BPID: A Benchmark for Personal Identity Deduplication

Runhui Wang[*], **Yefan Tao**[*], **Adit Krishnan**[*], **Chris Kong**[*], **Xuanqing Liu**, **Yuqian Deng**, **Yunzhao Yang**, **Henrik Johnson**, **Andrew Borthwick**, **Shobhit Gupta**, **Aditi Sinha**, **Davor Golac**

Amazon Web Services Inc.

{runhuiw,tayefan,aditkris,luyankon,xuanqing,yuqiand}@amazon.com

{yyunzhao, mauritz,andborth,sgg,aditisg,dgolac}@amazon.com

Abstract

Data deduplication is a critical task in data management and mining, focused on consolidating duplicate records that refer to the same entity. Personally Identifiable Information (PII) is a critical class of data for deduplication across various industries. Consumer data, stored and generated through various engagement channels, is crucial for marketers, agencies, and publishers. However, a major challenge to PII data deduplication is the lack of open-source benchmark datasets due to stringent privacy concerns, which hinders the research, development, and evaluation of robust solutions.

This paper addresses this critical lack of PII deduplication benchmarks by introducing the first open-source, high-quality dataset for this task. We provide two datasets: one with 1,000,000 unlabeled synthetic PII profiles and a subset of 10,000 pairs curated and labeled by trained annotators as matches or non-matches. Our datasets contain synthetic profiles built from publicly available sources that do not represent any real individuals, thus ensuring privacy and ethical compliance. We provide several challenging data variations to evaluate the effectiveness of various deduplication techniques, including traditional supervised methods, deep-learning approaches, and large language models (LLMs). Our work aims to set a new standard for PII deduplication, paving the way for more accurate and secure solutions. We share our data publicly at this link [1].

1 Introduction

Data deduplication is a field of study dedicated to removing duplicate records that belong to the same entity, and is an essential problem in natural language processing (NLP) and data mining (Rajaraman and Ullman, 2011; Getoor and Machanavajjhala, 2012a; Konda et al., 2016a). For instance, Grammarly's plagiarism checker detects plagiarism from billions of web pages and academic databases; Google News identifies all versions of the same news article from different sources for comprehensive coverage; and Amazon Web Services (AWS) has an Identity Resolution service for linking customer identifiers from various sources into a unified customer profile.

Personal identifiable information (PII) encompasses a wide range of data, including names, addresses, email addresses, social security numbers, and more, which can uniquely identify individuals. Handling PII presents unique challenges in data deduplication. Similar-looking strings in names, addresses, and other personal details sometimes represent different individuals and often require deep semantic understanding to accurately determine whether two records refer to the same individual. Simple string matching techniques are insufficient, as variations in spelling, abbreviations, and typographical errors can lead to incorrect conclusions. Accurate PII data deduplication requires sophisticated algorithms to understand and interpret these nuances.

The significance of accurate Personally Identifiable Information (PII) data deduplication is evident in its influence across government agencies and industries. Consumer records, as well as financial, criminal or property records are generated in various applications and engagement channels at a rapid pace (Wu et al., 2022a). Unifying and mapping this data enables agencies to identify individuals across different channels and personalize advertising and marketing campaigns. Traditionally, consumer records were unified using third-party cookies and device IDs. However, with the increasing deprecation of third-party cookies and device IDs to enhance consumer privacy, marketers and publishers must develop new consumer identity resolution capabilities. Organizations often invest considerable time creating customized solutions that link consumer identifiers, such as names,

[*] Equal contribution

[1] https://zenodo.org/records/13932202

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 538–546

emails, and phone numbers. These solutions are not only expensive to develop, but they also require continuous maintenance due to the diversity of consumer data. In the absence of a diverse benchmark dataset for testing, both rule-based and machine learning-based methods are likely to commit errors on real-world consumer data.

Despite the significant progress in entity resolution (ER) technologies, a major challenge persists in the lack of open-source benchmark datasets for PII entity resolution. While numerous existing datasets are available for non-PII entity resolution, such as products, academic papers, and music (Primpeli and Bizer, 2020a), no equivalent datasets exist for PII due to privacy concerns. Industry providers have been reluctant to release testing datasets, creating a barrier to the development and benchmarking of robust ER solutions. This gap restricts the ability to perform reliable evaluations and comparisons, which are crucial to advance the state-of-the-art in this field.

