Fill-in-the-Blanks: Automatic Generation and Evaluation of Language Models' Pseudonyms for English and Swedish Texts

Maria Irena Szawerna[*,1], **Jacob Lee Suchardt**[†,1,2]
[*]Språkbanken Text, SFS, University of Gothenburg
[†]Leipzig University & ScaDS.AI Dresden/Leipzig
maria.szawerna@gu.se, jacob.lee.suchardt@gmail.com

Abstract

While considerable effort has gone into developing solutions for detecting Personally Identifiable Information (PII) in linguistic data, less research has gone into automating the generation of appropriate pseudonyms and developing evaluation methods, both relevant for the creation of privacy-friendly language resources. We conduct pilot experiments using Masked and Generative Large Language Models to generate predictions for redacted PII-spans in a *cloze*-like fashion for English legal texts and parallel news articles in Swedish and English. Furthermore, we explore metrics for automatic evaluation of the generated pseudonyms in the legal data, and investigate the effect of part-of-speech constraints on performance. For the parallel, multilingual data, we contribute our manual PII-annotation and conduct a fine-grained error analysis across two of our pseudonym generation methods and a baseline. Our results illustrate the complexity of pseudonym evaluation and the particular challenge of automatic, at-scale evaluation as well as the models' tendency to predict prototypical and even stereotypical answers.

Keywords: privacy, pseudonymization, pseudonym generation, anonymization, privacy preservation, masked language models, large language models

1. Introduction

A major concern in the digital age is the protection of personal data and privacy, both from an ethical and a legal perspective. The European Union's General Data Protection Regulation (GDPR, Official Journal of the European Union, 2016) suggests pseudonymization as a way to mitigate privacy risks in research data. Therein, pseudonymization is defined as the systematic replacement of personal data with placeholders, where the mapping between the two is stored separately. The definition of pseudonymization provided by ISO/IEC 29100:2024 (en), in turn, describes it as the process of replacing Personally Identifiable Information (PII)[3] with an alias. Minimally invasive pseudonymization with minimal edits of the source data is necessary to allow privacy-friendly working with and sharing of language resources in many research fields (e.g. corpus linguistics, language documentation, L2 research, digital humanities, social sciences, research involving large quantities of medical and legal data, various NLP tasks, etc.; see e.g. Eder et al., 2019; Volodina et al., 2020; Blokland et al., 2020; Lison et al., 2021; Riabi et al., 2024). It can even potentially enable safe interactions with cloud-based Large Language Models (LLMs) (c.f. Hou et al., 2025).

Example 1, sourced from the Parallel Global Voices Corpus (Prokopidis et al., 2016), illustrates different types of placeholders that may substitute the PIIs (*Julie Smith*, *American*) in a sentence: e.g. numerical sequences (*0001*), placeholders derived from the removed PII's semantic category (*PERSON.01*), or a grammatically and semantically appropriate surrogate entity (*Jane Brown*). According to GDPR, these are valid pseudonymization methods, provided that the original data can be restored if needed. However, some sources only refer to the replacement with semantically and grammatically appropriate surrogates as pseudonymization (Eder et al., 2019), which is the kind of pseudonymization underlying our paper.[4]

(1) *Julie Smith, an American citizen [...]*
0001, an **0002** citizen [...]
PERSON.01, an **DEM.01** citizen [...]
Jane Brown, an **English** citizen [...]

Following Hou et al. (2025), the pseudonymization task can be separated into three discrete steps:

[1]These authors contributed equally to this work.
[2]The work was completed while the author was at FLoV, University of Gothenburg.
[3]The same source defines PII as "information that (a) can be used to establish a link between the information and the natural person to whom such information relates, or (b) is or might be directly or indirectly linked to a natural person" with a special subcategory of sensitive PII which consists of information of particularly sensitive or protected nature.
[4]Pseudonymization, anonymization, and de-identification are related terms whose interpretation varies between jurisdictions. See Lestyán et al. (2025) for a discussion of this issue.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1155–1169
11–16 May 2026.

i) PII detection in data, ii) PII surrogate (pseudonym) generation, and iii) substitution of the PII spans detected in the data with the pseudonyms.[5]

Some approaches eliminate this distinction, as in the case of Yermilov et al. (2023), where framing the task as a seq2seq translation was explored. Although much work has focused on PII detection (e.g. Szawerna et al., 2024; Ji et al., 2025; Savkin et al., 2025), the task of generating PII replacements and evaluating them has received less attention. For numerical or class-derived placeholders, generation is trivial once the PII spans and their semantic categories have been identified. In contrast, generating grammatically and semantically appropriate pseudonyms that preserve the integrity of the input texts – favorable in contexts where the text needs to remain readable, e.g. for machine learning purposes or qualitative research – has proven to be much more difficult (Lison et al., 2021) and is the focus of this paper.

Transformer-based Masked Language Models (MLMs), first introduced by Devlin et al. (2019), model language by learning to predict the most likely word from its context, making them uniquely suited for lexical substitution tasks (LST, Zhou et al., 2019; Arefyev et al., 2020). On the other hand, although Generative Large Language Models (GenLLMs), pioneered by Radford and Narasimhan (2018), rely on only unidirectional (causal) modeling to predict the next token, they excel in many tasks and have shown some promising results for LST (c.f. Shi et al., 2024; Dutilleul et al., 2024). LST is also highly reminiscent of *cloze* tests in language learning, and it is analogous to the pseudonym generation step in the presented pseudonymization framework once PIIs have been removed from the data and only gaps for pseudonyms remain.

Therefore, we propose to further explore MLMs' and GenLLMs' predictive capabilities to generate pseudonyms aimed at replacing PII as a dynamic, lightweight, and locally deployable solution, and compare them with rule-based baselines.

We conduct experiments in order to address the following research questions:

RQ1: When it comes to small (feasible to run locally) models, how good are MLMs and GenLLMs at predicting suitable pseudonyms from limited context and how do they compare with each other and naïve baselines?

RQ2: Can additional Universal Part-of-Speech (UPOS) information/constraints aid performance?

RQ3: How can the quality of generated pseudonyms be assessed automatically?

[5]The latter two steps are often conflated (e.g. Yermilov et al., 2023), while the first step can also be split into PII detection and PII labeling (cf. Volodina et al., 2020).

2. Prior Research

Rule- or dictionary-based pseudonym generation methods have been implemented for, e.g., medical texts in English (Sweeney, 1996; Simancek and Vydiswaran, 2024) and Swedish (Alfalahi et al., 2012; Dalianis, 2019; Vakili et al., 2024), English Wikipedia articles (Olstad et al., 2023), German e-mails (Eder et al., 2019), Swedish learner essays (Volodina et al., 2020), and Polish heterogeneous Internet material (Oleksy et al., 2021). However, rule-based approaches are inherently at least partially specific to the data for which they were developed, and therefore to the predefined entity types and most commonly contained semantic contexts. Furthermore, their manual design becomes increasingly challenging for PII extending beyond named entities (personal names, organizations) or numerical information (dates, phone numbers), such as specifically identifying events and descriptions of political opinions (Volodina et al., 2020), or highly context-dependent elements, such as family relations (Volodina et al., 2023). A comprehensive evaluation of the generated pseudonyms is rarely reported, as there are no agreed-upon evaluation metrics for this task. Eder et al. (2019) resorted to human judgments of grammaticality and acceptability, and a frequency analysis of the resulting vocabulary. Although positive results are reported, they are limited to the aforementioned structured PII categories and the approach is not easily scalable to other corpora.

