinking data on the Web [31]. Application domains of entity matching include e-commerce, where offers from different vendors are matched for example for price tracking, and financial data integration, where information about companies from different sources is combined [8]. While early matching systems relied on manually defined matching rules, supervised machine learning methods have become the foundation of entity matching systems [9] today. This trend was reinforced by the success of neural networks [3] and today most state-of-the-art matching systems rely on pre-trained language models (PLMs), such as BERT or RoBERTa [23, 34, 36, 48]. The major drawbacks of using PLMs

© 2025 Copyright held by the owner/author(s). Published in Proceedings of the 28th International Conference on Extending Database Technology (EDBT), 25th March-28th March, 2025, ISBN 978-3-89318-098-1 on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

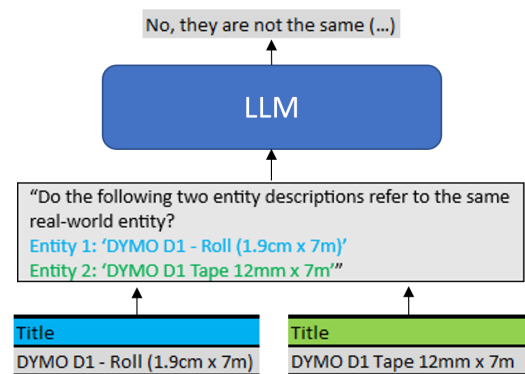


Figure 1: Example of prompting an LLM to match two entity descriptions.

for entity matching are that (i) PLMs need significant amounts of task-specific training examples for fine-tuning and (ii) they are not robust concerning unseen entities that were not part of the training data [1, 36].

Generative large language models (LLMs) [50] such as GPT, Llama, Gemini, or Mixtral have the potential to address both of these shortcomings. Due to being pre-trained on large amounts of textual data as well as due to emergent effects resulting from the model size [46], LLMs often show a better zero-shot performance compared to PLMs and are more robust concerning unseen examples [4, 50].

This paper investigates using LLMs for entity matching as a less task-specific training data dependent and more robust alternative to PLM-based matchers. We evaluate the models in a zero-shot scenario as well as a scenario where task-specific training data is available and can be used for selecting demonstrations, generating matching rules, or fine-tuning the LLMs. Our study covers hosted LLMs as well as open-source LLMs which can be run locally. Figure 1 shows an example of how LLMs are used for entity matching. The two entity descriptions at the bottom of the figure are combined with the question whether they refer to the same real-world entity into a prompt. The prompt is passed to the LLM, which generates the answer shown at the top of Figure 1.

**Contributions:** We make the following contributions:

- (1) **Range of prompts:** We experiment with a wider range of zero-shot and few-shot prompts compared to the related work [16, 30, 35, 45, 49]. This allows us to present a more nuanced picture of the strengths and weaknesses of the different approaches.
- (2) **No single best prompt:** We show that there is no single best prompt per model or per dataset but that the best prompt depends on the model/dataset combination.
- (3) **Prompt sensitivity:** We are first to investigate the sensitivity of LLMs concerning variations of entity matching

prompts. Our experiments show that the matching performance of many LLMs is strongly influenced by prompt variations while the performance of other models is rather stable.

- (4) **LLMs versus PLMs:** We show that GPT4 without any task-specific training data outperforms fine-tuned PLMs on 3 out of 4 e-commerce datasets and achieves a comparable performance for bibliographic data. We are the first to compare the generalization performance of LLM- and PLM-based matchers for unseen entities. PLM-based matchers perform poorly on entities that are not part of any pair in the training set [1, 36]. LLMs do not have this problem as they perform well without task-specific training data.
- (5) **Hosted versus open-source LLMs:** We show that open-source LLMs can reach a similar F1 performance as hosted LLMs given that a small amount of task-specific training data or matching knowledge in the form of rules is available.
- (6) **Fine-tuning:** We compare fine-tuning hosted and open-source LLMs for entity matching. Our results show that fine-tuning significantly improves the performance of the LLMs. GPT-mini retains strong generalization capability across datasets, whereas fine-tuning reduces generalizability for the Llama models.
- (7) **Explanations and automated error analysis:** We are first to use an LLM to generate structured explanations for matching decisions. We further demonstrate that the model can automatically identify potential causes of matching errors by analyzing explanations of wrong decisions.

**Structure:** Section 2 introduces our experimental setup. Section 3 compares prompt designs and LLMs in the zero-shot setting, while Section 4.1 investigates whether the model performance can be improved by providing demonstrations for in-context learning. In Section 4.2, we experiment with adding matching knowledge in the form of natural language rules. Section 4.3 compares fine-tuning LLMs to the previous approaches. Section 5 presents a cost and runtime analysis of the hosted LLMs. In Section 6, we use GPT4 to create structured explanations to gain insights into the model decisions. In Section 7 we demonstrate how to automatically discover error classes by analyzing structured explanations of wrong decisions. Section 8 presents related work, while Section 9 concludes the paper and summarizes the implications of our findings.

**Replicability:** All data and code used for the experiments presented in this paper is publicly available<sup>1</sup> meaning that all experiments can be replicated.

## 2 EXPERIMENTAL SETUP

This section provides details about the large language models, the benchmark datasets, the serialization of entity descriptions, and the evaluation metrics that are used in the experiments.

**Large Language Models:** We compare three hosted LLMs from OpenAI<sup>2</sup> and three open-source LLMs run on local GPUs:

- **gpt-4o-mini-2024-07-18 (GPT-mini):** This hosted LLM from OpenAI offers lower API usage fees compared to GPT-4 and GPT-4o. It has a context window of 128K tokens. The training data cutoff date is October 2023.

- **gpt4-0613 (GPT-4):** This version of OpenAI’s GPT-4 model from June 2023 has a context window of 8192 tokens. The training data cutoff date is September 2021.
- **gpt-4o-2024-08-06 (GPT-4o):** GPT-4o is OpenAI’s flagship model with a 128K context and a training data cutoff date of October 2023.
- **Llama-2-70b-chat-hf (Llama2):** This Llama2 version<sup>3</sup> from Meta has 70B parameters and has been optimized for dialogue uses cases. It has a context window of 4K tokens and the training data cutoff date is September 2022.
- **Meta-Llama-3.1-70B-Instruct (Llama3.1):** This Llama3.1 model from Meta has 70B parameters and is optimized for dialogue use cases. It has a context window of 128K tokens and a training data cutoff date of December 2023.
- **Mixtral-8x7B-Instruct-v0.1 (Mixtral):** Mixtral is an open-source model that consists of 8 smaller models. It is developed by Mistral AI<sup>4</sup> and has a context window of 32K.

The model names that are introduced in the brackets above are used in the following chapters to refer to the respective models. We use the langchain<sup>5</sup> library for interacting with the OpenAI API as well as for template-based prompt generation. We set the temperature parameter to 0 for all LLMs to reduce randomness. The temperature parameter adjusts the randomness of the model’s outputs by scaling the logits before applying the softmax function. We run the open-source LLMs on a local machine with an AMD EPYC 7413 processor, 1024GB RAM, and four NVIDIA RTX6000 GPUs.

**PLM Baselines:** We compare the performance of the LLMs to two PLM-based matchers:

- **RoBERTa:** We employ the RoBERTa-base [26] model for entity matching as the model has been shown to reach high performance in related work [23, 33, 48]. We fine-tune the model for entity matching using the respective development sets.
- **Ditto:** The Ditto [23] matching system is one of the first dedicated entity matching systems using PLMs. Ditto introduces various data augmentation and domain knowledge injection modules. We run Ditto using RoBERTa-base as internal model.

We select RoBERTa-base as a representative model for comparing PLMs to LLMs. We select Ditto as it combines a PLM with additional matching-specific functionality. Ditto also outperforms earlier matchers such as Deepmatcher [28], DeepER [14], and EmbDI [6], and it performs within a 2% F1 range compared to more complex methods such as SETEM [13] or HierGAT [47] on the datasets selected below.

**Benchmark Datasets:** We use the following benchmark datasets for our experiments [22, 36]:

- **WDC Products:** The WDC Products benchmark consists of product offers originating from thousands of different e-shops spanning product categories such as electronics, clothing, and tools for home improvement. We use the most difficult version of the benchmark including 80% corner-cases (see below). We use the following product attributes: *brand*, *title*, *currency*, and *price*.
- **Abt-Buy:** This benchmark dataset also contains product offers that need to be matched. The offers are from similar categories as those in WDC Products. The title attribute

<sup>1</sup><https://github.com/wbsg-uni-mannheim/MatchGPT/tree/main/LLMForEM>

<sup>2</sup><https://platform.openai.com/docs/models/>

<sup>3</sup><https://huggingface.co/meta-llama/Llama-2-70b-chat-hf>

<sup>4</sup><https://mistral.ai/>

<sup>5</sup><https://www.langchain.com/>

**Table 1: Statistics for all datasets. In-context example selection and fine-tuning are performed on the training and validation sets. Prompts are evaluated on the test sets.**

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

values in Abt-Buy are rather textual and describe various product features. Attributes used: *title*, and *price*.

- **Walmart-Amazon:** The Walmart-Amazon benchmark represents a slightly more structured matching task in the product domain. The types of products in this dataset are similar to WDC Products and Abt-Buy. Attributes used: *brand*, *title*, *model number*, and *price*.
- **Amazon-Google:** The Amazon-Google dataset consists of rather textual offers for software products, e.g. different versions of the Windows operating system or image/video editing applications. Attributes used: *brand*, *title*, and *price*.
- **DBLP-Scholar:** The task of this benchmark dataset is to match bibliographic entries from DBLP and Google Scholar. Attributes used: *authors*, *title*, *venue*, and *year*.
- **DBLP-ACM:** Similar to DBLP-Scholar, the task of DBLP-ACM is to match bibliographic entries between two sources. Attributes used: *authors*, *title*, *venue*, and *year*.