To address this lack of benchmarking data, we leverage publicly available data sources (which are free to share and distribute) to construct the first open-source, high-quality benchmark dataset for the personal identity resolution problem. We build upon these sources to generate diverse and challenging testing examples that include both matching and non-matching pairs of profiles. Our dataset comprises synthetic data that is curated by trained annotators to capture diverse potential PII data variations, and does not represent any real individual. This design ensures that our benchmark dataset challenges identity resolution services that rely solely on pre-existing consumer databases or simplistic heuristics.

In our benchmark release, we provide two datasets for evaluation. The first blocking dataset includes 1,000,000 synthetic personal profiles without labels connecting duplicate identities, offering a broad testing ground for various ER techniques. The second matching dataset is a subset of the first, containing 10,000 pairs of identities, each labeled as either a match or a no-match, to facilitate more detailed and supervised testing. We aim to set a new standard for PII entity resolution by introducing this benchmark dataset. This initiative is crucial for advancing ER technologies and ensuring they can meet the growing demands for accuracy, privacy, and security in handling personal identity information.

The rest of the paper is structured as follows: In Section 2, we will discuss related work in PII deduplication. Section 3 describes how we generate the datasets, including the data sources and the creation process for synthetic personal profiles. In Section 4, we evaluate various algorithms on this dataset, including traditional methods, deep learning approaches, and LLMs, and find that our dataset proves challenging even for cutting-edge methods.

2 Related Work

Personal identifiable information (PII) is a sensitive topic in real-world machine-learning applications due to legal, ethical, and regulatory restrictions. Data privacy and artifical/synthetic data creation are important topics in this context (Sei et al., 2022; Qinl et al., 2022). There are numerous advantages to synthetic PII data; First, they can be shared without privacy constraints, and second, their volume and characteristics can be controlled to diversify data variations, and accurately evaluate large scale systems (Christen and Pudjijono, 2009). Previous work has introduced corruption, noise, and distributional changes to synthetic PII data (Christen and Vatsalan, 2013) to test the robustness of machine-learning solutions.

In this work, we are particularly interested in evaluating deduplication for PII, broadly referred as Entity Resolution (ER) , the process of consolidating records that represent the same real-world entity (Getoor and Machanavajjhala, 2012b; Konda et al., 2016b). ER typically consists of two primary phases: blocking and matching. The blocking phase generates candidate pairs of entities, and the matching phase provides a final match/no-match decision for each candidate pair. Our benchmark provides two datasets, one to evaluate each of these two phases.

A considerable body of work has proposed deep-learning techniques for the matching phase (Kasai et al., 2019; Peeters et al., 2020; Li et al., 2021; Miao et al., 2021; Akbarian Rastaghi et al., 2022; Yao et al., 2022). Recent work proposes contrastive learning methods and/or labeled data for BERT-based models in ER tasks (Li et al., 2021; Wang et al., 2022; Peeters and Bizer, 2022), as well as cutting edge large language models (LLMs) (Peeters and Bizer, 2023a). We evaluate a representative set of methods in our experiments (section 4), and show that our dataset proves challenging even to

state-of-the-art methods.

However, we note that the blocking phase is critical to reduce the computational load of the matching system, since the number of candidate pairs potentially grows as the square of the dataset size. More recently, (Papadakis et al., 2023) and (Zeakis et al., 2023) have benchmarked blocking workflows. The challenge of the blocking phase is to achieve a minimal candidate set to reduce computation while maximizing the identification of true matches. To better evaluate blocking methods, we also provide a large blocking dataset of one million records. An ideal blocking method can correctly identify all matching pairs in our matching dataset, while reducing the overall candidate count.

