DARE: EFFICIENT MULTILINGUAL PII DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting personally identifiable information (PII) is a fundamental component of privacy-preserving natural language processing, requiring models that reliably identify sensitive information across languages and heterogeneous data distributions. However, current models either involve billions of parameters and achieve state-of-the-art performance, or involve millions of parameters and fall short in performance. In this work, we introduce DARE, an efficient bidirectional encoder for multilingual PII detection with 300M parameters, trained on a heterogeneous mixture of publicly available datasets spanning multiple languages and annotation schemes. DARE achieves strong detection performance across well-established multilingual privacy benchmarks and outperforms state-of-the-art open-weight PII detection baselines with up to $5\times$ more total parameters under a common evaluation protocol, while incurring up to $14.3\times$ lower inference overhead.

1 INTRODUCTION

Personally identifiable information (PII) is embedded throughout modern text: customer conversations, support tickets, operational logs, financial records, and documents collected for analysis or reuse can all contain names, contact details, addresses, account identifiers, credentials, and other information associated with individuals. Automatically detecting such information is therefore an important component of privacy-preserving text processing. At the same time, useful anonymization must balance disclosure risk against data utility: removing too little can expose sensitive information, whereas removing too much can destroy the value of the text for downstream use (Lison et al., 2021; Pilán et al., 2022).

Existing approaches span several modeling paradigms. Hybrid systems combine regular expressions, dictionaries, validation rules, and named entity recognition (NER) components (Data Privacy Stack, 2026), while neural sequence labelers predict entity tags from contextual representations (Lample et al., 2016; Devlin et al., 2019). Autoregressive approaches instead generate extraction outputs, using task formulations ranging from sequence-to-sequence prediction to instruction following (Yan et al., 2021; Lu et al., 2022; Wang et al., 2023; Zhou et al., 2024; Sainz et al., 2024). This formulation supports flexible schemas but requires sequential output generation. GLiNER (Zaratiana et al., 2024) shows that open-type NER can also be implemented non-autoregressively by matching entity types and textual spans with a bidirectional encoder.

PII detection is related to NER, but it is not simply general NER with a different label list. Privacy-oriented taxonomies combine conventional entities such as people and locations with structured identifiers, contact information, financial and government identifiers, personal attributes, and credentials. Whether an entity or attribute warrants redaction depends on its context and the application's privacy policy; information that does not directly identify a person can still contribute to identification (Lison et al., 2021; Pilán et al., 2022). A detector specialized for a fixed privacy taxonomy therefore addresses a different objective from an extractor designed to accept arbitrary entity types.

Recent work reinforces the relevance of bidirectional architectures for this setting. OpenAI Privacy Filter adapts a pretrained language model into a bidirectional token classifier with constrained span decoding (de Bourcy et al., 2026) and 1.5B total parameters. This motivates a central question: can we build a state-of-the-art model without billions of parameters and by training only on heterogeneous public data?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Learning from public corpora makes this question non-trivial. Available PII corpora differ in language coverage, annotation guidelines, and label granularity. The same span may be annotated as a generic *person* in one source and as a *first name* or *last name* in another. Simply merging these annotations can turn compatible labels into conflicting targets. Marginal-likelihood training provides a way to learn from disjoint or overlapping label sets (Greenberg et al., 2018). Learning when entities are missing altogether from the annotations is a related but distinct problem (Mayhew et al., 2019; Effland & Collins, 2021).

In this work, we introduce DARE, a compact bidirectional encoder with 300M parameters for multilingual PII detection that can run on consumer hardware. To learn from heterogeneous public datasets, we replace a single hard target with admissible sets of BIO (begin, inside, outside) tags (Ramshaw & Marcus, 1995) and optimize their marginal likelihood. Additionally, we apply a thresholded privacy decoder to further improve performance.

Across well-established multilingual privacy benchmarks, DARE achieves state-of-the-art performance compared to open-weight state-of-the-art privacy detection baselines, while being significantly faster compared to other baselines. Furthermore, on two PII-clean suites, DARE detects less PII characters across both suites compared to baseline competitors.

Our main contributions are as follows:

- We open-source DARE, a compact multilingual bidirectional encoder specialized for multilingual PII detection that achieves state-of-the-art performance and can run on consumer hardware.
- In our experiments, DARE outperforms competitor baselines by having lower inference overhead for various batch sizes.
- Our proposed approach detects less privately identifiable information across PII-clean benchmarks compared to baseline competitors.

2 Related Work

Sequence Labeling and Multilingual Encoders: Neural NER commonly combines contextual token representations with tokenwise or structured prediction (Lample et al., 2016; Ma & Hovy, 2016). BERT (Devlin et al., 2019) established bidirectional pretraining as a basis for such tasks , while XLM-R (Conneau et al., 2020) demonstrated cross-lingual transfer with a shared multilingual encoder. More recent work improves encoder efficiency and context handling through ModernBERT (Warner et al., 2025), with mmBERT (Marone et al., 2026) extending this architecture to multilingual pretraining. NuNER (Bogdanov et al., 2024) instead specializes encoder representations for entity recognition using LLM-annotated data. DARE builds on this encoder-based line, adapting an existing multilingual backbone to privacy detection with heterogeneous public annotations.

Autoregressive Information Extraction: A separate line formulates extraction as sequence generation. Yan et al. (2021) unify flat, nested, and discontinuous NER through a generative representation of entity spans. UIE (Lu et al., 2022) extends text-to-structure generation to multiple extraction tasks using a unified structural language and schema prompts. InstructUIE (Wang et al., 2023) uses multi-task instruction tuning for unified extraction, while UniversalNER (Zhou et al., 2024) distills LLM-generated annotations into models for open NER. Moreover, GoLLIE (Sainz et al., 2024) conditions extraction on annotation guidelines. GNER (Ding et al., 2024) further examines the role of non-entity tokens and prediction-to-input alignment in generative NER. These approaches include both encoder–decoder and decoder-only models, the relevant computational distinction is autoregressive output decoding. In contrast, DARE predicts token distributions in parallel within each input window and does not generate a serialized extraction output.

