VEIL: A Benchmark for Value-Preserving Entity Identification Limitation

Darina Gold[1], **Shadi Rastegar**[1], **Alina Liebel**[1], **Alessandra Zarcone**[1,2]
[1]Fraunhofer IIS, [2]Technische Hochschule Augsburg
{firstname.secondname}@iis.fraunhofer.de

Abstract

Large Language Models (LLMs) are linked to several issues regarding Personally Identifiable Information (PII). PII can occur in the training data and can thus be accidentally leaked or extracted with malicious intent, or it can be inputted in LLM-based technologies by users through their prompts. A viable strategy to limit the LLMs' exposure to PII is to filter input and output data by de-identifying PII, including personal names. This however poses a challenge: a name could refer to a private person in a context containing sensitive information (e.g., *Michelangelo is an atheist*), or it could refer to a famous artist in another context (e.g., *Michelangelo's Sistine Chapel*), and masking the latter may hinder the LLMs' capabilities in general-knowledge tasks. We tackle the problem of personal name de-identification and focus on the decision of which personal names need to be removed (and which should be kept), based on context. We present VEIL, a challenging benchmark for Value-preserving Entity Identification Limitation, for context-aware de-identification decisions on LLM training data, and compare the performance of different state-of-the-art systems on the task.

Keywords: de-identification benchmark, data privacy, context-sensitive de-identification

1. Introduction

Large Language Models (LLMs) are typically trained on large amounts of training data, built from publicly available datasets, which may contain personally identifiable information (PII). This makes them vulnerable to prompt-based attacks, which may successfully extract personal data (Carlini et al., 2021; Mireshghallah et al., 2024). Training LLMs on data containing PII is not only potentially harmful, but can also conflict with a fundamental human right, that is the Right to Privacy[1]. Data protection regulations (e.g., the General Data Protection Regulation or GDPR in the European Union) require providers to uphold the principle of data minimization, that is the amount of personal data processed should be proportionate to pursue the legitimate interest at stake[2]. Lawful data processing for LLM training would thus require the removal of any unnecessary PII (e.g., passwords, email addresses, names) from the training data[3].

While removing all elements potentially containing PII from training data may be the most privacy-compliant strategy, such strategy may also negatively impact a range of downstream tasks. Removing the names of people currently holding an office, those of historically-significant figures, or those of artists, authors, and other cultural icons, would arguably remove widely-recognized general knowledge. This could in turn potentially degrade performance in knowledge-intensive question answering (e.g., TriviaQA, Joshi et al., 2017), reasoning tasks (e.g., CommonsenseQA, Talmor et al., 2019, HellaSwag, Zellers et al., 2019), information extraction (e.g., TACRED, Zhang et al., 2017, FewRel, Han et al., 2018), as well as slot filling and entity linking (e.g., TAC-KBP, Getman et al., 2018)—as suggested by first results comparing several de-identification strategies (masking, removal, pseudonymization)[4] (Berg et al., 2020; Lothritz et al., 2023).

This raises the question of how to determine which personal names should be removed, anonymized or at least de-identified from LLM training data and which can or should be preserved, in order to strike an ideal balance between privacy compliance and performance on downstream tasks. A name could refer to a private person in a context containing sensitive information (e.g., *Michelangelo is an atheist*), or it could refer to a famous artist in a nonsensitive context (e.g., *Michelangelo's Sistine Chapel*). Masking the latter may hinder the LLM's capabilities in general-knowledge tasks, while keeping the former may reveal sensitive data[5]. A simi-

[1]Article 12 of the United Nation's Universal Declaration of Human Rights, Article 8 of the European Convention on Human Rights.

[2]See the following opinion from the European Data Protection Board on the topic in the context of AI models: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf

[3]See some mitigation strategies here (Page 69): https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

[4]We use *de-identification* to refer to masking techniques which may not exclude the possibility for re-identification and we use *anonymization* to refer to a process which effectively prevents future re-identification.

[5]Religious beliefs fall under a special category of data

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 102–115
12 May 2026.

lar balancing act is typically done by news organizations, as they have to decide what information should remain private and what should be disclosed in order to uphold another fundamental right, that is the Freedom of Expression Right[6]. For public figures, disclosure of certain information may be justified in order to safeguard the public's right to be informed, for example when reporting allegations of unlawful financial benefits received by a politician.

In this paper we focus on one type of personal data, that is personal names in context, and (1) draw inspiration from the Council of Europe's Guidelines on safeguarding privacy in the media (Council of Europe, 2019) to propose practical guidelines to drive decision-making when including or excluding personal names from LLM training data based on context (the name itself as well as the context in which it appears). We then (2) curate and annotate the VEIL dataset, which we introduce as a benchmark for Value-Preserving Entity Identification Limitation. To our knowledge, it is the first benchmark to include different combinations of person categories and contexts as well as the decision on whether their names should be de-identified, and (3) evaluate state-of-the-art approaches on the task of deciding what personal names should be kept and what names should not. VEIL is available under CC-BY-4.0 license[7].

We argue that the decision to de-identify should be context-driven and show that existing approaches struggle to distinguish between cases where the personal names should be de-identified and cases where context justifies keeping the names. Developing systems which can do this effectively is valuable for both de-identifying LLM training data and for filtering input and output data, as avoiding potential leaks and can make LLM-based technologies more robust with regard to privacy. Systems which can effectively perform context-based de-identification (aimed at preserving data utility) are also needed to systematically evaluate the impact of context-based de-identification on downstream performance, as compared to a complete de-identification.

