An Evaluation Study of Hybrid Methods for Multilingual PII Detection

Harshit Rajgarhia[1], Suryam Gupta[1], Asif Shaik[1], Gulipalli Praveen Kumar[1], Y Santhoshraj[1], Sanka Nithya Tanvy Nishitha[1], Abhishek Mukherji[1]
[1]Centific Global Solutions Inc.

Abstract

The detection of Personally Identifiable Information (PII) is critical for privacy compliance but remains challenging in low-resource languages due to linguistic diversity and limited annotated data. We present **RECAP**, a hybrid framework that combines deterministic regular expressions with context-aware large language models (LLMs) for scalable PII detection across 13 low-resource locales. **RECAP**'s modular design supports over 300 entity types without retraining, using a three-phase refinement pipeline for disambiguation and filtering. **Benchmarked** with `nervaluate`, our system outperforms fine-tuned NER models by 82% and zero-shot LLMs by 17% in weighted F1-score. This work offers a scalable and adaptable solution for efficient PII detection in compliance-focused applications.

1 Introduction

The exponential growth of user-generated content has created vast repositories where Personally Identifiable Information (PII) often remains exposed, posing significant privacy risks and compliance challenges [1, 2]. To address these threats, regulations like GDPR, HIPAA, and CCPA mandate strict safeguards and penalties for non-compliance [3, 4, 5]. Additionally, as Large Language Models (LLMs) are deployed into production systems, robust evaluation across the model becomes paramount, and is especially critical for sensitive tasks like PII detection, which demand high precision, reliability, and adaptability. Yet, existing PII annotation systems struggle with ambiguity, format variability, and scaling over low-resource languages.

In this work, we propose **RECAP** (**RE**gex and **C**ontext-**A**ware **P**rompting), that combines a regex-based deterministic solution with context-enriched LLM for PII detection, addressing key limitations of Named Entity Recognition systems, zero-shot LLMs, and rule-based methods.

Core Challenges Addressed: (1) **Low-Resource Performance Gap:** Existing entity recognition systems often perform poorly in low-resource locales [1] due to lack of annotated training data, limited linguistic resources, and the high computational cost associated with training models for each new language or domain; (2) **Scalability Bottleneck:** Pure regex methods lack semantic understanding, while transformer-based NER models suffer from limited PII type coverage. Standalone LLMs, though flexible, produce inconsistent outputs and are prone to hallucination; and (3) **Ambiguity and Variation:** PII entities exhibit both structural variation and semantic ambiguity across locales, making them difficult to classify accurately using traditional approaches, which often result in missed and conflicting labels, thereby reducing overall reliability.

Key Contributions: (1) To our knowledge, our work is the first to introduce a PII detection framework that spans across 13 diverse, low-resource locales, with support for 300+ PII types across six domains, without requiring any model training or fine-tuning; (2) Our novel solution combines deterministic regex patterns with context-aware LLMs to advance PII detection for low-resource locales; (3) **RECAP** implements a three-phased pipeline (multi-label disambiguation, span consolidation, contextual filtering) to systematically reduce ambiguity and false positives; (4)

[1]"Low-resource" locales have limited publicly available annotated data for training.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

We present detailed **benchmarking** results using `nervaluate` evaluation framework [6], where **RECAP** outperforms the state-of-the-art baselines, NER by 82% and LLM by 17% (weighted F1-score).

2 Related Work

Multilingual & Low-Resource PII Detection: Most work focuses on high-resource languages like English. Exceptions include deep learning for Luganda PII [7], and few-shot cross-lingual methods for clinical texts [8]. Datasets from MultiCoNER [9], AI4Privacy [10], BigCode [11] offer broader coverage but still lack low-resource representation — a gap our work addresses.

Regex-based Detection: Regular expressions (Regex) have been employed for general NER tasks like automated resume parsing [12] and high-risk PII detection [13]. However, this approach suffers from high false positives and performs poorly on unstructured formats.

Deep Learning Methods: Models like CASSED use BERT for structured data [14], while DTL-PIIE employs transfer learning for social media text [15]. BERT-based models show strong performance with balanced data [16] but struggle with multi-label entities and numeric false positives.

LLM-based Approaches: LLMs are used for both PII detection and synthetic data generation (e.g., SPY [17], ProgGen [18]). They perform well in domain-specific settings like education [19, 20] and chemistry [21]. Models like GPT-NER frame detection as a generation task [22], and strategic zero-shot prompting detects sensitive information across global contexts [23]. However, these approaches often suffer from over-redaction and hallucination, misidentifying non-PII.

