When Privacy Helps: Pseudonymisation as a Strategy for Improved Cyber Incident Classification

Loya C. Haughton, Eduardo Fidalgo, Rocío Alaiz-Rodríguez,
Manuel Castejón-Limas and **Laura Fernández-Robles**
Group for Vision and Intelligent Systems (GVIS), I4 Institute, Universidad de León, León, Spain
{loya.haughton, eduardo.fidalgo, rocio.alaiz,
manuel.castejon, l.fernandez}@unileon.es

Abstract

Organisations often rely on Cyber Threat Intelligence (CTI) for collective defence against evolving threats. Therefore, it is important that these shared reports are both compliant with data protection legislation and analytically useful. Pseudonymisation is a recognised data protection measure, yet its effect on downstream utility remains underexplored in cybersecurity. Unlike other domains where personal identifiers carry predictive value, personal data in cyber incident reports may act as noise rather than signal, suggesting pseudonymisation could improve classification accuracy. This hypothesis is evaluated in two steps. First, we derive three new datasets by applying Data Masking, Data Tokenisation and Data Substitution to a subset of the CECILIA-10C-900 dataset. We then evaluate 21 models, spanning traditional Machine Learning (ML) classifiers, encoder transformers and QLoRA fine-tuned Large Language Models (LLMs), on these datasets, for a CTI classification task based on the Spanish National Cybersecurity Institute's incident taxonomy. The RoBERTa-base model achieved the highest overall weighted F1-score of 87.35% when Data Tokenisation was applied, while Llama-3.1-8B demonstrated the largest gain (+12.63 pp) with Data Masking. These findings reframe pseudonymisation from only a compliance measure into a preprocessing step that may simultaneously protect privacy and improve classification in specific model-technique pairings.

1 Introduction

The global proliferation of cyberattacks (Abdullah et al., 2025) and expanding cyber incident reporting mandates (Schmitz-Berndt, 2023) have intensified the need for tools that extract recurring attack vectors and tactics from incident reports. Beyond formal reporting channels, incident reports also appear across blogs and social media in diverse formats and languages, complicating systematic analysis at scale (Sun et al., 2023).

To address this, organisations share and analyse cyber incident reports to support collective defence (Nitz et al., 2025). Large Language Models (LLMs) and encoder transformers have increasingly been used for such analysis including incident classification (Zhang et al., 2025). However, since cyber incident reports may contain personal data (victim names, locations and organisational affiliations), sharing such data, whether through off-premise LLMs or with other organisations, may introduce privacy risks (Albakri et al., 2019; Staab et al., 2024).

Pseudonymisation, recognised under General Data Protection Regulation Articles 25 and 32 as a data protection safeguard (European Parliament and Council of the European Union, 2016; European Network and Information Security Agency, 2019), offers a potential resolution, but its impact on downstream Natural Language Processing (NLP) tasks, such as text classification performance, remains without empirical evaluation in cybersecurity contexts. If cyber incident reports can be accurately classified *after* pseudonymisation, organisations gain two possible benefits simultaneously: privacy-preserving data handling and enhanced threat intelligence.

This paper, to the best of our knowledge, provides the first empirical evaluation of pseudonymisation as a performance-affecting preprocessing step for CTI classification. Our contributions are threefold:

1. We present a reproducible pseudonymisation methodology for cyber incident reports, that enables organisations to share CTI data while maintaining data protection and preserving (or improving) downstream classification utility.

2. We pre-process and annotate a six-class subset of the CECILIA-10C-900 dataset with eighteen entity types and derive three pseudonymised variants, CECILIA-6C-MAS (using Data Masking), CECILIA-6C-TOK (Data Tokenisation) and CECILIA-6C-SUB (Data Substitution).

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

3. We evaluate 21 models (traditional Machine Learning (ML) classifiers, encoder transformers and QLoRA fine-tuned LLMs) across all four dataset variants and show that pseudonymisation yields positive median F1 gains, with RoBERTa-base achieving 87.35% (+3.00 pp) and Llama-3.1-8B gaining 12.63 pp. Crucially, this work provides evidence that the optimal technique varies by architecture, implying that technique selection should be model-guided.

2 Related Work

This research is positioned at the intersection of three domains: Natural Language Processing (NLP) for cyber incidents, privacy-preserving text processing and the preservation of data utility. We review each in turn, highlighting the gap that our work addresses.

Natural Language Processing in Cybersecurity. Machine Learning models, including transformer architectures, have been successfully applied to NLP in cybersecurity (Sun et al., 2023; Zhang et al., 2025). This has led to the development of tools which automatically classify cyber incidents by attributes such as: threat actors (Mumtaz et al., 2023), attack patterns (Irshad and Siddiqui, 2023) and risk level (Lughbi et al., 2024). This trend has continued with the increasing adoption of Large Language Models, which can process unstructured threat reports at scale, enabling faster incident triage and more informed defence (Zhang et al., 2025).

However, the classification of CTI reports has remained a formidable challenge. Rabitti et al. (2024) highlight that cyber incident taxonomies used for classification are often fragmented and incomplete, with no globally standardised cyber incident taxonomy available. This makes cross-study comparison difficult, possibly limiting cumulative progress in the field. Delgado et al. (2024), whose CECILIA-10C-900 dataset this work builds on, adopted the Spanish National Cybersecurity Institute (INCIBE) taxonomy (Instituto Nacional de Ciberseguridad, 2020) for cyber incident classification. This taxonomy defines ten incident classes and closely follows the European Union Agency for Cybersecurity (ENISA) taxonomy (ENISA, 2018) recommended for use across the European Union.

Given the advances in NLP, there is an urgent need for privacy-by-design implementations, in which solutions are designed with data protection and privacy ingrained. For example, none of the classification approaches surveyed above explicitly incorporate preprocessing steps for privacy protection, a gap that this body of work directly addresses.