Schema-Conditioned Non-Autoregressive Extraction: Flexible entity schemas do not require autoregressive decoding. Machine-reading-comprehension approaches (Li et al., 2020) express entity types as queries and identify corresponding answer spans. USM (Lou et al., 2023) formulates universal extraction as semantic matching through token linking, while GLiNER (Zaratiana et al., 2024) matches requested entity-type representations with candidate spans using a bidirectional trans-

former. Unlike these schema-conditioned approaches, DARE uses a fixed BIO vocabulary and does not require entity descriptions at inference.

Learning from Heterogeneous Supervision: Entity corpora differ in label inventories, annotation granularity, and completeness. Greenberg et al. (2018) marginalize over compatible label sequences in a BiLSTM-CRF to learn from disjoint label sets. Missing annotations motivate different treatments: Mayhew et al. (2019) downweight likely false negatives, while Effland & Collins (2021) use an expected entity ratio loss for partially supervised NER. DARE applies tokenwise marginalization over admissible BIO tags to reconcile coarse and fine labels.

Privacy Detection and Evaluation: Privacy-oriented extraction has a long history in clinical de-identification, including recurrent models that identify protected information in patient notes (Dernoncourt et al., 2017). General-purpose tools such as Presidio (Data Privacy Stack, 2026) combine pattern recognizers, validation rules, and NER. GLiNER-PII (Zaratiana, 2026) specializes schema-conditioned span extraction for personal information, whereas OpenAI Privacy Filter (de Bourcy et al., 2026) converts a pretrained language model into a bidirectional token classifier with constrained span decoding. These systems differ not only in architecture but also in the privacy policies and entity inventories they implement. Text-anonymization research emphasizes the distinction between detecting annotated entities and preventing identification, as well as the trade-off between disclosure risk and retained utility (Lison et al., 2021; Pilán et al., 2022). REDACT and OpenPII provide multilingual resources for evaluating privacy-oriented detection (Vats et al., 2026; Ai4Privacy, 2026). Compared to other methods, DARE targets efficient detection under a fixed taxonomy.

3 Proposed Method

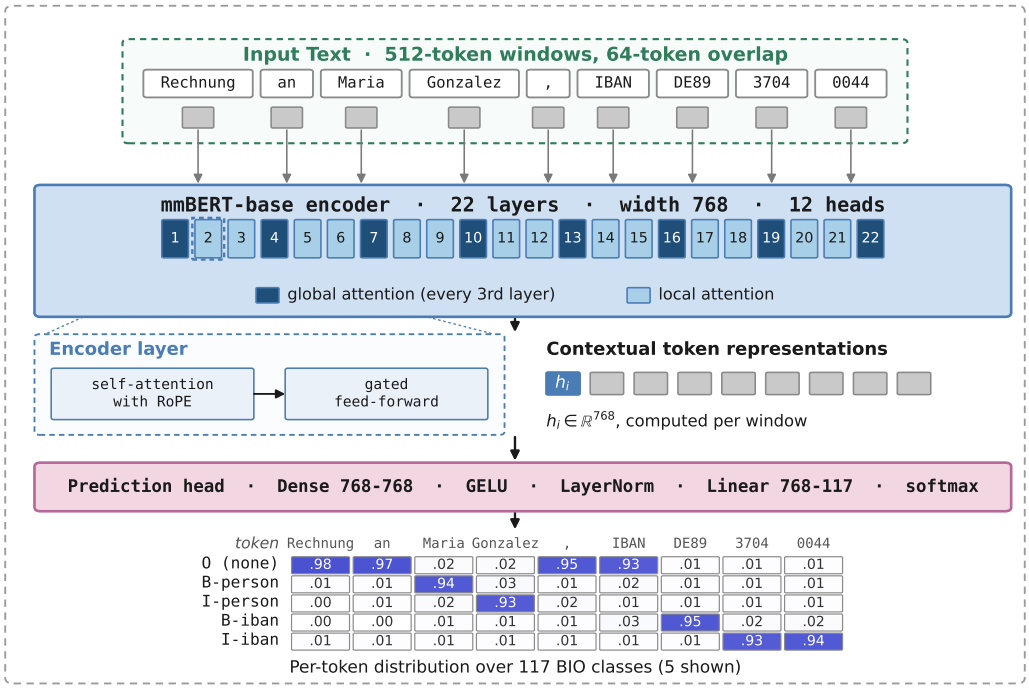

Figure 1: DARE architecture. Only five of the 117 classes are shown to improve clarity.

3.1 Architecture and Prediction Task

For an input document, the model predicts character spans with labels from a fixed set $\mathcal{T}$ of 58 sensitive types. The token label vocabulary is $\mathcal{Y} = \{\text{O}\} \cup \{\text{B-}t, \text{I-}t : t \in \mathcal{T}\}$, with 117 classes:

B-t marks the first token of a span of type t, I-t a subsequent token of that span, and O a token outside any sensitive span. The sensitive-type inventory includes person names, locations, contact information, financial and government identifiers, personal attributes, and credentials. Its breadth reflects the training sources.

We describe the token-level computation for a single input window. Let $x = (x_1, \ldots, x_n) \in \mathcal{V}^n$ denote its sequence of token IDs, where $\mathcal{V}$ is the token-ID vocabulary and n is the sequence length.

For our backbone, we use mmBERT-base (Marone et al., 2026), a multilingually pretrained encoder based on the ModernBERT (Warner et al., 2025) architecture. It consists of 22 layers, a hidden width of 768, 12 attention heads, and a gated feed-forward intermediate width of 1,152. The backbone employs global attention every third layer, while the remaining layers use a local attention window of 128 tokens.

The prediction head applies a dense transformation, a GELU activation, layer normalization, and a linear 117-class projection, followed by softmax:

$$h_i = \big[\mathrm{Encoder}_\theta(x)\big]_i, \qquad p_i = \mathrm{softmax}(W_2\,\mathrm{LN}(\mathrm{GELU}(W_1 h_i)) + b_2)\,. \tag{1}$$

Here, the encoder maps the input sequence to contextual representations in $\mathbb{R}^{n\times 768}$, and $[\cdot]_i$ selects the representation $h_i \in \mathbb{R}^{768}$ of the i-th input token. The head uses weight matrices $W_1 \in \mathbb{R}^{768\times 768}$ and $W_2 \in \mathbb{R}^{117\times 768}$, an output bias $b_2 \in \mathbb{R}^{117}$, and layer normalization denoted by LN. The resulting $p_i \in [0,1]^{117}$ is a probability distribution over $\mathcal{Y}$, where, $p_i(y)$ denotes the probability assigned to label y at token position i. We use θ to denote the model's trainable parameters collectively, including those of the backbone and prediction head.