3 BPID Dataset Construction

In this section, we describe our generation approach for our benchmark dataset. We include five universal personally identifiable attributes that are common to industry consumer records (Wu et al., 2022b) as well as governmental records - the name, any physical addresses, email addresses, phone numbers associated with the individual, and their date of birth, to match or not-match a pair of personal profiles. We first collect raw values for each of these five attributes from the below sources:

- **Name** : We generated artificial names by combining first names from the SSA popular baby names dataset[2] and the Census Bureau popular surnames dataset[3].
- **Physical Addresses** : We randomly choose physical addresses located in the United States from the USDOT National Address Database[4].
- **Email Addresses** : We generate realistic email addresses by combining parts of the names with additional keywords or numerical strings, and a randomly chosen domain name.
- **Phone Numbers** : We combine country or area codes with randomly generated phone numbers.
- **Date-of-Birth** : We select random dates-of-birth ranging from 1900 to 2024.

Real-world personal profile data is often incomplete. Typically, 20% or more attribute values are unavailable (Sei et al., 2022), which significantly impacts PII data deduplication efforts. To make our dataset representative of real-world usecases, we randomly set 20% of the attribute values to empty strings.

3.1 Synthetic Profile Construction

We generate synthetic individual profiles by combining randomly chosen values of the name, physical addresses, phone numbers, date-of-birth, and email addresses that exhibit similarities to the chosen name. We provide a synthetic sample profile constructed in this manner below:

```
SYNTHETIC PERSONAL PROFILE
    "fullname": "harold stickelman",
    "phonenumbers": ["9516784827", "9095194618"],
    "emailaddresses": ["stickelman2@verizon.net"],
    "addresses": ["4 Via Camp Comurieta CA 92562"],
    "birthdate": "1990-11-14"
```

We then select and manually generate modified versions of ten thousand of these profiles to construct the matching dataset, and ask human annotators to judge whether these ten thousand original and modified profile pairs represent the same individual, or two different identities. We detail the modification and annotation process for the matching dataset in the following subsections.

3.2 Profile Modification by Trained Human Annotators

Consider the following synthetic profile, which we provide to our human annotators,

```
ORIGINAL PERSONAL PROFILE PROVIDED TO ANNOTATORS
    "fullname": "harold stickelman",
    "phonenumbers": ["9516784827", "9095194618"],
    "emailaddresses": ["stickelman2@verizon.net"],
    "addresses": ["4 Via Camp Comurieta CA 92562"],
    "birthdate": "1990-11-14"
```

We instruct our annotators to introduce variations in one or more attribute values in the above personal profile. Annotators are permitted to modify values, insert new values, or delete existing values of each of the five attributes, while maintaining similarities to the original profile to optimize sample difficulty. Some sample variations generated by our annotators are as follows:

```
POSITIVE MODIFIED PROFILE GENERATED BY ANNOTATORS
(UNANIMOUS MATCH TO ORIGINAL)
    "fullname": "h stickel man",
    "phonenumbers": [],
    "emailaddresses": ["stickelman2@verizon.net"],
    "addresses": [],
    "birthdate": "1990 Nov"

NEGATIVE MODIFIED PROFILE GENERATED BY ANNOTATORS
(UNANIMOUS NOT A MATCH TO ORIGINAL)
    "fullname": "harriet m stickelman",
    "phonenumbers": ["9516784827", "9095194618"],
    "emailaddresses": ["stickelman2@verizon.net"],
    "addresses": ["4 Via Camp Comurieta CA 92562"],
    "birthdate": "1993 Jun 19"
```

[2]https://www.ssa.gov/oact/babynames/

[3]https://www.census.gov/topics/population/genealogy/data.html

[4]https://www.transportation.gov/gis/national-address-database

The above process enables us to generate challenging pairs of matching and non-matching personal profile pairs to test the accuracy of an identity deduplication system.

3.3 Automated Profile Modifications

This section describes the programmatic modification and variations of raw attribute values introduced by us in the benchmark dataset, in addition to the human-generated modifications.