Privacy-Preserving Text Processing. Scholars have repeatedly documented the scarcity of labelled cybersecurity datasets for over a decade (Abt and Baier, 2014; Nitz et al., 2025). Consequently, research on privacy-preserving text processing is underexplored in the cybersecurity domain; with most research focussing on the legal and clinical domains (Pépin and Zulkernine, 2022; Deußer et al., 2025). Generalising findings across these domains is likely non-trivial given differences in jargon, legislative context and the nature of personal data involved.

A further obstacle is terminological inconsistency: technical literature frequently conflates anonymisation and pseudonymisation. Individual technique names are unstandardised, with some works incorrectly equating pseudonymisation with a single technique such as Data Substitution (Lison et al., 2021). Though both terms fall under the umbrella of de-identification, under GDPR Article 4(5), pseudonymisation specifically refers to transforming personal data so it cannot be attributed to a data subject without additional information. This ambiguity not only hampers reproducibility in research but also risks misapplication of data protection safeguards, increasing the need for systematic evaluation of pseudonymisation techniques.

Though pseudonymised data is considered personal data under the GDPR, the regulation's risk-based framework means this classification is not monolithic (European Data Protection Board, 2025). By reducing re-identification risk, pseudonymisation may improve the compliance posture of an organisation, possibly providing defensible grounds for broader processing and data sharing, even in the absence of full anonymisation. Therefore, input-level pseudonymisation can be considered a necessary complement to model-level safeguards. These considerations make the choice of pseudonymisation technique consequential not just for privacy compliance but also for downstream performance.

Yet existing work has largely overlooked this latter dimension. The closest analogous work to ours is Vīksna and Skadina (2025), which develops a multilingual anonymisation tool supporting masking, suppression and substitution, but unlike our research, does not evaluate the utility of this data for downstream classification tasks. Omitting this evaluation leaves a gap in understanding utility preservation, which could affect the choice of

technique for a given task.

Utility of de-identified data in downstream tasks Researchers are divided on whether dataset de-identification harms or improves performance. Whereas Vakili et al. (2022) found that the downstream performance was maintained on de-identified clinical notes classification tasks (albeit in Swedish), Lothritz et al. (2023) found that de-identification degraded performance in classification tasks, especially with BERT models. Conversely, Yermilov et al. (2023) observed a slight improvement of around one percent in downstream tasks when using Data Substitution.

These divergent findings may suggest that de-identification can simultaneously reduce bias and spurious correlations while introducing semantic noise and eroding contextual specificity. The net effect may depend less on whether the data is de-identified, than on which technique is applied. Understanding these nuances is therefore crucial for selecting appropriate pseudonymisation techniques that balance privacy protection with task utility; this is a gap that our study directly addresses through per-model evaluation of pseudonymisation impact. Our results suggest that pseudonymisation should be treated as a model-dependent preprocessing choice rather than as a uniform compliance step. To our knowledge, no prior work systematically evaluates pseudonymisation techniques across multiple model families in CTI classification, nor evaluates pseudonymisation as a performance-enhancing preprocessing step in this context.

3 Methodology

3.1 Datasets

The CECILIA-10C-900 dataset (Delgado et al., 2024), contains 923 cyber incidents manually labelled according to the ten classes of the INCIBE Cyber Incident Taxonomy: Abusive Content, Information Gathering, Intrusion Attempts, Malicious Code, Intrusions, Availability, Information Content Security, Fraud, Vulnerable and Others.

The dataset is imbalanced, with the Information Content Security class comprising 50.4% of the dataset, while the four smallest classes together comprise 3.2% or 28 records. Therefore, samples from the four smallest classes (that is, Intrusion Attempts, Information Gathering, Abusive Content and Vulnerable) were deleted from our experimental dataset to avoid training and evaluating models on extremely low-support categories under a single-label setting.

This decision is motivated by the findings of Delgado et al. (2024) that the training samples available in these categories were insufficient for their best-performing model to learn discriminative representations across five-fold cross-validation. We believe that retaining these classes would introduce irreducible noise into the evaluation, obscuring the effect of the phenomenon under study[1].

Forty-three duplicated reports were also removed from the dataset (4.7%). This resulted in 852 remaining records. Both deduplication and the removal of minority classes resulted in minimal redistribution among the remaining six classes; for example, the dominant class (Information Content Security) shifted from 50.4% to 51.4% of the dataset.

Eighteen entity types were selected, each representing either a direct identifier or a quasi-identifier. The entities and their descriptions are provided in Appendix A.3. The dataset was annotated with these entities by a single annotator with domain expertise, using the LabelStudio (Tkachenko et al., 2020-2025) software. The limitations of this design choice are discussed in Section 6.

3.2 Pseudonymisation Techniques

Using the entity annotations, we applied the Data Masking, Data Tokenisation and Data Substitution pseudonymisation techniques to the entities of each cyber incident, individually. This is demonstrated in the graphical abstract shown in Figure 1.

This yielded three new datasets: CECILIA-6C-MAS, CECILIA-6C-TOK and CECILIA-6C-SUB, representing Data Masking, Data Tokenisation and Data Substitution, respectively. Pseudonymisation was executed by applying Python find-and-replace operations on the named entities identified during the Named Entity Recognition (NER) phase. Data Masking replaced entities with "xxxx" regardless of token length, while Data Tokenisation replaced the entities with their NER label type, such as Organization. In our Data Substitution pipeline, we first removed dates to avoid temporal contradictions, expanded acronyms using the SpaCy AbbreviationDetector (Neumann et al., 2019; Honnibal et al., 2020), followed by manual review for undetected cases. Subsequently, we generated fictitious entity replacements using the Python Faker library (Faraglia, 2024). These datasets were then divided into the train−test split described in Section 4.1.

[1]Notably, the creators of the CECILIA-10C-900 dataset, Delgado et al. (2024), had created their own six-category subset of the dataset by merging the four smallest classes into an existing class, Others. However, in this body of work, we are *deleting* the four classes, since merging these classes into an existing class could introduce semantic noise.