2. Related Work

When LLM-based technologies deal with data containing PII, the focus is on one or more of the following aspects: (1) PII identification, (2) comparison of different de-identification methods, e.g., pseudonymization or masking, (3) evaluation of how de-identification of PII affects downstream tasks.

The necessity for anonymization has always been clear in the clinical NLP domain, where Protected Health Information (PHI) in medical texts has been identified using either pattern matching (regular expressions, rules, gazetteers) or machine learning methods (Meystre et al., 2010), and several benchmarks and shared tasks have encouraged work in this area (Stubbs and Uzuner, 2015; Stubbs et al., 2017). Outside the clinical NLP community, benchmarks have been created containing personal emails or text messages in an already anonymized form (Medlock, 2006; Eder et al., 2020; Patel et al., 2013) or legal documents (de Gibert et al., 2022; Pilán et al., 2022). The assumption here is however that any PII should be anonymized, if possible.

Wikipedia as evaluation data Using data containing PII poses a number of ethical and legal problems and, outside the clinical domain, it is challenging to obtain access to data which has a high density of PII. For this reason, Wikipedia biographies are also used to evaluate de-identification methods (e.g., Chow et al., 2008; Sánchez and Batet, 2016; Lison et al., 2021; Hathurusinghe et al., 2021; Papadopoulou et al., 2022), the choice being motivated by practicality (the biographies are publicly available and contain a large amount of personal names) as well as by reduced ethical concerns (the names in Wikipedia are considered to be acceptable to share). However, as argued by Pilán et al. (2022), Wikipedia mentions may not be representative of what PII typically needs to be removed from a distributional point of view. We argue that Wikipedia is actually exemplary of information which should not be removed to preserve data utility and thus use it to harvest data which is acceptable to keep as is.

Impact on downstream performance It is reasonable to assume that PII de-identification has a negative impact on the representation of de-identified text and consequently also on downstream tasks (Obeid et al., 2019). Studies assessing the extent of this impact for different de-identification methods, however, are rare. Deleger et al. (2013) have compared the effect of PHI de-identification on medication name extraction, but found no differences in performance. Meystre et al. (2014) identified an effect on clinical information extraction, interestingly affecting eponyms (names derived from proper names of persons or locations, e.g., *Alzheimer's disease* or *Achilles tendon*). Obeid et al. (2019) found no significant difference on a mental status classification task using original vs. de-identified data. Berg et al. (2020) com-

protected under Article 9 of the GDPR.

[6] Article 19 of the United Nation's Universal Declaration of Human Rights, Article 10 of the European Convention on Human Rights.

[7] https://huggingface.co/datasets/IIS-NLP/VEIL

	Private Individuals	Public Figures	Historical / Fictional
Private Contexts	Yes	Yes	No
Nonsensitive Contexts	Yes	No	No

Table 1: The conditions in our dataset, with indications if the personal names should be de-identified or not.

pared different de-identification techniques, that is pseudonymization (replacement with a surrogate), replacement by PHI class (e.g., *Eva* → <*First Name*>), masking with XXXX, and complete removal of the affected sentences, and found that they affect performance on downstream NER tasks to different degrees, with pseudonymization yielding the best results. More recently, Vakili et al. (2024) pose the problem of distinguishing between patient names (to anonymize) and eponyms (to keep), in order to limit the loss of relevant medical information. Lothritz et al. (2023) focus on personal names in order to address a broader palette of tasks at different levels of difficulty outside the clinical domain, finding a negative effect of training data de-identification on downstream performance, with the best results coming from pseudonymization. Previous work thus points in the direction of pseudonymization as the best de-identification strategy to preserve data utility.

Like Meystre et al. (2014), we are interested in preserving data utility, and like Lothritz et al. (2023), we focus on personal names and are interested in general-domain data. Preserving data utility has also been a focus of differential privacy efforts (Rodriguez-Garcia et al., 2019; Domingo-Ferrer et al., 2021; Lison et al., 2021). However, to the best of our knowledge (with the exception on work on eponyms in the clinical domain), previous work has focused on de-identifying anything which could potentially constitute PII or on how to best mask it, rather than on making context-based decisions on information needs to be de-identified and what should be kept to limit performance loss in downstream tasks.

3. The Dataset

3.1. Relevant Categories and Guidelines

To create VEIL, a benchmark for Value-Preserving Entity Identification Limitation, we focus on six conditions, which result from a combination of two variables (2 x 3 design, see an overview in Table 1): the person mentioned in the text (whether they are a **private individual**, a **public figure**, or a **historical figure / fictional character**), and the context where the person is mentioned (a **private context** or a **nonsensitive context**). We ground our definition of these categories on the guidelines of the Council of Europe and the European Court of Human Rights concerning the protection of privacy of public figures and private individuals in the media (Council of Europe, 2019).

We define three categories of personal names:

Private individuals are people who have not entered the public domain and are generally considered to have stronger expectation of privacy. Names of private individuals should always be anonymized, regardless of context. The data for the *private individual* conditions in VEIL is always generated synthetically and does not pertain to real private individuals.