The model has 307.6M parameters, including 111.0M non-embedding parameters. Figure 1 provides a high-level view of the architecture.

3.2 Training mixture and label reconciliation

We train on a public data mixture comprising 53.5M tokens, 190.9M characters, and 1.5M source annotation records across 13 languages. To reconcile differences in annotation granularity, each source-annotated span is associated with a nonempty set $S \subseteq \mathcal{T}$ of admissible entity types. For a span covering token positions $a, \ldots, b$, we define the admissible BIO labels as

$$A_i = \begin{cases} \{\text{B-}t : t \in S\}, & i = a, \\ \{\text{I-}t : t \in S\}, & a < i \le b. \end{cases}$$

This construction retains the source annotation's beginning and continuation positions while allowing ambiguity in the entity type. When S contains a single type, the corresponding admissible BIO label sets are singletons. Unannotated token positions receive $A_i = \{\mathrm{O}\}$.

Let $\mathcal{M}$ denote the set of token positions contributing to the loss, excluding padding and special tokens. We minimize

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \log\left(\sum_{y\in A_i} p_i(y)\right). \tag{2}$$

The objective marginalises over admissible labels separately at each token; it does not impose BIO-transition constraints or require a shared entity type across a source span. BIO transition handling is performed at inference by the decoder described in Section 3.3.

Here, $p_i(y)$ is the probability assigned to label y at token position i. The inner sum is the total probability mass assigned to labels compatible with the source annotation. The loss encourages this mass to increase without imposing a unique target within A_i. When A_i is a singleton, the corresponding term reduces to standard cross-entropy.

Unannotated positions remain included in the loss with $A_i = \{\mathrm{O}\}$. The reported model does not apply heuristic masking to positions suspected of having missing PII annotations. The objective therefore accommodates the specified differences in label granularity but does not explicitly account for missing entity annotations.

Training windows are sampled with replacement. Let Λ denote the set of training languages and n_ℓ the number of training windows associated with language ℓ. We assign language ℓ sampling

probability

$$q(\ell) = \frac{n_\ell^\alpha}{\sum_{k \in \Lambda} n_k^\alpha}, \qquad \alpha = 0.5. \tag{3}$$

Each window in language ℓ is therefore sampled with probability $q(\ell)/n_\ell$. This reduces the sampling imbalance between languages relative to uniform sampling over all training windows, tempering the dominant English share.

3.3 Thresholded privacy decoder

At inference, we separate the decision to mark a token as sensitive from the choice of its BIO label. For each text token i in an input window, we compute the total non-background probability and apply a threshold:

$$\begin{aligned} s_i &= \sum_{y \in \mathcal{Y} \setminus \{\mathrm{O}\}} p_i(y) = 1 - p_i(\mathrm{O}), \\ \hat{y}_i &= \begin{cases} \arg \max\limits_{y \in \mathcal{Y} \setminus \{\mathrm{O}\}} p_i(y), & s_i \geq \tau, \\ \mathrm{O}, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

Here, $s_i \in [0, 1]$ is the aggregate probability of all sensitive BIO labels, $\tau \in [0, 1]$ is the decision threshold, and $\hat{y}_i$ is the token label before span reconstruction. The reported DARE configuration uses $\tau = 0.175$ as described in Appendix A.

A token is assigned its most likely non-background BIO label when the combined non-background probability reaches the threshold; otherwise, it is assigned O. Unlike an argmax over the full label vocabulary, this rule can select a sensitive label even when O is the most probable individual class. The threshold applies to the combined sensitive probability, not to the probability of the selected label. These decisions are made tokenwise, BIO transition handling is performed during the subsequent span reconstruction.

The decoder thresholding was applied to other baselines when applicable. For detailed results on how the individual components affect performance, we kindly point the reader to Appendix B, C.

4 Experiments

4.1 Experimental Protocol

Baselines. We compare against **Presidio** (Data Privacy Stack, 2026), an open-source PII detection framework that combines rule-based detection with an NER component. Next, **GLiNER-PII** (Zaratiana, 2026), is a release of the GLiNER bidirectional span architecture adapted to personally identifiable information. Moreover, we include **OpenAI Privacy Filter** (de Bourcy et al., 2026), a mixture-of-experts bidirectional token classifier that claims state-of-the-art performance. Lastly, we evaluate two compact open-weight general-purpose autoregressive language models: *i)* **Gemma 4 E2B-it** (Team et al., 2026), and *ii)* **Qwen3.5-2B** (Qwen Team, 2026).

Benchmarks. We evaluate on two multilingual privacy corpora. **REDACT** (Vats et al., 2026) is a multilingual benchmark for PII type classification under controlled variation which consists of 13.5K records, 324K entity annotations, 51 entity types, 4K surface-form patterns, and 25 languages in total. In this work, we report results on the 1000 locked stratified test set of REDACT, as recommended by the original authors. Moreover, we include 10K documents drawn randomly with stratification from the designated validation set of the **AI4Privacy** (Ai4Privacy, 2026) benchmark that covers 30 languages and 19 entity types. Lastly, we incorporate two internal evaluation suites **(Suite A and Suite B)** with 112 and 400 clean documents that investigate over-detection.

Metrics. Let G_d and $\hat{G}_d$ be the sets of gold and predicted character positions in document d. Since these are position sets, a character covered by two predicted spans is counted once. Moreover, since we work with character positions, the metrics include type-agnostic character coverage, which aligns well with models trained on different entity groups from the benchmarks. We micro-aggregate over documents as follows:

$$\mathrm{TP} = \sum_d |G_d \cap \hat{G}_d|, \qquad \mathrm{FP} = \sum_d |\hat{G}_d \setminus G_d|, \qquad \mathrm{FN} = \sum_d |G_d \setminus \hat{G}_d|, \tag{5}$$

so that $\mathrm{TP} + \mathrm{FP} = \sum_d |\hat{G}_d|$ is the number of characters redacted and $\mathrm{TP} + \mathrm{FN} = \sum_d |G_d|$ the number of gold sensitive characters, and report

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F_2 = \frac{5PR}{4P + R} = \frac{5\mathrm{TP}}{5\mathrm{TP} + \mathrm{FP} + 4\mathrm{FN}}. \tag{6}$$

Precision P is the share of redacted characters that were PII, while recall R is the share of PII characters that were redacted. We follow prior work (Ferrández et al., 2012; Pilán et al., 2022) and we adopt F_2 as a main metric on the grounds that recall is of paramount importance because protected information cannot be disclosed at any rate.

