De-Identification of Privacy Sensitive Information in Resumes with GPT-4: A Utility Analysis for Automated Job Role Classification

Sascha Löbner
Goethe University Frankfurt
sascha.loebner@m-chair.de

Daria Linke
Goethe University Frankfurt

Frédéric Tronnier
Goethe University Frankfurt
frederic.tronnier@m-chair.de

Abstract

As organizations face the challenge of managing large amounts of data, privacy concerns have become increasingly prevalent when sharing sensitive privacy information with machine learning experts. This paper addresses the fundamental issue of privacy-sensitive information de-identification by introducing in-prompt de-identification, an approach that exploits the capabilities of large language models. Existing de-identification techniques often struggle to ensure complete privacy, and methods with higher privacy often result in a loss of data utility. In contrast, in-prompt de-identification is capable of generating synthetic, human-readable data samples from given inputs and bridges the gap between privacy and utility. With this article, we contribute to the de-identification of real-world resume data using in-prompt de-identification based on OpenAI's GPT-4. Notably, our classification model, trained on GPT-4 generated data, shows no significant loss in performance compared to our baseline model trained on the original data.

Keywords: De-identification, privacy sensitive information, resume, generative ai, fairness.

1. Introduction

Handling tremendous amounts of job applications is a tedious task faced by companies around the world. More often than not, outdated or generic methods are used to reduce the applications to a manageable amount. This can deter highly qualified applicants, stuck in lengthy application procedures. But with the rise of AI, a huge variety of new techniques exist that are capable to improve and automate this process of job role classification (Roy et al., 2020). A huge problem for the implementation of such technologies by organizations are privacy concerns towards new AI-based technologies. Since resumes contain manifold privacy-sensitive information (PSI), privacy regulations, e.g., the General Data Protection Regulation (2016) (GDPR) limit the capabilities of those new AI-based approaches by restricting the available datasets and thus the model-performance (Raghavan et al., 2020). The absence of good training data makes it difficult for companies to train and deploy robust models. To create such high-quality data that complies with current regulations, careful consideration of PSI in CV is incremental, as even innocuous details in CVs can cause bias and discrimination in the hiring process (Kang et al., 2016). Common approaches to circumvent this problem are training on anonymized data where the data subject cannot be identified, and synthetically generated data (Bayer et al., 2022) reflecting reality in quality and structure. While synthetic resumes can serve as training data for machine learning models to ensure they are trained on diverse and unbiased examples, anonymized or pseudonymized data is required for automated job role classification tasks. For instance, scrubbing aims to generate text that maintains the original structure and content while excluding any identifiers or presumed sensitive details about the individuals involved (ISO/IEC 20889:2018). However, not all indirect identifiers can be detected using existing methods due to contextual variations, leaving significant risks of data breaches (Liu et al., 2021). Ensuring the utility of de-identified data, for natural language processing (NLP) tasks, requires replacing PSI with suitable surrogates, a process that demands extensive manual efforts (Stubbs et al., 2015).

In this article, we address the issue of de-identifying PSI in resumes by exploiting GPT-4's advanced large language model (LLM) capabilities of PSI-detection

URI: https://hdl.handle.net/10125/108944
978-0-9981331-8-8

(Löbner et al., 2023). In-prompt de-identification is an innovative solution for generating a private twin from a given text, thereby providing privacy without compromising the utility of the generated data. To this end, we pose the following research questions:

RQ1: Can GPT-4 effectively generate synthetic resumes that maintain content quality and structure similar to real resumes?

RQ2: To what extent do the synthetic resumes, generated by GPT-4, ensure privacy, compared to traditional anonymization methods?

RQ3: How do synthetic resumes impact the utility of classification models versus real or anonymized data?

Thus, with this work, we contribute to the field of data generation by exploring GPT-4's synthetic data generation capabilities by analyzing maintained quality and similarity. After eliciting an appropriate prompt, we find that in most cases, the generated de-identified resumes closely resemble the input resumes. We investigate to what extent synthetic resumes generated by GPT-4 ensure privacy compared to traditional anonymization methods. We effectively address privacy concerns by concealing most PSI, but face challenges with more subtle PSI. We analyze the impact of synthetic resumes on the utility of classification models compared to real or anonymized data. The results demonstrate that in-prompt de-identification preserves utility by providing qualified descriptions and generalizations of PSI.

2. Related Literature

In this section, we explore literature on synthetic data generation for privacy and its use in resume classification

Synthetic data generation has been explored in various contexts to balance privacy with data utility. Yue et al. (2023) test synthetic online review generation, using a Yelp dataset and GPT-2, aiming to obtain the same statistical properties as the original data. The work of Mattern et al. (2022) followed a similar approach for generating product reviews without revealing reviewers' personal details, using GPT-2. Their model was fine-tuned with a DP optimizer. Similarly, Al Aziz et al. (2021) applied DP fine-tuning to generate de-identified medical records. Particularly relevant to our work is the study by Skondras et al. (2023), who used GPT-generated synthetic resumes for classification, achieving substantial accuracy. This motivates to research GPT-4's capabilities for de-identification and the generation of usable synthetic data for resume classification without compromising privacy.