The motivation for selecting these datasets is threefold: (i) Measuring the performance of entity matching methods on benchmarks containing few matches leads to unstable results. Thus, we select WDC Products [36] and a subset of the dataset used in the DeepMatcher paper [28] which contain at least 150 matches in the test set (see Table 1). (ii) We select datasets that contain a decent amount of difficult to match corner case record pairs. Corner cases are matching and non-matching pairs that exhibit the property of resembling a pair of the respective other class due to very (dis-)similar surface forms [36]. (iii) The datasets should cover different topical domains (products and bibliographic data) in order to evaluate the cross-domain generalization of the models. WDC Products and Walmart-Amazon contain duplicates for some entities within the same dataset, representing a dirty-dirty matching scenario [10], while the other datasets represent a clean-clean matching scenario.

**Splits:** For WDC Products, we use the training/validation/test split of size small [36]. For the other benchmark datasets, we use the splits established in the DeepMatcher paper [28]. We perform a large number of experiments using OpenAI models. In order to keep the OpenAI API usage fees on an affordable level, we down-sample all test sets to approximately 1250 entity pairs. Table 1 provides statistics about the numbers of positive (matches) and negative (non-matches) pairs in the training, validation, and test sets of all benchmarks used in the experiments.

**Serialization:** For the serialization of pairs of entity descriptions (records) into prompts, we serialize each entity description into a single string by concatenating their attribute values using blanks as delimitator, e.g.  $serialize(e) := Val_{A1} Val_{A2} \dots Val_{An}$ . Figure 1 shows an example of this serialization practice for a pair of product offers. We apply the same serialization method for

the bibliographic data. We list the attributes that we use for each dataset as well as their order in the dataset descriptions above. All datasets contain textual attributes, e.g. the *title* of a product or publication, as well as numerical attributes like *price* and *year*. The decision was made not to add the names of the attributes themselves to the serialization string as this negatively affected performance in early experiments.

**Evaluation:** The responses that are generated by the models are natural language text. In order to decide whether a response indicates a positive matching decision, we apply lower-casing to the answer and subsequently parse for the word *yes*. In all other cases, we assume the model has decided against a match. This rather simple approach turns out to be surprisingly effective as shown by Narayan et al. [30]. For measuring model performance, we use the metrics *F1-score*, *precision*, and *recall* on the *matching* (*positive*) class following related work [3]. While the tables in the following sections report F1-scores, the precision and recall results of all experiments are available in the project repository.

### 3 SCENARIO 1: ZERO-SHOT PROMPTING

In the first scenario, we analyze the impact of different prompt designs on the entity matching performance of the LLMs. We further investigate the prompt sensitivity of the different models for the entity matching task, and finally compare the performance of the LLMs to the PLM baselines.

**Prompt Building Blocks:** We construct prompts as a combination of smaller building blocks to allow the systematic evaluation of different prompt designs. Each prompt consists of at least a task description and the serialization of the pair of entity descriptions to be matched. In addition, the prompts may contain a specification of the output format. We evaluate alternative task descriptions that formulate the task as a question using simple or complex wording combined with domain-specific or general terms. Our goal is to present the task in a simple and concise fashion (similar to [30]) while allowing for some variations in the structure and wording in order to assess performance spread and the prompt sensitivity of the models. The alternative task descriptions are listed below:

- **domain-simple:** "Do the two product descriptions match?" / "Do the two publications match?"
- **domain-complex:** "Do the two product descriptions refer to the same real-world product?" / "Do the two publications refer to the same real-world publication?"
- **general-simple:** "Do the two entity descriptions match?"
- **general-complex:** "Do the two entity descriptions refer to the same real-world entity?"

A specification of the output format may follow the task description. We evaluate two formats: *free* which does not restrict the answer of the LLM and *force* which instructs the LLM to "Answer with 'Yes' if they do and 'No' if they do not". The prompt continues with the entity pair to be matched, serialized as discussed in Section 2. Figure 1 contains an example of a complete prompt implementing the prompt design *general-complex-free*. Examples of all prompt designs are found in the accompanying repository. In addition to the prompts that we generate using these building blocks, we also evaluate the entity matching prompts proposed by Narayan et al. [30].

**Effectiveness:** Table 2 shows the results for each dataset separately. Table 3 shows the results of the zero-shot experiments averaged over all datasets. With regards to overall performance, the GPT4 model outperforms all other LLMs on all product datasets

**Table 2: Results (F1) of the zero-shot experiments for all datasets. Best results are set bold, second best are underlined.**

Prompt	WDC Products						Abt-Buy						Walmart-Amazon					
	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral
domain-complex-force	80.84	88.35	<b>87.64</b>	65.23	<b>83.67</b>	<b>53.37</b>	<u>90.95</u>	<u>95.15</u>	90.47	57.59	<b>89.84</b>	<b>82.20</b>	<u>86.36</u>	89.00	84.04	46.70	<b>84.85</b>	<b>70.44</b>
domain-complex-free	80.00	<b>89.61</b>	67.35	<b>69.09</b>	50.86	51.98	<b>91.93</b>	<b>95.78</b>	89.35	64.13	78.90	78.07	86.28	89.33	82.91	51.96	69.87	54.42
domain-simple-force	21.35	83.72	81.53	23.89	33.22	8.43	77.55	93.56	<u>91.77</u>	65.82	79.77	51.96	48.84	88.78	83.84	52.17	54.81	27.56
domain-simple-free	16.85	84.50	42.99	24.75	1.59	14.81	60.81	94.38	78.24	63.43	34.40	52.30	43.82	88.67	56.00	40.15	17.76	24.43
general-complex-force	78.80	85.83	<u>87.02</u>	66.02	<u>81.89</u>	42.04	89.88	94.40	90.67	56.70	88.11	79.02	<b>86.58</b>	<b>89.67</b>	83.67	44.69	<u>84.14</u>	50.56
general-complex-free	<b>81.15</b>	86.72	23.86	<u>67.59</u>	<u>67.81</u>	<u>52.22</u>	87.73	94.87	72.00	55.77	87.31	<u>81.27</u>	85.99	<u>89.45</u>	45.85	44.95	83.43	53.82
general-simple-force	20.71	77.39	82.48	46.54	62.57	9.89	80.67	93.23	<b>93.95</b>	<b>82.03</b>	88.72	54.04	46.33	86.41	<b>86.65</b>	<b>63.91</b>	72.05	21.20
general-simple-free	18.84	83.41	41.77	39.30	44.24	12.03	73.05	92.77	83.80	<u>77.21</u>	80.00	58.86	37.34	88.60	62.50	56.03	60.69	19.63
Narayan-complex	47.31	81.23	31.89	44.97	9.09	21.05	79.89	92.13	58.59	68.44	34.40	40.00	41.46	83.37	24.00	57.74	12.50	15.17
Narayan-simple	71.01	81.91	21.91	52.72	9.16	15.33	86.63	92.42	57.34	73.99	35.06	37.80	69.28	84.72	20.28	<u>63.32</u>	18.69	11.65
Mean	51.69	84.27	56.84	50.01	44.41	28.12	81.91	93.87	80.62	66.51	69.65	61.55	63.23	87.80	62.97	52.16	55.88	34.89
Standard deviation	27.96	3.42	25.66	16.25	28.83	18.31	9.22	1.17	13.01	8.50	23.24	16.32	20.44	2.08	24.39	7.69	27.56	19.39

Prompt	Amazon-Google						DBLP-Scholar						DBLP-ACM					
	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral
domain-complex-force	<u>70.98</u>	75.61	<b>73.56</b>	<b>57.93</b>	<b>73.99</b>	<b>40.98</b>	<b>86.11</b>	88.44	<b>89.76</b>	<b>85.46</b>	84.29	<u>75.40</u>	<u>96.51</u>	96.90	96.53	86.69	92.59	87.33
domain-complex-free	<b>72.18</b>	75.57	61.17	<u>56.29</u>	60.59	24.91	86.06	<u>89.78</u>	84.03	<u>85.11</u>	80.87	<b>77.75</b>	95.97	96.71	<b>97.06</b>	<u>91.57</u>	91.24	85.66
domain-simple-force	24.55	75.32	58.00	29.86	19.33	13.39	41.51	<u>77.21</u>	84.35	84.07	77.17	60.39	88.65	98.03	96.85	<u>87.62</u>	<b>98.81</b>	88.15
domain-simple-free	19.48	74.51	35.85	17.16	5.76	17.69	7.63	88.20	73.79	81.53	67.69	59.67	53.30	97.28	94.29	75.35	94.41	<b>90.32</b>
general-complex-force	65.25	74.91	<u>70.98</u>	53.59	<u>69.36</u>	<u>31.65</u>	85.82	87.22	<u>87.54</u>	79.67	<u>85.97</u>	66.15	94.66	96.60	90.25	82.67	90.09	87.63
general-complex-free	64.09	74.38	20.44	49.48	68.45	29.45	85.66	87.50	76.85	76.42	<b>86.32</b>	68.01	94.16	94.16	95.87	82.18	89.93	84.23
general-simple-force	25.53	53.60	56.28	49.32	33.87	11.24	54.19	78.26	85.65	66.37	80.27	31.65	89.87	97.85	<u>96.86</u>	65.71	<u>98.41</u>	73.50
general-simple-free	17.56	66.67	32.19	37.79	25.35	12.70	40.75	81.47	72.20	51.06	73.99	38.59	85.40	97.47	95.58	55.21	95.98	74.88
Narayan-complex	18.45	<b>76.38</b>	13.90	39.63	10.40	3.36	56.34	<b>89.82</b>	78.82	42.99	55.27	27.40	93.31	97.27	96.67	84.26	95.32	85.26
Narayan-simple	42.24	<u>75.70</u>	5.71	48.71	6.58	2.51	84.12	88.37	72.82	70.47	59.40	35.06	<b>97.60</b>	<b>98.41</b>	<u>95.77</u>	<b>97.62</b>	95.04	83.30
Mean	42.03	72.27	42.81	43.98	37.37	18.79	62.82	85.63	80.58	72.32	75.12	54.01	88.94	96.97	95.57	80.89	94.18	84.03
Standard deviation	22.39	6.76	23.16	12.23	26.51	11.99	25.87	4.53	6.15	14.07	10.43	18.01	12.43	1.19	1.94	11.87	3.01	5.29

**Table 3: Average F1-scores over all datasets for the zero-shot experiments.**

Prompt	All Datasets (Average F1)					
	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral
domain-complex-force	85.29	88.91	<b>87.00</b>	66.60	<b>84.87</b>	<b>68.29</b>
domain-complex-free	<b>85.40</b>	<b>89.46</b>	80.31	<b>69.69</b>	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	<u>85.02</u>	63.89	<b>83.26</b>	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	<u>67.81</u>	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

by at least 1% F1 achieving an absolute performance of 89% or higher on 5 of 6 datasets without requiring any task-specific training data. On the publication datasets, GPT-4o achieves nearly the same performance (0.1-1% F1). This gap increases on the product datasets to 1-3% F1 making the more recent model marginally worse than GPT4. GPT-mini performs up to 6% F1 worse than GPT-4o with only marginal performance difference on 4 of 6 datasets. Among the open-source LLMs, Llama3.1 consistently outperforms Llama2 by 1-21% F1. Llama3.1’s performance is comparable to GPT-mini on all datasets. The Mixtral model performs less effectively on this task, lagging behind the other open-source models by 7-16% on 4 datasets. In summary, the results indicate that locally run open-source LLMs can perform similarly to OpenAI’s GPT-mini model given that the right prompt is selected. However, if maximum performance is desired, none of the other LLMs can match GPT-4 in a zero-shot setting. GPT-4o offers a more cost-effective alternative to GPT-4 (see Section 5), though its performance is slightly lower. The GitHub repository provides additional results for the models GPT3.5-turbo, SOLAR, and StableBeluga2.