Training. We use AdamW (Loshchilov & Hutter, 2019) as optimizer, with a learning rate of 3×10^{-5}, combined with a weight decay of 0.01. We use a batch size of 16, with gradient accumulation over two batches, gradient clipping at 1.0, and a linear schedule with 6% warmup. Lastly, we use early stopping with 5 rounds of patience. All experiments are run on a single NVIDIA H100 GPU and are seeded. The evaluation protocol is the same for all methods.

We open-source the DARE weights and the corresponding train/val and test benchmark splits for reproducibility[1].

4.2 Zero-shot Multilingual PII Detection

Table 1: Zero-shot performance evaluation of DARE and all competitor baselines. Best-performance per metric is in bold, while, the second best performance is underlined. P represents precision, while R represents recall. Values inside brackets represent 95% bootstrap CI across 40K resamples with a sample size equal to the benchmark evaluation set. The same bootstrapped sets are used for all methods.

Method	REDACT				AI4Privacy			
	P	R	F_2	40K ΔF_2 (DARE vs Method)	P	R	F_2	40K ΔF_2 (DARE vs Method)
Presidio	39.26 [37.72, 40.84]	51.09 [49.87, 52.32]	48.19 [47.10, 49.27]	+27.27 [26.24, 28.34]	50.28 [49.76, 50.82]	63.78 [63.35, 64.21]	60.53 [60.17, 60.90]	+33.59 [33.22, 33.94]
GLiNER-PII (290M)	63.61 [61.92, 65.27]	58.61 [57.31, 59.97]	59.54 [58.37, 60.78]	+15.91 [14.68, 17.12]	77.41 [76.94, 77.87]	79.95 [79.54, 80.37]	79.43 [79.07, 79.78]	+14.69 [14.32, 15.06]
Qwen3.5 2B	64.05 [61.70, 66.39]	40.77 [38.65, 42.93]	43.97 [41.96, 45.99]	+31.49 [29.32, 33.60]	85.97 [84.99, 86.93]	69.79 [69.23, 70.32]	72.52 [72.03, 72.99]	+21.60 [21.12, 22.10]
Gemma4E2B-it	**73.68** [71.60, 75.68]	60.97 [59.30, 62.59]	63.15 [61.65, 64.60]	+12.31 [10.96, 13.73]	**95.53** [95.06, 95.96]	76.76 [76.26, 77.27]	79.90 [79.46, 80.34]	+14.22 [13.77, 14.68]
OpenAI Privacy Filter 1.5B	<u>66.02</u> [64.11, 68.00]	<u>66.88</u> [65.42, 68.35]	<u>66.71</u> [65.42, 67.98]	+8.75 [7.76, 9.79]	90.71 [90.25, 91.13]	<u>80.81</u> [80.34, 81.27]	<u>82.62</u> [82.21, 82.99]	+11.50 [11.13, 11.89]
DARE (300M)	65.29 [63.17, 67.47]	**78.51** [77.39, 79.66]	**75.46** [74.42, 76.49]	–	<u>94.35</u> [94.15, 94.54]	**94.06** [93.88, 94.24]	**94.12** [93.97, 94.27]	–

Initially, we measure the zero-shot performance of DARE and all competitor baselines in the REDACT and AI4Privacy benchmarks. We present the results in Table 1, where as observed, despite its smaller size, DARE manages to outperform all baselines and achieves better performance across both benchmarks. To quantify uncertainty, for the benchmark experiments we use non-parametric bootstrapping (Efron, 1992) with 40K resamples and a sample size equal to the benchmark evaluation set. Each resample contains N documents drawn with replacement from the N evaluated documents. We recompute micro-aggregated metrics within each resample and calculate between-method differences using the same resampled indices.

For the REDACT benchmark, Gemma4E2B-it achieves the highest precision with 73.68%, however, it loses in terms of recall, incurring a lower F_2 value. While, DARE manages to achieve the highest recall value of 78.51%, combined with a well-balanced precision value of 65.29%, which in turn yields the best F_2 value in the benchmark of 75.46%, 8.75% points higher compared to the second best competitor, OpenAI Privacy Filter, a model approximately $5\times$ its size.

The results are consistent for the AI4Privacy benchmark, Gemma4E2B-it manages to achieve the highest precision of 95.53%, followed closely by DARE, with only a 1.18% point difference. However, as for REDACT, DARE manages to win on recall and achieve an F_2 value of 94.12%, followed

[1] https://anonymous.4open.science/r/DARE-D41D/

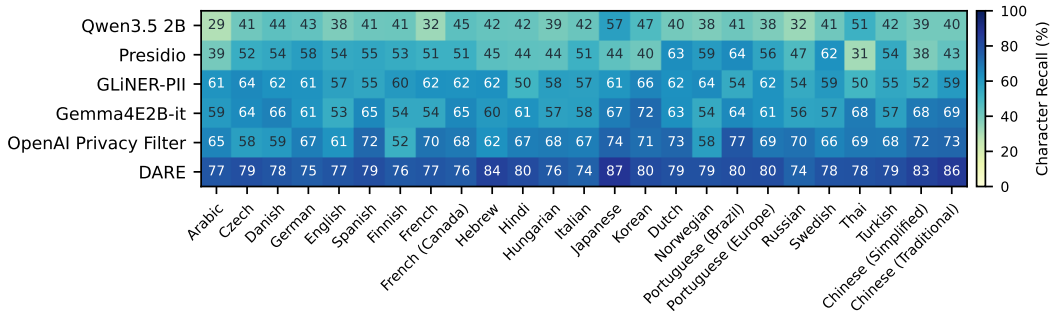

Figure 2: Character recall on the REDACT benchmark for DARE and all the competitor baselines on a per-language basis.

by the larger model OpenAI Privacy Filter 1.5B, which comes in second with a difference of 11.5% points.