Regarding fine-tuning and prompting, advancements in prompt-tuning methods are of particular interest. Kurakin et al. (2023) demonstrated that prompt-tuning with differential policies could significantly improve utility in classification tasks, offering a more parameter-efficient alternative to full model fine-tuning. While White et al. (2023) provides best practices, Liu et al. (2023b) provides suggestions to improve prompt quality, e.g. using prompt paraphrasing.

Due to strict regulations, the generation of synthetic text data has especially been investigated for healthcare applications. Kweon et al. (2023) generated a substantial clinical note dataset, which served as the foundation for training a domain-specific LLM with Llama. Tang et al. (2023) use synthetic data for fine-tuning pre-trained models. They find GPT performing better in generating synthetic health data than in downstream tasks where it lags behind fine-tuned models and poses privacy risks.

The anonymization of personal data remains a critical concern. Weitzenboeck et al. (2022) investigate the compliance of anonymization with GDPR and analyze a risk-based and a strict approach. Pilán et al. (2022) compare text anonymization approaches and named entity recognition (NER) models for PSI detection. They motivate the use of generalization to keep utility high.

For resume classification, many studies have already investigated the performance of machine learning techniques. Zaroor et al. (2017) utilized a matching algorithm to enhance accuracy. Javed et al. (2015) designed a job role classification system for several taxonomies using an SVM-kNN cascade. Pal et al. (2022) explored SVM, Random Forrest and Naive Bayes to improve automated job-role classification. Roy et al. (2020) confirmed the efficacy of using cosine similarity and SVM for this task, achieving an accuracy of 0.78

To the best of our knowledge, privacy implications of using GPT-4 specifically for synthetic data generation and resume classification have not been investigated yet.

3. Methodology

To assess in-prompt de-identification with GPT-4, we examine the privacy-utility trade-off in job classification. We use GPT-4 for its high performance and 32,000-token context window, enabling the handing of long resumes in few-shot learning. To comply with the organization's requirements we were restricted to using their Microsoft Azure hosted GPT-4 and could not test other models.

In *RQ1* we investigate different prompting techniques, crucial for a good model performance (White et al., 2023). Existing strategies differ in sample quantity and structure, from none to several, and simple instructions to complex chains of thoughts (Wei et al., 2022). The exploration for the optimal prompt uses a subset of 20 stratified samples to represent all PSI types. To examine the similarity between generated resumes and raw data, we use the

similarity metric BERTscore (0 to 1). It assesses the evolution across different temperature settings, while a score of 1 indicates semantically identical data (Zhang et al., 2019), a substantial privacy risk. Sampling temperature (0 to 2), can affect model output and privacy. It determines how much the model favors more likely words over less likely words. High temps. produce varied results, while low ones are more deterministic. To avoid privacy/utility outliers, we use temps. 0.5, 1.0, 1.5.

In *RQ2*, assessing privacy preservation within synthetic datasets presents a complex challenge, characterized by the absence of standardized measures (Mattern et al., 2022). Determining the privacy implications of generated data often requires human evaluation (Al Aziz et al., 2021), which is very time-intensive. Thus, we approximate privacy preservation by measuring the similarity of synthetic and original data. This comparison can be approached in two ways: by scrutinizing the data sets' distributions, or by directly examining the synthetic-original pairs. Scrutinizing (Kurakin et al., 2023; Mattern et al., 2022) is not applied because its support at Huggingface (HF) is currently limited to English text. Thus, we decided for a one-to-one comparison using the BERTscore.

Rather than mandating complete dissimilarity between synthetic samples and their originals, we focus on identifying potential privacy violations (Yue et al., 2023). The process involves matching an array of annotated private entities for the original sample against the text of its synthetic twin. Each occurrence is counted separately if the same private entity is repeated multiple times within the generated resumes. This is due to the task's need to identify and annotate all private entities, including repeats. To account for variations in the number and length of annotations, the privacy metric is calculated by dividing the summed length of all matches $M[m]$ in the generated resume by the summed length of all annotations $P[p]$ in the original, with M_k as the array containing all private annotations in sample k and P_k as the array containing all private annotations for sample k:

$$Privacy_k = 1 - \frac{\sum_{m=1}^{|M_k|} length(M[m])}{\sum_{p=1}^{|P_k|} length(P[p])} \ .$$

For *RQ3*, we define utility as the retained usefulness of de-identified data, or the degree of how well generated synthetic data mimics the original data (Tomashchuk et al., 2019; Yue et al., 2023). We compare the utility of GPT-4 generated resumes against our baseline and scrubbing model. Classifying resumes by job roles offers insights into the privacy and utility balance. To evaluate and compare our classification results, we use precision, recall and F1-score (Goodfellow et al., 2016). For robust results, we use 5-fold cross-validation with stratified sampling to maintain label proportions in each fold.