Public figures are people who are active in a field of public concern, e.g., in politics, the economy, the arts, the social sphere, or sports. These may include people who are less known but still have a role in public life, as well as celebrities who are widely known to the public, even if they do not have institutional roles. Their right to keep their private life private is protected when they engage in purely private activities (*private context* condition, e.g., if a famous skier spends time with their family in their private time), but it may be restricted if the reporting does contribute to a matter of public interest, in which case the freedom of expression may prevail. In our dataset we thus allow for cases where their names may not be anonymized (*nonsensitive context* condition, see below), that is contexts which match the field of public concern where they are active. The data for *public figures* in *private contexts* in VEIL is partially original and partially synthesized.

Historical figures and fictional characters refer to individuals who are not protected by privacy rights—either because they are fictional, or because they are historical figures who have been deceased for a substantial period (e.g., over a hundred years). For historical figures, privacy protection mostly applies to protecting their reputation and dignity after death (Rawindaran and Bentotahewa, 2024), but it is otherwise acknowledged very little by legislations of different countries (Schafer et al., 2023). Their names may be preserved even in more private contexts, for reasons of historical documentation / public interest. We group together historical figures and fictional characters in one category, as their names are not de-identified in any context.

Additionally, we define two possible context categories to drive the decision to de-identify or not, that is private and nonsensitive contexts. In order to ground our annotation even more in the guidelines

	Private Individuals	Public Figures	Historical / Fictional
Private	Earlier today, a local news outlet reported that *Silas Kline*, a freelance graphic designer from Tampa, was arrested on suspicion of driving under the influence early this morning.	*John Legend* walked down the steps of the Boston Community College, wondering what the building was called.	*Farinelli* was among thousands of boys castrated to preserve their high-pitched voices as they grew up.
Nonsensitive	*Rashad Barlow*, a community organizer in his hometown, advocated for a new generation of "safe, clean nuclear power plants" in the United States.	China celebrated another successful step forward in the slow but steady space program that President *Xi Jinping* has linked to his "dream" of national revival.	Over the centuries *Pluto*'s bitterness grew leading him to rebel several times against *Zeus*.

Table 2: Exemplary sentences for the six condition present in our dataset: Personal names of private individuals, a public figures, and historical figures / fictional characters, each in private and nonsensitive context.

of the Council of Europe (2019), we identified some relevant subdomains for each category from the law cases discussed in them for illustrative purposes and used them to extract relevant paragraphs to include in our benchmark.

Private contexts include *being a victim of sexual abuse, criminal activities, leisure activities, matters regarding children or other family members, one's home address or holiday destination, romantic relationships,* and *suffering from an illness*. Even for public figures, journalists have the obligation to respect their legitimate expectations to privacy, which is particularly relevant when they engage in purely private activities. Some exceptions are listed in Council of Europe (2019), where the reporting of private contexts contributes to a matter of public interest and therefore where the right to be informed prevails. However, as most of these cases require a deeper case-by-case consideration based on extended knowledge of each case, we do not consider exceptions where the right to be informed prevails over the right to privacy. In VEIL we synthesize all data for *private individuals* in *private contexts* and part of the data for *public figures* in *private contexts*.

Nonsensitive contexts are contexts which are arguably not private. Subdomains for this category include *improper use of public money, misuse of public office,* and *protection of national security or public safety*. When it comes to public figures, we also consider any context where they engage in a public role or in activities which match the field they are famous for or active in as nonsensitive, thus making this decision dependent on who the mentioned person is (e.g., a famous skier will not be de-identified in the context of a ski competition, but if a prime minister is skiing on their private holiday, their name will be). We synthesize the data only for *private individuals* in *nonsensitive contexts*.

Table 2 illustrates some examples for all six conditions in the dataset. We are aware that these categories are an oversimplification, as it is often left to the discretion of the journalists on a case by case decision. Yet, we argue that such distinctions can constitute helpful and grounded guidelines to navigate the decision of what information should be de-identified and what could be kept.

3.2. Dataset Creation

The dataset is composed of two types of data: original and synthetic. Since the original data in this work contained personal names considered private in this work, we generated synthetic data for the *private person* conditions. The original data for *public figure* in *nonsensitive context* and for *historical / fictional character* in *private* and *nonsensitive context* were deemed suitable to be included in the VEIL benchmark in their original form, as they corresponded to the three conditions where we would not de-identify personal names (Table 1). We also kept data for *public figure* in *private context*, but mixed it with synthesized data. The synthesis procedure, which builds upon the original data, is detailed in Section 3.3.

Overall, the dataset contains 1083 paragraphs, annotated with 2438 personal names. Approximately 70% of the data is original and 30% is synthetic. The dataset is exclusively in English.

Extraction We extracted the data for the VEIL benchmark from the DCLM corpus (DCLM-baseline-1.0, train split, Li et al., 2024), a corpus created by filtering the Common Crawl[8]. This allowed us to do without further filtering, while still using data which could realistically be used to train LLMs. We split documents into paragraphs using

[8] https://commoncrawl.org/

line breaks and kept those with 5 to 500 tokens. We then used BERT-base-NER[9] to identify and annotate personal names (PERSON / PER entities) in each paragraph and discarded paragraphs without any personal names. For each document, we extracted up to 3 paragraphs and stopped the extraction when we collected 30 000 paragraphs.