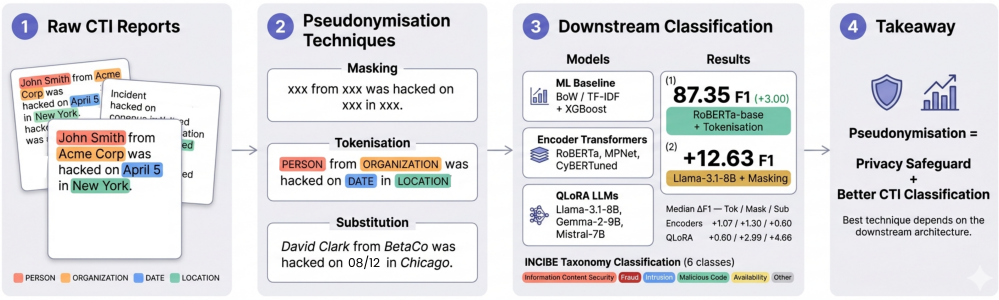

Figure 1: A graphical abstract highlighting the CTI Reports from the CECILIA-10C-900 subset being annotated then pseudonymised, followed by the creation of an ML baseline, and the key results of transformer and LLM finetuning - of which the best performing models were included.

3.3 Models

Traditional Machine Learning Classifiers. We evaluated five classifiers: Logistic Regression (Hosmer Jr et al., 2013), Support Vector Machine (Hearst et al., 1998), Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016) and Multinomial Naïve Bayes (McCallum and Nigam, 1998). Each model was paired with two vectorisers: Bag-of-Words (also known as Count Vectoriser) and Term Frequency-Inverse Document Frequency (TF-IDF), resulting in ten pipelines that served as a baseline for subsequent experiments.

Transformers. Eleven encoder-based transformer models were also fine-tuned and evaluated: BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2020), SecBERT (jackaduma, 2022), SecRoBERTa (jackaduma, 2021), MPNet (Song et al., 2020), DeBERTa-v3 (He et al., 2021), SecureBERT (Aghaei et al., 2023), SecureBERT2.0 (Aghaei et al., 2025) and CyBERTuned (Jang et al., 2024).

LLM. Finally, five LLMs were evaluated: Llama-Primus-Merged (Yu et al., 2025), Llama-3.1-8B-Instruct (Llama Team and AI Meta Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Foundation-Sec-8B-Instruct (Kassianik et al., 2025) and Gemma-2-9B-Instruct (Riviere et al., 2024).

Label language. ML classifiers and encoder transformers represent labels as numeric vectors, making label language irrelevant. LLMs, however, interpret labels as raw text within their prompt. Since CECILIA-10C-900 uses Spanish labels, the LLMs were first evaluated with both Spanish and English labels to determine which yielded stronger performance; the better-performing language was used for subsequent LLM experiments.

4 Experimentation and Results

4.1 Experimental Settings

All experiments were conducted on a Dell PowerEdge R7525 server with 96 CPU cores, 1 TB RAM and three NVIDIA A100-PCIE-40GB GPUs (CUDA 12.8, Driver 570.124.06).

The subset of the CECILIA-10C-900 dataset and its pseudonymised variants were divided using stratified sampling with a fixed seed of 42. Traditional ML classifiers were evaluated using five-fold stratified cross-validation, while transformer and LLM-based models used a 70/30 train–test split. Cross-validation was retained for traditional ML models because it provides a more robust estimate of performance on smaller models. However, applying this evaluation procedure to LLMs would be computationally prohibitive.

Both LLMs and transformers were evaluated using the same 70/30 split, to ensure methodological consistency and facilitate direct comparison between the two approaches. For all models, performance was reported using weighted precision, recall and F1-score, supplemented by 95% bootstrap confidence intervals over 1,000 iterations.

Traditional ML classifiers were implemented with Scikit-Learn (Pedregosa et al., 2011) with default hyperparameters, except for Support Vector Machine (linear kernel, balanced class weights) and Random Forest (100 estimators, Gini criterion, balanced weights). Transformer models were sourced from the HuggingFace Transformers library and fine-tuned with a batch size of 8, 10% warmup steps and 0.01 weight decay. For the LLM-based experiments, prompts (included in Appendix A) were executed with a temperature of 0.1. QLoRA fine-tuning was configured with an r parameter of 16 (Dettmers et al., 2023). For few-shot learning, three examples were included for each class.

4.2 Results

ML Classifiers Baseline XGBoost was the best-performing traditional ML classifier, as seen in Table 1. XGBoost achieved a weighted F1-score of 80.06% with the Bag of Words Vectoriser. On the other hand, Logistic Regression with TF-IDF Vectoriser performed the worst with a weighted F1-score of 63.62%.

Therefore, XGBoost establishes a strong performance baseline for traditional ML on this task and creates a baseline for subsequent experimentation.

Classifier	Vec.	P	R	F1	F1 95% CI
XGB	**BoW**	**79.63**	**81.08**	**80.06**	**[76.18, 83.95]**
XGB	TF-IDF	79.21	80.26	79.30	[75.34, 83.25]
LR	BoW	77.65	78.50	77.49	[74.78, 80.20]
SVM	BoW	76.93	78.50	76.82	[74.20, 79.44]
SVM	TF-IDF	76.83	77.21	76.00	[70.98, 81.02]

Table 1: The five best-performing traditional ML classifiers on a subset of CECILIA-10C-900 (support = 256). Vec = vectoriser, P = weighted precision, R = weighted recall, F1 = weighted F1-score, CI = Confidence Interval. All values in %. The best performance is **bolded**. XGB = XGBoost; LR = Logistic Regression; SVM = Support Vector Machine; BoW = Bag of Words. See Section 3.3 for citations.