For all *job role classifications*, a pre-trained BERT-based model is chosen for its adaptability and high performance in various domains (García-Pablos et al., 2020). BERT analyzes text bi-directionally which provides an enhanced contextual understanding improving overall classification accuracy. For the training, we use PyTorch for deep learning and HF's resources, including pre-trained models and tools like *AutoModelForSequenceClassification* for BERT fine-tuning. We employ the *gbert-large* model for its efficacy with German texts. Despite BERT's 512 token limit, it remains competitive against full-length models (Park et al., 2022).The Longformer model slightly improves F1-scores by 0.05 but lacks a German version on HF. We mitigate this issue by listing recent jobs and relevant projects first. The hyperparameters are adjusted to minimize errors and overfitting beyond ten epochs (Goodfellow et al., 2016). Small learning rates are more effective for regularisation and performance (Smith, 2018). We use *AdamW* with a decreasing learning rate and weight decay for regularisation, *max length* of 512, *epochs* 10, *batch size* 2, *learning rate* 5×10^{-6} and *scheduler type* linear. Named entity recognition (NER) is the task of detecting and classifying tokens in unstructured text into defined categories, here PSI. To ensure a robust baseline detection of private tokens, it is important to compare the effectiveness in identifying PSI data with NER and GPT-4's in-prompt de-identification method. We use HF's *BertForTokenClassification* with the pre-trained *gbert-large* model for NER tasks. The text is split into 512-token chunks with 20-token overlap for context. We set *epochs* 10, *batch* 1, *learning rate* 5×10^{-6}, *scheduler type* linear, *optimizer* AdamW. To evaluate privacy scores and BERTscores at different temps., generation procedures are viewed as treatments, and each resume is considered a block observation. In this paired data scenario, each resume can be matched with its counterpart in the generated population. Comparing classification results across anonymization levels, the respective folds serve as observations. Since privacy scores are likely to be highly skewed, and the sample size is limited to five folds, non-parametric tests are preferred over parametric tests for their suitability in non-normalized contexts (Pett, 2015). We use the Friedman test, to compare multiple treatments. For significant results (p-value < 0.05), the post-hoc Conover test is used for pairwise comparison.

Compliant and ethical data processing: The following steps were taken to ensure the GDPR-compliant and ethically appropriate processing of

resume data: First, informed consent for the processing of the personal data was collected from the data subjects, by the organization, for research purposes. Second, using de-identified data instead of real data was considered but withdrawn as it would have had an impact on usability of the synthetically generated data. The use of third-party data was initially considered but the processing of such data, although freely available online, remains prohibited, even for research purposes, under the GDPR, and is not in line with our research values. Third, participants could withdraw their data at any time during the analysis. Fourth, the use of personal data was minimized to the creation of synthetic data generation and the baseline model. Fifth, where applicable, direct identifiers were removed, e.g. outside the free text of the resumes, to ensure the highest degree of pseudonymity, even before the actual de-identification. Sixth, all data was processed entirely within the organization, under the organization's restrictions for processing, protected storing and accessing of the data. GPT-4 instances were used only through the organization's Microsoft Azure to keep the data within the organization. BERT models were processed locally. The data processing was further restricted to a minimal number of researchers within the research project to prevent unauthorized secondary use or access. We aim to evaluate synthetic data generation to circumvent ethical and regulatory issues.

Data: The data set comprises a total of 282 resumes collected with informed consent and provided by an organization that will remain anonymous by the agreement between the authors and the organization. The resume data predominantly features individuals with a professional background in IT consulting. The resumes are project-focused and thus more detailed than "regular" resumes, following a standardized structure. The majority of resumes has between 700-1.500 words.

Text length is crucial when choosing a GPT model, due to limitations on token count. Approximately 100 tokens are estimated by OpenAI (2023) to be equivalent to 75 words. Direct and indirect identifiers encompass the following data types (count, percentage): *firm (2880, 38.89 %)* including current and past employers, *university (Uni) (437, 6.07 %)* including different types of knowledge institutes, *location (947, 13.15 %), gender (2594, 36.03 %), birthdate (66, 0.8 %), name (20, 0.2 %)*. The identifiers were elicited from GDPR Art. 4, Art 9, and related literature (Löbner et al., 2023). All remaining entities are collected in the miscellaneous (misc) (315, 4.37 %) category (Trienes et al., 2020).

The resumes contain role suggestions that indicate an individual's suitability for specific projects, from an IT-consulting perspective. These annotations, created by management personnel, are used for the classification tasks. The roles are: *project managers (25 %), scrum masters (18 %), software developers (18 %), business analysts (25 %), testers (0.06 %), software architects (0.06 %), security experts (0.01 %)*, and *UX designers (0.01 %)*. Work experience is labeled into senior (40 %), senior expert (24 %), expert (22 %) and junior (14 %). Gender is split into male 77 %, female 23 %, other 0 %, what is unfortunately not unusual for the tech market.

Necessary adjustments to the dataset included annotating and removing job roles from resumes to prevent model memorization and addressing class imbalance. Especially, the classes of "security expert" and "UX designer" were removed due to a lack of instances. For the remaining six classes, weights were introduced to the loss function to mitigate prediction bias towards frequent classes These weights, based on the distribution of the training set label, were integrated into the *CrossEntropyLoss* function. Since BERT has its own pre-procession layer, additional steps are not necessary.