Moreover, in Table 1 we present the bootstrapped difference in F_2 (ΔF_2) between DARE and a baseline m given as:

$$\Delta F_2^{(b)}(m) = F_2\left(\text{DARE}; \mathcal{D}^{(b)}\right) - F_2\left(m; \mathcal{D}^{(b)}\right). \tag{7}$$

The results in Table 1 indicate that the 95% CI of the bootstrap difference (presented in brackets) shows a consistent improvement observed across 40K bootstrap resamples.

Lastly, in Figure 2, we investigate the recall per-language for DARE and all the considered baselines, where, as observed DARE shows excellent recall for all languages in the benchmark. DARE has the highest recall on all of 25 languages; with the highest recall in the Japanese language and the lowest recall in the Italian/Russian language. In Appendix D, Figure 6 we present the language recall for the AI4Privacy benchmark where the results are consistent and DARE still provides the best recall per-language compared to the other baselines.

In conclusion, the results show that **DARE achieves excellent zero-shot performance on the considered benchmarks, outperforming state-of-the-art methods up to 5 times it's size.**

4.3 Supervised Fine-tuning Multilingual PII Detection

Next, we investigate how DARE performs compared to other PII detection baselines[2], when supervised fine-tuning is included for the downstream target tasks.

Initially, for the REDACT benchmark we sample 1K instances with stratification from the training set. 50% of the instances are used for training, while the rest are used for validation. We evaluate the performance of DARE and PII baselines on the same locked evaluation set as designed by the original authors (Vats et al., 2026). We repeat the same experiment for the AI4Privacy benchmark, sampling 1K instances with stratification, 50% split to train and 50% split to validation. The test set is taken from the designated evaluation set from the AI4Privacy benchmark (Ai4Privacy, 2026). Both evaluation sets for REDACT and AI4Privacy have no overlap with our training/validation set and are equivalent to the evaluation sets used in Section 4.2.

We present the results of our experiments in Table 2, where, as observed, DARE manages to outperform all PII detection baselines after supervised fine-tuning on the respective downstream dataset.

Investigating the REDACT benchmark results, DARE after supervised fine-tuning reaches the best precision score of 66.94%, falls slightly behind in terms of recall with 0.53% points to GLiNER-PII, while achieving the best overall F_2 score. For the AI4Privacy benchmark, DARE achieves the best precision, recall, and F_2 score across all baselines.

[2]Presidio does not support fine-tuning.

Table 2: Performance after supervised fine-tuning for DARE and the competitor PII baselines. Best performance per-metric is reported in bold, while the second-best performance is underlined. Mean value is reported combined with the standard deviation across 5 seeds. P stands for precision, R stands for recall, pre stands for the zero-shot performance in Table 1.

	REDACT				AI4Privacy			
Method	P	R	F_2	pre	P	R	F_2	pre
DARE (300M)	**66.94** ±0.99	92.95 ±0.40	**86.24** ±0.06	75.46	**93.23** ±2.39	**99.56** ±0.02	**98.22** ±0.54	94.12
GLiNER-PII (290M)	65.40 ±0.62	**93.48** ±0.26	86.09 ±0.05	59.54	89.97 ±5.44	**99.56** ±0.06	97.43 ±1.28	79.43
OpenAI Privacy Filter (1.5B)	64.17 ±0.05	78.97 ±0.95	75.49 ±0.71	66.71	66.93 ±1.25	96.52 ±0.22	88.68 ±0.58	82.62

Following supervised fine-tuning, all methods improve from the original zero-shot performance scores, as can be observed from the difference to the *pre* column in Table 2. In conclusion, **DARE is the leading method and manages to outperform all competitors in the benchmarks we consider, even when the methods are fine-tuned on the downstream target task.**

4.4 Inference efficiency

In this section, we measure the inference wall-clock times by running DARE and all competitor baselines on the REDACT locked evaluation set. We investigate two scenarios: *i)* one where we use batches of a single test example for inference, and *ii)* one where we use batches of 8 examples. We present the results in Figure 3, where, as observed, DARE manages to outperform all baselines by incurring a lower total inference time for both scenarios included.

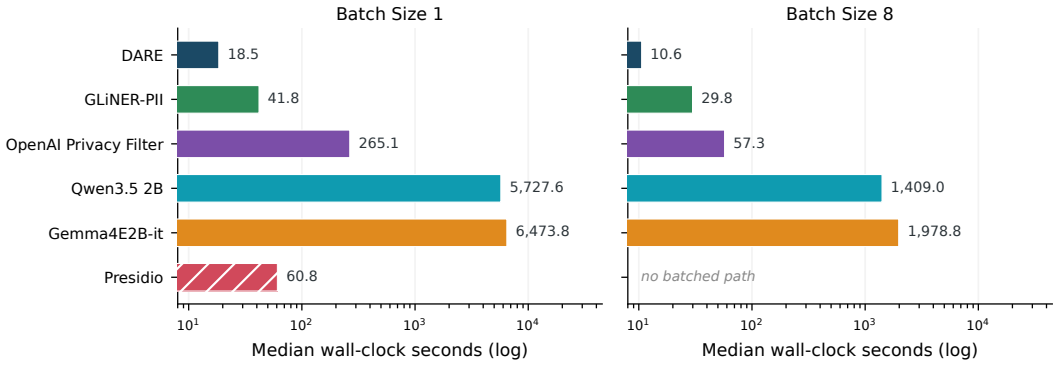

Figure 3: Inference median wall-clock time costs for DARE and all competitor baselines measured on the REDACT evaluation locked set. **Left)** Batched prediction with 1 example, **Right)** Batched prediction costs with 8 examples. Measurements calculated on a single NVIDIA H100 GPU.

In the case of performing inference with batches of size 1, DARE has a lower inference cost compared to both GLiNER-PII, Presidio[3], and OpenAI Privacy Filter by a factor of $\mathbf{2.26}\times$, $\mathbf{3.3}\times$, and $\mathbf{14.3}\times$ respectively. As expected autoregressive language model architectures have higher inference time costs compared to DARE, by a factor of $\mathbf{309.6}\times$ for Qwen3.5 2B and $\mathbf{350}\times$ for Gemma4E2B-it respectively, given their sequential prediction.