Encoder-Based Transformer Classification Results The upper panel of Table 2 reports encoder-based transformer performance on a subset of CECILIA-10C-900. A general-purpose model, RoBERTa-base achieves the highest weighted F1-score of 84.35%, outperforming the best ML baseline by 4.82 percentage points.

Generally, the domain-specialised models underperformed their general-purpose counterparts. CyBERTuned, a model fine-tuned on cybersecurity corpora, was ranked second overall, with a weighted F1-score of 82.96%. Notably, SecureBERT2.0, the newest transformer model, demonstrated the most selective or conservative classification behaviour, resulting in the largest precision/recall difference among the models.

The worst-performing model, DeBERTa-v3 (68.74% F1-score) narrowly outperformed the worst-performing traditional ML classifier, Logistic Regression with TF-IDF (F1-score = 63.62%).

LLM Prompt-based Evaluation. In zero-shot settings, Spanish labels substantially outperformed their English counterparts (median F1 of 42.46% compared with 19.48%). Under few-shot prompting, however, this gap narrowed considerably, with the two conditions yielding comparable performance (median F1 of 60.90% and 58.60%, respectively). Given the consistent advantage observed in the zero-shot regime, Spanish labels were retained for all subsequent QLoRA fine-tuning experiments. An investigation of the mechanisms underlying the superior performance of Spanish-language labels falls outside the scope of the present study; nonetheless, this finding constitutes a promising avenue for future research.

QLoRA fine-tuning evaluation The model with the most parameters (9 billion) achieved the highest F1-score: Gemma-2-9B. Meanwhile, the worst-performing model, Mistral-7B, has the least parameters. Notably, none of the LLMs outperformed the ML Classifier baseline.

Among LLMs, domain specification appears to improve performance compared to transformers. The two domain-specialised models achieved the second and third-best performances. However, the best-performing LLMs also do not surpass the best-performing encoder-based transformers, which achieve the highest overall weighted F1-scores on the original dataset.

A common trend among the Llama variants assessed (that is, Foundation-Sec-8B, Llama-Primus-Merged and Llama-3.1-8B) is that they demonstrate higher recall than precision, thereby prioritising the detection of all incidents in a class, even if some are false positives.

4.3 Impact of Pseudonymisation

Table 2 presents the central finding of this study: pseudonymisation can improve classification performance under specific model–technique combinations.

Encoder-based transformers. Three of eleven encoder-based transformers show positive deltas across all techniques: MPNet, SecureBERT2.0 and SecROBERTa.

More generally, the median ΔF1-score is positive across all three techniques, with gains of +1.07pp for Data Tokenisation, +1.30pp for Data Masking, and +0.60pp for Data Substitution. However, the variance is substantial. RoBERTa-base, the best-performing model, most benefits from Data Tokenisation (+3.00pp) yet, while using that same technique, SecBERT exhibits a sharp performance decline (-8.84pp), the largest decline observed across all models. Though the median effect of Data Tokenisation is positive (+1.07pp),

Model	P	R	F1	F1 95% CI	ΔTokenisation	ΔMasking	ΔSubstitution
Encoder Transformers							
RoBERTa-base	**85.69**	**83.98**	**84.35**	[79.98, 88.40]	+3.00	+1.30	−0.93
CyBERTuned ×	83.87	82.81	82.96	[77.98, 87.13]	+1.07	+0.01	−2.00
BERT-base	82.11	83.59	82.70	[78.12, 87.20]	−2.48	−4.22	−5.69
SecBERT ×	82.32	82.42	82.12	[76.74, 86.56]	−8.84	−4.55	+0.60
MPNet †	80.96	81.64	80.96	[75.78, 85.45]	+1.85	**+4.08**	+0.21
SecureBERT ×	80.06	80.47	79.66	[74.28, 84.54]	−6.25	−5.36	+1.64
SecureBERT2.0 ×	83.85	77.73	79.11	[73.84, 83.55]	+2.35	+3.54	+0.80
DistilBERT	76.24	80.86	78.30	[72.68, 84.03]	−0.53	+1.46	+0.85
SecRoBERTa ×	76.83	77.73	77.13	[71.70, 82.32]	+3.23	+2.30	**+2.79**
ALBERT	74.84	78.12	76.10	[70.54, 81.62]	+0.84	−1.61	+2.59
DeBERTa-v3	67.97	70.31	68.74	[62.71, 74.76]	**+4.85**	+3.43	−1.69
Encoder Median					*+1.07*	***+1.30***	*+0.60*
QLoRA Fine-tuned LLMs							
Gemma-2-9B	**78.63**	**77.73**	**78.03**	[72.72, 83.23]	−3.82	−4.24	−4.61
Foundation-Sec-8B ×	68.18	69.14	68.45	[62.14, 74.60]	−1.14	−2.50	+4.66
Llama-Primus-Merged ×	66.33	67.19	66.13	[59.72, 72.35]	+0.60	+6.32	+9.95
Llama-3.1-8B	64.54	67.58	65.99	[59.89, 71.98]	**+8.73**	**+12.63**	**+12.31**
Mistral-7B	65.33	65.23	64.98	[59.22, 70.85]	+3.11	+2.99	+1.95
QLoRA Median					*+0.60*	*+2.99*	***+4.66***

Table 2: Model performance (%) on CECILIA-10C-900, and the impact of pseudonymisation. RoBERTa-base with Data Tokenisation achieved the highest overall F1 score (87.35%, i.e., 84.35 + 3.00 pp), while Llama-3.1-8B showed the largest gain from pseudonymisation (+12.63 pp with Data Masking). P, R, and F1 are weighted precision, recall and F1 scores on the Original (non-pseudonymised) variant; F1 95% CI is the F1 score confidence intervals from bootstrap (1,000 iterations). Δ columns show F1 score change in percentage points. Best F1 score per model family (encoder transformers vs fine-tuned LLM) in **bold**. × denotes a domain-specialised model; † highlights the best classifier from the original CECILIA-10C-900 research article (Delgado et al., 2024). All other citations are found in Section 3.3.

this technique produces the most inter-model variability, with a range of 13.69 percentage points (−8.84pp for SecBERT to +4.85pp for DeBERTa-v3).