Scrubbed data was created by generalising all dates to year. Names and locations within Germany are removed. Locations abroad are generalized by country. Given the diverse nature of the misc category, finding substitutes is complex. For gender, several strategies are used. Job titles ending with 'in' (female German identifier) are translated into gender-neutral English, since IT Job titles often use English terms. The applicant's pronouns and salutations are removed. To preserve more contextual information, universities and companies were classified into subgroups. Within the university group, four distinct categories were created to identify the type of university, e.g., University of Applied Sciences. Similarly, the attribute organization was categorized in, i.e. the consulting sector or the public sector. service provider sector, public sector, manufacturing sector, and non-profit organizations. Each sub-category contained more than ten entries. Entities are replaced by a label or substituted in a permutation process that involved shuffling within the subcategory while avoiding self-replacement.

4. Results

RQ1: In-Prompt De-Identification: First, to generate effective synthetic resumes with GPT-4 we have to find an appropriate prompt. Best practices include clear and concise instructions, and adopting a persona (White et al., 2023). We prompted GPT-4 to assume the role of an employee at a data protection officer and provided clear and specific instructions on how to create the synthetic resume. GPT-4 was tasked with solving labor-intensive problems only, as tasks like generalizing dates are easily handled by regular expressions, and thus excluded.

Table 1: Token recognition: GPT-4, PSI detection: NER

	GPT-4			NER		
Label	**Prec.**	**Rec.**	**F1.**	**Prec.**	**Rec.**	**F1**
Firm	0.94	0.96	0.95	0.77	0.83	0.80
Loc	0.76	0.96	0.85	0.72	0.84	0.77
Uni	0.88	0.90	0.89	0.60	0.70	0.65
Gender	0.70	0.47	0.56	0.73	0.89	0.80
Birthdate	0.84	0.92	0.88	1.00	1.00	1.00
Name	0.71	0.94	0.81	0.00	0.00	0.00
Misc	0.02	0.31	0.04	0.66	0.18	0.28
Macro avg	0.69	0.78	0.71	0.64	0.63	0.61
Weighted avg	0.81	0.81	0.80	0.74	0.81	0.77

Using prompt paraphrasing (Liu et al., 2023a) it took several rounds of evaluation to find an optimized prompt. Specific requirements like gender-neutral language were omitted after observing GPT-4's challenges in implementation. Generated resumes were partially gender-neutral, while others remained unchanged. Also, the requirement to substitute a similar industry was challenging due to unknown companies. To improve privacy performance, the emphasis on anonymization tasks in the prompt was increased. Finally, the prompt below produced least privacy leakage. All prompts are in German as, despite explicit requests for German resumes, English prompts resulted in English resumes. For readability, we translated the prompt in English:

Please anonymize the following resume and replace all specific organizations and related terms with more general, non-identifiable synonyms. Also replace all geographical details. Personal details, e.g., date of birth, name, and nationality should also be anonymized.

We also tested fine-tuning and few-shot learning (Liu et al., 2023a) with unsatisfactory results as the synthetic texts still contained the original PSI. To later remove PSI and compare the performance with the NER, we used the paraphrasing technique (Liu et al., 2023a) tasking GPT-4 to retrieve and categorise words into identifier classes:

Can you please extract these categories from the following resume: Firm: current or past employers, clients or public bodies. Uni: educational institutions, Name: Name of the person, Birthdate: the person's birthday, Loc: geographical information, Gender: Gender of the person, terms such as project manager or businessman, Misc: for words or phrases that are specific and unique to a particular company or organization. These can be specific project names, abbreviations or internal identification numbers that serve to uniquely identify the respective company or organisation. Please use this JSON format: "firm": ["org1","org2"..], "uni": [] ...

Retrieved and categorised entities, are compared to ground truth annotations (table 1). GPT-4 does a fair job, identifying tokens related to the *firm*. However, about 10 % of *uni* were not retrieved. Regarding *loc* and *name*, GPT-4 produces many false positives. For *gender* and *misc*, GPT-4 struggles to distinguish between tokens indicating gender and references to organizations, especially with abbreviations or projects. Bad results are due to many false positives, with GPT-4 annotating 5,602 *misc* entities without capturing the correct ones.

Generated Resumes with GPT-4: The generated resumes demonstrate coherence and maintain a high degree of semantic similarity to their original counterparts. With increasing temperature, the generated resumes tend to become shorter (figure 1). This trend may be due to fewer projects being included within the resumes (figure 2). Additionally, projects generated at temp. 1.5 are sometimes included but often lack details like activities or tools. Sometimes, GPT-4 presents a few projects and cuts off all subsequent ones with the note "Rest of CV follows the above format, with anonymized information on projects, clients, and locations".

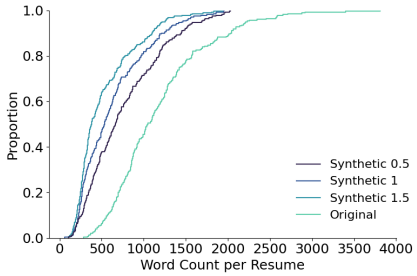


Figure 1: Distribution of counted words of original and synthetic resumes at different temperature levels.

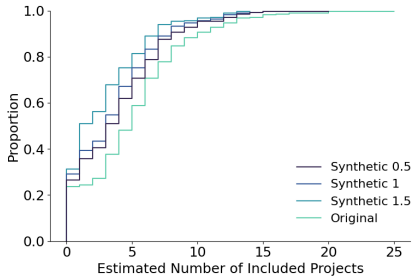


Figure 2: Estimation of included projects based on keyword-count marking the start of project descriptions, e.g., relevant project experience.