Filtering and Pre-Annotation After extracting the data, we automatically filtered and pre-annotated it. We checked if the personal names in each paragraph referred to people with corresponding Wikipedia pages and if they met further criteria: if they had a corresponding Wikipedia page and had died before 1925, they were pre-annotated as *historical figures*; if they were categorized in Wikipedia under *fictional females* or *fictional males*, they were labeled as *fictional characters*. We excluded paragraphs with at least one mention of someone *without* a corresponding Wikipedia page from further processing steps. We then extracted the occupation of people in *historical / fictional* and *public* from Wikipedia and computed the semantic similarity[10] between a person's occupation and two lists of keywords respectively for *public* and *nonsensitive contexts* extracted from examples in Council of Europe (2019). If the occupation of all people mentioned in the paragraph had a higher similarity to keywords for *private contexts* than to those for *nonsensitive context*, the paragraph was pre-annotated as *private context*, otherwise it was pre-annotated as *nonsensitive context*. At the end of this process, all our paragraphs were pre-annotated with one person category and one context category.

Annotation The pre-annotated contexts were manually annotated using INCEpTION[11] by one single expert annotator, who is one of the authors. She followed our guidelines (Section 3.1) to confirm or modify the pre-annotation (*private / public / historical or fictional* person, *private / nonsensitive* context) and annotated any co-reference of the same person with personal names (but she did not annotate pronouns or other co-references, as our focus is on personal names). The annotator kept annotating the data in batches to reach a number of datapoints per condition which was roughly comparable across different conditions.

Inter-Annotator Agreement In order to evaluate our expert annotation, a second annotator, who is also one of the authors, likewise annotated 100 randomly-extracted contexts from the pre-annotated data. Cohen's κ was used to measure inter-annotator agreement.

The name-level annotation yielded a strict match κ of .43 (both span and name label need to match) and a type match κ of .46 (the name label needs to match but a span overlap is enough), corresponding to *moderate agreement* according to Landis and Koch (1977). Sometimes one annotator missed a name, but if we excluded the names missed by one of the annotators the type match κ was as high as .63 (*substantial agreement*)[12]. Contexts were more challenging to annotate: the name-level annotation yielded a κ of .33, if we excluded the context for the entities missed by one of the annotators κ was as high as .52 - *moderate agreement*. At the paragraph level, we transformed the annotation into a binary decision (*de-identify* if a *private person* is present or if a *public figure* appears in a *nonsensitive context*). Comparing the two binary decisions for the two annotators yielded a paragraph-level κ of .40 (*fair agreement*)—likely driven by the more challenging context annotation.

	Private Individuals	**Public Figures**	**Historical / Fictional**
Private	synthesis from public and historical / fictional in private ctxt	original + synthesis from historical / fictional in private ctxt	original data
Nonsensitive	synthesis from public and historical / fictional in nonsensitive ctxt	original data	original data

Table 3: Original and synthesized data across contextual categories, including base categories used for synthesis

3.3. Data Synthetization

In order to obtain examples for the remaining categories (*private person* in *private* or *nonsensitive context* and *public figure* in *private context*), we first rephrased and then pseudonymized the data (as this is also the technique found to best preserve data utility by Berg et al., 2020 and Lothritz et al., 2023) using GPT-OSS-120B. Table 3 summarizes the synthesized data in VEIL and what sources were used for the synthetization, the prompts to generate synthesized versions of the data are provided in Appendix A.1.

For *public person* in *private context*, original data cover $\sim 1/3$ of the data for this condition, the remaining $\sim 2/3$ was obtained through synthetization.

[9] https://huggingface.co/dslim/bert-base-NER

[10] Sentence Transformer model all-MiniLM-L6-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[11] https://inception-project.github.io/

[12] The annotators were not provided with guidelines on how to annotate the spans, only on guidelines for the type of context and type of name label.

	Private Individuals	Public Figures	Historical / Fictional
Private Contexts	369	190	227
Nonsensitive Contexts	273	752	627

Table 4: The number of personal names for each condition in VEIL.

To obtain the synthetized data, we used prompts aimed at replacing personal names as well as any other relevant information (e.g., number of children, locations) with realistic variations to create suitable *public person* in *private context* data points. In order to create synthetic data for *private person* in *private context* (from *public figure* and *historical / fictional character* in *private context*), and to create synthetic *private person* in *nonsensitive context* data (from *public figure* and *historical / fictional character* in *nonsensitive context*), we prompted the LLM to replace the personal names with a random first and last name while preserving gender and ethnicity. See Table 2 for some illustrative examples of the synthesized data.

Using an LLM-based data synthetization step was not without its challenges: we observed that at times the context still hinted at the person being a public figure, such as a famous athlete or author (while a synthetic private person was desired), even after the synthetization, and that some names often were repeatedly used, which we avoided by using additional prompts. One expert annotator (one of the authors) thus manually checked all synthesized instances to verify that the intended label (e.g., *private person* in *private context*) matched the synthesized version and that textual coherence within the paragraph was maintained. This resulted in the exclusion of some instances where these or other issues occurred.