3.3.1 Positive Name Variations

A name variant is an alternative of a name that is considered to be equivalent to that name but which differs from the name in its particular external form. In other words, the two names are considered somehow equivalent and can be substituted for the other in most cases. Name variants occur for many reasons including spelling variations (e.g., Geoff and Jeff), nicknames (e.g., Bill for William), abbreviations (e.g., GPE for Guadalupe), cognates or translations (e.g., Peter for Pierre), cultural differences (e.g., Michael in English vs Michel in French), abbreviations and ordering (e.g., JPR Shields from Roberts Pierre John Shields) and common typographical errors (e.g., Chad vs. Cjad).

3.3.2 Negative Name Variants

This includes names that look similar at first glance, but are likely to refer to different individuals. For instance, "Jon" is often a short form of "Jonathan" while "John" is a standalone name. Despite their similar appearance and pronunciation, they are distinct. Similarly, "Marc" and "Mark" are both given names typically pronounced the same way, potentially referring to different people. Gender variations are male-female name pairs that share letters or phonetic sounds, making them appear similar, while clearly referring to individuals of different genders. For instance, Daniel (male) and Danielle (female) share a base ("Dan"), Jon (male), and Jen (female) are typographically similar, and Paula (female) is the feminine form of Paul (male).

3.3.3 Variant Generation

To generate positive and negative name variations, we select similar candidate pairs of first names (e.g., *Mary* vs *Mark*) from our raw name values and ask three annotators to judge each pair as a match, no-match, or ambiguous. We then include name pairs where all annotators agree to a match or no-match decision in our benchmark dataset. Note that matching pairs that include a positive name pairing (e.g., Larry vs. Lawrence) can be matched by the annotators depending on the other attribute values, while those with a negative pairing (e.g., *Mary* vs. *Mark*) will not be matched since they denote different individuals.

3.3.4 Physical Address Variations

Our benchmark encompasses format variations to the address, replacement, or deletion of different parts of each attribute (e.g., zip code, street address, city, or state in the address attribute), introduction of contradicting information (e.g., 101 Lincoln Ave Chicago IL vs. 101 Lincoln Ave Seattle WA), and semantic variations such as one hundred and fourth vs. 104th (positive) or Lombard Ave vs. Lombard Avenue (positive) or Lincoln Street North vs. Lincoln Street South (negative).

3.3.5 Date-of-Birth, Phone Number and Email Variations

Our benchmark contains format variations and incomplete or partial dates. Analogously, we introduce variations to the phone numbers and email addresses, where we drop or retain parts of the entry such that the altered version still indicate the same underlying value.

3.4 Match/No-Match Annotation Process

We perform the pairwise match/no-match annotation process as follows. We first choose a raw personal profile, and a modified version that is generated by combining the approaches described in section 3.2 and section 3.3. We then present both the original and the modified versions of the profile to three trained human annotators. Each annotator is asked to fill in the below details.

1. Evaluate the extent of match between each of the attribute values in the original and modified profile. Use the neutral assessment to indicate cases where there is insufficient information to make a match/no-match decision.
    - Name - match / neutral / no-match
    - Email - match / neutral / no-match
    - Phone - match / neutral / no-match
    - Address - match / neutral / no-match
    - DOB - match / neutral / no-match
2. Provide an overall assessment between the original and modified profiles - match / neutral / no-match.

	Profile 1	Profile 2
Name	Lucian Duke Long	Duke Lucien
Email	[]	['dukelucien@company.xyz','dukeslucien@academic.edu']
Phone	['1250649013']	['125 064 1924']
Addr	['TX 76693 1876']	['TX 76693 1876']
DOB	1972-06-02	1976-12-29

Table 1: A False Positive by Sudowoodo because of high string similarity on the profile level.

	Profile 1	Profile 2
Name	Stern Concetta	Stern Salisbury Concetta
Email	['stern.concetta@personalmail.org', 'sternconcetta@personalmail.org']	['sconcetta@govtportal.gov', 'salisburyconcetta@mywork.biz']
Phone	['6467364713', '8210541872']	['445 601 4713']
Addr	[]	[]
DOB	26-06-1964	jun 16tue 1964

Table 2: A False Positive by Claude3-Sonnet because of high string similarity on the profile level.