Figure 3 illustrates the distribution of BERTscores, assessing the semantic similarity between original resumes and each population of generated data, for different temperature settings. Consistent with the findings on word count and project divergence, the difference between original and synthetic data is lowest at a temperature of 0.5 and highest at 1.5, when GPT-4 becomes more creative. This finding is supported by the statistical significance analysis for the F1-score. The Friedman test indicated a significant difference with a test statistic of 12.523 and a p-value of 0.002. The Conover post-hoc test results showed no significant difference between temp. = 0.5 and temp. = 1.0 ($p = 0.421$), but a significant difference was observed between temp. = 0.5 and temp. = 1.5 ($p < 0.001$), and between temp. = 1.0 and temp. = 1.5 ($p = 0.01$). Analyzing the outliers with the highest divergence shows that the absence of *projects*

Figure 3: Distribution of BERTscore metrics precision, recall and F1-score at temperatures 0.5, 1.0 and 1.5 following (Zhang et al., 2020).

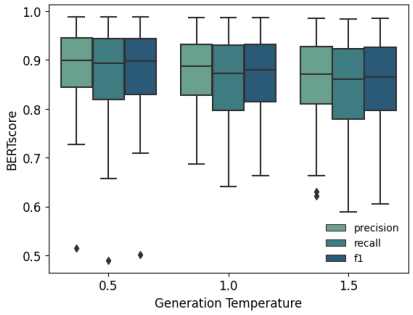
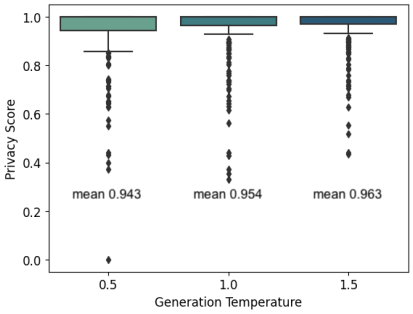


Figure 4: Distribution an mean of privacy scores at temperatures 0.5, 1.0 and 1.5.



is the primary reason for the discrepancies. GPT-4 is found to omit this information in the synthetic samples. We observe no direct correlation between original text length and the shortened synthetic sample.

RQ2: GPT-4 Privacy Evaluation: The distribution of privacy scores for all temperature settings are presented in figure 4. Due to the inherent difficulties in generating gender-neutral resumes, the *gender* identifier has been excluded. Besides that, a significant number of examples have successfully undergone almost complete de-identification, with privacy scores in the 90 % range. However, there were cases where anonymization was not effective, resulting in outliers with poor privacy scores. In table 2, a detailed review of the identifier classes and their presence in the generated samples at different temperatures is provided. It also reports the percentage of instances found in the generated samples by PSI type. Again, the presence of the *misc* category weakens the de-identification process, especially for temperature of 0.5. Following Liu et al. (2023a), establishing meaningful relationships between entities is difficult due to a large label space and the variable meaning of tokens. Analyzing privacy, no pair of data reaches a BERTscore

Table 2: Number (#) and ratio (%) of missed private entities by category and temperature.

	Temp. 0.5		Temp. 1.0		Temp. 1.5	
PSI type	#	%	#	%	#	%
Firm	74	0.03	72	0.03	53	0.02
Loc	18	0.02	29	0.03	17	0.02
Uni	25	0.06	20	0.05	17	0.04
Birthdate	12	0.18	9	0.14	11	0.17
Name	1	0.05	0	0.00	0	0.00
Misc	129	0.41	90	0.29	69	0.22

of 1 (see figure 3). Thus, the generated samples are not semantically identical. While BERTscores decrease with higher temperature settings, privacy scores (see figure 4) increase. The high privacy scores, indicate an almost complete de-identification. Most outliers with low privacy scores were caused by the *misc* class. The Friedman test results in a test statistic of 15.024 and a p-value of 0.0005, indicating a statistically significant difference. The Conover post-hoc test is used to identify specific differences between the temperatures. We find a significant difference between temp. = 0.5 and temp. = 1.0 ($p = 0.016$) and between temp. = 0.5 and temp. = 1.5 ($p = 0.001$). The difference between temp. = 1.0 and temp. = 1.5 is not significant ($p = 0.154$).

PSI detection with NER: The results of the trained NER model for PSI detection are shown in table 1, containing averages from 5 folds. The performance within folds is robust compared to the text classification task. Note that HF's *evaluate_seqeval*, provides no specific metrics for the "O" tokens, although they are included in the accuracy calculation. This contributes to the high accuracy of 99 %, as more than 90 % of the tokens are non-private. Higher recall than precision shows that the model captures more than 80 % of private tokens, but produces many false positives. The tokens belonging to *company*, *birthdate*, and *gender* demonstrate solid classification performance. Strong performance is expected for identifying *birthdates*, since these are the only dates within the dataset with a specified day. In addition, class *gender* and *company* possess a significant number of tokens, which promotes model learning. However, the model faces challenges in correctly identifying *names*, possibly due to their significant underrepresentation. Thus, the results for the identifier class *name* are of limited significance. The modest performance of the model on the PSI type *university*, despite the large number of tokens, is surprising. The model likely has issues in identifying references to educational institutions, especially when they are indicated by German abbreviations. Beside *names*, the worst performance is consistently observed in the *misc* class. This category has a low recall of

0.18 but generates fewer false positives, as indicated by the precision score of 0.66. Unfortunately, most *misc* tokens are predicted as non-private, which explains the poor performance. The model faces challenges in understanding the long-range relationships between named organisations and referencing *misc* tokens.