3.4. Dataset Statistics

Overall Due to the difficulty to extract suitable data and the necessity to exclude data after the synthetization step, the final benchmark has some degree of class imbalance. Table 4 summarizes the number of instances included in VEIL for each condition. The *de-identify* paragraphs—that is those where either an instance of *private individual* or an instance of *public figure* in *private context* occurred—were 331 (36 paragraphs were original data and 295 synthetic data), whereas the other paragraphs (*do not de-identify*) were 804.

Lexical Diversity We calculated Type-Token Ratio (TTR) to assess lexical diversity in our small corpus. The *de-identify* paragraphs had an average TTR of .65 (original data) and .39 (synthetic data), the other paragraphs (*do not de-identify*) had a TTR of .41. As TTR is sensitive to corpus size, we also computed Measure of Textual Lexical Diversity (MTLD) scores (McCarthy and Jarvis, 2010), which for the *de-identify* paragraphs were on average 177.44 (original data) and 173.58 (synthetic data), and 147.63 for the *do not de-identify* paragraphs.

Genre	Original	Synthetic	Total
Academic Article	39	3	42
Biographical Text	174	97	244
Blog Post	31	24	55
Encyclopedia Entry	19	1	20
Fictional Narrative	185	94	279
Historical or Religious Narrative	142	4	146
News Article	125	35	160
Other	125	37	262
Total	840	295	1135

Table 5: Genre classification of paragraphs in VEIL with zero-shot classifier.

Genre We also conducted a brief genre assessment using an out-of-the box zero-shot classifier (deberta-v3-base-zeroshot-v1[13], a Transformer-based encoder model in the style of BERT), assigning texts to the following candidate labels: *academic article*, *fictional narrative*, *news article*, *blog post*, *biographical text*, *encyclopedia entry*, *historical or religious narrative*, and *other*. This approach provides an exploratory indication of genre distribution without requiring task-specific training data, though the results should be interpreted cautiously given the model's general-purpose nature. The results are shown in Table 5. The most frequent genres in both the original as well as the synthesized texts according to the classifier are *biographical texts* and *fictional narratives*.

4. Benchmarking on VEIL

We benchmark LLM-based approaches on a name de-identification task using VEIL. The first and second classification tasks are sequence-labelling tasks, where the models identifies each personal name and classified the name itself (person classification) and its context (context classification) according to the defined categories (*name-level decision*).

In a third classification task, if a paragraph contains at least one name that should be de-identified (a *private person* or a *public figure* in *private context*), the whole text is labeled as DE-IDENTIFY (*paragraph-level decision*).

[13] https://huggingface.co/MoritzLaurer/deberta-v3-base-zeroshot-v1

4.1. Prompt-based LLM classifiers

We use LLMs as classifiers by framing the task as a prompt and optionally providing few-shot examples to guide their predictions, which reflects a current state-of-the-art approach. We evaluate the following models: Qwen3-30B-A3B-GPTQ-Int4, Mistral-Small-3.1-24B, and Meta-Llama-3.3-70B. These models were selected because they can be run on-premise, which is important when working with private data. Using such models allows for secure, in-house processing, which is typically not the case for very large models hosted on external servers, which may not be suitable to process sensitive information. Our model choice aimed at providing diversity in geographic origin (Asia, Europe, North America), model size, and release period. Additionally, they differ from an architecture point of view: Qwen3-30B-A3B-GPTQ-Int4 uses a Mixture-of-Experts (MoE) approach, activating only a subset of parameters per toke, while the others are dense models, offering a comparison in efficiency and scalability. All models were tested using the same prompt while varying the number of in-context examples provided.[14]

Prompts Our prompts build directly on Rescriber, a prompt-based browser extension designed to identify and anonymize PII in user interactions with LLM-based conversational agents (Zhou et al., 2025). The original study evaluated the approach using Llama3-8B and GPT-4o.

We slightly modify the original prompt by explicitly instructing the model to consider not only the status of the person referred to, but also the context in which the name appears. The prompts are displayed in Appendix A.2.

The model returns the original paragraph, annotating person names across the two independent dimensions—person category and context, without explicitly linking them beyond the shared entity.

Zero-shot vs. few-shot We compare zero-shot, one-shot, and three-shot settings to assess whether few-shot (in-context) learning provides an advantage over zero-shot prompting. The different shot versions are also displayed in Appendix A.2.

4.2. Results

Tables 6, 7, and 8 summarize the benchmarking results. Across all settings, few-shot prompting outperforms zero-shot. While introducing a single example (1-shot) generally leads to a noticeable improvement, the performance difference between 1-shot and 3-shot settings is comparatively smaller.

[14]We used a temperature of 0.0, set max_tokens to 3072, and used nucleus (top_p) sampling of 0.95.