Nikiforova et al. (2020) compared Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), and BERT-based language models' ability to predict the next word from a cloze-style context in Russian against a human-like model trained on human responses. Their findings indicate that BERT-based models achieve the highest lexical and Part-of-Speech (POS) accuracy. Other evaluation methods included object-verb-functional-modifier level accuracy and lexical overlap, while cosine similarity of centroid vectors across vocabularies, Kullback-Leibler divergence between de-cased model vocabularies, and Kolmogorov-Smirnov testing on the probability distributions of predictions were used to compare the outputs of different generation methods without their linguistic contexts.

Yermilov et al. (2023) tested rule-based replacement and GenLLMs for pseudonym generation in English. The task was restricted to `PERSON`, `LOCATION`, and `ORGANIZATION` categories and framed in three ways: i) Named Entity Recognition (NER)-detection plus pseudonyms from Wikidata entities with similar features, ii) sequence-to-sequence translation with BART, and iii) one-shot prompting GPT-3 to detect PIIs, followed by one-shot pseudonymization using GPT-3.5 and the extracted

entities. The baseline comprised off-the-shelf NER systems combined with rule-based replacements. The quality of the generated pseudonyms was evaluated using text syntheticity detection, i.e. training binary classifiers to distinguish between original and pseudonymized data, with low accuracy taken as an indication of high-quality pseudonymization. Although GPT-based pseudonymization pipelines performed best, NER approaches were the most consistent at obfuscating the original information.

Vats et al. (2024) explored pseudonym generation for English with Top-1 and Top-K (K=10) sampling strategies using off-the-shelf and in-domain fine-tuned BERT and RoBERTa, as well as fine-tuned Llama2 (only Top-1). PII definition and detection utilized i) a handcrafted list, ii) masking rarely appearing tokens, or iii) using a NER system, which leaves the extent of the pseudonymized classes underspecified. Baselines consisted of Oracle (trained on ground truth) and two pretrained (WikiText-103 corpus) language models that were further trained on the masked ground truth where the mask token was i) a special token, or ii) had zero weight in the loss function). To assess performance, the pretrained language model was fine-tuned on the pseudonymized data of three (text-based) datasets and perplexity was scored on an unmasked test set. Across three PII detection heuristics, scores were in favor of RoBERTa (especially fine-tuned) and Top-K selection.

Hou et al. (2025) compared multiple methods and combinations for PII detection, pseudonym generation, and replacement across various English datasets for PII entities of type `PERSON`, `LOCATION`, and `ORGANIZATION`. Two pseudonym generation methods were explored: i) Random sampling from the list of previously detected PII entities of the same semantic category,[6] and ii) prompt-based generation via Qwen2.5-1.5B-Instruct as a locally deployed, small-scale instruction-tuned LLM. The evaluation here is dataset-dependent (e.g. Rouge-1/2/L for XSum, accuracy for MNLI GLUE). Moreover, the resulting scores encompass the entire detect-generate-replace pipeline at once, and different trends emerge depending on the underlying dataset, limiting comparability to other approaches.

Notably, a recent paper by Madaan et al. (2026)[7] is quite similar to our work in its approach. In that paper, a multi-token completion system based on MLMs for English is proposed. The evaluation consists of calculating the accuracy of the NER labels predicted for the spans against the original annotation and exploring model perplexity. The authors also experiment with downstream tasks: text summarization and language modeling to evaluate the usefulness of the data with the inserted pseudonyms. They additionally report human judgments of PI leakage and which of their models' output was deemed to be the best and also assess domain generalizability. While their approach is much more focused on, and appropriate for, handling multi-token expressions using MLMs than our single-token predictions, their evaluation does not provide the same level of insights into the semantic and grammatical acceptability of the generated replacements, nor does it extend beyond English; finally, in contrast to the present paper, no comparison to GenLLMs or any baseline is provided.

In summary, automatic evaluation of pseudonym generation in isolation is rare, and there are no canonical evaluation methods for this step. Instead, the methods presented in, e.g., Yermilov et al. (2023); Vakili et al. (2024); Hou et al. (2025) hinge on the assumption that a PII detection subtask was involved in creating the pseudonymized texts underlying evaluation. As such, measures related to privacy-preservation make an appearance, too: Yermilov et al. (2023) measure the percentage of original entities that were leaked in the pseudonymized texts; Gardiner et al. (2024) introduce a "residual risk score" based on hand-annotating a sub-sample of the de-identified corpus with the entities missed during PII detection and quantifying how revealing they are, which could also be applied to information leaked during pseudonym generation.

Crucially, however, evaluation methods which presuppose PII detection as a subtask are not suitable for our research, which focuses exclusively on pseudonym generation. Comparison is further limited when results were obtained only for a specific subset of pseudonym categories, especially named entities, in contrast to less concrete demographic attributes or miscellaneous personal information (as in Yermilov et al., 2023; Eder et al., 2019). Lastly, in addition to human grammaticality and acceptability judgments for natural language pseudonyms (Eder et al., 2019), our interests include exploring automatic evaluation methods to assess performance at scale.

3. Resources

3.1. Data

We draw data from the Text Anonymization Benchmark corpus (TAB, Pilán et al., 2022, MIT License) and Parallel Global Voices (PGV, Prokopidis et al., 2016, CC BY 4.0). The TAB corpus consists of court proceedings from the European Court of Hu-

[6] Exclusively reusing PII entities, even outside of the original context, may not sufficiently negate re-identification risks.

[7] As of the camera-ready deadline for the present article it is only available as a pre-print and due to be published in the proceedings of EACL 2026.

man Rights in English, annotated with PII entities as text spans, with the goal of protecting one person per text. PIIs fall into one of 9 categories (Pilán et al., 2022): `PERSON` (names), `CODE` (numbers and identification codes), `LOC` (places and locations), `ORG` (organization names), `DEM` (sociodemographic attributes), `DATETIME` (dates, times, or durations), `QUANTITY` ("meaningful" quantities), `MISC` (other potentially sensitive information not included in previous categories). We use the "quality checked" subset of the corpus' training split, from which we excluded 24 samples due to inconsistencies in annotation, yielding a total of 306 texts (15,494 PIIs; 13,051 sentences). On average, TAB documents consist of 43 sentences (range 12–121; SD: 25.6) and include 51 PII spans (range 10–165; SD: 32.6).