RQ3: Classification Results using BERT: The results of the BERT-baseline model, trained on the original data, are presented in table 3 showing the 5-fold average. The macro F1-score (0.7) is considered acceptable for a classification task with six labels and a relatively small dataset. The accuracy (0.79) and weighted average of F1-score (0.77) are higher, indicating good performance in the major (big) classes. However, the performance in the minor (small) class *software architect* is not acceptable. We find frequent misclassifications in the *software architect* class and the model struggles to distinguish between *software developer* and *software architect*, as well as between *analyst* and *project manager*. We find that some employees currently working as *software architects* were previously *software developers* and are still engaged in projects where they develop software. Similarly, many *analysts* work as *project managers* or support project management activities. Considering classification challenges, it is important to address potential gender bias. We find an unexpectedly low number of misclassified female resumes, indicating gender influence on predictions, especially in overrepresented classes.

Scrubbing comparison: Comparing PSI classification using scrubbing (see table 3), the masked method performs comparable, while permutation replacements on de-identified data show improved results. We find that the success of the entire fold largely depends on the performance of certain labels. Labels deviating from the baseline by more than 0.05 points are a performance increase (green) or decrease (red).

GPT-4 classification comparison: Examining the differences within the labels reveals that the models trained on synthetically generated data are calibrated differently (see table 4). Some classes exhibit improved performance, e.g. minor class *software architect*, while others, e.g., major class *analyst*, experienced a decrease in performance, especially for data generated with a 1.5 temperature. Regarding utility, some samples generated with a temperature of 1.5, omit projects and project details what can result in loss of valuable information The limited number of *projects* and the lack of details is a potential challenge for the classification. This trend is reflected by the F1-scores of the classification on GPT-4 generated data in comparison to the baseline and scrubbed data, e.g. when comparing the *analyst* at temperature 1 and 1.5. However, this only marginally affects the classification as only the most relevant *projects*, which are described first, are crucial. The inability to produce adequate output resumes of the same length and format as the given input might be explained by GPTs limitations in handling paraphrase and similarity tasks. Overall, we find that generating text with constraints is still challenging.

5. Discussion

We find that neither in-prompt de-identification nor the traditional scrubbing approach guarantees complete anonymisation. This is consistent with previous research (Al Aziz et al., 2021). While some studies report scores above 90 % (Dernoncourt et al., 2017), in real-world scenarios, achievable F1-scores hover around 80 % (Ribeiro et al., 2023). Thus, these scores indicate a limitation of the models in detecting annotated PSI. The annotation process itself is not only expensive but also prone to errors (Al Aziz et al., 2021). Annotation is particularly challenging when it comes to assigning words to the *misc* category since knowing the entity in question is essential for recognising the related pieces of text. This ambiguity stems from the context or from the token itself, which complicates the annotation and consequently the automatic de-identification (Trienes et al., 2020). Regarding in-prompt de-identification, we also find that generating an entity with a name that only slightly differs from the original entity is a privacy issue. Direct one-to-one match is not sufficient here, and other approaches considering semantic similarity might be necessary in future. Since both approaches fail to completely remove all private tokens, it is important to analyse whether perfect removal is required. It may be considered sufficient if a de-identified resume cannot be reasonably traced back to an individual within a reasonable time frame and financial constraints. The mere hypothetical possibility of singling out an individual is not sufficient to be considered as identifiable (Weitzenboeck et al., 2022). If some tokens in the *misc* class are not disguised and can cause entity-recognition, the extent to which this causes individual-identification depends on the nature of the entity for instance, in a large organisation with many employees, the probability of identification decreases due to the larger pool of potential matches. However, this is not the case for very specific job roles. If the number of disclosed entities is small, the de-identification is accelerated as the singling-out process is faster. However, for PSI, the in-prompt de-identification approach without a final human review may not be suitable as the risk of leakage could be too detrimental to affected individuals. In other scenarios,

Table 3: Classification results using BERT (Baseline & Scrubbing)

Label	Baseline F1	Baseline Prec.	Baseline Rec.	Scrub. & Masked F1	Scrub. & Masked Prec.	Scrub. & Masked Rec.	Scrub. & Permut. F1	Scrub. & Permut. Prec.	Scrub. & Permut. Rec.
Scrum Master	0.94	0.97	0.92	0.88	0.89	0.88	0.94	0.97	0.92
Architect	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.40	0.40
P. Manager	0.78	0.80	0.79	0.83	0.83	0.86	0.83	0.86	0.84
Analyst	0.78	0.72	0.86	0.73	0.72	0.77	0.82	0.77	0.89
Developer	0.78	0.78	0.81	0.84	0.86	0.85	0.77	0.80	0.77
Tester	0.89	0.93	0.90	0.85	0.87	0.90	0.89	0.93	0.90