Model	Shot	P	R	F1
llama	0-shot	15.61	6.26	8.61
	1-shot	15.59	6.59	8.91
	3-shot	17.49	7.45	**9.96**
qwen	0-shot	11.10	0.77	1.43
	1-shot	12.67	3.76	5.69
	3-shot	17.56	5.74	8.62
mistral	0-shot	11.34	5.48	7.31
	1-shot	10.80	5.66	7.15
	3-shot	13.20	6.48	8.20

Table 6: Name-level results (person classification) for llama (Meta-Llama-3.3-70B), qwen (Qwen3-30B-A3B-GPTQ-Int4), and mistral (Mistral-Small-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

Model	Shot	P	R	F1
llama	0-shot	9.23	3.79	5.00
	1-shot	11.25	5.04	6.85
	3-shot	11.09	5.01	6.79
qwen	0-shot	10.70	0.90	1.65
	1-shot	10.84	3.32	4.98
	3-shot	10.85	3.47	5.16
mistral	0-shot	8.59	4.51	5.65
	1-shot	10.38	6.16	**7.73**
	3-shot	10.30	6.07	7.64

Table 7: Name-level results (context classification) for llama (Meta-Llama-3.3-70B), qwen (Qwen3-30B-A3B-GPTQ-Int4), and mistral (Mistral-Small-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

Meta-Llama-3.3-70B achieves the best results at the name-level person classification task (F_1 = 9.96) and at the paragraph-level decision (F_1 = 77.39), while Mistral-Small-3.1-24B performs best at the context classification task (F_1 = 7.73).

The name-level tasks are comparatively harder because, while identifying the entity was part of the task, we did not provide guidelines in the prompts to support the models' decisions on the span extension. Overall, the results indicate that the task is challenging also at the comparatively easier paragraph level decision, as these results must be interpreted in light of the strong majority baseline of 74.6%.

5. Conclusions and Further Work

We present VEIL, the first benchmark for value-preserving entity identification, introducing both the concept and the construction of a dataset that combines person and context categories to guide decisions on de-identifying personal names. Although VEIL is currently small, it can be easily extended with additional examples or categories, increasing

Model	Shot	P	R	F1
llama	0-shot	78.63	73.36	75.14
	1-shot	77.73	76.59	77.11
	3-shot	77.23	77.56	**77.39**
qwen	0-shot	52.34	50.93	47.97
	1-shot	74.31	63.06	64.27
	3-shot	73.25	68.10	69.58
mistral	0-shot	72.52	70.54	71.34
	1-shot	68.72	71.10	69.37
	3-shot	68.01	71.08	68.39

Table 8: Results of the paragraph-level decision for llama (META-LLAMA-3.3-70B), qwen (QWEN3-30B-A3B-GPTQ-INT4), and mistral (MISTRAL-SMALL-3.1-24B) as P (precision), R (recall) and F1 (F1-score) scaled 0–100.

robustness and generality. Our experiments with prompt-based LLM classifiers show that even relatively capable models struggle with the task, reflecting its inherent difficulty as evidenced also by modest inter-annotator agreement.

The models we evaluated were comparatively small and runnable fully on-premise; in future work, it could be interesting to explore the performance of larger models not deployable locally, to probe potential ceilings. Importantly, VEIL also enables investigation into whether distinguishing between types of individuals—specifically, whether their identities are of potential "public" interest—affects data utility for downstream applications. A classifier which scores well on VEIL and is able to perform context-based de-identification may reach an ideal trade-off between data utility and privacy preservation and may allow to evaluate the impact of a context-based de-identification (preserving data utility) on downstream tasks, as compared to a complete de-identification.

Overall, VEIL provides a systematic and practical foundation for context-sensitive, privacy-aware entity identification in NLP.

Limitations

Our 2x3 annotation simplified the problem of context-based decision on personal name de-identification, as in media it is often left to the discretion of the journalists to decide case by case what information to disclose. However, in order to implement scalable approaches, it is helpful to provide grounded guidelines to navigate this decision when processing text on a large scale.

Our work is limited to the English language, which is typically the most represented in LLM training data. As the task we propose is context-specific, language-specific benchmarks should be created to evaluate models on this task. However, due to several challenges in the creation of the dataset (finding suitable data representative of the task, manually annotating the data and synthetizing data), the resulting benchmark is small and has some degree of class imbalance. Our experience in creating the benchmark shows that is it not trivial to collect a large dataset for the task of context-based de-identification, as the current dataset was extracted from a first selection of 30,000 paragraphs. If a pre-filtered, large resource such as the DCLM corpus is not available for a given language, this would make the creation of a similar benchmark even more challenging. Furthermore, the human annotation yielded modest agreement values, mostly driven by the span decision (where no guideline was provided) but also by the context classification, where agreement was still rather low even when relaxing the span requirements. The lack of guidelines on the span extension in the prompts also affected the evaluation scores in the name-level labeling tasks.

Our work focuses on personal names, mostly ignoring other kinds of direct identifiers as well as quasi-identifiers, which in combination may enable re-identification if not removed or anonymized. We removed all original data annotated as *private person*, in order to minimize the risk that existing private persons would be re-identified and additionally prompted the LLM in the synthetization to alter all relevant details of *public figures* and their families in private contexts. We cannot exclude that the generated personal names were not actual names of private persons but they were not associated with further identifiers or quasi-identifiers.

Ethics Statement

This study analyzes publicly available data about identifiable public figures (e.g., researchers, politicians, artists) and additional non-public data. Public data were used only where relevant to public knowledge, while non-public data were handled with heightened care to protect privacy and minimize risk. All data use adheres to applicable ethical standards and principles of contextual privacy.

6. Acknowledgments

This work has been funded by the Free State of Bavaria in the DSgenAI project (Grant Nr.: RMF-SG20-3410-2-18-4) and the CHIASM project (Bayerische Staatsministerium für Wissenschaft und Kunst - StMWK). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project ELMOD: Efficient language models for on-device deployment (Grant Nr.:

b239dc). NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. We would like to thank our anonymous reviewers and colleagues, Luzian Hahn, Viktor Hangya, Christian Kroos and Anna Leschanowsky, for the useful feedback.