**Sensitivity:** Small variations in prompts can have a large impact on the overall task performance [25, 30, 51]. We measure this prompt sensitivity as the standard deviation (SD) of the F1 scores of a model over all 10 prompt designs and list this standard deviation in the lower section of Tables 2 and 3. Comparing the prompt sensitivity of the models, the GPT4 model is most invariant to the wording of the prompt (mean standard deviation 2.26) while also achieving high results with most of the prompt designs. Comparing the sensitivity of GPT4 to all other models shows that they have a significantly higher prompt sensitivity (standard deviation 6.18 to 18.54 in Table 3).

**Prompt to Model Fit:** The best result for each model is set bold in Table 2, the second best result is underlined. This highlighting shows that there is no prompt design that performs best for most models. As a result, a general statement of how to design a prompt for the entity matching task cannot be made. While the presented analysis is not exhaustive regarding all possible prompt designs, the results indicate that the best prompt depends on the model/dataset combination. While a good performing prompt can be found by testing a set of pre-defined prompts (as we did), automated approaches for prompt tuning and evolution could still further improve the results [18, 43].

**Comparison to PLM Baselines:** We compare the zero-shot performance of the LLMs to the performance of two PLM-based matchers: a fine-tuned RoBERTa model [26] and Ditto [23], an entity matching system which also relies on domain-specific training data. Table 4 shows the overall best results for each LLM in comparison to the two PLM-based matchers on all datasets. For three out of the six datasets, GPT4 achieves higher performance than the best PLM baseline (2.65-4.71% F1), while the performance for the other three datasets is 3.69, 4.49 and 0.73% F1 lower. This shows that GPT4 without using any task-specific training data is able to reach comparable results or even outperform PLMs that were fine-tuned using thousands of training pairs (see Table 1). The reliance on large amounts of task-specific training data to achieve good performance is one of the main shortcomings of fine-tuned PLMs.

**Table 4: Comparison of F1 scores of the best zero-shot prompt per model with PLM baselines. The "unseen" rows correspond to training on the dataset named in the column and applying the model to the WDC Products test set.**

	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	<b>89.61</b>	<b>95.78</b>	<b>89.67</b>	76.38	89.82	98.41
GPT-4o	<u>87.64</u>	<u>93.95</u>	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	<u>87.02</u>	<u>79.27</u>	<u>93.88</u>	<b>99.14</b>
Ditto	84.90	91.31	86.39	<b>80.07</b>	<b>94.31</b>	<u>99.00</u>
$\Delta$ best LLM/PLM	<b>4.71</b>	<b>4.47</b>	<b>2.65</b>	<b>-3.69</b>	<b>-4.49</b>	<b>-0.33</b>
RoBERTa unseen	-	55.52	36.46	31.00	29.64	16.25
Ditto unseen	-	48.74	31.55	33.12	32.82	29.00
$\Delta$ RoBERTa unseen	-	<b>-22.01</b>	<b>-41.07</b>	<b>-46.53</b>	<b>-47.89</b>	<b>-61.28</b>
$\Delta$ Ditto unseen	-	<b>-36.16</b>	<b>-53.35</b>	<b>-51.78</b>	<b>-52.08</b>	<b>-55.90</b>

**Generalization:** Another shortcoming of PLM-based matchers is their low robustness to out-of-distribution entities, e.g. entities that are not part of any training pair [1, 36]. In another set of experiments, we apply each of the previously fine-tuned RoBERTa and Ditto models — excluding the models fine-tuned on WDC Products — to the WDC Products test set, which contains a different set of products which are thus unseen to these fine-tuned models. We report the results of these experiments in the "RoBERTa unseen" and "Ditto unseen" rows at the bottom of Table 4. Compared to fine-tuning directly on the WDC Products development set (84.90% F1 for Ditto), the transfer of fine-tuned models leads to large drops in performance ranging from 36 to 56% F1 for Ditto and 22 to 61% F1 for RoBERTa. All LLMs achieve at least 8% F1 higher performance than the best transferred PLM while GPT4 outperforms the best PLM by 40% to 68%. These results indicate that LLMs have a general capability to perform entity matching, while PLM-based matchers are closely fitted to the entities within the fine-tuning dataset.

## 4 SCENARIO 2: WITH TRAINING DATA

Task-specific training data in the form of matching and non-matching entity pairs can be used to (i) add demonstrations to the prompts, (ii) learn textual matching rules, and (iii) fine-tune the LLMs. In this section, we explore whether and how our zero-shot results can be improved by using task-specific training data.

### 4.1 In-Context Learning

For the in-context learning experiments, we provide each LLM with a set of task demonstrations [24] as part of the prompt in order to guide the model’s decisions. The demonstrations are followed in the prompt by the entity description pair for which the model should generate a matching decision. Figure 2 shows an example of an in-context learning prompt containing a single positive and a single negative demonstration. We vary the amount of demonstrations in each prompt from 6 to 10 with an equal amount of positive and negative examples. For the selection of the demonstrations, we compare three different heuristics:

<b>Task Desc.</b>	Do the following two product descriptions match?
<b>Demonstrations</b>	Product 1: ‘Title: DYMO D1 19 mm x 7 m’ Product 2: ‘Title: Dymo D1 (19mm x 7m – BoW)’
<b>Answer</b>	Yes.
<b>Task Desc.</b>	Do the following two product descriptions match?
<b>Demonstrations</b>	Product 1: ‘Title: DYMO D1 Tape 24mm’ Product 2: ‘Title: Dymo D1 19mm x 7m’
<b>Answer</b>	No.
<b>Task Desc.</b>	Do the following two product descriptions match?
<b>Task Input</b>	Product 1: ‘Title: DYMO D1 – Roll (1.9cm x 7m)’ Product 2: ‘Title: DYMO D1 Tape 12mm x 7m’

**Figure 2: Example of a prompt containing a positive and a negative demonstration before asking for a decision.**

- **Random:** As baseline heuristic, task demonstrations are drawn randomly from the training set of the respective benchmark.
- **Related:** Related demonstrations are selected from the training set of the respective benchmark with the idea of presenting correct matching decisions on highly similar products. This is done by calculating the Generalized Jaccard<sup>6</sup> similarity between the string representation of the pair to be matched and all positive and negative pairs in the corresponding training set. Afterwards, the pairs are sorted by similarity and the most similar positive and negative pairs are selected as demonstrations.
- **Hand-picked:** The hand-picked demonstrations were selected by a data engineer with the goals of being diverse and potentially helpful for corner case decisions. For the four datasets in the product domain, these examples are drawn from the WDC Products training set and were chosen to represent various product categories, as well as pairs where different attributes are important for the matching decision. For the two datasets from the publication domain, the examples are selected from the pool of training examples of DBLP-Scholar covering a range of distinct venues, publication years, and research areas.

**Effectiveness:** Table 6 shows the averaged results of the in-context experiments in comparison to the best zero-shot baselines. Table 5 shows the results for each dataset separately. Depending on the model/dataset combination the usefulness of in-context learning differs. The GPT4 model, which is the best performing model in the zero-shot scenario, only improves significantly on Amazon-Google (9%) with marginal improvements on two datasets (0.6-1.5%) when supplying related demonstrations and an improvement of 2% on DBLP-Scholar with handpicked demonstrations. GPT4’s performance on WDC Products and Abt-Buy drops irrespective of the demonstration selection method, meaning that the model does not need the additional guidance in these cases. The GPT-4o model on the other hand sees improvements on all datasets when supplying demonstrations closing the gap to GPT-4 compared to zero-shot and even outperforming it for WDC Products. GPT-mini and Mixtral are not capable of using the in-context information as both models lose between 4 and

<sup>6</sup>[https://anhaidgroup.github.io/py\\_stringmatching/v0.3.x/GeneralizedJaccard.html](https://anhaidgroup.github.io/py_stringmatching/v0.3.x/GeneralizedJaccard.html)