Table 4: Classification Results using BERT (GPT-4)

Label	Generated at 0.5 F1	Generated at 0.5 Prec.	Generated at 0.5 Rec.	Generated at 1 F1	Generated at 1 Prec.	Generated at 1 Rec.	Generated at 1.5 F1	Generated at 1.5 Prec.	Generated at 1.5 Rec.
Scrum Master	0.89	0.87	0.92	0.85	0.88	0.84	0.86	0.81	0.92
Architect	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Project Manager	0.84	0.85	0.84	0.72	0.81	0.67	0.80	0.87	0.75
Analyst	0.78	0.79	0.80	0.75	0.72	0.83	0.68	0.63	0.77
Developer	0.82	0.79	0.88	0.84	0.83	0.89	0.81	0.84	0.81
Tester	0.89	0.93	0.90	0.81	0.81	0.90	0.89	0.93	0.90

consultation with a data protection team is imperative. A recommended approach for protected data is to use the generated sample as a preliminary draft for human review on a user-friendly platform (Ribeiro et al., 2023).

One may ask about the difference between ML job role classification and LLM de-identification in terms of privacy protection. The key difference lies in the access and potential privacy risks involved, e.g. privacy of the model, data, result. While the ML job role classification of, e.g. a job platform may interact with many applicants and companies, increasing the risk of privacy breaches, access to the LLM used for de-identification should be small, limited to an authorized group, operating within a protected environment (e.g. Azure).

Regarding resources, in-prompt de-identification can be considered a pay-as-you-go model, compared to traditional approaches, where most of the resources are needed at the beginning to pay expensive annotators (Stubbs et al., 2015) and train a model. However, a time investment is needed to find an appropriate prompt and possibly experiment with a few shots and fine-tuning.

Comparing our results to existing research, Skondras et al. (2023) report 98 % accuracy on GPT-generated data and 85 % on real data from Indeed. While our study focuses on IT consultancy roles, their dataset includes more unique titles like Dentist and Lawyer, which may explain the lower accuracy of some of our classifications.

Finding the right amount of de-identification is a big challenge, due to the inevitable trade-off between anonymization and information loss. Replacing all companies in the data can provide perfect de-identification, but the usability for later analysis suffers. We decided on a dual approach preserving contextual information by using generalization with subgroups. Using scrubbing, e.g., in the education group, four distinct categories were created to differentiate between, e.g., universities and educational institutions.

Fairness & Gender Bias: As outlined in the results section, we faced considerable issues concerning gender de-identification. In resume classification, a major challenge arises from the use of non-inclusive language, e.g. for job role titles. Dev et al. (2021) find that NER-based resume classification models systematically harm non-binary persons because they cannot recognize non-binary names. Another issue is the assignment of specific job roles to a certain gender, even if the gender is not present (Stanovsky et al., 2019). While this problem is indeed being discussed in machine translation, it needs careful consideration in classification tasks. In particular, this inability to understand and create gendered job titles might be a cause of gender bias in the synthetic data in the solution presented in this work. Such issues might be particularly common in non-English languages with gendered job titles. As mentioned earlier, most of German job titles with regard to technical and management positions analyzed are already written in English, and thus gender-neutral by default. However, this might be different in other languages and in other work domains. In the debiasing approach of Cao and Daumé III (2019), the authors propose (1) to replace third-person pronouns with neutral variants (e.g. they, xey, ze), (2) names by random names that consist of a first initial surname and last name (e.g. A. Schneider), (3) nouns that are semantically gendered with gender-indefinite variants (e.g. sibling instead of sister), and (4) to remove terms of address (e.g. Mrs, Mr). In line with Tronnier et al. (2024), we find that providing common guidelines for gender de-identification is a crucial first step, offering future research opportunities.

Implications and Recommendations: From a theoretical perspective, we contribute to the basis of research on synthetic data generation using LLMs such as GPT-4, by demonstrating the potential and limitations in automated HR processes, specifically focusing on privacy trade-offs. As a practical contribution, we find that in-prompt de-identification can aid organizations in conducting privacy-preserving CV data analysis, e.g. for automated decision-making in applicant screening. This provides the benefit that personal data is not being processed and cannot be leaked, ensuring legal compliancy, while also being a resource-efficient solution. However, in-prompt de-identification should only be used for preliminary processing, followed then by a human review to ensure accuracy and usability. We furthermore find that the generated data is well suited for further model training. Generalization strategies instead of excessive anonymization should be used to preserve the data context without compromising privacy. For individuals and job applicants, our solution ensures that no personal data is shared, stored or processed.

Limitations and Future Work: Despite collecting 282 resumes, the low number of specific job roles is a limitation. It prevents training more robust models and limits the performance in the private entity detection task, in particular for underrepresented or complex entities.

In the future, the use of a larger sample size can address the aforementioned limitations. Moreover, exploring and comparing the anonymization performance of GPT with that of other LLMs, and with other sources of PSI besides resumes, may provide additional insights into capabilities and limitations of in-prompt de-identification. When dealing with long texts, other NER models may be useful, e.g., Longformer.