Notably, the results of Data Masking on transformer models are the opposite of those of the ML pseudonymisation baseline. With ML classifiers, seven of the ten pipelines experienced performance degradation when Data Masking was applied, while seven transformer models (not the same models) saw increased performance.

These increases can be partially attributed to the Fraud class, which was the most volatile and pseudonymisation-sensitive. On the original dataset, BERT-base and ALBERT score 0.00 F1-score on Fraud. With Data Masking, BERT-base increases to 33%, likewise with Data Tokenisation, ALBERT scores 18%. This trend continued with CyBERTuned, which scores 53% on the non-pseudonymised Fraud class but 76% with the same class in CECILIA-6C-MAS and 83% on CECILIA-6C-TOK.

QLoRA fine-tuned LLMs. The lower panel of Table 2 reveals a striking pattern. QLoRA fine-tuned models generally benefit more from pseudonymisation than encoders, with median deltas of +0.60 (Data Tokenisation), +2.99 (Data Masking), and +4.66 (Data Substitution). There is also a notable increase in Data Substitution effectiveness when compared with encoder transformers. While encoders show a near-neutral median response to synthetic substitution, QLoRA models show a clear positive trend.

Llama-3.1-8B exhibited the strongest gains across all three strategies, despite its low baseline on original data (F1-score = 65.99%). On a more granular level, this is because Data Masking reduced the likelihood that the Intrusion class would be the default prediction class. After Data Masking, the accuracy of the Malicious Code class predictions more than doubles to 65.6%, Intrusion improves to 58.3% and the majority class rises from 86.4% to 91.7%. This is shown in Figure 2.

Conversely, Gemma-2-9B, the strongest QLoRA model on the unpseudonymised data (78.03%), shows consistent degradation across all strategies. This is reflected in per-class analysis where the model classified 69.4% of the Intrusion class correctly, and then after Data Masking was applied, experienced a 19% performance degradation in this class.

Robustness of findings The robustness of these findings varies by model family. For QLoRA fine-tuned models, confidence intervals computed on individual variant F1 scores provide strong corroborating evidence: the lower bounds of the 95% Confidence Intervals (CIs) for Llama-3.1-8B's Data Masking [0.7336, 0.8328] and Data Substitution [0.7295, 0.8336] both exceed the upper bound of the CECILIA-10C-900 subset's baseline CI [0.5989, 0.7198], indicating non-overlapping intervals and a robust effect. Encoder transformer deltas are more modest; for smaller gains the CIs overlap substantially, and these should be interpreted as directional rather than definitive. Overall, the strongest benefits of pseudonymisation are observed in QLoRA-based LLMs.

5 Discussion

Using pseudonymisation to remove noise, not signal. The finding that pseudonymisation can improve classification adds nuance to the prevailing assumption that privacy preservation degrades utility. We hypothesise that personal data (such as names, dates and locations) may function as noise for incident categorisation. It is possible that these tokens do not determine whether a report describes Fraud or an Intrusion; instead they introduce lexical variation to which models may overfit. Therefore, when entities are removed or replaced, models may be more likely to focus on the cybersecurity content (attack descriptions, techniques, impacts) that actually determines the incident class. This is distinct from healthcare settings, where patient demographics may be clinically relevant.

The Fraud class offers the clearest evidence. Models that scored a 0.00 F1 score on Fraud in the original dataset recovered substantially after pseudonymisation, and even models that already classified Fraud saw further gains: CyBERTuned, for instance, improved from 53% to 83% after Data Tokenisation. One possible explanation is that entity tokens were pulling models toward higher-support classes, and removing them allowed the Fraud class to become more distinguishable. Notably, the strength of this effect varies by model family. As part of our future work, we intend to confirm this hypothesis with per-entity ablation.

The optimal technique depends on the model family. In our experimentation, pseudonymisation was correlated with an increased median weighted F1-score. However, the technique that works best differs across model families attributable to how each architecture processes input.

For encoder transformers, Data Masking yields the strongest median improvement (+1.30pp). This is consistent with the noise-removal hypothesis: the more aggressively entities are obscured, the more the model attends to structural cues. Encoders are also pre-trained with masked language modelling, so masked input is a familiar format for them.

Traditional ML classifiers show the opposite pattern. Seven of ten pipelines degraded under Data Masking, possibly because collapsing distinct entity tokens into a single placeholder destroys the token-frequency signal that bag-of-words and TF-IDF representations depend on.

In our experimentation, among QLoRA fine-tuned LLMs, Data Substitution is the most effective technique, even surpassing performance on non-pseudonymised data. We posit three complementary explanations: substitution eliminates spurious associations between real-world entities and incident categories encoded in the model's pre-training corpus; it mitigates potential interference from safety guardrails that recognised threat indicators may activate; and it preserves natural language fluency. This last point may be architecturally significant. Unlike encoders, autoregressive LLMs process text left-to-right and depend on natural-sounding input for comprehension. Therefore, masked tokens like "xxxx" may disrupt this flow, whereas substituted text reads like an ordinary cyber incident report, possibly making it a more compatible input format.

Why individual models diverge. Within each model family, the magnitude and even direction of pseudonymisation's effect varies considerably. Two factors may explain most of this variation: baseline strength and entity reliance.

In our experimentation, models with weaker baselines tend to benefit most. Llama-3.1-8B gained up to +12.63pp from a baseline of 65.99%. Per-class analysis reveals why: on original data, 40.6% of Malicious Code samples were misclassified as Intrusion. After Data Masking, Malicious Code accuracy more than doubled, a pattern which may mean that the entity tokens had biased the model toward the Intrusion class, an effect that masking appears to have reduced.