**Table 5: Results (F1) of the few-shot and rule-based experiments. Best result is bold, second best is underlined.**

Prompt	Shots	WDC Products						Abt-Buy						Walmart-Amazon					
		GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral
Fewshot-related	6	51.59	85.71	89.96	59.64	77.43	37.45	90.96	93.83	94.35	66.78	90.72	49.77	74.32	91.19	90.54	56.23	81.31	50.78
	10	57.87	86.45	<b>91.74</b>	53.23	<b>84.04</b>	41.93	<b>92.62</b>	<u>94.35</u>	94.87	63.78	92.80	52.44	74.85	<b>91.24</b>	<b>90.63</b>	56.15	86.89	54.70
Fewshot-random	6	67.48	86.55	87.69	57.25	77.53	53.80	88.19	94.12	95.26	62.99	92.65	55.75	75.00	88.89	88.62	60.68	86.34	53.68
	10	73.23	86.37	87.40	50.83	80.17	50.22	90.31	93.21	<b>95.55</b>	68.28	<b>93.75</b>	49.80	76.78	89.00	88.78	<b>67.44</b>	88.00	46.94
Fewshot-handpicked	6	55.56	<u>87.23</u>	88.28	58.26	75.45	48.12	88.06	93.36	<u>95.49</u>	71.56	89.23	47.79	70.59	88.84	88.94	<u>66.52</u>	86.65	50.64
	10	59.73	86.72	87.89	46.96	79.39	44.86	89.58	93.62	<u>93.95</u>	<b>82.21</b>	92.80	42.02	74.38	87.89	90.34	65.86	<b>88.53</b>	45.56
Hand-written rules	0	73.76	85.71	86.49	53.77	80.84	<u>69.81</u>	90.50	94.15	87.69	41.93	89.86	86.18	<u>87.60</u>	89.16	84.55	33.65	<u>88.16</u>	<b>83.08</b>
Learned rules	0	<u>78.68</u>	87.06	85.24	59.33	80.17	<b>70.25</b>	89.43	93.40	92.91	48.38	88.19	<b>88.83</b>	<b>87.94</b>	86.21	88.40	43.27	86.47	<u>80.11</u>
Mean	-	64.74	86.48	86.48	54.91	79.38	52.06	90.03	93.81	93.76	63.24	91.25	59.07	77.68	89.05	89.05	56.23	86.54	58.19
Standard deviation	-	9.25	0.52	0.52	4.22	2.44	11.38	1.48	0.40	0.39	11.95	1.89	16.83	6.04	1.54	1.54	11.30	2.13	13.83
Best zero-shot	0	<b>81.15</b>	<b>89.61</b>	87.64	<b>69.09</b>	<u>83.67</u>	53.37	<u>91.93</u>	<b>95.78</b>	93.95	<u>82.03</u>	89.84	82.20	86.58	89.67	86.65	63.91	84.85	70.44
$\Delta$ Few-shot/zero-shot	-	-7.92	-2.38	4.10	-9.45	0.37	0.43	0.69	-1.43	1.60	0.18	3.91	-26.45	-9.80	1.57	3.98	3.53	3.68	-15.74
$\Delta$ Rules/zero-shot	-	-2.47	-2.55	-1.15	-9.76	-2.83	16.88	-1.43	-1.63	-1.04	-33.65	0.02	6.63	1.36	-0.51	1.75	-20.64	3.31	12.64

Prompt	Shots	Amazon-Google						DBLP-Scholar						DBLP-ACM					
		GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral	GPT-mini	GPT-4	GPT-4o	Llama2	Llama3.1	Mixtral
Fewshot-related	6	61.58	84.27	82.95	62.48	65.28	39.11	75.83	88.00	86.44	69.11	80.18	53.82	88.25	98.41	<b>98.21</b>	78.37	97.78	72.12
	10	63.52	<b>85.21</b>	<b>83.46</b>	62.36	71.43	47.89	77.93	88.52	88.33	64.29	82.38	55.57	92.54	<b>99.01</b>	98.20	76.34	97.58	66.95
Fewshot-random	6	57.37	78.08	78.67	59.78	74.32	48.99	82.56	90.21	<u>90.16</u>	67.04	<u>86.43</u>	55.80	96.55	<u>98.81</u>	<u>98.21</u>	76.22	98.40	76.21
	10	60.56	78.76	78.61	59.92	<b>80.46</b>	46.17	<u>84.98</u>	89.30	<b>90.54</b>	69.63	<b>87.17</b>	56.13	97.19	97.66	98.21	77.64	98.80	74.39
Fewshot-handpicked	6	55.19	76.92	77.25	<b>64.65</b>	73.93	45.92	68.86	<u>90.98</u>	89.19	81.34	85.34	67.70	<b>98.61</b>	94.34	97.47	80.78	98.62	86.36
	10	58.49	76.57	77.22	<u>63.89</u>	<u>79.52</u>	<u>49.18</u>	62.96	<b>91.81</b>	90.06	73.56	86.13	55.61	<u>98.42</u>	95.97	97.66	<u>86.96</u>	<b>99.21</b>	68.97
Hand-written rules	0	66.37	72.47	73.83	46.39	71.90	<b>58.31</b>	76.74	87.34	89.30	53.72	84.82	<b>83.55</b>	93.95	97.09	96.30	77.88	97.85	<b>93.26</b>
Learned rules	0	<u>68.28</u>	73.50	72.39	59.77	71.88	48.55	84.00	89.42	78.59	8.43	81.95	78.38	95.42	90.25	92.25	46.21	95.97	81.04
Mean	-	61.42	78.22	78.22	59.91	73.59	48.02	76.73	89.45	89.45	60.89	84.30	63.32	95.12	96.44	96.44	75.05	98.03	77.41
Standard deviation	-	4.19	4.27	4.27	5.41	4.51	4.95	7.15	1.41	1.41	2.34	11.03	3.25	2.76	2.76	11.38	0.94	8.39	
Best zero-shot	0	<b>72.18</b>	76.38	73.56	57.93	73.99	40.98	<b>86.11</b>	89.82	89.76	<b>85.46</b>	86.32	77.75	97.60	98.41	97.06	<b>97.62</b>	<u>98.81</u>	<u>90.32</u>
$\Delta$ Few-shot/zero-shot	-	-8.66	8.83	9.90	6.72	6.47	8.20	-1.13	1.99	0.78	-4.12	0.85	-10.05	1.01	0.60	1.15	-10.66	-0.01	-3.96
$\Delta$ Rules/zero-shot	-	-3.90	-2.88	0.27	1.84	-2.09	17.33	-2.11	-0.40	-0.46	-31.74	-1.50	5.80	-2.18	-1.32	-0.76	-19.74	-0.96	2.94

**Table 6: Mean results for the in-context learning.**

Prompt	Shots	All Datasets (Mean F1)					
		GPT4-mini	GPT4	GPT4o	Llama2	Llama3.1	Mixtral
Fewshot-related	6	73.76	90.24	<u>90.41</u>	65.44	82.12	50.51
	10	76.56	<b>90.80</b>	<b>91.21</b>	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	<b>88.06</b>	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	<u>87.60</u>	51.03
Hand-written rules	0	81.49	87.65	86.36	51.22	<u>85.57</u>	<b>79.03</b>
Learned rules	0	<u>84.14</u>	86.64	84.96	44.23	84.11	<u>74.53</u>
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	<b>85.51</b>	89.95	88.10	<b>76.01</b>	86.25	69.18
$\Delta$ Few-shot/zero-shot	-	-5.00	0.85	3.10	-5.49	1.81	-11.42
$\Delta$ Rules/zero-shot	-	-1.37	-2.29	-1.74	-24.78	-0.68	9.86

26% performance on most datasets. For all other LLMs providing in-context examples usually leads to performance improvements, while the size of the improvements varies widely.

In summary, in-context learning improves the performance of the LLMs for approximately 61% of the model/dataset combinations that we tested (see row  $\Delta$  *Few-shot/zero-shot* in Table 5). Providing demonstrations was not helpful for GPT4 which does not need the additional guidance on two datasets as well as for the smaller models GPT-mini and Mixtral which suffer large performance drops on many datasets. As a result, the usefulness of in-context learning cannot be assumed but needs to be determined experimentally for each model/dataset combination.

**Comparison of Selection Methods:** The best demonstration selection method also varies depending on the dataset. The open-source LLMs generally reach the best performance when random or handpicked demonstrations are provided. In contrast, GPT-4 and GPT-4o achieve the highest scores on most datasets using related demonstrations, suggesting that these models are better able to understand and apply specific patterns from closely

related examples to the current matching decision. The handpicked demonstrations, while not helpful for the Llama models on their source dataset WDC Products, lead to improvements on all other product datasets. The same effect is visible for the handpicked demonstrations transferred to DBLP-ACM.

## 4.2 Learning Matching Rules

In the next set of experiments, we provide a set of textual matching rules in the prompt in order to guide the model to select the correct solution. We differentiate between two kinds of rules (i) handwritten and (ii) learned rules. Handwritten rules are a set of binary rules created by defining which attributes need to match for the given domain to signify a match. The rules also inform the model of potential heterogeneity in these attributes, such as slight differences in surface form or value formats. For the learned rules, we pass the set of handpicked in-context pairs to GPT4 and ask the model to automatically generate matching rules from these examples. Similar to the handwritten rules, they refer to specific attributes that should be matching and potential sources of heterogeneity that the GPT4 model extracted from the provided examples. A subset of these handwritten and learned rules for the product domain is depicted in Figure 3. The full list of learned rules is available in the project repository.

**Effectiveness:** Table 5 shows the results of providing matching rules in comparison to the best zero-shot prompt and the in-context experiments. The results show that GPT4 with matching rules does not improve over its best zero-shot performance and instead loses 1% to 3% F1 on all datasets. All other models see improvements on some datasets of 0.3% to 17% F1 over zero-shot depending on the model/dataset combination. Especially the Mixtral LLM, which has comparatively low performance compared to all other LLMs in the zero-shot and few-shot settings, significantly improves with the provision of rules on all datasets,

<b>Task Desc.</b>	The following rules regarding product features need to be observed:
<b>Hand-Written Rules</b>	<ol style="list-style-type: none"> <li>1. The brand of matching products must be the same if available.</li> <li>2. Model names of matching products must be the same if available.</li> <li>3. Model numbers of matching products must be the same if available.</li> <li>4. Additional features of matching products must be the same if available.</li> <li>5. Matching attributes may not have the exact same surface form due to different case, typos, value formats.</li> <li>6. If an attribute is missing for one description, it is likely still a match if the existing attributes match.</li> </ol>
<b>Task Desc.</b>	Do the following two product descriptions match? Answer with 'Yes' if they do and 'No' if they do not.
<b>Task Input</b>	Product 1: 'Title: DYMO D1 – Roll (1.9cm x 7m)' Product 2: 'Title: DYMO D1 Tape 12mm x 7m'
<b>Learned Rules</b>	<ol style="list-style-type: none"> <li>1. The product titles match if they both refer to the same core product, which means they have the same primary brand and product model ...</li> <li>2. Order of words and usage of separators (like hyphen, comma, slash, parentheses) does not matter ...</li> <li>3. The product titles match if they both contain the same key information, even if one product title contains additional details ...</li> <li>4. Exact matching of all specifications is not necessary. Titles match as long as the critical specs like model number, series, and brand are the same ...</li> <li>5. The product titles do not match if they refer to different models of the same brand ...</li> <li>6. ...</li> </ol>

Figure 3: Example of a prompt containing handwritten matching rules for the product domain. A subset of the learned rules is depicted below.

gaining from 3 to 17% F1. In summary, the provision of matching rules can be helpful, especially for the open-source LLMs with Mixtral achieving its highest scores on all datasets using rules but providing task demonstrations generally leads to higher performance gains than providing matching rules for all other models.