The PGV corpus contains news articles from the Global Voices website in multiple languages, including 336 pairs of parallel texts in Swedish and English, some of which contain information about specific individuals. From these, we chose 10 pairs (henceforth PGV-PII, Szawerna and Suchardt, 2025) containing numerous PII-like elements, which we further hand-annotated using the *doccano* (Nakayama et al., 2018) annotation software, amounting to a total of 720 PII spans for the Swedish texts, and 708 for English. Our annotation followed the TAB corpus annotation scheme but does not include information about the degree of information sensitivity, and only includes entities deemed strictly necessary to mask. On average, each of the 10 PII-annotated PGV documents in Swedish consists of 59 sentences (range 28-113, SD: 25.9) with 71 PII spans (range 31-153, SD: 32.9), whereas their English counterparts average 53 sentences (range 27-89, SD: 22.3) and 72 PII spans (range 32-155, SD: 33.1).

3.2. Models

Following the findings of Nikiforova et al. (2020) and Vats et al. (2024), we focus on the RoBERTa-large (henceforth **RoBERTa**) architecture to explore the suitability of **MLMs** for pseudonym generation with maximal comparability between our two object languages. For our bilingual data, the AI-Sweden-Models/roberta-large-1160k (AI Sweden) model is equivalent to the English FacebookAI/RoBERTa-large (Liu et al., 2019) in terms of architecture (355M parameters) and model weights upon initialization. The English model was trained on 160GB of text; the Swedish model was trained on a subset of the Nordic Pile (1.3TB) with 414M text samples (subset size not provided) and the tokenizer was retrained from scratch. We employ the models from HuggingFace with the Transformers' `fill-mask` pipeline (Wolf et al., 2020). At this point, we rule out the option of fine-tuning the models since domain-specific fine-tuning data may not be available, and fine-tuning on in-house data risks leakage of sensitive data during deployment. In addition to data, both computational resources and technical knowledge are prerequisites for fine-tuning. By using vanilla MLMs, we aim to explore their capabilities as "universal" and light-weight models for pseudonym generation, which would facilitate access for, e.g., non-computational linguists and digital humanities researchers.

To assess the capabilities of small, locally deployed **GenLLMs**, we focus on Qwen/Qwen2.5-1.5B-Instruct-GGUF (henceforth **Qwen**, Qwen et al., 2025). This model was also chosen in Hou et al. (2025) and had been the most reliable during our initial performance assessment which included two similarly sized models, unsloth/gemma-3-1b-it-GGUF and unsloth/Llama-3.2-1B-Instruct-GGUF. Qwen was accessed via Ollama (without fine-tuning, for the aforementioned reasons) and was prompted through the same Python code that utilized RoBERTa with the LangChain framework's `langchain_ollama` package. Qwen is marketed as multilingual and was thus used for both English and Swedish data after initial testing.

4. Methods

We explore pseudonym generation capabilities of MLMs compared to GenLLMs and random baselines in a cloze-like setting on data from the TAB (English) and PGV corpora (English and Swedish).[8] To obtain the input contexts for the models, documents are split into sentences and the pre-annotated PII spans are replaced with a mask token. Annotated PII spans are naïvely replaced with a singular mask token, although some spans (77% in TAB; 46.4% (English) and 45% (Swedish) in PGV) encompass multi-word PIIs (MW-PII, e.g. "Mr André Borgers").

We construct two baselines from the vocabulary of word forms retrieved from Universal Dependencies' English and Swedish PUD (Nivre et al., 2020) to contextualize the language models' performance. For a given PII, the **random baseline** returns a pseudonym in the form of a random entry from its inventory, while the **UPOS baseline** samples an entry which has the same UPOS as the original PII. UPOS tags are given in PUD; otherwise, we tag with Stanza (Qi et al., 2020). UPOS tags for MW-PII are based on their right-most constituent. For pseudonym generation with language models, the MLMs and GenLLM receive a target sentence with at least one mask token, as well as the pre- and succeeding two sentences as context, since preliminary experiments showed that larger context improves semantic coherence. Given the masked

[8] We provide the code and the PII-annotated PGV corpus (under CC BY 4.0) in Appendix A.

input, retrieving the pseudonyms from RoBERTa's output is trivial. For Qwen, mask tokens additionally receive a numeration in order of appearance. The model is prompted to fill all the gaps and return a JSON mapping between the mask tokens and the generated pseudonyms.[9] In the cases where i) Qwen fails to return a valid JSON after several attempts, or ii) the returned JSON does not contain a key that matched the numbered `mask` token in question, a `[FAILED]` token is entered instead.

We explore two factors during pseudonym generation with language models: First, different **PII masking strategies** for handling the PII spans in the input context are implemented: i) all PII spans are masked (maskL), or ii) the model's previous predictions are used to replace the PII spans in the context to the left of the currently to be pseudonymized mask token (repL). In the maskL condition, a target sentence with multiple mask tokens can be pseudonymized in one step by either model but during repL RoBERTa receives a novel input after each pseudonym is selected, since we update the left hand context with the previous prediction (example in Appendix B). Similarly, the GenLLM is re-prompted with the updated input after each step. Second, we explore enriching the pseudonym generation with **UPOS information**: The Top-k (k=10) MLM predictions are filtered for the first candidate with the desired UPOS tag. When no suitable candidate was found or UPOS filtering is disabled, the Top-1 prediction is chosen instead. For Qwen, the UPOS information is supplied as a tag on the numerated mask token, e.g. `[MASK.NOUN.5]`. This yields 4 pseudonym generation configurations per model (2 masking strategies x 2 UPOS settings).

Experiment 1 Using the baselines, English RoBERTa-large, and Qwen, we generate replacements for the spans tagged as requiring masking in the TAB corpus and perform a quantitative analysis of their linguistic diversity. We also explore the following as automated evaluation metrics: i) PII overlap (partial/absolute string match between pseudonym and original PII, excluding numerals), ii) UPOS accuracy, iii) word and iv) sentence cosine vector similarity of the original word/sentence and pseudonymized version with all-MiniLM-L6-v2, v) text classification after Yermilov et al. (2023) using google-bert/bert-base-cased, and vi) Rouge-1/2/L/Lsum (Lin, 2004) on automatic text summaries of the original and pseudonymized texts, generated with SEBIS/legal_t5_small_summ_en.

[9]While restricting the input size was necessitated by RoBERTa's predefined maximum input size, the process was adopted for Qwen as well to increase comparability and because the GenLLM frequently failed to produce appropriate outputs when one-shot prompted to pseudonymize entire documents at once.

Tag	Error source
NO_ERROR	None (acceptable pseudonym)
OVERLAP	Partial or complete PII leakage
GRAMMAR	Sentence rendered ungrammatical
POS	Part of speech
STRUCTURAL	Punctuation, word segmentation, or replacement of long spans
SEM:CATEGORY	Semantic category
SEM:CONTRADICTORY	Contradicts established narrative or common knowledge
SEM:MERGED-ENTITY	Collapses two original entities
FAILED	Qwen model failure

Table 1: Manual error annotation tagset.