Gemma-2-9B is an outlier. Despite being the best-performing QLoRA model (78.03%), it is the only LLM that degrades under all three techniques. Its architecture, which features Grouped-Query Attention and a substantially larger vocabulary (256k versus $\sim$128k for the Llama-3.1 variants), may enable it to encode named entities as fine-grained lexical features correlated with incident type. When pseudonymisation removes these tokens, Gemma

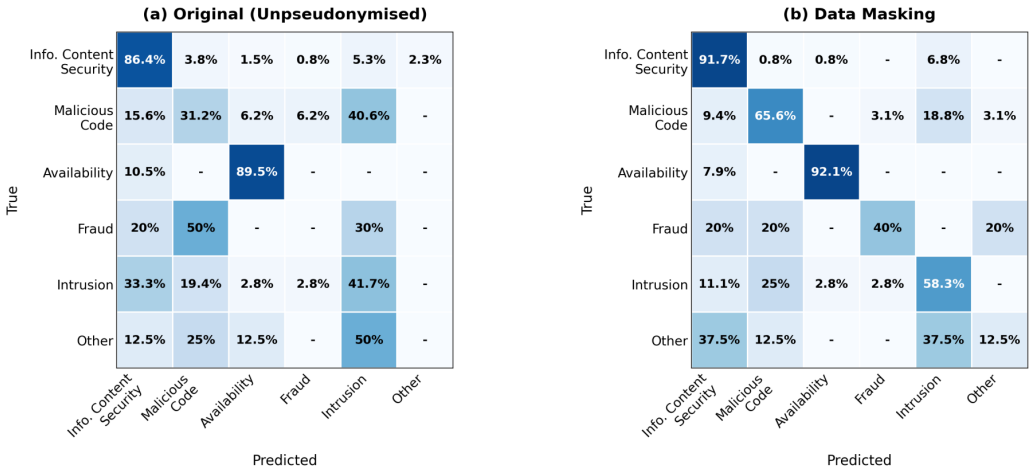

Figure 2: Llama-3.1-8B displayed improved performance on the CECILIA-6C-MAS dataset on the right, in comparison to the original (non-pseudonymised) subset of CECILIA-10C-900 on the left.

loses signal that its architecture was uniquely positioned to exploit. Intrusion accuracy, for example, drops from 69.4% to 50.0% after Data Tokenisation. One interpretation could be that whereas Llama-3.1-8B was misled by entities, Gemma was leveraging them. The fact that pseudonymisation improves the performance of Llama-3.1-8B but degrades that of Gemma is consistent with this hypothesis.

Domain-specialised models, whether encoder based (SecBERT, SecureBERT) or autoregressive (Llama-Primus-Merged), follow a similar pattern to Gemma, though for different reasons. Domain-specific pre-training appears to encourage reliance on entity tokens as classification signal. Consequently, masking and tokenisation disrupt in-distribution representations, while substitution, which preserves surface-level plausibility, is better tolerated. This is particularly evident in the Malicious Code class, where both SecBERT and SecureBERT lose 18pp under Data Masking, a class rich in the structured technical indicators that domain vocabularies are optimised to capture.

Finally, the finding that three encoders show positive deltas across all techniques (MPNet, SecureBERT2.0 and SecRoBERTa) suggests that some architectures are inherently resilient to pseudonymisation. However, this robustness is not exclusive to a single architecture, since while the latter two models are based on RoBERTa-base, MPNet is not. This finding may warrant further investigation.

Implications for CTI pipelines. These findings carry a practical implication: pseudonymisation technique selection should be guided by the downstream model rather than applied uniformly. Before deploying pseudonymisation in production, practitioners should assess whether their chosen model relies on entity-level features (for instance through attention analysis or ablation), since removing those features may degrade performance for architectures capable of leveraging them.

6 Conclusion

This work suggests that pseudonymisation may enhance automated cyber incident classification, under specific model–technique pairings. Across eleven transformer and five LLM models, positive median deltas were observed for all three pseudonymisation strategies, with RoBERTa-base achieving an 87.35% weighted F1-score, a 3.00 percentage point gain over its baseline.

Crucially, our experimentation suggested that the optimal technique varies by architecture: encoder transformers benefit most from Data Masking, QLoRA fine-tuned LLMs from Data Substitution and domain-specialised models are sensitive to technique choice. Furthermore, the best improvements emerged in models that had the weakest baselines on original data.

These findings suggest that pseudonymisation technique selection should be model-guided, a consideration not typically addressed in current data protection guidance. For organisations navigating the tension between cyber incident reporting mandates and data minimisation obligations, these findings may reframe pseudonymisation from a compliance obligation to a preprocessing step with measurable utility benefits in the settings we evaluated. Ultimately, this could provide a practical path to satisfying both incident reporting obligations and data minimisation principles.

Limitations

CECILIA-10C-900 dataset. First, the CECILIA-10C-900 dataset is imbalanced and small-scale (852 unique records) which limits generalisation to minority classes. Second, as incidents in this dataset were sourced publicly, the personal data present is likely limited in scope, which could affect the generalisability of the findings. Likewise, all incidents evaluated are in English, though cyber incident reports exist in many languages.

Finally, the multilabel nature of cyber incidents is not addressed; therefore, single-label evaluation may penalise models when incident categories inherently overlap.

Contextual Utility. The downstream utility gains reported in this body of work are specific to automated classification of incident type. However, pseudonymisation reduces contextual utility for downstream *analytical* tasks. In operational CTI settings, entity-level details (such as targeted organisation names, geographic locations and sector affiliations) can be the primary signals for threat attribution and strategic intelligence. By replacing these with generic tokens, classification accuracy may improve, but the tactical nuance an analyst requires for attribution and situational awareness (the "who" and "where") is diminished. This trade-off is inherent to pseudonymisation and highlights that the utility gains reported here are specific to automated classification; complementary approaches for preserving contextual utility in analyst-facing workflows remain an open challenge.