**Sensitivity:** We measure the prompt sensitivity of the LLMs as the standard deviation of the F1-scores across all few-shot and rule experiments. We list this standard deviation in the lower part of Tables 5 and 6. Comparing the prompt sensitivity of the models to the zero-shot deviations across different prompt formulations, the average deviation from the mean has decreased for all models, suggesting that the additional guidance in the form of demonstrations and rules leads to more robust results.

### 4.3 Fine-Tuning

In the next set of experiments, we fine-tune the GPT-mini model via the OpenAI API as well as the Llama2 and Llama3.1 models using local hardware. We use the training and validation sets of each dataset to train a fine-tuned model with the *domain-simple-force* prompt and subsequently apply the fine-tuned models with this prompt to all datasets. We fine-tune GPT-mini for 10 epochs using the default parameters suggested by OpenAI. For the Llama models, we fine-tune using 4-bit quantization to manage the high VRAM requirements of the 70B models. We employ Low-Rank Adaptation (LoRA) and also train for 10 epochs.

**Effectiveness:** The results of the fine-tuned LLMs are shown in Table 7. The lower part of the table restates the best zero-shot and GPT4 results for comparison. When comparing the fine-tuning results to the best zero-shot performance (Section  $\Delta$  best zero-shot in Table 7), we observe a substantial improvement of 1% to 26% F1 depending on the dataset for all models. Only the Llama models on WDC Products do not profit from fine-tuning. On four out of six datasets, the best fine-tuned Llama3.1 and GPT-mini models exceed the performance of zero-shot GPT4 by 1 to 10% F1 (See Section  $\Delta$  best GPT4 in Table 7).

Table 7: Results for fine-tuning LLMs and subsequent transfer to all datasets. Left-most column shows the dataset used for fine-tuning.

		WDC	A-B	W-A	A-G	D-S	D-A
WDC Products	Llama2	66.81	75.98	72.83	54.77	41.74	28.86
	Llama3.1	72.05	83.47	76.92	63.97	65.25	80.91
	GPT-mini	<u>88.89</u>	92.49	88.61	77.78	86.82	97.28
Abt-Buy	Llama2	58.79	92.15	81.84	68.61	84.12	95.31
	Llama3.1	77.87	93.60	84.85	74.49	79.86	94.07
	GPT-mini	83.66	94.17	88.83	76.63	86.03	97.85
Walmart-Amazon	Llama2	49.71	88.13	90.57	66.50	64.57	87.74
	Llama3.1	51.12	89.92	91.01	73.76	82.62	95.41
	GPT-mini	72.64	<u>94.94</u>	<u>92.99</u>	<u>82.06</u>	89.31	97.85
Amazon-Google	Llama2	59.50	77.16	63.21	76.19	78.59	88.70
	Llama3.1	61.28	85.00	82.35	78.67	70.49	87.73
	GPT-mini	64.75	90.23	82.98	<b>87.11</b>	87.02	96.88
DBLP-Scholar	Llama2	29.41	49.48	59.52	46.55	<u>92.80</u>	97.45
	Llama3.1	35.11	74.27	77.15	58.59	<u>92.37</u>	97.84
	GPT-mini	57.70	84.07	84.02	73.33	<b>93.95</b>	97.64
DBLP-ACM	Llama2	6.15	29.96	29.60	16.22	66.84	99.20
	Llama3.1	15.33	49.82	41.90	25.00	83.37	<b>99.60</b>
	GPT-mini	31.54	85.25	67.99	49.43	89.70	<u>99.40</u>
Best zero-shot	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
	GPT-mini	81.15	91.93	86.58	72.18	86.11	97.60
$\Delta$ best zero-shot	Llama2	<b>-2.28</b>	<b>+10.12</b>	<b>+26.66</b>	<b>+18.26</b>	<b>+7.34</b>	<b>+1.58</b>
	Llama3.1	<b>-5.80</b>	<b>+3.76</b>	<b>+6.16</b>	<b>+4.68</b>	<b>+6.05</b>	<b>+0.79</b>
	GPT-mini	<b>+7.74</b>	<b>+3.01</b>	<b>+6.41</b>	<b>+14.93</b>	<b>+7.84</b>	<b>+1.80</b>
$\Delta$ best GPT4	Llama2	<b>-22.8</b>	<b>-3.63</b>	<b>+0.90</b>	<b>-0.19</b>	<b>+2.98</b>	<b>+0.79</b>
	Llama3.1	<b>-11.74</b>	<b>-2.18</b>	<b>+1.34</b>	<b>+2.29</b>	<b>+2.55</b>	<b>+1.19</b>
	GPT-mini	<b>-0.72</b>	<b>-0.84</b>	<b>+3.32</b>	<b>+10.73</b>	<b>+4.13</b>	<b>+0.99</b>
Best GPT4	-	<b>89.61</b>	<b>95.78</b>	<u>89.67</u>	76.38	89.82	98.41

In summary, fine-tuning the models leads to improved results compared to the zero-shot version of the model rivaling the performance of the best GPT4 prompts with the much cheaper GPT-mini model and consistently improving the performance of the Llama models by 1-26% F1 on 5 out of 6 datasets leaving Llama 3.1 only slightly behind GPT-mini on 4 datasets. Furthermore, the experiments show that the fine-tuned Llama models reach a similar performance or outperform GPT4 on 4 out of 6 datasets.

**Generalization:** We observe a generalization effect for the GPT-mini model fine-tuned on one dataset to datasets from related domains and across domains. Transferring models between related product domains leads to improved performance over the best zero-shot prompts for many combinations of datasets. The effect is especially visible for the combinations WDC Products, Abt-Buy and Walmart-Amazon which contain similar products. The transfer to Amazon-Google results in better performance than zero-shot for all of the mentioned product datasets. Conversely, the reverse transfer from Amazon-Google does not yield improved results. Furthermore, all GPT-mini models fine-tuned on the datasets from the product domain exhibit good generalization to the publication domain, resulting in improvements of 1-3% F1 over the best zero-shot. Transferring fine-tuned models within the publication domain shows the same effect. The transfer does not work in the other direction as transferring a model fine-tuned for the publication domain leads to lower performance on the product datasets. For the Llama models this effect is only visible for some inter-product transfers mostly for Llama2.

**Table 8: Costs for hosted LLMs on WDC Products. Best performing prompts are selected for the analysis for each scenario.**

	Zeroshot			6-Shot			10-Shot			Rules (written)			Rules (learned)			Fine-tune	
	GPT-mini	GPT-4	GPT-4o	GPT-mini	GPT-4	GPT-4o	GPT-mini	GPT-4	GPT-4o	GPT-mini	GPT-4	GPT-4o	GPT-mini	GPT-4	GPT-4o	Train	Inference
F1 (Best prompt)	81.15	89.61	87.64	67.48	87.23	89.96	73.23	86.72	91.74	73.76	85.71	86.49	78.68	87.06	85.24	-	88.89
Mean # Tokens prompt	76	77	93	633	639	641	992	942	1,009	213	214	213	815	817	815	97	88
Mean # Tokens completion	89	40	1	2	2	2	2	2	2	1	1	4	1	1	3	1	1
Mean # Tokens combined	166	117	94	635	641	643	994	944	1,011	214	215	217	816	818	818	98	89
Token increase to ZS	-	-	-	3.8x	5.5x	6.8	6x	8.1x	10.8x	1.3x	1.8x	2.3x	4.9x	7x	8.7x	0.6x	0.5x
Cost per prompt	0.006€	0.474€	0.024€	0.01€	2.056€	0.162€	0.015€	3.037€	0.254€	0.003€	0.649€	0.057€	0.012€	2.458€	0.207€	0.280€	0.003€
Cost increase to ZS (GPT-mini)	-	73x	3.8x	1.5x	319x	25x	2.4x	470x	39x	0.5x	101x	9x	1.9x	381x	32x	43x	0.5x
Cost increase per $\Delta$ F1 to ZS	-	8.7x	0.6x	0.1x	52x	3x	0.3x	84x	3.7x	0.07x	22x	1.7x	0.8x	64x	7.8x	-	0.06x

**Table 9: Runtime in seconds per prompt (request) for all LLMs using the best prompts from the previous sections on the WDC Products dataset. Runtimes marked with \* are for the quantized version of the model used for fine-tuning.**

Model	Zeroshot	6-Shot	10-Shot	Rules (written)	Rules (learned)	Fine-Tune (Inference)
GPT-mini	1.54 s	0.46 s	0.51 s	0.47 s	0.47 s	0.46 s
GPT4	2.19 s	0.75 s	0.78 s	0.68 s	0.76 s	-
GPT4-o	0.51 s	0.48 s	0.53 s	0.48 s	0.49 s	-
Llama2	22.62 s	7.15 s	7.82 s	23.16 s	24.51 s	*0.30 s
Llama3.1	0.54 s	1.70 s	2.36 s	0.67 s	1.70 s	*0.30 s

## 5 COST AND RUNTIME ANALYSIS

Apart from pure matching performance there are additional considerations such as data privacy requirements and the cost of using hosted LLMs which may result in the decision to use a less performant but cheaper hosted LLM or to run an open-source LLM on local hardware. The cost analysis presented in the following gives an overview of expected costs for hosted models. The purpose of the analysis is to give the reader general guidance of what to expect with regards to the cost dimension. We leave a more in-depth analysis of costs including acquisition costs for GPUs and electricity for the open-source models to future work.