3 Solution Architecture and Workflow

Our **RECAP** architecture (Figure 1) is designed to tackle the inherent challenges of multilingual PII detection by combining the precision of rule-based methods for structured PIIs [2], with the semantic understanding of LLMs for unstructured PIIs. The system employs a modular, locale-aware design where each of the 13 supported locales has a dedicated detector containing its specific regex patterns and optimized prompts. The core of our approach is a three-phase refinement pipeline that progressively improves detection quality from an initial hybrid baseline to a final refined output, as summarized in Table 1 and evidenced by the performance gains across locales in Table 3. Sample sizes with locale information are shown in Table 2.

I. Baseline Hybrid Detection: The process begins by receiving a text sample and its associated locale, invoking the corresponding locale-specific detector. The text is processed by a comprehensive set of regular expressions to detect structured PIIs. These can be categorized into two types: (a) universal patterns for entities like `IP_ADDRESS` that follow global formats, and (b) locale-specific patterns for entities like national IDs (e.g., India's `AADHAAR_IN` vs. Belgium's `SSN_BE`), which require custom regex patterns per country. In parallel, the entire text is passed to an LLM (GPT-4o) using a carefully engineered zero-shot prompt (Listing 1) to detect unstructured PIIs (`NAME`, `ADDRESS`, `USERNAME`, and `PASSWORD`). This hybrid baseline provides broad coverage in terms of detecting PII entities, but introduces three key challenges: (1) **Multi-labeling** from semantically different but syntactically similar regex patterns; (2) **Span overlaps**, where one entity is fully or partially contained within another, leading to redundant and inconsistent labeling; and (3) **Contextual false positives** on short numeric sequences like `CVV` (Card Verification Value) or `AGE` that appear in non-sensitive contexts.

II. Context-based Multi-label Resolution: This phase focuses on resolving ambiguity in cases where a single entity span is assigned multiple candidate labels by the baseline. These multi-labeled outputs typically arise from identical syntactic patterns across numerically formatted entities. While the baseline regex method effectively identify entities, they may be unable to determine which label is most contextually appropriate. Our resolution module identifies all entities assigned multiple labels. For each, the original text, the character span, and the candidate labels are passed to the LLM with a custom prompt that instructs it to analyze the surrounding context and select the single most appropriate label (Figure 2, Top). This leverages the LLM's semantic understanding to resolve ambiguities that are intractable for rules alone and ensures consistent, context-aware labeling, significantly boosting precision and recall.

III. Ambiguity Resolution and Entity Consolidation: The final phase applies two targeted filters to produce a clean, coherent set of predictions. (1) **Entity Span Overlap Resolution**: A deterministic

[2] We refer to *Structured PIIs* as those which have a syntactically regular, well-defined formats and are usually represented by numerical patterns, such as `SSN` (a unique ID number in the US). In a similar fashion, *Unstructured PIIs* refers to entities which have semantically variable and arbitrary patterns, such as `ADDRESS`.