Experiment 2 Following the evaluation of Experiment 1, we limit our generation methods for the PII-annotated PGV corpus in Experiment 2 to the most promising configurations per category – UPOS baseline, RoBERTa repL + UPOS, and Qwen maskL + UPOS – to allow for a more fine-grained analysis with a manual error annotation. The error annotation tagset is shown in Table 1 and the corresponding annotation guidelines in Appendix C. We report relative frequencies of error categories per model and discuss selected examples.

5. Results and Discussion

5.1. Experiment 1

Table 2 reports on the word/phrase form frequencies of the pseudonym vocabularies. Measures based on the original data are also included and show that the TAB documents include many unique spans (8074), i.e. word forms/phrases which appear only once across the collection of all PII spans. Nevertheless, most unique spans occur at least twice — likely a byproduct of narrative cohesion. In contrast, our pseudonym generation methods lead to a lower rate of unique spans: The random baseline shows the highest diversity, followed by Qwen, RoBERTa and finally the UPOS-constrained baseline. UPOS-constraints increase diversity in RoBERTa, whereas for Qwen and the baseline they lower it. Manual inspection confirms that the models tend to output more generic replacements which are re-used across the texts, e.g. *Monday*, *appeal*, or numbers (*2, 20, 1994*) or function words (*the, after*), while the most frequent original PII (though with a much lower count) is *Istanbul*.

Figure 1 illustrates the ratios and mean of unique pseudonym spans to the total number of PII spans per document, with 1 indicating that all replacement spans are unique and 0 that none of them are. At ∼0.8, original TAB documents typically exhibit a higher uniqueness ratio than ones pseudonymized by the language models but they do not reach the extreme values of the

		Span form frequency				unique
		avg.	std	min	max	spans
Original		2	3	1	124	8074
Baseline	random	3	2	1	14	5346
	UPOS	32	49	1	326	483
RoBERTa	maskL	11	74	1	1647	1423
	+UPOS	10	63	1	1510	1485
	repL	11	63	1	1587	1459
	+UPOS	10	48	1	1087	1503
Qwen	maskL	4	28	1	938	3485
	+UPOS	6	41	1	1289	2684
	repL	6	25	1	603	2507
	+UPOS	7	44	1	1120	2107

Table 2: PII and pseudonym word/phrase form diversity in the TAB corpus/Experiment 1.

random baselines (∼1); yet, there are many outliers. RoBERTa pseudonymization strategies average around the 0.6 mark, whereas typically Qwen ranges from 0.6 (repL + UPOS) to 0.8 (maskL) – closely resembling the original. The larger interquartile ranges of the language model configurations suggest that some documents received much less diverse pseudonyms, likely contributing to the high max. frequency counts seen in Table 2 previously. However, the discussion here must take into account that the TAB annotation, and thus our MLM pipeline design, treats MW-PIIs as a singular span and predicts only one token as its replacement, which reduces the variety that the configurations here could produce. In contrast, 30–50% of Qwen pseudonyms can be considered MW-pseudonyms (maskL: 46%, maskL + UPOS: 28%, repL: 42%, repL + UPOS: 29%).

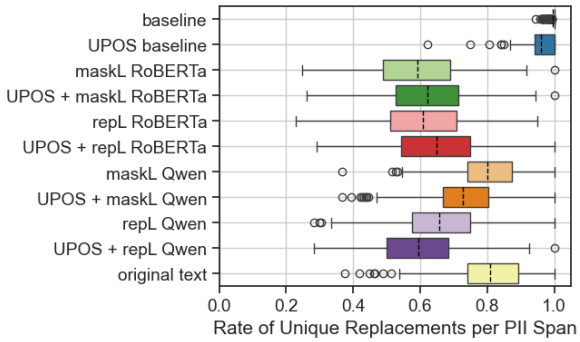

Figure 1: Rate of unique PII and replacements per pseudonymization configuration on document level.

Table 3 displays the UPOS accuracy, PII overlap rate (partial/absolute), and cosine vector similarity (word/sentence) for the baselines and language model configurations on the TAB data. In terms of UPOS accuracy, all language model configurations improve on the random baseline (11.5%) by typically 15–45 points, but a distance to the UPOS baseline's performance (93.4%, missing 7% due to tagger inconsistencies/errors) remains.

Model	Mode	UPOS	PII overlap %		Similarity	
		acc.%	part.	abs.	word	sent.
Baseline	random	11.5	**0.07**	**0**	0.152	0.730
	upos	**93.4**	1.15	0.02	0.229	0.740
RoBERTa	maskL	26.7	7.18	1.30	0.295	**0.805**
	+UPOS	50.2	7.99	1.55	0.320	0.800
	repL	28.1	6.48	1.09	0.295	0.803
	+UPOS	55.3	7.60	1.37	0.330	0.798
Qwen	maskL	34.2	1.70	0.20	0.324	0.785
	+UPOS	67.9	17.06	0.12	0.286	0.783
	repL	39.6	2.56	0.20	**0.340**	0.793
	+UPOS	62.3	13.55	0.41	0.308	0.787

Table 3: UPOS accuracy (acc.), absolute (abs.)/partial (part.) overlap, and similarity scores for Experiment 1.

Qwen maskL + UPOS (67.9%) and RoBERTa repL + UPOS (55.3%) pseudonyms align most frequently with the original PII's UPOS tag. Among configurations without UPOS constraints, Qwen repL (39.6%) leads. The remaining RoBERTa and Qwen configurations reach around 30% UPOS accuracy (range 26.7–34.2%). UPOS constraints also raise the likelihood of partial PII overlap: Although Qwen maskL predictions exhibit partial overlap with the original PII string in just 1.7% of cases (close to the UPOS baseline of 1.15%), maskL + UPOS reaches 17.06%. RoBERTa configurations exhibit partial overlap in 6–8% of cases. However, the prediction of pseudonyms identical to the original PII (absolute overlap) is generally rare (<1.6%) across all conditions.

Word similarities show diverging trends between generation methods: UPOS filtering increases similarity for RoBERTa and the baselines, but diminishes it for Qwen. For both language models, repL still leads to higher similarity scores than the respective maskL conditions. In the UPOS condition, RoBERTa repL achieves the highest overall word similarity (0.33) while the unconstrained Qwen repL (0.34) reaches the overall highest word similarity. Although UPOS accuracy and word similarity appear to positively correlate for baseline and RoBERTa configurations, Qwen reveals a contradictory trend. In comparison, the sentence similarity scores are generally higher but exhibit a smaller range due to the inherently high overlap of the sentences' non-PII tokens. Here, UPOS constraints lower similarity scores for language models (avg. -0.005 points) but not the baseline (+0.01 points). RoBERTa conditions without UPOS constraints achieve the highest sentence similarities (maskL: 0.805). While both language models outperform the baselines, RoBERTa also yields higher sentence similarity scores than Qwen, overall. Notably, the Qwen configurations with a higher MW-pseudonym rate, i.e. those without UPOS constraints, do not seem to be at a particular disadvantage in this metric.