For the second scenario where inference is done in batches of 8, DARE has a lower inference cost compared to both GLiNER-PII and OpenAI Privacy Filter by a factor of $\mathbf{2.8}\times$ and $\mathbf{5.4}\times$ respectively. Moreover, it outperforms autoregressive language model architectures by a factor of $\mathbf{133}\times$ for Qwen3.5 2B, and $\mathbf{186.7}\times$ for Gemma4E2B-it. For detailed results per-method we kindly refer the reader to Table 8 in Appendix E.

The results confirm that **DARE not only achieves the best performance across baselines in zero-shot and supervised fine-tuning scenarios, but it additionally is the fastest method across all baselines.**

[3]Presidio has no inference improvement when run on a GPU.

4.5 False Detection on PII-Clean Suites

In this section, we investigate overdetection for DARE and competitor baselines. To that aim, we evaluate on two internal suites that contain no privately identifiable information. Suite A and Suite B contain: *i)* 112 documents (12,740 characters), and *ii)* 400 documents (50,398 characters).

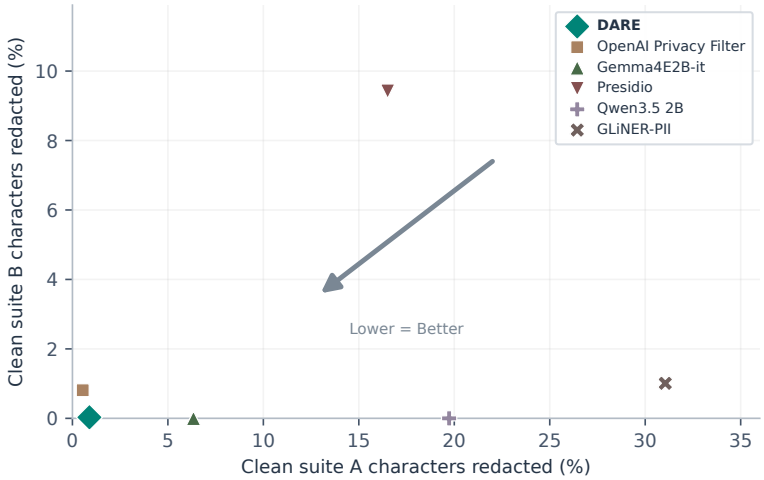

Figure 4: Characters redacted per clean suite for DARE and all competitor baselines. The lower the value the better for every axis.

We present the results in Figure 4, where, as observed, DARE redacts the lowest amount of characters on the combined benchmarks compared to other baselines.

For the Suite A benchmark, OpenAI Privacy Filter achieves the lowest value of 0.54% of characters redacted, closely followed by DARE with 0.89% of characters redacted. While, for the Suite B benchmark, both autoregressive models Gemma4E2B-it and Qwen3.5 2B achieve 0% of characters redacted, closely followed by DARE with 0.03%. For detailed results regarding the performance of individual methods on both suites, we refer to Table 9 in Appendix E.

Across both benchmarks, **DARE is the only method that achieves the combined lowest value of characters redacted, as shown in Figure 4.**

5 Limitations

Although DARE's backbone was pretrained in over 1800 languages and was trained on public data that spans 13 languages, as with other models, the generalization performance can degrade on languages that are not present in our training corpora. Furthermore, the generalization performance of our model might not transfer to additional PII entities that are specific per country/language and are not present in our training data. Lastly, while our work focuses on PII detection, successful detection does not always correspond to the input being safe in terms of identification, as the context, even with the sensitive details redacted might still lead to identification.

6 Conclusion

In this work we introduce DARE, a compact bidirectional encoder that can run on consumer hardware. DARE a 307M model manages to achieve state-of-the-art performance on public established benchmarks among open-weight models up to $5\times$ it's size, while, being significantly faster in inference compared to other baselines.

Reproducibility statement

The baselines and benchmarks that we incorporate in our work are publicly available. Our model weights, the splits for training, validation, and testing are present at the anonymous repository link provided in Section 4.1.

Our training details are described in Section 4.1 and Appendix A. The experiments are seeded and the seeds are additionally provided in the anonymous repository. Following acceptance we will fully open-source our code and training corpora.

AI Disclosure

In this work, we used generative AI tools for generating the synthetic datasets of Suite A and Suite B related to Section 4.5. We have not used generative AI tools for other tasks with required disclosure. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts.

References

Ai4Privacy. Openpii 1.5m: Multilingual pii masking dataset with asia pacific extension, 2026. URL https://huggingface.co/datasets/ai4privacy/pii-masking-openpii-1.5m.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne P Bernard. NuNER: Entity recognition encoder pre-training via LLM-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11829–11841. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.660. URL https://aclanthology.org/2024.emnlp-main.660/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Data Privacy Stack. Presidio: Data protection and de-identification sdk, 2026. URL https://github.com/data-privacy-stack/presidio. Accessed 24 September 2026.

Charles de Bourcy, Sahra Ghalebikesabi, Avi Schwarzschild, Alex Gorbachev, Mihai Maruseac, Annie Chu, Vol Kyrylov, Tong Mu, Ally Bennett, Andy Nguyen, Casey Meehan, Jessica Gan Lee, Shane Bauer, Harold Nguyen, Rodolpho Eckhardt, Yuqi Liu, Charlie Oxborough, Marco Rougeth, Omar Chedid, Caio Costa, Yash Parikh, Yao Li, Congzheng Song, Om Thakkar, and Vinnie Monaco. Model card for openai privacy filter, 2026. arXiv:2608.18274.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017. doi: 10.1093/jamia/ocw156. URL https://doi.org/10.1093/jamia/ocw156.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Bowen Yan, and Min Zhang. Rethinking negative instances for generative named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3461–3475. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.206. URL https://aclanthology.org/2024.findings-acl.206/.

Thomas Effland and Michael Collins. Partially supervised named entity recognition via the expected entity ratio loss. *Transactions of the Association for Computational Linguistics*, 9:1320–1335, 2021.

Bradley Efron. *Bootstrap Methods: Another Look at the Jackknife*, pp. 569–593. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_41. URL https://doi.org/10.1007/978-1-4612-4380-9_41.

Oscar Ferrández, Brett R South, Shuying Shen, F Jeffrey Friedlin, Matthew H Samore, and Stéphane M Meystre. Evaluating current automatic de-identification methods with veteran's health administration clinical documents. *BMC Medical Research Methodology*, 12(1):109, 2012. doi: 10.1186/1471-2288-12-109.