NER Annotation Methodology. NER annotations were produced by a single annotator, therefore increasing the likelihood of bias. Additionally, the pseudonymisation applied in this study relies on manual entity annotations, effectively simulating an oracle (or idealised) NER pipeline. In production, automated NER systems will inevitably miss entities (false negatives), meaning that personal data may still reach third-party processors. This has a direct privacy implication since the real-world privacy guarantee of our solution is effectively bounded by NER recall.

Furthermore, NER errors such as precision degradation, boundary mismatches and varying performance across entity types could introduce noise absent from our controlled setting, potentially reducing the classification gains observed here. Quantifying this sensitivity (for instance, by evaluating classification performance under degraded NER recall) remains an important direction for future work.

Evaluation Protocols. For few-shot learning, the first three samples per training class were selected as the examples to be used in the prompt. While this design choice promotes reproducibility, it may also introduce selection bias. Future work could systematically compare different few-shot example selection strategies (e.g., random sampling and prototype-based selection) to assess the sensitivity of results to this choice.

Additionally, we reported aggregated metrics across all entity types rather than consistently isolating the performance of individual entities, such as persons, dates or organisations. Consequently, the specific influence of different entity types on overall model utility remains unexamined. Conducting a per-entity-type ablation study represents a critical next step to clarify these granular dynamics.

Finally, the encoder transformer and LLM evaluations relied on a single train-test split, on which we reported bootstrap confidence intervals. However, paired significance tests across variants and multi-seed evaluation would further strengthen robustness claims. Likewise, we reported weighted metrics; however, incorporating macro-F1 and full per-class results across all variants would provide additional granularity, particularly for minority classes.

Future Work. There is an opportunity to conduct further research on: (i) scaling to larger, multilingual datasets; (ii) integrating data augmentation with pseudonymisation; (iii) scaling to multilabel taxonomies such as the MITRE ATT&CK framework or multilevel taxonomies, for example, incorporating both class and subclass levels of the INCIBE taxonomy; (iv) expanding the suite of LLMs to include larger models, since all LLMs used had 9 billion or fewer parameters, (v) combining pseudonymisation with differential privacy for formal guarantees, (vi) exploring the use of LLMs for automated NER and dataset pseudonymisation; and (vii) conducting per-entity ablation studies for both NER recall sensitivity and entity-type contribution.

Ethics Statement

This research aims to improve privacy in cybersecurity incident handling. We note that pseudonymised data remains personal data under the GDPR, and organisations deploying this approach must implement appropriate technical and organisational safeguards, including secure storage of any re-identification keys.

We caution that the pseudonymisation applied in this study was performed using manual entity annotations, as described in Section 3.1. Deploying automated pseudonymisation in production carries the risk of incomplete entity detection, which could result in the unintended disclosure of personal data to third-party processors. Practitioners adopting pseudonymisation as a preprocessing step should therefore not treat it as a substitute for comprehensive data protection impact assessments and should evaluate the recall of their NER systems before processing sensitive data through external services.

Finally, per the methodology of Delgado et al. (2024), the CECILIA-10C-900 dataset was sourced from publicly available cyber incident reports. Therefore, no human subjects were recruited.

Acknowledgements

This work is funded by the European Union through the Secure Interactive Environments for SensiTive data Analytics (SIESTA) Project (Horizon Europe) under grant number 101131957. The authors also acknowledge the contributions of the GVIS Annotation Team.

References

Muhammad Abdullah, Muhammad Munib Nawaz, Bilal Saleem, Maila Zahra, Effa binte Ashfaq, and Zia Muhammad. 2025. Evolution Cybercrime—Key Trends, Cybersecurity Threats, and Mitigation Strategies from Historical Data. *Analytics*, 4(3):25.

Sebastian Abt and Harald Baier. 2014. A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, pages 37–45, Scottsdale Arizona USA. ACM.

Ehsan Aghaei, Sarthak Jain, Prashanth Arun, and Arjun Sambamoorthy. 2025. SecureBERT 2.0: Advanced language model for cybersecurity intelligence.

Ehsan Aghaei, Xi Niu, Waseem Shadid, and Ehab Al-Shaer. 2023. SecureBERT: A domain-specific language model for cybersecurity. In *Security and Privacy in Communication Networks: 18th EAI International Conference, SecureComm 2022*, pages 39–56. Springer.

Adham Albakri, Eerke Boiten, and Rogério De Lemos. 2019. Sharing Cyber Threat Intelligence Under the General Data Protection Regulation. In *Privacy Technologies and Policy*, pages 28–41, Cham. Springer International Publishing.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Juan José Delgado, Eduardo Fidalgo, Enrique Alegre, Andrés Carofilis, and Alicia Martínez-Mendoza. 2024. CECILIA: Enhancing CSIRT effectiveness with transformer-based cyber incident classification. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 186–195, Lancaster, UK.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. 2025. A Survey on Current Trends and Recent Advances in Text Anonymization. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9.

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.

ENISA. 2018. Reference incident classification taxonomy. Technical report, European Union Agency for Cybersecurity (ENISA). Produced by the Reference Security Incident Taxonomy Working Group (RSIT WG).

European Data Protection Board. 2025. Guidelines 01/2025 on pseudonymisation. Technical Report 01/2025, European Data Protection Board. Adopted 16 January 2025; subject to public consultation.

European Network and Information Security Agency. 2019. *Pseudonymisation techniques and best practices: recommendations on shaping technology according to data protection and privacy provisions.* Publications Office, LU.

European Parliament and Council of the European Union. 2016. 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 (General Data Protection Regulation). OJ L 119, 4.5.2016.

Daniele Faraglia. 2024. Faker: A Python package that generates fake data. Accessed: 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing.

Marti A. Hearst, Sue T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*, 3rd edition. John Wiley & Sons.