Table 4 reports the average Rouge scores based

Model	Mode	ATS Rouge			
		1	2	L	Lsum
Baseline	random	0.4172	0.2497	0.3689	0.368
	upos	0.4261	0.2735	0.3796	0.3796
RoBERTa	maskL	0.4151	0.2485	0.3655	0.3655
	+UPOS	0.4453	0.2859	0.3967	0.3961
	repL	0.4256	0.2591	0.3716	0.3715
	+UPOS	**0.4547**	**0.2951**	0.4056	**0.4061**
Qwen	maskL	0.3986	0.2315	0.3534	0.3541
	+UPOS	0.435	0.268	0.3851	0.3849
	repL	0.4369	0.2763	0.3871	0.3867
	+UPOS	0.4538	0.2945	**0.4063**	0.4056

Table 4: Average Rouge-1/2/L/Lsum scores of baselines and selected RoBERTa/Qwen configurations.

on the automatic text summarization of the pairs of original and pseudonymized texts. Our initial assumption was that this could be a measure of semantic coherence of the produced text, as semantically inappropriate pseudonyms may sufficiently disrupt the text to significantly affect the generated summary. Overall, UPOS constraints appear to have a positive effect relative to their unconstrained counterparts, and repL conditions perform better than maskL for both RoBERTa and Qwen. However, the differences between the baselines and pseudonymization conditions appear to be minute and Qwen maskL is even supposedly outperformed by the random baseline.

We cannot directly compare these results to those obtained by Madaan et al. (2026), who treated summarization of pseudonymized texts as a way of evaluating downstream usefulness of the data, as they likely calculated the score against a ground truth summary. Nevertheless, we observe similar (or even higher) scores – which, interestingly, also applies to our baseline. Given what we discussed above, we must conclude that this assessment method does not appear to be sufficiently sensitive to the differences in generated pseudonyms to serve as an evaluation metric or a way to test downstream usability.

Distinguishing original and pseudonymized texts was trivial for our binary classifier (100% accuracy). We do not attribute this outcome to overfitting, since comparable results were obtained when training for as little as one epoch. Rather, due to the partially parallel structure of the documents, there are likely some definitive giveaways in the pseudonymized documents. For example, the leading legal case number was frequently replaced with *1* or *2* by RoBERTa, and *no.* (without UPOS restriction), or *0*, *1*, *20* (with UPOS restriction) by Qwen – much less complex tokens than the original ones, e.g. *18407/91*. While Yermilov et al. (2023) also report very high accuracy scores for some of their approaches, their LLM-based solutions seemed harder to detect. However, unlike Yermilov et al. (2023), we did not only pseudonymize PERSON, LOCATION, and ORGANIZATION, but all of the PII categories according to Lison et al. (2021), including more nebulous categories such as DEM or MISC, increasing the difficulty of the task of convincingly replacing the PII spans.

Out of the explored evaluation methods, UPOS accuracy, PII overlap, and cosine word similarity seem to be most informative, and better grammaticality and semantic acceptability measures are needed assess pseudonym quality. We therefore focused on the the first three measures (specifically *absolute* PII overlap) alongside the linguistic diversity (as per Figure 1) when selecting the most promising models. Second, the inherent time cost of RepL conditions (especially for Qwen) also had to be taken into consideration.

5.2. Experiment 2

A summary of the relative frequencies of acceptable pseudonyms and various types of errors per language and selected model used for pseudonymization of PGV-PII can be found in Figure 2. Contrary to the stronger resemblance of Qwen maskL + UPOS pseudonyms to the original texts in terms of linguistic diversity and its higher UPOS accuracy in Experiment 1, the RoBERTa repL + UPOS approach displays a noticeably higher rate of acceptable pseudonyms than Qwen in Experiment 2: RoBERTa oscillates around the 30% mark for both English and Swedish, while Qwen reaches 20.6% for English and just 16.4% in Swedish – similar to the UPOS baseline's performances (∼15.5%).

The most common error category for all approaches is SEM:CATEGORY. This error type was assigned to ∼50% of the pseudonyms suggested by the baseline, by far exceeding the rate present for RoBERTa and Qwen (25-30%). In turn, though POS errors are generally less common due to the shared constraint, the baseline produces these the least. The OVERLAP error type continues to be almost non-existent in the baseline and even the Qwen conditions now. While it remains a concern for RoBERTa, especially for English (around 5%), additional safeguards in the form of filters can be implemented for this model easily. RoBERTa also struggles the most with semantic errors of the types CONTRADICTORY and MERGED-ENTITY, indicating that this approach does not track entities reappearing in the text on its own and may benefit from such an additional constraint or an increase in supplied context. To produce a MERGED-ENTITY error, an entity must be named at least twice. Thus, this type is hardly produced by the baseline, which is in line with the linguistic diversity observations from Figure 1 where both baselines were unlikely to predict a given pseudonym more than once per document. For all approaches, issues with grammatical-

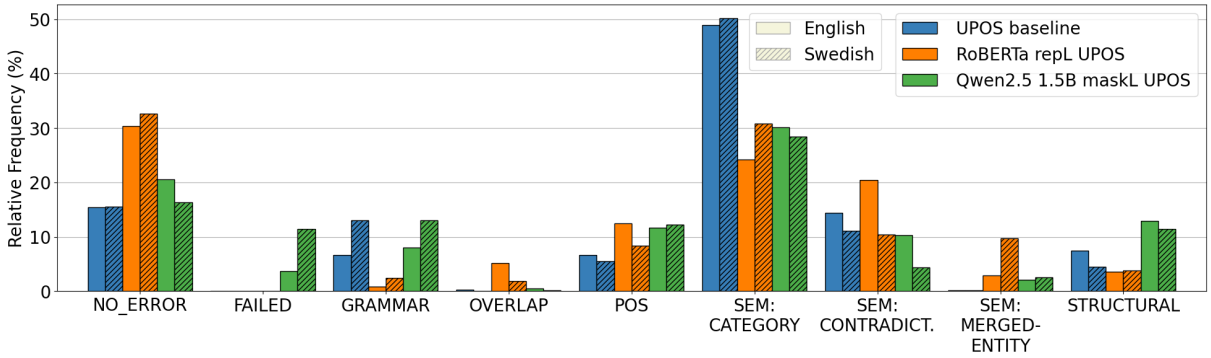

Figure 2: Relative frequencies of pseudonym annotation types per generation method and languages.

ity are more frequent in Swedish than English (e.g. "In several *remarkably countries it has been prohibited", Qwen #395/gv-eng-swe-66_ENG.json). Possibly, this is due to the more complex morphological agreement constraints of Swedish where, e.g., adjectives and nouns must agree in gender, number, and definiteness. Grammatical errors are least frequent for RoBERTa and roughly equally common for the baseline and Qwen (English: ∼7%, Swedish: 13%). While Qwen et al. (2025) do not specify whether the model was trained on Swedish data or not, it was able to generate text and pseudonyms in Swedish during initial testing; nevertheless, in comparison, its performance for that language appears worse than for English. This can also be noted in the case of `NO_ERROR` pseudonyms, where Qwen most noticeably underperforms in Swedish relative to English. Qwen also produces more `STRUCTURAL` errors than RoBERTa or the UPOS baseline, especially due to instances where it provides unrelated spans found earlier in the input text as suggested replacements. Moreover, Qwen lead to a unique error case of `FAILED`, where ≤5 (English) to ≥10% (Swedish) of PII spans are not provided a pseudonym by the model.