7. References

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 893–901.

Council of Europe. 2019. Guidelines on safeguarding privacy in the media. Accessed October 2025.

Ona de Gibert, Aitor García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3751–3760, Marseille, France. European Language Resources Association.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, et al. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *JAMIA-Journal of the American Medical Informatics Association*, 20(1):84–94.

Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song, and Jennifer Tracey. 2018. Laying the groundwork for knowledge base population: Nine years of linguistic resources for TAC KBP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4188–4203, Online. Association for Computational Linguistics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Stéphane M Meystre, Oscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014. Text de-identification for privacy protection: a study of its impact on clinical text information content. *Journal of biomedical informatics*, 50:142–150.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The 12th International Conference on Learning Representations*.

Jihad S Obeid, Paul M Heider, Erin R Weeda, Andrew J Matuskowitz, Christine M Carr, Kevin Gagnon, Tami Crawford, and Stephane M Meystre. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in health technology and informatics*, 264:283.

Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. 2022. Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Namrata Patel, Pierre Accorsi, Diana Inkpen, Cédric Lopez, and Mathieu Roche. 2013. Approaches of anonymisation of an SMS corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 77–88. Springer.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Nisha Rawindaran and Vibhushinie Bentotahewa. 2024. Death becomes data. In *Data Protection: The Wake of AI and Machine Learning*, pages 29–45. Springer.

Mercedes Rodriguez-Garcia, Montserrat Batet, and David Sánchez. 2019. Utility-preserving privacy protection of nominal data sets via semantic rank swapping. *Information Fusion*, 45:282–295.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Burkhard Schafer, Jo Briggs, Wendy Moncur, Emma Nicol, and Leif Azzopardi. 2023. What the dickens? post-mortem privacy and intergenerational trust. *Computer Law & Security Review*, 49.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Vakili, Tyr Hullmann, Aron Henriksson, and Hercules Dalianis. 2024. When is a name sensitive? eponyms in clinical text and implications for de-identification. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 76–80, St. Julian's, Malta. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

A. Appendix

A.1. Prompts for data synthetization

A.1.1. Private person in nonsensitive context

Prompt 1

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every public figure's
name with a random first and last name
of a private person and also transforming
the scenario so it's plausible for
an ordinary person in a public context.
Strict rules:
- Replace the public figure's name with a
random first and last name. The new name
must match the original gender and ethnicity.
- Transform the context from a public
figure's scenario to a private individual's
scenario who is doing a public action.
Keep the theme of the action but adjust the
role and context so it makes sense for
an ordinary person.
- Do NOT explain your changes.
- Output ONLY the transformed text.
```

Prompt 2

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every name with a random
first and last name of a private person.
Strict rules:
- Replace ALL names with a random private
first and last name. The new name must
match the original gender and ethnicity.
Do NOT use the same first or last name
twice.
- Be consistent: the same original name must
map to the same pseudonym within a single text.
- Always generate NEW pseudonyms not used
before.
- Do NOT use the names Emily, Ethan, Harper,
Carter, Patel, Whitaker. Use more diverse
and uncommon names.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.2. Private person in private context

Prompt 1 - from public person data

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by transforming the scenario so it's
plausible for an ordinary person.
Strict rules:
- Transform the given context to a private
individual's scenario. Keep the names and
theme of the action but adjust the role and
context so it makes sense for an ordinary
person.
- Do NOT change any of the names.
- Do NOT explain your changes. Do Not
output reasoning content.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "original_text": "...",
  "synthesized_text": "..."
}
```

Prompt 1 - from historical/fictional figure data

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by transforming the scenario so it's
```

```
plausible for an ordinary person.
Strict rules:
- Transform the context from a historical
scenario to a private individual's scenario.
Keep the names and theme of the action but
adjust the role and context so it makes
sense for an ordinary person.
- Do NOT change any of the names.
- Do NOT explain your changes. Do Not
output reasoning content.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "original_text": "...",
  "synthesized_text": "..."
}
```

Prompt 2

```
You are a strict data synthesization
assistant.
Your task is to synthesize new
data by replacing every person name
with a random first and last name of a
private person.
Strict rules:
- Replace ALL names with a random private
first and last name. The new name must
match the original gender and ethnicity.
Do NOT use the same first or last name
twice.
- Be consistent: the same original name
must map to the same pseudonym within a
single text.
- Always generate NEW pseudonyms not used
before.
- Do NOT use the names Emily, Ethan, Harper,
Carter, Patel, Whitaker. Use more diverse
and uncommon names.
- Do NOT change anything else.
- Do NOT add explanations.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.3. Public person in private context

```
You are a strict data synthesization
assistant. Your task is to synthesize new
data by replacing every historical and
fictional name with a public person's name
and also transforming the scenario so it's
plausible for a non-historical context.
Strict rules:
- Replace the historical and fictional
names with a unique public name. The new name
must match the original gender and ethnicity.
- Names must be unique across the entire
dataset synthesization. If you used a name
once, do NOT use it again for other texts.
- Do not reuse any public person name that
has appeared in previous generations. Do not
use Emma Watson.
- Transform the context from a historical
scenario to a more modern and yet private
individual's scenario. Keep the theme of the
action but adjust the role and context so it
makes sense for the new public person.
- Do NOT explain your changes. Do Not output
reasoning content.
Return ONLY valid JSON in this format:
{
  "text": "...",
  "names_map": {
      "Original Name": "New Name"
  }
}
```