INCIBE Instituto Nacional de Ciberseguridad. 2020. Guía nacional de notificación y gestión de ciberincidentes.

Ehtsham Irshad and Abdul Basit Siddiqui. 2023. Cyber threat attribution using unstructured reports in cyber threat intelligence. *Egyptian Informatics Journal*, 24(1):43–59.

jackaduma. 2021. SecRoBERTa: A RoBERTa-based model pre-trained on cybersecurity text. `https://huggingface.co/jackaduma/SecRoBERTa`. HuggingFace model repository. Accessed: 2025.

jackaduma. 2022. SecBERT: A pretrained language model for cyber security text. `https://huggingface.co/jackaduma/SecBERT`. Accessed: 2026.

Eugene Jang, Jian Cui, Dayeon Yim, Youngjin Jin, Jin-Woo Chung, Seungwon Shin, and Yongjae Lee. 2024. Ignore me but don't replace me: Utilizing non-linguistic elements for pretraining on the cybersecurity domain. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 29–42. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, et al. 2025. Llama-3.1-foundationai-securityllm-base-8b technical report.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Pierre Lison, Ildikó Pilán, David Sánchez, 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 (ACL-IJCNLP)*, pages 4188–4203. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Jingfei Du Naman Goyal, Mandar Joshi, Omer Levy Danqi Chen, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Llama Team and AI Meta Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Huda Lughbi, Mourad Mars, and Khaled Almotri. 2024. CybAttT: A dataset of cyberattack news tweets for enhanced threat intelligence. *Data*, 9(3).

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. *AAAI Workshop on Learning for Text Categorization*, 752:41–48.

Gohar Mumtaz, Sheeraz Akram, Muhammad Waseem Iqbal, M. Usman Ashraf, Khalid Ali Almarhabi, Ahmed Mohammed Alghamdi, and Adel A. Bahaddad. 2023. Classification and prediction of significant cyber incidents (sci) using data mining and machine learning (dm-ml). *IEEE Access*, 11:94486–94496.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Lasse Nitz, Mehdi Akbari Gurabi, Milan Cermak, Martin Zadnik, David Karpuk, Arthur Drichel, Sebastian Schäfer, Benedikt Holmes, and Avikarsha Mandal. 2025. On Collaboration and Automation in the Context of Threat Detection and Response with Privacy-Preserving Features. *Digital Threats*, 6(1):5:1–5:36.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ian Pépin and Farhana Zulkernine. 2022. A Comparative Study of De-identification Tools to Apply to Free-Text Clinical Notes. In *Proceedings of the 32nd Annual International Conference on Computer Science and Software Engineering*, CASCON '22, pages 181–186. IBM Corp.

Giovanni Rabitti, Amir Khorrami Chokami, Patrick Coyle, and Ruben D. Cohen. 2024. A taxonomy of cyber risk taxonomies. *Risk Analysis*.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sandra Schmitz-Berndt. 2023. Defining the reporting threshold for a cybersecurity incident under the NIS Directive and the NIS 2 Directive. *Journal of Cybersecurity*, 9(1):tyad009.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nan Sun, Ming Ding, Jiaojiao Jiang, Weikang Xu, Xiaoxing Mo, Yonghang Tai, and Jun Zhang. 2023. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. *IEEE Communications Surveys & Tutorials*, 25(3):1748–1774.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Rinalds Vīksna and Inguna Skadina. 2025. Anonymise: A Tool for Multilingual Document Pseudonymisation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1327–1332, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*, pages 232–241. Association for Computational Linguistics.

Yao-Ching Yu, Tsun-Han Chiang, Cheng-Wei Tsai, Chien-Ming Huang, and Wen-Kwang Tsao. 2025. Primus: A pioneering collection of open-source datasets for cybersecurity llm training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10402–10424. Association for Computational Linguistics.

Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. 2025. When LLMs meet cybersecurity: A systematic literature review. *Cybersecurity*, 8(1):55.

A Appendix

A.1 Zero-Shot Prompt

```
You are a cybersecurity analyst.
Classify the following cyber
incident report into exactly ONE
of these categories:
{class_list}

Cyber incident report:
{text}

Respond with ONLY the exact
category name from the list above.
Do not add any explanation,
punctuation, or extra text.
```

A.2 Few-Shot Prompt

```
You are a cybersecurity analyst.
Classify cyber incident reports
into exactly ONE of these
categories: {class_list}

Here are examples:
{example_block}

Now classify this report:
{text}

Respond with ONLY the exact
category name. No explanation.
```

Note: In the above prompt, *example_block* is the list of eighteen representative cyber incidents (three per class) and their respective class label.

A.3 Entities used to annotate CECILIA-10C-900

Table begins on the next page.

#	Entity Label & Description
1	**Person** — names of persons
2	**NORP** — nationalities, religious or political groups
3	**Facilities** — buildings, airports, motorways, bridges, etc.
4	**GPE_Location** — countries, cities, states, etc.
5	**Location** — mountains, rock formations, seas, rivers, etc.
6	**Organization** — organizations, companies, institutions, etc.
7	**Products** — objects, vehicles, meals (services excluded)
8	**Events** — sporting, social, or atmospheric events; wars, battles
9	**Law** — legal documents
10	**Date** — absolute dates, relative dates, and time periods
11	**Time** — any period of time less than one day
12	**Art** — titles of books, songs, or other literary works
13	**Money** — monetary values with unit (e.g. $1000, 500€)
14	**Language** — any language
15	**Group** — groups excluding nationalities/religious/ political (e.g. hacker groups)
16	**Quantities** — measurements such as weight, distance, etc.
17	**Numbers_O** — ordinal numbers (first, second, third, etc.)
18	**Numbers_C** — other numbers not captured by Date, Time, Numbers_O, or Quantities

Table 3: CECILIA-10C-900 entity labels and their descriptions. Adapted from internal communication of the GVIS Research Group.