Certain prominent trends pertaining to the predictions are highly relevant from an ethical viewpoint. For instance, for RoBERTa, both in English and in Swedish texts, *Trump* is a likely prediction, especially in contexts that involve Twitter or presidency, e.g. *ber President Trump om förklaringar* 'asks President Trump for explanations' or *On Twitter, Donald Trump (realDonaldTrump) was enraged by the news*. The surname itself appears 30 times among the 1428 pseudonyms generated by this approach, creating more monotonous narratives and introducing a US-centric bias. We presume the cause lies in the model's training data.

Another major issue is featuring certain information or attributes in a context that can be stereotypical or otherwise undesirable by the respective group. For example, gender stereotypes are perpetuated in RoBERTa instances like *hennes pappa är taxichaufför på deltid och hennes mamma hemmafru* 'her father is a part-time taxi driver and her mother is a housewife' being pseudonymized with *läkare* 'doctor' and *sjuksköterska* 'nurse'. In the case of Qwen, the corresponding replacements *försvarsmann* '*army man, *defender' and *kommunikator* 'communicator' are misspelled (*försvarsman*, *kommunikatör*), and the former, while possible from the derivational perspective, is not an attested Swedish lexeme describing a profession. These implied meanings could still be perceived as adhering to gender stereotypes. This is less prominent in English, where RoBERTa suggests *teacher* and *nurse*, whereas Qwen — *engineer* and *doctor*. The Swedish examples here not only underline the need to keep these issues in mind when using either language model for pseudonym generation, but also serves to probe and indicate — perhaps unsurprisingly — the existence of certain biases in the models. In both instances, replacements from the baseline are predictably nonsensical (*fred* 'peace', *process* 'process'; *levies*, *office*).

6. Conclusions and Future Work

We have implemented a simple pseudonym generation pipeline with MLMs and LLMs, and tested it on a large corpus of English legal texts (TAB) as well as on a small parallel corpus of news articles in English and Swedish (PGV-PII) using the respective RoBERTa-large models, Qwen2.5-1.5B-Instruct, and a random baseline, with and without UPOS constraints. For PGV-PII, we selected texts from PGV and contribute our own manual PII annotations. While methods for pseudonymization are sought after, automatic and scalable evaluation of their outcomes is still underdeveloped. To this end, we analyzed the linguistic diversity of the English pseudonyms generated for TAB in Experiment 1 and explore automated evaluation metrics. Additionally, using the pseudonyms generated for Swedish and English PGV-PII by RoBERTa, Qwen, and the UPOS baseline, we undertook a qualita-

tive analysis akin to the judgments in Eder et al. (2019), but prioritized understanding the nature of error rather than the degree of bad fit.

In Experiment 1, we found that most of the models seem to generate less diverse pseudonyms compared to the original PIIs, both at the corpus- and document-level, with certain GenLLM setups approaching the original diversity, and the naïve baselines far exceeding it. To enhance diversity and reduce stereotypical replacements, the introduction of uniqueness criteria or the refinement of our treatment of MW-PII could be considered. Although scores for the quantitative metrics in Experiment 1 were similar, and the Qwen configuration selected for Experiment 2 had the highest UPOS accuracy with a seemingly more appropriate linguistic diversity, the qualitative analysis in Experiment 2 was in favor of RoBERTa: The chosen RoBERTa configuration generated reasonable pseudonyms around 30% of the time, while Qwen and the UPOS baselines performed worse (15-20%). Semantic vector similarity measures may thus be a better indicator of pseudonym quality than UPOS accuracy measures. Generally, inclusion of UPOS constraints improved both the diversity measurements, and the quantitative metrics compared to the respective unconstrained counterparts. Regarding the increased risk of PII overlaps and lowered sentence similarity, the former can easily be avoided with an additional filter on the Top-K RoBERTa predictions, while sentence similarity may not necessarily be a strong indicator of pseudonym quality.

Other possibilities to improve our pipeline include enforcing or providing the replacement's original PII category type or adding co-reference resolution to improve semantic coherence. Combining our findings with the multi-word expression capabilities of Madaan et al. (2026) could significantly improve the effectiveness of the approach. Due to the competitive performance, predictable format of outputs, and consistency in behavior, future work may focus on MLMs in particular as a promising but relatively lightweight solution for automatic pseudonym generation which allows for controllable integration of such additional heuristics. Nevertheless, the role of GenLLMs could be investigated in future research as a more resource-heavy yet potentially better-performing alternative for some languages. While their performance was not outstanding in our experiments, it is worth noting that we only tested a small (1.5B) GenLLM, and larger models' capabilities are left unexplored for future research.

On the matter of evaluation measures, we note the lack of automatic equivalents for semantic acceptability judgments and highlight the need for testing grammaticality as well. We further posit that sentence similarity cosine scores and comparing automatic summaries of the original texts against the pseudonymized texts were not informative.

Finally, we emphasize the necessity of addressing the issues of the models' demonstrated proclivity towards displaying e.g. gender bias in subsequent work.

Acknowledgments

The work of the first author was conducted within the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* with the funding number 2022-02311 for the years 2023-2029. She was also supported by *Språkbanken*, which is jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161).

Limitations

As mentioned in section 4, our pseudonym generation pipeline makes a number of naïve assumptions, limiting its potential performance — we do, however, still consider it worth testing as a proof of concept in our pilot experiments.

A quantitative analysis akin to Experiment 1 was not conducted for the pseudonyms from Experiment 2 due to the limited size of this second dataset. In turn, this was a consequence of the time-intensive nature of both manual PII annotation and manual error annotation. For the qualitative error analysis, 20 texts had to be annotated for three generation conditions (3 x 720 (Swedish), 3 x 708 (English) pseudonyms). Since PII-annotated corpora are rare to begin with, no pre-existing parallel corpus was available to us. Although the TAB corpus had previously been automatically translated to Spanish, with the annotations projected back (Sierro et al., 2024), the authors of that experiment obtained better results after a manual correction of the projected annotation. Thus, manual effort became a prerequisite and our preference was not to rely on machine-generated translations. However, our restricted corpus size still limits comparability, and the results may not necessarily be generalizable, especially given that only one domain (news) was tested.