**Costs:** Table 8 lists the costs associated with the hosted LLMs across all experimental scenarios for the WDC Products dataset. The cost of using a hosted LLMs is dependent on the length of the respective prompts, measured by the amount of tokens, and the current prices of the respective model. Thus, the results we present here are only a snapshot as of August 2024 as the prices are subject to change. We compare the costs of all OpenAI models. The prices for using the models were as follows for 1 million prompt/completion tokens: \$0.15/\$0.60 for GPT-mini, \$30.00/\$60.00 for GPT-4, and \$2.50/\$10.00 for GPT-4o.

Table 8 shows that the in-context learning (6-shot, 10-shot) and the rule-based approaches (hand-written, learned) from Section 4 require between 1.3 and 11 times the amount of tokens per prompt compared to basic zeroshot prompting (see row *Token increase to ZS* in Table 8). For all of them this is due to longer prompts, either because of the inclusion of few-shot demonstrations or rules. The fine-tuning approach on the other hand requires less tokens than zero-shot as the prompt we chose for fine-tuning uses the restricted output format *force* (see Section 3) whereas the best zero-shot prompt for GPT-mini uses the *free* format which allows the model to answer more verbosely. From a cost perspective, the in-context learning and the rule-based approaches increase the costs by 1.5 to 470 times compared to the cost of the zero-shot GPT-mini model. While GPT-mini is the cheapest model in this lineup, the GPT-4o model achieves

significantly higher performance, often approaching or even surpassing GPT-4, at a fraction of GPT-4’s cost. If many training examples are available, fine-tuning the GPT-mini model results in comparably high performance for for a fraction of the cost of even GPT-4o.

**Runtime:** Table 9 lists the average runtime per prompt for all LLMs. The selected prompts and the used numbers of tokens are the same as in Table 8. If the prompt allowed free form answering, this leads to much longer runtimes compared to forcing the model to answer shortly. The large difference in runtimes between zero-shot Llama2 and Llama3.1 in Table 9 is an example of this. The runtimes of the hosted models are a snapshot of the API performance in August 2024 and may change at any time. Prompting GPT4 generally takes around 50% longer than the other two OpenAI models which have comparable runtimes if the answering scheme is the same. The locally hosted open-source LLM Llama2 requires the largest amount of time for most scenarios on our hardware (see Section 2), particularly when generating freely in zero-shot and rule-based cases, where its runtime is 10 to 33 times longer than that of GPT-4. On the other hand, the Llama3.1 model achieves a comparable runtime to the GPT models in most setups.

## 6 EXPLAINING MATCHING DECISIONS

Understanding the decisions of a matching model is important for users to build trust towards the systems. Explanations of model decisions can further be used for debugging matching pipelines. The size and structure of deep learning models make explaining their decisions a challenging task, which has led to a dedicated line of research in the field of entity matching [2, 12, 32, 33]. Instead of relying on external explainability methods, LLMs can directly be queried for explanations of their decisions. In this Section, we use GPT4 to generate structured explanations for its decisions and show how to aggregate these explanations to derive global insights about matching decisions.

### 6.1 Generating Explanations

For the generation of explanations, we first prompt the LLM to match a pair of entities and subsequently ask the model for an explanation of its decision using a second prompt. If we do not pose any restrictions on the format of the explanation, the model would answer with natural language text describing the different aspects that influenced its decision [29]. Instead of allowing free-text explanations, we ask the model to organize its explanations into a fixed structure which will later allow us to parse and aggregate the explanations. Figure 4 shows examples of complete conversations for generating structured explanations of matching decisions for pairs from the Walmart-Amazon and DBLP-Scholar datasets. After prompting for and receiving a decision in the



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. <b>Entity 1:</b> night owl night owl zeus-810 16-channel h. 264 dvr kit with 1tb hard drive 773.82 <b>Entity 2:</b> 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 <b>explain your decision in a structured format</b> , 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 <b>importance of the attribute</b> 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 <b>similarity score for the attribute values</b> . 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'
AI	<b>attribute=brand</b> - importance=0.30 - values=Night Owl###Night Owl Optics - similarity=0.90 <b>attribute=model</b> - importance=0.40 - values=Zeus-810 16-channel H. 264 DVR kit###Zeus-810 16-channel 8-camera H. 264 DVR surveillance kit - similarity=0.85 <b>attribute=price</b> - importance=0.20 - values=773.82###676.63 - similarity=0.80 <b>attribute=hard drive</b> - importance=-0.10 - values=1TB###missing - similarity=0.00
Actor	Prompt (DBLP-Scholar)
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. <b>Entity 1:</b> b george , j haritsa secure bufering in firm real-time database systems vldb j, 2000 <b>Entity 2:</b> b george , j haritsa secure transaction processing in firm real-time database systems sigmod conference , 1997.0"
AI	No.
User	Now explain your decision ... (same as in example above)
AI	<b>attribute=authors</b> - importance=0.30 - values=b george , j haritsa###b george , j haritsa - similarity=1.00 <b>attribute=title</b> - importance=-0.70 - values=secure bufering in firm real-time database systems###secure transaction processing in firm real-time database systems - similarity=0.60 <b>attribute=year</b> - importance=-0.50 - values=2000###1997.0 - similarity=0.00 <b>attribute=conference</b> - importance=-0.50 - values=vldb j.###sigmod conference - similarity=0.00

Figure 4: Conversation instructing the model to match an entity pair and asking for a structured explanation of the decision. Top: Walmart-Amazon, bottom: DBLP-Scholar.

first exchange with the model, we continue the conversation by passing a second prompt (the second user prompt in Figure 4). Specifically, we ask for a structured format of the explanation that includes all attributes of both product offers that were used for the matching decision. Each attribute should be accompanied by an importance value as well as a similarity value for the compared attributes. The sign of the importance values should be negative if the attribute comparison contributed to a non-match decision and vice versa.

The generated structured explanation of the product pair from Walmart-Amazon is shown in the second blue AI row in Figure 4. The explanation shows that the model is capable of extracting various attributes from the serialized strings. The highest positive importance is assigned to the attribute *model* followed by *brand* and *price*. Although none of the extracted attribute values perfectly match, they are very similar and the model correctly assigns them a high similarity and positive importance value and considers them indications for matching product offers. Interestingly, the model extracted the *hard drive* size from the first offer which is missing in the second offer and assigned due to this circumstance a low negative importance score. As the size of the hard drive is an important piece of information for matching, the model may be accounting for this uncertainty by reducing its confidence in this specific case. The explanation for the DBLP-Scholar pair is shown in the 4th blue AI row in Figure 4. The values of the *authors* attribute match perfectly, which the model recognizes as relevant evidence for a match by assigning a positive importance of 0.3. The model further correctly assigns a high negative importance to *year* and *conference* which are reasonably different to support a non-match decision. Here it is interesting that while the *title* overlaps in all but two words, the model still uses this as the most important evidence for predicting non-match.

To evaluate the meaningfulness of the similarity values created by the model in the structured explanations, we calculate their Pearson correlation with the well known string similarity metrics Cosine and Generalized Jaccard. We apply the latter metrics to each of the extracted attributes found in the explanations and calculate the correlation between them and the generated similarities. We find that the model generated similarities exhibit a strong positive correlation with Cosine similarity and Generalized Jaccard similarity, ranging between 0.75–0.85 and 0.73–0.83, respectively, across all datasets. These results point to the general meaningfulness of the GPT4 created similarity values.

We subsequently generate structured explanations for all pairs in the test sets of both datasets using the best-performing zero-shot prompt. A sample of the generated explanations was manually verified against the corresponding model decisions, confirming the connection between the explanations and the model’s decisions. All explanations are available in the project repository to enable the further analysis of their quality.

## 6.2 Aggregating Explanations

The structured explanations can easily be parsed to extract the attributes, importance scores, and similarity values. We aggregate the extracted values by attribute and calculate average importance scores for all attributes deemed relevant by the model for its decisions. Examples of five of these aggregated average importance scores are shown in Table 10 for both datasets. We can see that the model frequently assigned a high importance to

**Table 10: Global insights about the importance of different attributes for matching and non-matching decisions for the DBLP-Scholar and Walmart-Amazon datasets.**

Attribute	Matches			Non-Matches		
	Freq.	Mean Import.	St.Dev.	Freq.	Mean Import.	St.Dev.
DBLP-Scholar						
title	0.96	<b>0.59</b>	0.40	0.95	-0.40	0.38
authors	0.78	<b>0.65</b>	0.40	0.68	<b>-0.66</b>	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	<b>0.78</b>	0.34	0.99	-0.04	0.34
price	0.92	-0.03	0.27	0.86	-0.16	0.25
model	0.81	<b>0.63</b>	0.51	0.82	<b>-0.77</b>	0.37
color	0.24	0.23	0.31	0.35	-0.06	0.23
product type	0.12	<b>0.64</b>	0.48	0.11	-0.42	0.50

*brand* and *model* for the matches while the *price* was not considered relevant for these decisions on average. For non-matches, the model instead focuses on the *model* attribute and assigns a nearly neutral average importance to the *brand* attribute. For DBLP-Scholar, GPT4 focuses on differences and similarities of the *title* and *author* attributes of the publications for both matches and non-matches, while the attributes *conference* and *year* only contribute to a lesser extent to the matching decisions.

After the aggregation there are in total 81 attributes for DBLP-Scholar with seven of them being used in at least 10% of decisions while the remaining 76 make up the long-tail. 28 of 81 attributes have a mean importance, positive or negative, of at least 30%. For Walmart-Amazon there are 181 attributes with seven of them used in at least 10% of decisions. 64 of 181 have a mean importance of at least 30% towards the decision. The aggregation of the structured explanations for the DBLP-Scholar and the Walmart-Amazon datasets has demonstrated that global insights about a model’s decisions can be derived from the local explanations.