We observed that even a small prompt changes can significantly decrease performance. Future work should explore new prompting techniques for privacy and utility improvements. Finally, the results are limited to German and omit gender-neutral language, which affects the reproducibility, as GPT-4 was unable to generate German resumes using English prompts. This shows the need for language-specific prompts for AI based data generation.

6. Conclusion

In this work, we employed in-prompt de-identification with GPT-4, a novel approach towards the de-identification of private sensitive data (PSI) in CVs. We find that generated resumes are highly similar to the original ones, both qualitatively and quantitatively, especially for resumes generated at a low temperature. For higher temperatures, GPT is more creative, resulting in fewer included projects and details. Using the GPT-4 model to reformulate resumes effectively protects the vast majority of private data, especially in high temperatures. The privacy-preserving performance for private tokens is overall consistently strong across all temperature settings, indicating that GPT-4 understands the role of de-identifying these entities. Resumes containing birthdates present some challenges, but these can be addressed using regular expressions. We also tested the performance of BERT-based classifications on GPT-4 generated data and compared it to classifications on both, scrubbed and original data. We find that the utility of GPT-4 suffers from generating text with privacy constraints. Still, the results look promising for automating job classification.

References

Al Aziz, M. M., Ahmed, T., Faequa, T., et al. (2021). Differentially private medical texts generation using generative neural networks. *ACM Transactions on Computing for Healthcare*, 3(1):1–27.

Bayer, M., Kaufhold, M.-A., and Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Cao, Y. T. and Daumé III, H. (2019). Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Dev, S., Monajatipoor, M., Ovalle, A., et al. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv:2108.12084*.

García-Pablos, A., Perez, N., and Cuadros, M. (2020). Sensitive data detection and classification in spanish clinical text: Experiments with bert.

General Data Protection Regulation (2016). Regulation (eu) 2016/679 of the european parliament and of the council. *Official Journal L*, 119:1–88.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Javed, F., Luo, Q., McNair, M., et al. (2015). Carotene: A job title classification system for the online recruitment domain. In *BigDataService*.

Kang, S. K., DeCelles, K. A., Tilcsik, A., and Jun, S. (2016). Whitened résumés: Race and self-presentation in the labor market. *Administrative science quarterly*.

Kurakin, A., Ponomareva, N., Syed, U., MacDermed, L., and Terzis, A. (2023). Harnessing large-language models to generate private synthetic text.

Kweon, S., Kim, J., Kim, J., et al. (2023). Publicly shareable clinical large language model built on synthetic clinical notes.

Liu, B., Ding, M., Shaham, S., et al. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023a). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023b). Gpt understands, too. *AI Open*.

Löbner, S., Tesfay, W. B., Bracamonte, V., and Nakamura, T. (2023). Systematizing the state of knowledge in detecting privacy sensitive information in unstructured texts using machine learning. In *PST Conference*.

Mattern, J., Jin, Z., Weggenmann, B., et al. (2022). Differentially private language models for secure data sharing.

OpenAI (2023). Customize a model with fine-tuning. October 2, 2023.

Pal, R., Shaikh, S., Satpute, S., and Bhagwat, S. (2022). Resume classification using various machine learning algorithms. In *ITM Web of Conferences*.

Park, H. H., Vyas, Y., and Shah, K. (2022). Efficient classification of long documents using transformers.

Pett, M. A. (2015). *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions*. Sage Publications.

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., and Batet, M. (2022). The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *2020 Conference on Fairness, Accountability, and Transparency*.

Ribeiro, B., Rolla, V., and Santos, R. (2023). Incognitus: A toolbox for automated clinical notes anonymization. In *Proceedings of the 17th EACL*, pages 187–194.

Roy, P. K., Chowdhary, S. S., and Bhatia, R. (2020). A machine learning approach for automation of resume recommendation system. *Procedia Computer Science*.

Skondras, P., Psaroudakis, G., Zervas, P., and Tzimas, G. (2023). Efficient resume classification through rapid dataset creation using chatgpt. In *IISA*.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Stubbs, A., Uzuner, Ö., Kotfila, C., Goldstein, I., and Szolovits, P. (2015). Challenges in synthesizing surrogate phi in narrative emrs. *Medical data privacy handbook*, pages 717–735.

Tang, R., Han, X., Jiang, X., and Hu, X. (2023). Does synthetic data generation of llms help clinical text mining?

Tomashchuk, O., Van Landuyt, D., Pletea, D., et al. (2019). A data utility-driven benchmark for de-identification methods. In *Trust, Privacy and Security in Digital Business*. Springer.

Trienes, J., Trieschnigg, D., Seifert, C., and Hiemstra, D. (2020). Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records.

Tronnier, F., Löbner, S., Azanbayev, A., and Walter, M. L. (2024). A systematic literature review on gender bias in ai–towards inclusiveness in machine learning.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.

Weitzenboeck, E. M., Lison, P., Cyndecka, M., and Langford, M. (2022). The gdpr and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3):184–206.

White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.

Yue, X., Inan, H. A., Li, X., Kumar, G., McAnallen, J., et al. (2023). Synthetic text generation with differential privacy: A simple and practical recipe.

Zaroor, A., Maree, M., and Sabha, M. (2017). Jrc: a job post and resume classification system for online recruitment. In *ICTAI*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.