We acknowledge that by varying both the language of the texts (Swedish vs. English) and the model (Swedish RoBERTa vs. English RoBERTa), it cannot be ruled out that differences in performance could be inherent to the models and not necessarily language-related. Nevertheless, since these particular models were trained with these languages in mind, we considered it more beneficial to use their specialized semantic representations rather than to only use one multilingual model. The same critique applies to the GenLLM, however, no

similarly equivalent Swedish model was known to us.

The results reported here were not directly compared to those from prior work. As mentioned in section 1, to our knowledge, no common evaluation metrics for pseudonym generation exists. The judgments of Eder et al. (2019) are the closest to our evaluation of Experiment 2, but focus more on degrees of grammaticality and acceptability, whereas we were more interested in the trends concerning why a certain pseudonym is not a good replacement, resulting in considerable differences between the approaches.

Ethical Considerations

One of the major concerns when it comes to using larger language models for tasks is how resource-heavy they are (see e.g. Bender et al. (2021) or Rillig et al. (2023)). With that in mind, we have to acknowledge that the pseudonym generation with MLMs constitutes a slow process, the speed of which decreases with the size of the model used and the size of the input and its context. Generation using a GenLLM as small as the one we used (1.5B) is slower still. Especially in the case of generating pseudonyms one at a time and replacing the PII in the left context with the previously generated pseudonym (repL).

Any topic related to handling personal information is inextricably connected to ethical considerations. Even though our experiments presuppose a stage where all of the PII has been detected and labeled, the models are sometimes still able to leak back the original PII. In this case, especially synonyms, which are more difficult to detect, may be dangerous. Therefore, we highly recommend employing additional measures and maintaining a human-in-the-loop approach to ensure no accidental information leaks.

We have also shown that the models are capable of generating highly stereotypical content, potentially undesireable from the perspective of the group in question. This is another hurdle that needs to be overcome if this approach is to be used for pseudonymizing texts on a larger scale, though it is not specific to this application. Simply selecting the most likely prediction as the pseudonym is also expected to reinforce existing biases or even potentially introduce new ones — as for the latter case, we have observed that the models consistently situate the narratives in the Swedish (Swedish RoBERTa) or US American (both models) context; while on the document level this is not problematic, it reduces the diversity and representation on the corpus level if left unmanaged. Additionally, it means that on a larger scale, PII appearing in very similar or identical contexts, e.g. as part of structured documents created from a template, will be pseudonymized identically, with no room for diversity.

7. Bibliographical References

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012*.

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. A comparative study of lexical substitution approaches based on neural language models.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Rogier Blokland, Niko Partanen, and Michael Rießler. 2020. A pseudonymisation method for language documentation corpora: An experiment with spoken Komi. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–8, Wien, Austria. Association for Computational Linguistics.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Dutilleul, Mathis Debaillon, and Sandeep Mathias. 2024. ISEP_Presidency_University at MLSP 2024 shared task: Using GPT-3.5 to generate substitutes for lexical simplification. In *Proceedings of the 19th Workshop on Innovative*

Use of NLP for Building Educational Applications (BEA 2024), pages 605–609, Mexico City, Mexico. Association for Computational Linguistics.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria. INCOMA Ltd.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. 2024. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 64–75, St. Julian's, Malta. Association for Computational Linguistics.

Shilong Hou, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. 2025. A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation.

ISO/IEC 29100:2024(en). 2024. Information technology — Security techniques — Privacy framework. Standard, International Organization for Standardization, Geneva, Switzerland.

Zilyu Ji, Yuntian Shen, Kenneth R. Koedinger, and Jionghao Lin. 2025. Enhancing the de-identification of personally identifiable information in educational data. *Journal of Educational Data Mining*, 17(2).

Szivia Lestyán, William Letrone, Ludovica Robustelli, and Gergely Biczók. 2025. Anonymity-washing. ArXiv:2505.18627 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Pulkit Madaan, Krithika Ramesh, Lisa Bauer, Charith Peris, and Anjalie Field. 2026. Multi-token completion for text anonymization.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Anastasia Nikiforova, Sergey Pletenev, Daria Sinitsyna, Semen Sorokin, Anastasia Lopukhina, and Nick Howell. 2020. Language models for cloze task answer generation in Russian. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 28–37, Marseille, France. European Language Resources Association.

Official Journal of the European Union. 2016. Consolidated text: Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). *Official Journal*, (Document 02016R0679-20160504).

Marcin Oleksy, Norbert Ropiak, and Tomasz Walkowiak. 2021. Automated anonymization of text documents in Polish. *Procedia Computer Science*, 192:1323–1333.

Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of replacement options in text sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300, Tórshavn, Faroe Islands. University of Tartu Library.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

Matthias C. Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A. Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466. PMID: 36821477.

Maksim Savkin, Timur Ionov, and Vasily Konovalov. 2025. SPY: Enhancing privacy with synthetic PII detection dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*

4: Student Research Workshop), pages 236–246, Albuquerque, USA. Association for Computational Linguistics.

Ning Shi, Bradley Hauer, and Grzegorz Kondrak. 2024. Lexical substitution as causal language modeling. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 120–132, Mexico City, Mexico. Association for Computational Linguistics.

Maria Sierro, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of two different annotation approaches. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 18–24, St. Julian's, Malta. Association for Computational Linguistics.

Dalton Simancek and VG Vinod Vydiswaran. 2024. Handling name errors of a BERT-based de-identification system: Insights from stratified sampling and Markov-based pseudonymization. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 1–7, St. Julian's, Malta. Association for Computational Linguistics.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 333–7.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024. Detecting personal identifiable information in Swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63, St. Julian's, Malta. Association for Computational Linguistics.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical bert models. *BMC Med Inform Decis Mak*, 24(162).

Arpita Vats, Zhe Liu, Peng Su, Debjyoti Paul, Yingyi Ma, Yutong Pang, Zeeshan Ahmed, and Ozlem Kalinli. 2024. Recovering from privacy-preserving masking with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10771–10775.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Elena Volodina, Simon Dobnik, Therese Lindströ m Tiedemann, and Xuan-Son Vu. 2023. Grandma Karl is 27 years old – research agenda for pseudonymization of research data. pages 229–233. IEEE Computer Society.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

8. Language Resource References

AI Sweden. *AI-Sweden-models/Roberta-large-1160K*.

Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.

Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajič, Jan and Manning,

Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel. 2020. *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*. European Language Resources Association.

Pilán, Ildikó and Lison, Pierre and Øvrelid, Lilja and Papadopoulou, Anthi and Sánchez, David and Batet, Montserrat. 2022. *The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization*. MIT Press.