A.1.4. Public person in private context

```
You are a data synthesization
assistant.Your task is to synthesize new
data by replacing every historical and
fictional name with a public person's name
and also transforming the scenario so it's
plausible for a non-historical context.
Strict rules:
- Replace the historical and fictional names
with a unique public name. The new name must
match the original gender and ethnicity.
- Names must be unique across the entire
dataset synthesization. If you used a name
once, do NOT use it again for other texts.
- Do not reuse any public person name that
has appeared in previous generations. Do not
use Emma Watson.
- Replace each with a mononymous star name.
- Transform the context from a historical
scenario to a more modern and yet private
individual's scenario. Keep the theme of the
action but adjust the role and context so
it makes sense for the new public person.
- Do NOT explain your changes. Do Not
output reasoning content.
- Output ONLY the transformed text.
```

A.2. Prompts for classification

A.2.1. Prompt for detecting personal names

```
You are an expert in cybersecurity and
data privacy. Detect personal names of
public and non-public people.
```

```
Furthermore, treat historic characters
(died no later than 1925) and fictional
characters separately.

Return ONLY this JSON format:
{{
  "entity_type":
    "public_or_private_or_
     historic_or_fictional",
  "text": "NAME"
}}

{self.shot_examples[
  str(self.shots)
]["detect_pii"]}
```

A.2.2. Prompt for clustering personal names

```
Return only a JSON object. No explanation.
Cluster the a list of human personal names
from a text referring to one person.

Return ONLY this JSON format:
{{
    "NameVariant1":
    ["NameVariant1", "NameVariant2"]
}}

{self.shot_examples[
    str(self.shots)]
    ["cluster_pii"]}

        {prompt_body}
```

A.2.3. Prompt for classifying context

```
You are an expert in privacy and public
information classification.

Return ONLY this JSON:
{{
    "person": "{person_name}",
    "context": "PrivateContext"
                OR
                "PublicContext"
}}

{self.shot_examples[
    str(self.shots)]
    ["classify_context"]}

Text:
{text}

Person:
{person_name}
```

A.2.4. The different shot versions

```
{
"0": {
  "detect_pii": "",
  "cluster_pii": "",
  "classify_context": ""
},

"1": {
  "detect_pii": "\n        Example:\n
- \"Simone Biles won gold at the Olympics\"
  →\n
  {\"entity_type\":
    \"public\",
    \"text\": \"Simone Biles\"}\n",

  "cluster_pii": "\n        Example:\n
- \"Simone Biles won gold at the Olympics.
  Simone's the best gymnast.\"
  →\n
  {\"Simone Biles\":
 [\"Simone Biles\", \"Simone\"]}\n     ",

 "classify_context": "\n      Example:\n
  - \"Simone Biles\",
 \"Simone Biles won gold at the Olympics.\"
    →\n
    {\"person\": \"Simone Biles\",
 \"context\":\"PublicSetting\"}\n     "
},

"3": {
  "detect_pii":
  "Example:
- \"Simone Biles won gold at the Olympics.\"
  →
  {\"entity_type\":
    \"public\",
    \"text\": \"Simone Biles\"}
 - \"Marie Curie had two daughters,
 Irène and Ève.\"
 → {\"entity_type\":
    \"public\",
    \"text\": \"Marie Curie\"}
   {\"entity_type\":
    \"public\",
    \"text\": \"Irène\"}
  {\"entity_type\":
    \"public\",
    \"text\": \"Ève\"}
- \"John Doe's phone number is 555-1234.\"
 →
 {\"entity_type\":
    \"private\",
    \"text\": \"John Doe\"}",

  "cluster_pii":
  "Example:
```

```
  - \"Simone Biles won gold at the Olympics.
    Simone's the best gymnast.\"
    →
    {\"Simone Biles\":
    [\"Simone Biles\", \"Simone\"]}

    - \"Maria Salomea Skłodowska, or
   Madame Curie, earned two Nobel Prizes for
    her work in physics and chemistry, later
   becoming Prof. Curie and inspiring future
    scientists.\"
    →
    {\"Maria Salomea Skłodowska\":
    [\"Maria Salomea Skłodowska\",
    \"Madame Curie\", \"Prof. Curie\"]}

   - \"John Doe's phone number is 555-1234.
    She called him.\"
    →
    {\"John Doe\":
    [\"John Doe\"]}",

    "classify_context":
    "Example:
    - \"Simone Biles\",
  \"Simone Biles won gold at the Olympics.\"
    →
    {\"person\":\"Simone Biles\",
     \"context\": \"PublicSetting\"}

    - \"Marie Curie\", \"Marie Curie had two
      daughters, Irène and Ève.\"
      →
      {\"person\": \"Marie Curie\",
      \"context\": \"PrivateSetting\"}

   - \"John Doe\", \"John Doe's phone number
      is 555-1234.\"
      →
      {\"person\": \"John Doe\",
      \"context\": \"PrivateSetting\"}"
  }
}
```