## 7 AUTOMATED ERROR ANALYSIS

The analysis of wrongly matched entity pairs may lead to insights on how to improve the matching pipeline. The analysis of matching errors requires a decent understanding of the application domain, e.g. products or publications, and profound knowledge about the entity matching task. Error analysis usually involves manually inspecting the errors made by matching systems and subsequently deriving a set of error classes for categorizing these errors. The task of deriving the error classes is not mechanical but a rather creative task requiring reasoning capabilities and background knowledge. This section demonstrates that GPT4-turbo can automate this creative task and derive meaningful error classes from the errors and associated explanations that were created in Section 6. The machine-generated error classes can be helpful for data engineers as they widen the scope of their analysis.

### 7.1 Discovery of Error Classes

For the automatic discovery of error classes, we select all wrong decisions together with their structured explanations from the sets of explanations that we generated for the DBLP-Scholar and Walmart-Amazon datasets in Section 6. Afterwards, we pass

Actor	Prompt
User	<p>The following list contains false positive and false negative product pairs from the output of a product matching classification system. <b>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.</b></p> <p><b>False Negatives:</b></p> <p>FN1 {Entity 1} {Entity 2} {Explanation} {additional FNs}</p> <p><b>False Positives:</b></p> <p>FP1 {Entity 1} {Entity 2} {Explanation} {additional FPs}</p>
AI	<p>Based on the provided examples of false negatives (FN) and false positives (FP), we can identify several error classes that explain why the classification system fails on these examples.</p> <p><b>### Error Classes for False Negatives:</b></p> <ol style="list-style-type: none"> <li><b>Model Number Mismatch:</b> The system fails when there are slight differences in model numbers or product codes, even when other attributes match closely.</li> <li><b>{additional error classes}</b></li> </ol> <p><b>### Error Classes for False Positives:</b></p> <ol style="list-style-type: none"> <li><b>Overemphasis on Matching Attributes:</b> The system might give too much weight to matching attributes like brand or model number, leading to false positives even when other important attributes differ.</li> <li><b>{additional error classes}</b></li> </ol>

**Figure 5: Prompt used for the automatic generation of error classes given false positives and false negatives.**

a prompt to GPT4-turbo asking the model to synthesise error classes for false positive and false negatives cases separately. In the second part of this prompt, we include the selected erroneous pairs together with their GPT4 created explanations. This are 26 false positives and 26 false negatives for the DBLP-Scholar test set and 26 false positive and 15 false negatives for the Walmart-Amazon test set. Figure 5 shows the prompt that we use for the automatic generation of the error classes, as well as part of the answer of the LLM.

Table 11 and Table 12 show the generated error classes for both datasets and for each class the number of errors that fall into these classes. The latter are manually annotated by three domain experts. For DBLP-Scholar, three additional error classes were created but for the sake of presentation are not listed in the table. The full set of created error classes, as well as the false positives and false negatives used to create them are found in the accompanying repository. The counts in the # errors columns of Table 11 and Table 12 show that the automatically created error classes are relevant and cover not only frequent errors but also rarer errors. For example, for the DBLP-Scholar dataset, the first error class of the false positives refers to putting too much emphasis on the similarity of publication *titles* which is deemed

**Table 11: Generated error classes for the DBLP-Scholar dataset and manually annotated number of errors.**

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

correct by a human annotator for 15 of the 26 errors, while the third error class is relevant only for 5 of the errors, namely those where the model seemed to put too much emphasis on matching *year* and *venue* information in the pairs while ignoring crucial difference in the other attributes. After manual inspection, all of the created error classes are relevant for the errors being made and support a deeper understanding of what causes these errors. Some of the error classes also point at actions that could be taken to improve the matching pipeline. For example, the heterogeneity of how publication *venues* are listed in the DBLP-Scholar dataset (Table 11, error class 2 for false negatives) could prompt the user to improve the normalization of these values.

## 7.2 Assignment of Errors to Error Classes

In this final experiment, we investigate whether GPT4-turbo is capable of categorizing errors into the created error classes. Such a categorization allows data engineers to drill down from the error classes to concrete example errors which might give them hints on how to address the problem. For categorizing errors, we use the prompt shown in Figure 6. After instructing the model about the task, the prompt lists all error classes together with their descriptions. Subsequently, the prompt contains the entity pair to be categorized together with its correct as well as predicted label and the structured explanation of the matching decision. The model is asked to pick all error classes that apply to the pair and to provide a confidence value for each of its predictions.

Table 13 shows the accuracy values the GPT4-turbo model reaches on this task. From these values we can see that the model on average achieves a mean accuracy of over 80% for most error types (see row *Mean* in Table 13). Only the mean accuracy on Walmart-Amazon’s false positives is lower which is caused by the low accuracy of the first error class *Overemphasis on Matching Attributes* as the domain experts did not agree with the models classification in the first error class, more specifically the model rarely assigned this class while the domain experts considered it relevant in 23 out of 26 cases. Apart from this disagreement, the model is capable of correctly categorizing the errors with a high accuracy.

**Table 12: Generated error classes for the Walmart-Amazon dataset and manually annotated number of errors.**

False Negatives (15 overall)	# errors
<b>1. Model Number Mismatch:</b> The system fails when there are slight differences in model numbers or product codes, even when other attributes match closely.	9
<b>2. Attribute Missing or Incomplete:</b> When one product listing includes an attribute that the other does not, the system may fail to recognize them as a match.	9
<b>3. Minor Differences in Descriptions:</b> Small differences in product descriptions or titles can lead to false negatives, such as slightly different wording or the inclusion/exclusion of certain features.	11
<b>4. Price Differences:</b> Even when products are very similar, significant price differences can lead to false negatives, as the system might weigh price too heavily.	12
<b>5. Variant or Accessory Differences:</b> Differences in product variants or accessories included can cause false negatives, especially if the system does not adequately account for these variations being minor.	7
False Positives (26 overall)	# errors
<b>1. Overemphasis on Matching Attributes:</b> The system might give too much weight to matching attributes like brand or model number, leading to false positives even when other important attributes differ.	23
<b>2. Ignoring Minor but Significant Differences:</b> The system fails to recognize important differences in product types, models, or features that are significant to the product identity.	21
<b>3. Misinterpretation of Accessory or Variant Information:</b> Including or excluding accessories or variants in the product description can lead to false positives if the system does not correctly interpret these differences.	8
<b>4. Price Discrepancy Overlooked:</b> The system might overlook significant price differences, assuming products are the same when they are not, particularly if other attributes match closely.	14
<b>5. Condition or Quality Differences:</b> Differences in the condition or quality of products (e.g., original vs. compatible, new vs. refurbished) are not adequately accounted for, leading to false positives.	2

**Table 13: Accuracy of GPT4 for classifying errors.**

Error class	Walmart-Amazon		DBLP-Scholar	
	FP	FN	FP	FN
1	34.62	86.67	92.31	96.15
2	84.62	73.33	76.92	92.31
3	84.62	73.33	76.92	73.08
4	76.92	100	100	88.46
5	84.62	86.67	92.31	88.46
Mean	73.08	84.00	87.69	87.69

The presented methods for the automated creation of error classes and the classification of errors into these classes by an LLM can support data engineers in the analysis and debugging of specific combinations of models, prompts and datasets. The methods can also be used for the detailed comparison of different combinations of models, prompts and datasets. For example, the errors from all experiments presented in this paper could be classified into the classes presented in Tables 11 and 12 allowing the fine-grained comparison of the strengths and weaknesses of each combination. As this analysis goes beyond the scope of this paper, we leave it to future work.

## 8 RELATED WORK

**Entity Matching:** Entity matching [3, 8, 15] has been researched for over 50 years [17]. Early approaches involved domain experts hand-crafting matching rules [17]. Over time, advancements were made with unsupervised and supervised machine learning techniques resulting in improved matching performance [9]. By the late 2010s, the success of deep learning in areas such as natural language processing and computer vision paved the way for

Actor	Prompt
User	<p>Given the following error classes for a product matching classification system, please <b>classify the following product pair into all error classes</b> by their number <b>if they are relevant for this pair</b> and its explanation. Please give a short explanation of every decision as a list first. Finally, also provide a confidence score for each classification adhering to the JSON format of the following example: {"2": "90", "4": "30", "5": "75"}</p> <p>Error classes:</p> <ol style="list-style-type: none"> <li>1. Model Number Mismatch: The system fails when there are slight differences in model numbers or product codes, even when other attributes match closely (e.g., FN2, FN7, FN8).</li> <li>2. {remaining error classes}</li> </ol> <p>Now classify this pair: Original Label: Match Predicted Label: Non-Match {Entity 1} {Entity 2} {Explanation}</p>
AI	<p>Based on the provided information and the error classes, here's the classification for the product pair:</p> <ol style="list-style-type: none"> <li>1. Model Number Mismatch: This error class is relevant because ...</li> <li>2. {additional classifications}</li> </ol> <p>{ "1": "95", "2": "60", "3": "50", "4": "70" }</p>

Figure 6: Prompt used for the classification of errors.

early applications in entity matching [28, 37]. The Transformer architecture [41] and pre-trained models like BERT [11] and RoBERTa [26] revolutionized natural language processing, which has led the data integration community to also turn to these language models for entity matching [5, 23, 33, 42, 47, 48]. More recent work delved into the application of self-supervised and supervised contrastive losses [7, 19, 21] in combination with PLM encoder networks for entity matching [34, 44]. Other studies have explored graph-based methods [20, 47] and the application of domain adaptation techniques for entity matching [1, 27, 39, 40].

**LLM-based Entity Matching:** Narayan et al. [30] were the first to experiment with using an LLM (GPT3) for entity matching as part of a wider study also covering data engineering tasks such as schema matching and missing value imputation. In [35], we employ ChatGPT for entity matching and test different prompt designs on a single benchmark dataset. Fan et al. [16] experiment with batching multiple entity matching decisions together with in-context demonstrations to reduce the cost of in-context learning. Wang et al. [45] go beyond binary matching and apply LLMs to select matching records from a set of candidate matches. Zhang et al. [49] experimented with fine-tuning a Llama2 model for several data preparation tasks at once and include entity matching as one of their fine-tuning tasks. In [38], we experiment with fine-tuning Llama and GPT models for entity matching using different example representations, including free text and structured explanations.