We asked three different annotators to judge if each modified profile with the variations should be matched to the original profile, or be considered a different individual. We obtained an agreement rate of 83%, and only included the pairs where all three annotators unanimously decided a match or a no-match overall assessment. We excluded pairs that received inconsistent assessments from the annotators.

3.5 Benchmark Dataset Statistics

Our final benchmark consists of two datasets. Our blocking dataset contains one million synthetic profiles including both, the raw profiles constructed in Section 3.1 as well as the augmented profiles from Sections 3.2, 3.3. The blocking dataset has an missing rate of 17.8% over all five attributes. Names and dates-of-birth are present in 82.5%, 83.4% of the profiles respectively. We observe the average number of phone numbers, addresses, and emails per profile to be 1.2, 1.3, and 1.2 respectively.

Second, our matching dataset contains ten thousand pairs of personal profiles from Sections 3.2, 3.3 marked as matches or no-matches as detailed in Section 3.4. We also include every profile in the matching dataset as part of the blocking dataset to evaluate blocking methods. A perfect blocking method should correctly identify every pair in our matching dataset from the blocking dataset (i.e., have a high recall), while a perfect matching method should correctly classify each pair in the matching dataset as a match or no-match (i.e., have a high accuracy). The matching dataset contains ten thousand pairs of profiles with an attribute missing rate of 17.5%, names and dates-of-birth present in 82.9%, 83.5% of the profiles, and an average of 1.2 phone numbers, 1.3 addresses, and 1.2 emails per profile. Our annotators unanimously judged 4333/10000 pairs as matches and 5667/10000 as no-match pairs.

4 Evaluations

In this section, we provide results for both the blocking dataset (which contains one million unannotated profiles) and the matching dataset, which contains ten-thousand profile pairs with match/no-match human annotations. Note that the set of profiles in the matching set is a subset of the blocking set. We evaluate state-of-the-art entity matching and blocking methods over our benchmark datasets. Our matching methods include a traditional supervised **Random Forest** (Primpeli and Bizer, 2020b), pre-trained language model based methods, **Ditto** (Li et al., 2021) and **Sudowoodo** (Wang et al., 2022), and **LLM-based methods** (Peeters and Bizer, 2023b). We designed a LLM prompt (provided in appendix A) to determine if two profiles represent the same individual, and applied the prompt to various cutting-edge LLMs including Anthropic model **Claude3-Sonnet** [5], Meta AI **Llama3-70B-instruct** (Meta, 2024), Mistral AI **Mistral large** [6], and OpenAI **GPT4 turbo** (Achiam et al., 2023) [7]. We use the F1 score on the annotated matching dataset as our evaluation metric.

We selected **Sudowoodo** (Wang et al., 2022), **NLSHBlock** (Wang et al., 2024), **Sparkly** (Paulsen et al., 2023), and **Contriever** (Izacard et al., 2021)

[5]https://www.anthropic.com/news/claude-3-family

[6]https://mistral.ai/news/mistral-large/

[7]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

	Profile 1	Profile 2
Name	Burnette Joyce	Herbert Burnette
Email	['burnette_joyce@business.net']	['herbert_b@email-service.io', 'burnette_joyce@personal-email.net']
Phone	['017 769 0655']	[]
Addr	['10384 65th Avenue Northwest Montana 2380361']	['328 Kirkwood Circle Penn Yan DE 15842', '10384 65th Avenue Northwest 57 Mc Crory MT 2380361', '2523 cr 124 Greig CA 11590']
DOB	24 april	19910424

Table 3: A False Positive by both the best LLM and PLM-based method.

for the blocking task. We measure the recall (the percentage of the matched pairs in the matching dataset that be retrieved from the blocking dataset) and the candidate set size generated by each method. Note that a higher recall and lower candidate size are preferred for blocking methods. These methods cover pre-trained language model based solutions, traditional TF-IDF based solution, and dense information retrieval solutions.

We use an AWS EC2 P4d instance in our experiments for blocking and matching.

4.1 Entity Matching results

Table 5 shows the F1 scores of various methods on the matching set. The results indicate that our dataset proves challenging for methods of all categories - even highly capable LLMs do not achieve a satisfactory F1 score with the prompt in Appendix A on our benchmark dataset.