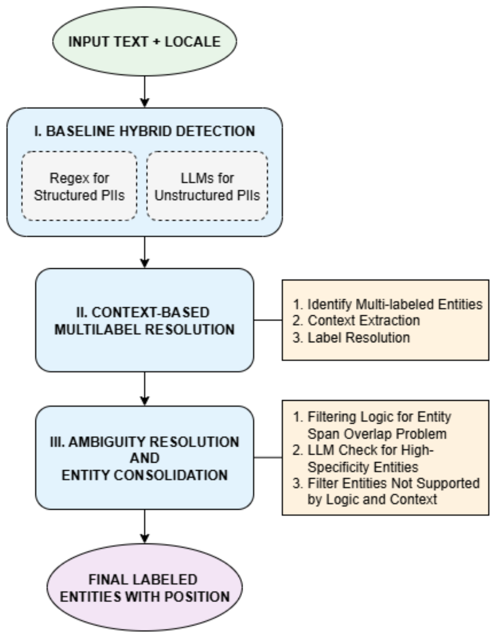

Figure 1: **RECAP** Architecture

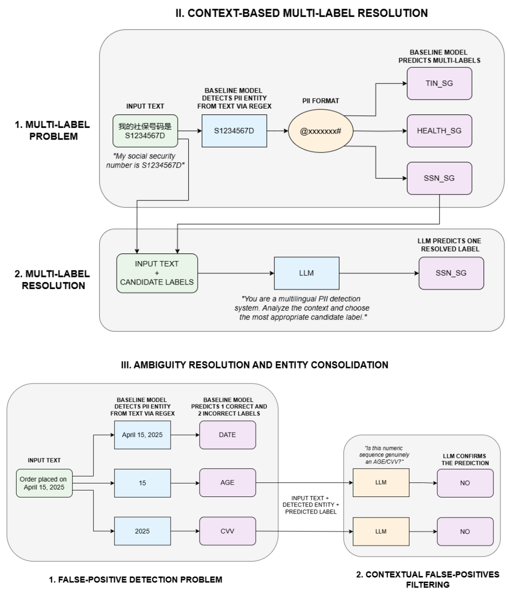

Figure 2: Multi-labeling (top) and False Positives (bottom) detection problem and resolution

algorithm processes all entities sorted by their start position, pre-defined label priority, and span length. It removes entities that are fully contained within a longer span if they have a lower priority (e.g., filtering out `AGE`="24" when it is contained within a correctly identified `ADDRESS`="24 Lincoln Avenue, NY"). (2) **Contextual False Positive Filtering**: For high-specificity, short numeric entities (`AGE`, `CVV`), a local context window (one sentence before and after the entity) is extracted. This context is submitted to the LLM to verify whether the numeric value is semantically plausible as the predicted PII type. An entity is retained only upon LLM confirmation (Figure 2, Bottom), drastically reducing false positives that arise from numeric coincidences in non-PII settings, thereby improving overall precision while maintaining recall.

Table 3 and 5 highlights consistent F1-score improvements across the three phases of **RECAP** architecture. Most locales showed steady gains, with `SV`_SE and `PT`_BR achieving notable increases of 77.53% and 47.76%, respectively. While `NL`_BE showed marginal fluctuations due to high initial performance, the overall trend highlights the efficacy of each refinement in enhancing PII detection.

Table 1: Performance progression across different phases

Phases	**Baseline Hybrid Detection**	**Context Based Label Resolution**	**Ambiguity and Entity Consolidation**
Solution Approach	Regex detection for structured entities and zero-shot LLM for unstructured ones	Multi label disambiguation using LLM with context-aware resolution	Suppression of overlap spans and false positives using logic filtering and LLM check
Weighted F1 Score (Gain)	0.511 (-)	0.585 ($\Delta \approx 14.48\%$)	0.657 ($\Delta \approx 12.30\%$)
Impact	Establishes initial coverage across 13 locales using structured rules and semantic generalization	Reduces incorrect or overly generic labels to improve precision and semantic alignment	Filters short numeric patterns and improves consistency by resolving labels in overlap spans

4 Benchmark Results and Comparative Evaluation

Benchmark Design: We evaluate **RECAP** against two strong baselines: (1) **transformers-based NER**: We select the best available HuggingFace model for each locale (see Appendix, Figure 4), representing traditional fine-tuned entity recognition models. These models are typically limited to generic labels (`PER`, `ORG`, `LOC`, `MISC`). (2) **Zero-shot LLM**: We use GPT-4o with a natural language prompt for PII extraction. While adaptable to multiple languages and label types, these models exhibit high variability in outputs, inconsistent formatting, and require careful prompt design.

The evaluation dataset was created by experts, who authored text samples and injected synthetic PII across six domains (Finance, Travel, Healthcare, IT, CPG, Media). This allows for a control

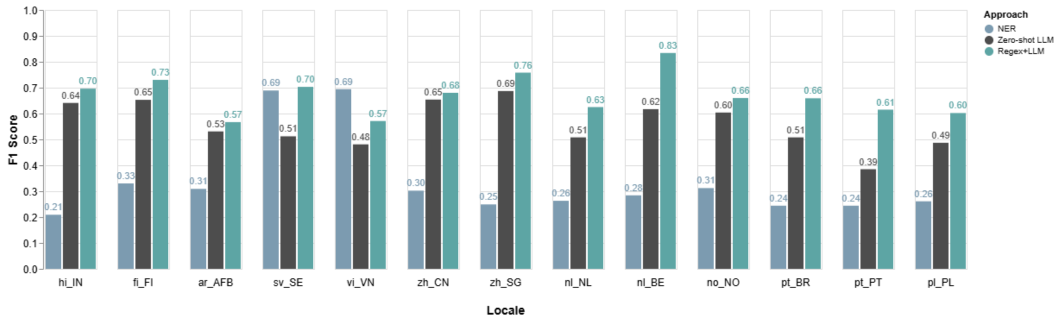

Figure 3: F1 Scores by Approach and Locale

over entity types and distributions. Text length varied from short (<21 words), to medium (21–240 words), large (240–1000 words), and extra-large (up to 4500 words), and were uniformly selected to ensure robustness. We use the `nervaluate` library with an **Exact Evaluation Method**, which requires a predicted entity's character span to match the gold span exactly to be counted as correct. Since our pipeline resolves label ambiguity at a later stage, we focused on evaluating span accuracy independently. This rigorous method emphasizes precise boundary detection, which is critical for PII redaction tasks. The primary evaluation metrics used across all comparisons were Accuracy, Precision, Recall, and F1 Score for label imbalance.