Prokopidis, Prokopis and Papavassiliou, Vassilis and Piperidis, Stelios. 2016. *Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories*. European Language Resources Association (ELRA).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Qwen and Yang, An and Yang, Baosong and Zhang, Beichen and Hui, Binyuan and Zheng, Bo and Yu, Bowen and Li, Chengyuan and Liu, Dayiheng and Huang, Fei and Wei, Haoran and Lin, Huan and Yang, Jian and Tu, Jianhong and Zhang, Jianwei and Yang, Jianxin and Yang, Jiaxi and Zhou, Jingren and Lin, Junyang and Dang, Kai and Lu, Keming and Bao, Keqin and Yang, Kexin and Yu, Le and Li, Mei and Xue, Mingfeng and Zhang, Pei and Zhu, Qin and Men, Rui and Lin, Runji and Li, Tianhao and Tang, Tianyi and Xia, Tingyu and Ren, Xingzhang and Ren, Xuancheng and Fan, Yang and Su, Yang and Zhang, Yichang and Wan, Yu and Liu, Yuqiong and Cui, Zeyu and Zhang, Zhenru and Qiu, Zihan. 2025. *Qwen2.5 Technical Report*. arXiv. ArXiv:2412.15115 [cs].

Maria Irena Szawerna and Jacob Lee Suchardt. 2025. PGV-PII.

Type	Input	Prediction
left context	The case originated in an application (no. **1**) against the Republic of Poland lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by **Polish** national, **named** ("the applicant"), on **request**.	
source sentence	1) The applicant was represented by <MASK>, a lawyer practising in <MASK>. 2) The applicant was represented by **X**, a lawyer practising in <MASK>.	1) X 2) Y
right context	The Polish Government ("the Government") were represented by their Agent, <MASK>, of the Ministry of Foreign Affairs.	

Table 5: Brief configuration example for repL (replacement of PII in left-hand context) with a total input size of 3 sentences. Model's previous predictions in bold.

A. Appendix: Data and Code

The TAB corpus can be accessed from https://github.com/NorskRegnesentral/text-anonymization-benchmark/tree/master.

Our code can be accessed from https://github.com/mormor-karl/fill-in-the-blanks.

The PGV-PII minicorpus can be accessed from https://spraakbanken.gu.se/en/resources/pgv-pii.

B. Appendix: Example for repL Condition

We provide an example for a repL configuration in Table 5 with context size=3 (i.e. +1 sentence to the left and right side of the core sentence) to illustrate. We use <MASK> to symbolize the masking token for which the prediction should be made. The entire section is fed as input to the model to generate the pseudonym for the current mask token. The example is taken from document 001-86681 from the training set of the TAB corpus (Pilán et al., 2022).

C. Appendix: Annotation Guidelines

The annotation of both the subset of the PGV (Prokopidis et al., 2016, CC BY 4.0) corpus and its pseudonymized version obtained in Experiment 2 was conducted by the first author of the present paper, a fluent speaker of the main languages featured in the dataset.

For the first stage of annotating the texts from the PGV corpus, we used the guidelines provided by Pilán et al. (2022) for annotating PII entities, but instead of annotating all potential entities with the type and adding additional information on whether they should be masked, we only annotated the ones we believe should be masked.

The second stage of annotation was concerned with the pseudonyms generated by MLMs. The categories presented below were used to annotate the machine-generated pseudonyms in context. The annotation was conducted manually. As far as various error categories are concerned, there is a hierarchy of precedence, meaning that despite one pseudonym possibly being the wrong part of speech and semantically distant from the original, only one of those categories is assigned; this order corresponds to their order of appearance below.

1. **NO_ERROR**: the pseudonym matches the original piece of personal information in terms of part of speech and semantic category in a reasonable fashion. It must also match preceding pseudonyms in terms of coreference and has to be reasonable according to your real-world knowledge.

2. **FAILED**: Only applicable to Qwen and used either in the case of a model response timeout (which resulted in a `[FAILED]` token being inserted instead of a generated replacement) or in the case of the suggested replacement being a span that occurring in the context of the masked piece of personally identifiable information.

3. **OVERLAP**: the pseudonym overlaps at least in part with the original piece of personally identifiable information. Also used in the case of synonyms (e.g. *dad* vs. *father*) or leaking an important piece of information from elsewhere (e.g. using the real surname in the place of the real given name).

4. **GRAMMAR**: the pseudonym causes the sentence or clause to become ungrammatical (in terms of agreement, syntax, or morphology) from the perspective of a proficient speaker.

5. **POS**: the pseudonym belongs to the wrong part of speech category. Note that unsolicited punctuation marks are a `STRUCTURAL` error. Nouns and proper nouns are treated as the same category (and therefore replacing *Mark*

	English			Swedish		
	UPOS	RoBERTa	Qwen	UPOS	RoBERTa	Qwen
NO_ERROR	0.1542	0.3042	0.2056	0.1554	0.3263	0.1638
FAILED	0.0000	0.0000	0.0375	0.0000	0.0000	0.1144
GRAMMAR	0.0667	0.0083	0.0806	0.1299	0.0240	0.1299
OVERLAP	0.0028	0.0514	0.0056	0.0000	0.0184	0.0014
POS	0.0667	0.1250	0.1167	0.0551	0.0833	0.1229
SEM:CATEGORY	0.4889	0.2417	0.3014	0.5014	0.3079	0.2839
SEM:CONTRADICTORY	0.1444	0.2042	0.1028	0.1116	0.1045	0.0438
SEM:MERGED-ENTITY	0.0014	0.0292	0.0208	0.0014	0.0975	0.0254
STRUCTURAL	0.0750	0.0361	0.1292	0.0452	0.0381	0.1144

Table 6: Relative frequencies used to construct Figure 2.

with *man* is a case of a `SEMANTIC:CATEGORY` error).

6. **`STRUCTURAL`**: the pseudonym significantly disrupts the structure of the sentence, clause, or text. Most often used in the case of a punctuation mark being selected instead of a word, but also applicable in the case of inappropriate replacements of URLs or longer segments of text deemed as sensitive. This category is also used when the pseudonym builds a compound word together with one of the adjacent tokens, e.g. *amerikanska städer* 'American cities' being pseudonymized as *småstäder* 'small cities (towns),' whether the space is there or not.

7. **`SEMANTIC:CATEGORY`**: the pseudonym belongs to an incorrect semantic category, e.g. a city name is replaced with a country name or a profession is replaced with an ethnicity. Note that in this case, replacing a full name and surname entity with just one given name or surname is not counted as an error.

8. **`SEMANTIC:CONTRADICTORY`**: the pseudonym contradicts the truth previously established in the text or real-world knowledge, e.g. an entity previously pseudonymized as *Mark* is now referred to as *Michael* or *Paris* is presented as the capital of *Germany*.

9. **`SEMANTIC:MERGED-ENTITY`**: the pseudonym removes the distinction between two separate entities, e.g. *Mark* and *Tobias* are both pseudonymized as *Michael*.

D. Appendix: Additional Tables and Figures

In Table 6 we provide the relative frequencies of various types of pseudonym annotations per model and language.