To verify the quality of the dataset, we report the performance of the best Random Forest that we trained using a series of features including various string similarity metrics for each attribute. As shown in Table 5, the Random Forest achieves a 0.608 F1 score, which indicates that rule-based decisions are insufficient in our dataset.

We also note that the Sudowoodo method is the best non-LLM method and Claude3-Sonnet is the best LLM method based on our evaluation. To understand the challenge of matching PII profiles, we conduct a case study. Table 3 shows a false positive match from Sudowoodo. The names are very similar, and the addresses belong to the same area. The phone numbers are very similar, but the last four digits are completely different. In summary, these two profiles are likely to belong to different people in the same area, but Sudowoodo likely matches them because of the high similarity on the profile level.

Table 2 shows a false positive match from the LLM that is the best performing, Claude3-Sonnet. The names can be considered as the same and the email addresses are also similar due to the same name. However, their phone numbers are different and they have different birthdates, which clearly indicates that the profiles represent different people who share the same name.

We note that there are other challenges in this problem, including properly handling name variations, missing values for some attributes, and multiple values for emails, phone numbers, and addresses. Especially for multiple values in one attribute, usually when two lists have a common element, we need to consider them as a matching attribute even if all other elements are different. We also notice that due to the nature of tokenization in LLMs, they struggle to correctly quantify the number of different digits for two phones. This challenge limits LLMs' ability to ignore very minor typos in phone numbers.

Method	Recall	CSS	Blocking time
Sudowoodo	<u>0.682</u>	10M	6min
NLSHBlock	0.612	10M	13min
Sparkly	0.629	10M	4min
Contriever	**0.711**	10M	33min

Table 4: Recalls, Candidate Set Sizes (CSS), and blocking runtimes for each method. (M=10^6)

Method 1	Precision	Recall	F1
Random Forest	0.653	0.609	0.629
Ditto	0.746	**0.804**	<u>0.752</u>
Sudowoodo	0.774	<u>0.802</u>	**0.788**
Claude3-Sonnet	0.660	0.656	0.658
Llama3-70B-instruct	0.707	0.753	0.729
Mistral large	**0.784**	0.491	0.604
GPT-4-turbo	<u>0.780</u>	0.613	0.687

Table 5: F1 scores on evaluated matching methods. Bold font indicates the best performance and the underline indicates runner-up.

4.2 Blocking results

Table 4 shows the blocking performance of the evaluated methods on three metrics, recall, candidate set size (CSS), and compute time to retrieve

Borderline Example 1	**Profile 1**	**Profile 2**
Name	Sarah Webber	S
Email	['s_adams@gov.us']	['susan.j@govt.gov','susan@123.com', 's_adams@email.org']
Phone	['4923544915', '4923542364']	['88 4923342364']
Addr	['5665 Encino Cove 324 Cape Fair TX 75119', 'Berclair TX 75109']	['Navajo Dam New Jersey']
DOB	1	1972-1-20
Borderline Example 2	**Profile 1**	**Profile 2**
Name	Greenwood F	Greenwood Francesca
Email	['gw1993@personal.info']	['gw1993@academic.org']
Phone	['0920435370']	['1027531458', '57 455 043 1113']
Addr	[]	['Oakbrook 7196', 'Truscott TX 76626']
DOB	27 september	1973-09-27
Borderline Example 3	**Profile 1**	**Profile 2**
Name	B Esquivel Esquivel	Esquivel Brendon
Email	['bsmith123@university.edu', 'b.s.m.i.t.h@govt.gov', 'bsmith_professional@email-service.co.uk']	['jesquivel@myschool.edu']
Phone	[]	['3864445702', '8241613926']
Addr	['TX 76693 Carolina 105 5720 Private Road 64106', '811 Olympic Drive 64106 TX 76693 Carolina']	[W 17 Dr Montana USA]
DOB	1971-02	february 1971

Table 6: Annotators disagreed on these three examples. For the first example, the addresses belong to different states, and while the phone numbers indicate a match, none of the other attributes provide a strong connection. For the second example, none of the other attributes provide a strong connection. The name appears fairly common, and the date of birth is not conclusive. In the third example, the name and DOB have some similarities, but the email ids indicate that the two individuals may not be the same, and the addresses indicate different states. We also note some applications may prefer to match these cases depending on the precision and recall requirements, or other prior knowledge about their specific data sources.

the candidates. We note that none of the evaluated methods achieves over 80% recall at a reasonable CSS, which indicates it is challenging to solve PII profile deduplication.