Comparative Results and Analysis: Figure 3 presents comparative results achieved by each approach across all locales (full results in Table 4). In low-resource settings such as Polish (PL_PL), **RECAP** (F1=0.60) achieves a 130.77% relative improvement over the NER baseline (F1=0.26) and a 22.45% improvement over the zero-shot LLM (F1=0.49). The **RECAP** framework consistently outperforms both traditional NER models and zero-shot LLMs across most locales, with ZH_SG and NL_BE gaining F1 scores as high as 0.76 and 0.83, respectively.

In sensitive PII detection tasks, recall is particularly important, as false negatives - missed detections - can lead to serious privacy breaches. While false positives may cause over-redaction, false negatives risk direct data exposure. **RECAP** achieves a weighted recall of 0.605, compared to 0.362 for NER (+67.13%) and 0.437 for zero-shot LLMs (+38.44%), demonstrating significantly stronger detection.

It is important to note that this is not a one-to-one comparison across approaches. Our architecture is explicitly designed to detect a fixed schema of 300+ predefined PII types across 13 locales. While LLMs were applied to detect this full set, their inherent variability often resulted in inconsistent formatting and hallucinated entities outside the defined set. In contrast, NER baselines are restricted to a narrow set of generic entity types, and do not encompass the broader set of sensitive identifiers commonly required in PII detection tasks. As such, while we report unified F1 scores, these results reflect fundamentally different label scopes and should be interpreted accordingly.

5 Conclusion

We presented **RECAP**, a hybrid PII detection architecture that combines regex patterns with prompt-based LLMs. **RECAP** addresses structural variation and low-resource challenges by leveraging rule-based precision and LLM-based contextual reasoning. Its modular design enables per-entity and per-locale customization without retraining. Benchmarked against a fine-tuned NER system and a zero-shot LLM, **RECAP** outperforms both across 13 diverse locales, particularly on complex and region-specific entities.

6 Limitations and Future Work

Limitations: (1) Use of different NER models per locale limits consistency in cross-locale baseline comparisons; (2) Reliance on a single LLM (GPT-4o); other models may offer complementary strengths; (3) Synthetic benchmark data may not capture the full complexity of real-user text; (4) Label set mismatch between broad PII types (300+) and narrow NER types (e.g., PER, ORG) complicates direct evaluation; (5) Domain coverage in our work (6 domains) may not represent all production text types (e.g., legal, social media).

Future Work: (1) Investigate RL for automatic prompt optimization per label/locale; (2) Design perturbation-based evaluation for robustness testing; (3) Explore LLMs for automated regex generation [24]; (4) Apply knowledge distillation for on-device inference; (5) Develop active learning with expert feedback to refine regex and models.

References

[1] Andrew Brzezicki. 64 user-generated content statistics to know in 2024. 2023.

[2] Andrew Brzezicki. Pii cybersecurity: Top threats and solutions. URL https://bigid.com/blog/pii-cybersecurity-top-threats-and-solutions/.

[3] Richie Koch. Pii cybersecurity: Top threats and solutions. URL https://gdpr.eu/eu-gdpr-personal-data/.

[4] Aleksandra Szczepańska and Małgorzata Poddębniak. Phi and pii: How they impact hipaa compliance and your marketing strategy. *Piwik PRO Blog*, May 2025. URL https://piwik.pro/blog/phi-and-pii-hipaa-compliance-and-marketing/.

[5] California consumer privacy act (ccpa), March 2024. URL https://oag.ca.gov/privacy/ccpa.

[6] David Batista and Matthew Antony Upson. nervaluate, 2025. URL https://github.com/davidsbatista/NER-Evaluation.

[7] Byamugisha Africano, Gideon Mpungu, Daudi Jjingo, and Ggaliwango Marvin. Pii detection in low-resource languages using explainable deep learning techniques. In *Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*, pages 94–103, 2024.