**Explaining Entity Matching:** The prevalence of PLMs over recent years in the field of entity matching has led to research

into the explainability of these matching systems [2, 12, 32, 33]. Most methods [12, 33] for explaining the matching decisions of PLMs provide local explanations for single entity pairs, e.g. as importance score of single tokens. Paganelli et al. [32] present an approach for explaining matching decisions by analyzing the attention scores of PLM-based matchers. The WYM [2] system is an example of an intrinsically interpretable system that was recently proposed based on the idea of finding important decision units among entity descriptions for PLM-based matchers. To the best of our knowledge, none of the existing methods automates the discovery error classes and generates human-interpretable descriptions of these error classes like the ones we presented in Section 7.

## 9 CONCLUSION

This paper has investigated using LLMs as a more robust and less task-specific training data dependent alternative to PLM-based matchers. We can summarize the high-level implications of our findings concerning the selection of matching techniques in the following rules of thumb: For use cases that do not involve many unseen entities and for which a decent amount of training data is available, PLM-based matchers are a suitable option which does not require much compute due to the smaller size of the models. For use cases that involve a relevant amount of unseen entities and for which it is costly to gather and maintain a decent size training set, LLM-based matchers should be preferred due to their high zero-shot performance and ability to generalize to unseen entities. If using the best performing hosted LLMs is not an option due to their high usage costs, fine-tuning a cheaper hosted model is an alternative that can deliver a similar F1 performance. If using hosted models is no option due to privacy concerns, using an open-source LLM on local hardware can be an alternative given that task-specific training data or domain-specific matching rules are available. Still, this approach is expected to result in a slightly lower F1 performance. We demonstrated that GPT4 can generate structured explanations of matching decisions and that we can automatically aggregate these explanations to gain global insights into the models decisions. Finally, we have shown that GPT4-turbo can perform the creative task of automatically deriving error classes from the explanations. This automation of the error analysis can save data engineers time and can point them at issues that they might have otherwise overlooked.

## ACKNOWLEDGMENTS

The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

## REFERENCES

- [1] Mehdi Akbarian Rastaghi, Ehsan Kamaloo, and Davood Rafiei. 2022. Probing the Robustness of Pre-trained Language Models for Entity Matching. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3786–3790.
- [2] Andrea Baraldi, Francesco Del Buono, Francesco Guerra, Matteo Paganelli, and Maurizio Vincini. 2023. An Intrinsically Interpretable Entity Matching System. In *Proceedings 26th International Conference on Extending Database Technology, Ioannina, Greece, March 28-31, 2023*. 645–657.
- [3] Nils Barlaug and Jon Atle Gulla. 2021. Neural Networks for Entity Matching: A Survey. *ACM Transactions on Knowledge Discovery from Data* 15, 3 (2021), 52:1–52:37.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- [5] Ursin Brunner and Kurt Stockinger. 2020. Entity Matching with Transformer Architectures - a Step Forward in Data Integration. In *Proceedings of the International Conference on Extending Database Technology*. 463–473.

- [6] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. 2020. Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD '20)*. 1335–1349.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*. 1597–1607.
- [8] Peter Christen. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer-Verlag, Berlin Heidelberg.
- [9] Vassilis Christophides, Vasilis Efthymiou, Themis Palpanas, George Papadakis, and Kostas Stefanidis. 2020. An Overview of End-to-End Entity Resolution for Big Data. *Comput. Surveys* 53, 6 (2020), 127:1–127:42.
- [10] Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis. 2015. *Entity Resolution in the Web of Data*. Springer International Publishing, Cham.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. 4171–4186.
- [12] Vincenzo Di Cicco, Donatella Firmani, Nick Koudas, Paolo Merialdo, and Divesh Srivastava. 2019. Interpreting Deep Learning Models for Entity Resolution: An Experience Report Using LIME. In *Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*. 8:1–8:4.
- [13] Huahua Ding, Chaofan Dai, Yahui Wu, Wubin Ma, and Haohao Zhou. 2024. SETEM: Self-ensemble Training with Pre-trained Language Models for Entity Matching. *Knowledge-Based Systems* 293 (June 2024), 111708.
- [14] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed representations of tuples for entity resolution. *Proc. VLDB Endow.* 11, 11 (jul 2018), 1454–1467.
- [15] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1 (2007), 1–16.
- [16] Meihao Fan, Xiaoyue Han, Ju Fan, Chengliang Chai, Nan Tang, et al. 2024. Cost-effective in-context learning for entity resolution: A design space exploration. In *2024 IEEE 40th International Conference on Data Engineering*. IEEE, 3696–3709.
- [17] Ivan P. Fellegi and Alan B. Sunter. 1969. A Theory for Record Linkage. *J. Amer. Statist. Assoc.* 64, 328 (1969), 1183–1210.
- [18] Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution. In *Forty-first International Conference on Machine Learning*.
- [19] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6894–6910.
- [20] Congcong Ge, Pengfei Wang, Lu Chen, Xiaozhe Liu, Baihua Zheng, et al. 2021. CollaborER: A Self-supervised Entity Resolution Framework Using Multi-features Collaboration. *arXiv:2108.08090 [cs]* (Sept. 2021).
- [21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, et al. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 18661–18673.
- [22] Hanna Köpcke, Andreas Thor, and Erhard Rahm. 2010. Evaluation of Entity Resolution Approaches on Real-World Match Problems. *Proceedings of the VLDB Endowment* 3, 1-2 (Sept. 2010), 484–493.
- [23] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep Entity Matching with Pre-Trained Language Models. *Proceedings of the VLDB Endowment* 14, 1 (2020), 50–60.
- [24] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, et al. 2022. What Makes Good In-Context Examples for GPT-3?. In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Association for Computational Linguistics, 100–114.
- [25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, et al. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9, Article 195 (2023), 35 pages.
- [26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]* (2019).
- [27] Michael Losler, Ioannis Koumarelas, and Felix Naumann. 2021. Knowledge Transfer for Entity Resolution with Siamese Neural Networks. *Journal of Data and Information Quality* 13, 1 (Jan. 2021), 2:1–2:25.
- [28] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, et al. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data*. 19–34.
- [29] Navapat Nananukul, Khanin Sisaengsuwanchai, and Mayank Kejriwal. 2024. Cost-Efficient Prompt Engineering for Unsupervised Entity Resolution in the Product Matching Domain. *Discover Artificial Intelligence* 4, 1 (2024), 56.
- [30] Avaniika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* 16, 4 (2022), 738–746.
- [31] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. 2017. A Survey of Current Link Discovery Frameworks. *Semantic Web* 8, 3 (jan 2017), 419–436.
- [32] Matteo Paganelli, Francesco Del Buono, Andrea Baraldi, and Francesco Guerra. 2022. Analyzing How BERT Performs Entity Matching. *Proceedings of the VLDB Endowment* 15, 8 (June 2022), 1726–1738.
- [33] Ralph Peeters and Christian Bizer. 2021. Dual-Objective Fine-Tuning of BERT for Entity Matching. *Proceedings of the VLDB Endowment* 14, 10 (2021), 1913–1921.
- [34] Ralph Peeters and Christian Bizer. 2022. Supervised Contrastive Learning for Product Matching. In *Companion Proceedings of the Web Conference 2022*. 248–251.
- [35] Ralph Peeters and Christian Bizer. 2023. Using ChatGPT for Entity Matching. In *New Trends in Database and Information Systems (Communications in Computer and Information Science)*. Springer Nature Switzerland, Cham, 221–230.
- [36] Ralph Peeters, Reng Chiz Der, and Christian Bizer. 2024. WDC Products: A Multi-Dimensional Entity Matching Benchmark. In *Proceedings of the 27th International Conference on Extending Database Technology, Paestum, Italy, March 25 - March 28, 2024*. 22–33.
- [37] Kashif Shah, Selcuk Kopru, and Jean David Ruvini. 2018. Neural Network Based Extreme Classification and Similarity Models for Product Matching. In *Proceedings of the 2018 Conference of the Association for Computational Linguistics, Volume 3*. 8–15.
- [38] Aaron Steiner, Ralph Peeters, and Christian Bizer. 2024. Fine-tuning Large Language Models for Entity Matching. *arXiv:cs.CL/2409.08185*
- [39] Mohamed Trabelsi, Jeff Hefflin, and Jin Cao. 2022. DAME: Domain Adaptation for Matching Entities. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1016–1024.
- [40] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Chengliang Chai, et al. 2022. Domain Adaptation for Deep Entity Resolution. In *Proceedings of the 2022 International Conference on Management of Data*. 443–457.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.
- [42] Jin Wang, Yuliang Li, and Wataru Hirota. 2021. Machamp: A Generalized Entity Matching Benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4633–4642.
- [43] Pengfei Wang, Xiaocan Zeng, Lu Chen, Fan Ye, Yuren Mao, et al. 2022. PromptEM: Prompt-tuning for Low-resource Generalized Entity Matching. *Proceedings of the VLDB Endowment* 16, 2 (2022), 369–378.
- [44] Runhui Wang, Yuliang Li, and Jin Wang. 2023. Sudowoodo: Contrastive Self-supervised Learning for Multi-purpose Data Integration and Preparation. In *2023 IEEE 39th International Conference on Data Engineering*. 1502–1515.
- [45] Tianshu Wang, Hongyu Lin, Xiaoyang Chen, Xianpei Han, Hao Wang, et al. 2024. Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching. *arXiv preprint arXiv:2405.16884* (2024).
- [46] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).
- [47] Dezhong Yao, Yuhong Gu, Gao Cong, Hai Jin, and Xinqiao Lv. 2022. Entity Resolution with Hierarchical Graph Attention Networks. In *Proceedings of the 2022 International Conference on Management of Data*. 429–442.
- [48] Alexandros Zeakis, George Papadakis, Dimitrios Skoutas, and Manolis Koubarakis. 2023. Pre-trained embeddings for entity resolution: An experimental analysis. *Proceedings of the VLDB Endowment* 16, 9 (2023), 2225–2238.
- [49] Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. 2024. Jellyfish: Instruction-Tuning Local Large Language Models for Data Preprocessing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [50] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, et al. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023).
- [51] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*. 12697–12706.