4.3 Borderline Cases

In this section, we list some examples in which the annotators did not agree on a conclusion (Table 6). These examples provide insights into annotator considerations for matching or not matching a specific pair of profiles. In Example 1, Table 6, the annotators disagreed, since the phone numbers indicate a match, but the addresses belong to different states, and the other attributes do not provide a strong connection. We note that some applications may prefer to match this case, depending on their precision and recall requirements or data sources.

In Example 2, Table 6, two annotators preferred to match, but the third annotator noted that none of the attributes provides a strong connection. The name is fairly common, and the date of birth is inconclusive. Similarly, in Example 3, the name and DOB have similarities, however the email ids and addresses indicate a mismatch. We instruct our annotators to maintain consistency in their judgments to avoid contradicting conclusions. However, we note that some ambiguous cases may be present in our matching dataset despite our best efforts. Label judgments should be made on a case by case basis for ambiguous examples depending on the target application requirements.

5 Conclusion

In this paper, we introduce the first fully public benchmark dataset to facilitate the evaluation of data deduplication methods for personally identifiable information (PII). Our dataset is meticulously designed to provide a rigorous and challenging testbed, surpassing the limitations of simplistic rules or heuristic techniques. Even state-of-the-art large language models (LLMs) exhibit non-trivial error rates on our dataset, underscoring the complexity of the task and setting a high bar for evaluation. Through this benchmark, we aim to foster advancements in PII data deduplication, promoting the development of innovative methods that prioritize privacy and data security while also enabling effective data management and analysis.

6 Ethical Considerations

We acknowledge that Personal Identity Deduplication is a sensitive task because of the potential

involvement of personally identifiable information (PII) to specific individuals or consumers. We note that our benchmark is entirely synthetic. The profiles constructed in our dataset do not represent any real-world individuals, since they are fictional combinations of random attribute values. In our profile modification process by trained human annotators, the annotators do not have access to any PII data representing real individuals. Therefore, our benchmark dataset does not leak any real personal information. Further, our benchmark enables the safe comparison of deduplication services and methods in future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mehdi Akbarian Rastaghi, Ehsan Kamalloo, 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*, pages 3786–3790.

Peter Christen and Agus Pudjijono. 2009. Accurate synthetic generation of realistic personal information. In *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13*, pages 507–514. Springer.

Peter Christen and Dinusha Vatsalan. 2013. Flexible and extensible generation and corruption of personal data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1165–1168.

Lise Getoor and Ashwin Machanavajjhala. 2012a. Entity resolution: Theory, practice & open challenges. *PVLDB*, 5(12):2018–2019.

Lise Getoor and Ashwin Machanavajjhala. 2012b. Entity resolution: theory, practice & open challenges. volume 5, pages 2018–2019. VLDB Endowment.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *ACL*, pages 5851–5861.

Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, and et al. 2016a. Magellan: Toward building entity matching management systems. *PVLDB*, 9(12):1197–1208.

Pradap Konda, Sanjib Das, Paul Suganthan G. C., AnHai Doan, and et al. 2016b. Magellan: Toward building entity matching management systems. volume 9, pages 1197–1208.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2021. Deep entity matching with pre-trained language models. *PVLDB*, 14(1):50–60.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Zhengjie Miao, Yuliang Li, and Xiaolan Wang. 2021. Rotom: A meta-learned data augmentation framework for entity matching, data cleaning, text classification, and beyond. In *SIGMOD*, pages 1303–1316.

George Papadakis, Marco Fisichella, Franziska Schoger, George Mandilaras, Nikolaus Augsten, and Wolfgang Nejdl. 2023. Benchmarking filtering techniques for entity resolution. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 653–666. IEEE.