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2824–2829, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1306. URL https://aclanthology.org/D18-1306/.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pp. 260–270, 2016.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5849–5859. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.519. URL https://aclanthology.org/2020.acl-main.519/.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 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)*, pp. 4188–4203, 2021. doi: 10.18653/v1/2021.acl-long.323.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Universal information extraction as unified semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13318–13326, 2023. doi: 10.1609/aaai.v37i11.26563. URL https://ojs.aaai.org/index.php/AAAI/article/view/26563.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5755–5772. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.395. URL https://aclanthology.org/2022.acl-long.395/.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pp. 1064–1074, 2016.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. mmBERT: A modern multilingual encoder with annealed language learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=iJDJCO4mji.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. Named entity recognition with partially annotated training data. In *Proceedings of CoNLL*, pp. 645–655, 2019.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, December 2022. doi: 10.1162/coli_a_00458. URL https://aclanthology.org/2022.cl-4.19/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Third workshop on very large corpora*, 1995.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47083–47107, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cda04d7ea67ea1376bf8c6962d8541e0-Paper-Conference.pdf.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Guneesh Vats, Anubha Agrawal, Shikha Singhal, Ajita Dash, Praison Selvaraj, Vidhan Jhawar, Ranga Prasad Chenna, and Bharadwaj Y M G. Redact: A systematically controlled multilingual benchmark for personal information detection. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2026.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. A unified generative framework for various NER subtasks. 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)*, pp. 5808–5822. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.451. URL https://aclanthology.org/2021.acl-long.451/.

Urchade Zaratiana. GLiNER multi-PII v1. https://huggingface.co/urchade/gliner_multi_pii-v1, 2026. Model release; accessed 2026-09-17.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5364–5376. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.300. URL https://aclanthology.org/2024.naacl-long.300/.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *International Conference on Learning Representations*, pp. 12276–12294, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/34678d08b36076de986df95c5bbba92f-Abstract-Conference.html.

A Training Setup

The model uses a 256K token multilingual vocabulary. Attention, embedding, MLP, and classifier dropout are zero in the saved configuration. AdamW uses its implementation defaults for betas $(0.9, 0.999)$ and epsilon 10^{-8}. The pinned training runtime is Python 3.9, PyTorch 2.8.0, Transformers 4.57.6, and Tokenizers 0.22.2.

Table 3: Training Hyperparameters

Parameter	Value
Backbone / total parameters	mmBERT-base / 307,620,213 (111,011,445 non-embedding)
Encoder / head	22 layers, width 768, 12 heads / 117 BIO classes
Optimizer / learning rate / weight decay	AdamW / 3×10^{-5} / 0.01
Batch / accumulation / effective batch	16 / 2 / 32 windows
Schedule / nominal budget	Linear, 6% warmup / 3 sampled epochs
Training chunks / overlap / token cap	1,700 characters / 220 characters / 512
Language sampling / seed	$\alpha = 0.5$ / 20270115
Checkpoint evaluation / patience	Every 500 steps / 5
Selected / terminal update	1,000 / 3,500
Inference window / token overlap	512 / 64
Decoder threshold / boundaries	0.175 / ASCII word completion
Precision / attention in reported inference	Float32 / SDPA

B Design Ablations

Initially, we ablate the balanced language sampling. Table 4 compares the released DARE checkpoint with a model trained by the same recipe except that windows are sampled in proportion to their frequency ($\alpha = 1$) rather than with $\alpha = 0.5$. We compare both models on the REDACT benchmark evaluation locked set where the two are indistinguishable (Table 4): the paired F_2 difference is $+0.12$, with an interval spanning zero.

However, the cost of removing balancing appears on clean text in Table 5. At the same threshold, the unbalanced model removes 1.92% of Suite A's characters and touches 51 of 112 documents, against 0.87% and 22 for DARE with balanced language sampling. On Suite B it removes 0.11% and touches 9 of 400 documents, against 0.03% and 1.

Because of the aforementioned reasons, we keep language-balanced sampling because it leaves REDACT unchanged and cuts the over-redaction of text containing no PII by more than half on Suite A and by about three quarters on Suite B, which matters for a redaction system applied to documents that are mostly non-sensitive.

Table 4: Effect of language-balanced sampling on the REDACT evaluation set. F_2 intervals are 95% percentile intervals from 40K document resamples; ΔF_2 is the paired difference against DARE, computed on the same resample in every replicate. P stands for precision, R for recall.

Method	τ	P	R	F_2 [95% CI]	ΔF_2 [95% CI]
DARE	0.175	**65.29**	78.51	75.46 [74.39, 76.48]	–
DARE without language balancing	0.175	65.15	**78.72**	**75.57** [74.52, 76.60]	-0.12 [-0.57, $+0.32$]

Table 5: Clean-text over-detection for DARE with and without language balancing during training.

		Suite A (112)		Suite B (400)	
System	τ	Chars (%)	Docs affected	Chars (%)	Docs affected
DARE	0.175	**0.89**	22/112	**0.03**	1/400
DARE without language balancing	0.175	1.92	51/112	0.11	9/400

Next, we investigate how the thresholded decoder affects performance. Table 6 presents DARE on REDACT at several decoder thresholds. Arg-max decoding labels a token as PII only when one PII

class outscores background; thresholding the pooled score $s_i = 1 - p_i(\mathrm{O})$ instead lets probability mass spread over several PII classes count. At the fitted $\tau = 0.175$ this adds 7.23 F_2 points over arg-max, trading 5.37 points of precision for 10.86 of recall, which is the trade F_2 is designed to reward. Lower thresholds score higher on this sample, up to 78.79 at $\tau = 0.02$, but only by dropping precision to 56.23.

Table 6: Effect of the decoder threshold for the released DARE checkpoint on the locked REDACT evaluation set. ΔF_2 is relative to arg-max decoding. The threshold is fitted on the 500-document selection split as the largest recall with precision at least 65.62%. P stands for precision, R for recall.