Derek Paulsen, Yash Govind, and AnHai Doan. 2023. Sparkly: A simple yet surprisingly strong tf/idf blocker for entity matching. *Proceedings of the VLDB Endowment*, 16(6):1507–1519.

Ralph Peeters and Christian Bizer. 2022. Supervised contrastive learning for product matching. *arXiv preprint arXiv:2202.02098*.

Ralph Peeters and Christian Bizer. 2023a. Using chatgpt for entity matching. In *European Conference on Advances in Databases and Information Systems*, pages 221–230. Springer.

Ralph Peeters and Christian Bizer. 2023b. Using chatgpt for entity matching. In *European Conference on Advances in Databases and Information Systems*, pages 221–230. Springer.

Ralph Peeters, Christian Bizer, and Goran Glavas. 2020. Intermediate training of BERT for product matching. In *DI2KG@VLDB*.

Anna Primpeli and Christian Bizer. 2020a. Comperbench: A collection of 21 complete benchmark tasks for entity matching.

Anna Primpeli and Christian Bizer. 2020b. Profiling entity matching benchmark tasks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3101–3108.

Xuedi Qinl, Chengliang Chai, Nan Tang, Jian Li, Yuyu Luo, Guoliang Li, and Yaoyu Zhu. 2022. Synthesizing privacy preserving entity resolution datasets. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2359–2371. IEEE.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

Yuichi Sei, J Andrew Onesimu, Hiroshi Okumura, and Akihiko Ohsuga. 2022. Privacy-preserving collaborative data collection and analysis with many missing values. *IEEE Transactions on Dependable and Secure Computing*, 20(3):2158–2173.

Runhui Wang, Luyang Kong, Yefan Tao, Andrew Borthwick, Davor Golac, Henrik Johnson, Shadie Hijazi, Dong Deng, and Yongfeng Zhang. 2024. Neural locality sensitive hashing for entity blocking. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 887–895. SIAM.

Runhui Wang, Yuliang Li, and Jin Wang. 2022. Sudowoodo: Contrastive self-supervised learning for multi-purpose data integration and preparation. *arXiv preprint arXiv:2207.04122*.

Ningning Wu, Robinson Tamilselvan, and Talha Tayyab. 2022a. A study on personal identifiable information exposure on the internet. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 813–818. IEEE.

Ningning Wu, Robinson Tamilselvan, and Talha Tayyab. 2022b. A study on personal identifiable information exposure on the internet. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 813–818. IEEE.

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*, pages 429–442.

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):2225–2238.

A Appendix

We use the following prompt to evaluate LLMs.

`Instruction` Task description: You are a profile annotator and you need to evaluate the similarity between two profiles with the following guideline. Each profile consists of 5 different attributes: *phone*, *email*, *fullname*, *addresses*, *birthdate*. You need to carefully compare each attribute and make the match / no-match decision on the given profile pair.
Keep the these principles in mind when making a decision.
`Principle 1`. Allow slight string variations of "common sense"/"human error", including upper/lower case, swapped positions of words in a string, absence/errors of country code, simple typos, date format, string synonyms. However, if two names have different first names, consider them as different people.
`Principle 2`. One element match between lists is considered as a match for the attribute: three attributes contain(phonenumbers, emailaddresses, addresses) a list of values, meaning one person can have more than one phonenumber/email/address. If two profiles have any email/phone/address in common, that means a "match" in this attribute. For example: sim([chris.paul@gmail.com, cp3@yahoo.com], [nba_champ123@amazon.com, chris.paul@gmail.com]) = match. This is because both email lists have "chris.paul@gmail.com".
`Principle 3`. Ignore invalid attribute values or values without enough information.
`Principle 4`. Make the decision based on holistic evaluation over all valid attributes. Only match two profiles if there is sufficient evidence.
Here are two personal profiles, please strictly follow the above guideline and use the below template to answer whether they are the same person:

Analysis: [reasons for final decision about whether these two profiles should be matched or not]

Answer: [Yes or No]

Here are the two profiles: [**Profile 1**], [**Profile 2**].
Analyze step by step in plain text.