Decoding	P	R	F_2	ΔF_2
No threshold (arg-max)	70.66	67.65	68.23	–
$\tau = 0.5$	69.99	69.41	69.53	+1.30
$\tau = 0.3$	67.54	74.55	73.04	+4.81
$\tau = 0.175$ (fitted, reported)	65.29	78.51	75.46	+7.23
$\tau = 0.05$	60.16	84.48	78.16	+9.93
$\tau = 0.02$ (best on this sample)	56.23	87.57	78.79	+10.56

C ZERO-SHOT RESULTS WITHOUT CALIBRATION

Table 7 reports every method on the REDACT locked evaluation set at its uncalibrated default.

Presidio and the two prompted models do not offer calibration and are unchanged. DARE with plain arg-max decoding still has the highest F_2 (68.23), ahead of Privacy Filter's default decoder (64.95) and Gemma4E2B-it (63.15). Calibration adds 7.23 points for DARE, but the ranking at the top does not depend on it. The system most affected by calibration is GLiNER-PII, which loses 12.90 points at the library's default $\tau = 0.5$.

On clean text (Figure 5), uncalibrated DARE redacts 0.16% of Suite A and none of Suite B, less than at its fitted threshold (0.89% and 0.03%), which is the lowest Suite A rate of any system and ties the two prompted models at zero on Suite B.

Table 7: Zero-shot results on REDACT on the locked evaluation set without calibration. Each system's F_2 at its reported operating point, and Δ is uncalibrated minus calibrated. Best value per-metric is presented in bold, and the second best value is underlined.

Method	Decoding	P	R	F_2	Calibrated F_2	Δ
Presidio	no threshold	39.26	51.09	48.19	48.19	+0.00
GLiNER-PII	$\tau = 0.5$ (default)	<u>72.40</u>	42.83	46.64	59.54	−12.90
Qwen3.5 2B	prompted	64.05	40.77	43.97	43.97	+0.00
Gemma4E2B-it	prompted	**73.68**	60.97	63.15	63.15	+0.00
OpenAI Privacy Filter	default decoder	65.62	<u>64.78</u>	<u>64.95</u>	66.71	−1.76
DARE	arg-max	70.66	**67.65**	**68.23**	75.46	−7.23

D PLOTS

E TABLES

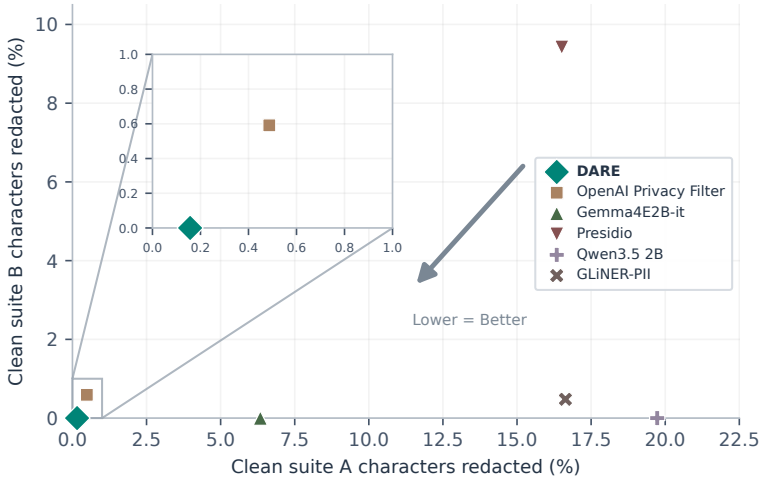

Figure 5: Clean-text over-detection (Suite A and Suite B) of DARE and every method at its uncalibrated default.

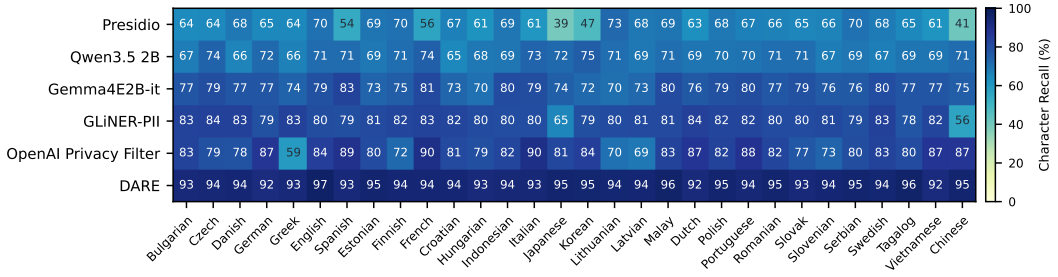

Figure 6: Character recall on the AI4Privacy benchmark for DARE and all the competitor baselines on a per-language basis for the zero-shot scenario.

Table 8: Inference throughput on the REDACT locked evaluation set for DARE and all the considered baselines. Experiments run on a single NVIDIA H100 GPU. Best value per-metric is reported in bold, while, the second best value is underlined.

Method	Batch	Docs	Median s	Tokens/s	Chars/s	Docs/s
DARE	1	1,000	<u>18.5</u>	47,278	<u>102,370</u>	<u>54.16</u>
DARE	8	1,000	**10.6**	**82,333**	**178,275**	**94.32**
GLiNER-PII	1	1,000	41.8	20,870	45,190	23.91
GLiNER-PII	8	1,000	29.8	29,248	63,329	33.51
OpenAI Privacy Filter	1	1,000	265.1	3,293	7,130	3.77
OpenAI Privacy Filter	8	1,000	57.3	15,238	32,995	17.46
Qwen3.5 2B	1	1,000	5,727.6	152	330	0.17
Qwen3.5 2B	8	1,000	1,409.0	619	1,341	0.71
Gemma4E2B-it	1	1,000	6,473.8	135	292	0.15
Gemma4E2B-it	8	1,000	1,978.8	441	955	0.51
Presidio	1	1,000	60.8	14,356	31,084	16.45

Table 9: Suite A and Suite B overdetection for DARE and all competitor baselines. Best value is reported in bold per-metric, while the second-best value is underlined.

	Suite A (112)		Suite B (400)	
Method	Chars (%)	Docs affected	Chars (%)	Docs affected
DARE	<u>0.89</u>	22/112	<u>0.03</u>	1/400
Presidio	16.51	97/112	9.42	214/400
GLiNER-PII	31.05	107/112	1.01	45/400
OpenAI Privacy Filter	**0.54**	10/112	0.81	58/400
Gemma4E2B-it	6.34	85/112	**0.00**	0/400
Qwen3.5 2B	19.73	69/112	**0.00**	0/400