[8] Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Wixted, Alejandro Garcia-Rudolph, Catalina Martínez-Costa, and Guenter Neumann. Few-shot cross-lingual transfer for coarse-grained de-identification of code-mixed clinical texts. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 200–211, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.20. URL https://aclanthology.org/2022.bionlp-1.20/.

[9] Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.334/.

[10] AI4Privacy. pii-masking-200k (revision 1d4c0a1) [dataset]. https://huggingface.co/datasets/ai4privacy/pii-masking-200k, 2023. Accessed: 2025-05-15.

[11] BigCode. bigcode-pii-dataset [dataset]. https://huggingface.co/datasets/bigcode/bigcode-pii-dataset, 2023. Accessed: 2025-05-15.

[12] Thatavarthi Giri Sougandh, Nithish Sagar Reddy, Meena Belwal, et al. Automated resume parsing: A natural language processing approach. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 1–6. IEEE, 2023.

[13] Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, 2023.

[14] Vjeko Kužina, Ana-Marija Petric, Marko Barišić, and Alan Jović. Cassed: context-based approach for structured sensitive data detection. *Expert systems with applications*, 223:119924, 2023.

[15] Yizhi Liu, Fang Yu Lin, Mohammadreza Ebrahimi, Weifeng Li, and Hsinchun Chen. Automated pii extraction from social media for raising privacy awareness: A deep transfer learning approach. In *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE, 2021.

[16] Luca Mainetti and Andrea Elia. Detecting personally identifiable information through natural language processing: A step forward. *Applied System Innovation*, 8(2):55, 2025.

[17] Maksim Savkin, Timur Ionov, and Vasily Konovalov. Spy: Enhancing privacy with synthetic pii detection dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 236–246, 2025.

[18] Yuzhao Heng, Chunyuan Deng, Yitong Li, Yue Yu, Yinghao Li, Rongzhi Zhang, and Chao Zhang. ProgGen: Generating named entity recognition datasets step-by-step with self-reflexive large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15992–16030, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.947. URL https://aclanthology.org/2024.findings-acl.947/.

[19] Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S Baker. De-identifying student personally identifying information with gpt-4. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 559–565, 2024.

[20] Yuntian Shen, Zilyu Ji, Jionghao Lin, and KR Koedginer. Enhancing the de-identification of personally identifiable information in educational data. *arXiv preprint arXiv:2501.09765*, 2025.

[21] Yue Zhang, Dionisios G Vlachos, Dongxia Liu, and Hui Fang. Rapid adaptation of chemical named entity recognition using few-shot learning and llm distillation. *Journal of Chemical Information and Modeling*, 65(9):4334–4345, 2025.

[22] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. GPT-NER: Named entity recognition via large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.239. URL https://aclanthology.org/2025.findings-naacl.239/.

[23] Atul Anand. Sensitive data detection in structured datasets using large language models. *Authorea Preprints*, 2024.

[24] Xi Ye, Qiaochu Chen, Xinyu Wang, Isil Dillig, and Greg Durrett. Sketch-driven regular expression generation from natural language and examples. *Transactions of the Association for Computational Linguistics*, 8:679–694, 2020.

[25] Davlan. distilbert-base-multilingual-cased-ner-hrl. URL https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl. Accessed: 2025-07-03.

[26] KB. bert-base-swedish-cased-ner. URL https://huggingface.co/KB/bert-base-swedish-cased-ner. Accessed: 2025-07-03.

[27] NlpHUST. ner-vietnamese-electra-base. URL https://huggingface.co/NlpHUST/ner-vietnamese-electra-base. Accessed: 2025-07-03.

[28] Julian-Schelb. roberta-ner-multilingual. URL https://huggingface.co/julian-schelb/roberta-ner-multilingual. Accessed: 2025-07-03.

[29] Babelscape. wikineural-multilingual-ner. URL https://huggingface.co/Babelscape/wikineural-multilingual-ner. Accessed: 2025-07-03.

A Appendix

NER Model Codes

```
NERd:  Davlan/distilbert-base-multilingual-cased-ner-hrl [25]
NERk:  KB/bert-base-swedish-cased-ner [26]
NERn:  NlpHUST/ner-vietnamese-electra-base [27]
NERj:  julian-schelb/roberta-ner-multilingual [28]
NERb:  Babelscape/wikineural-multilingual-ner [29]
```

Figure 4: Pretrained model codes used for multilingual NER.

Locale	Samples
sv_SE (Swedish – Sweden)	150
vi_VN (Vietnamese – Vietnam)	150+
zh_CN (Chinese – China)	105
zh_SG (Chinese – Singapore)	45
pt_BR (Portuguese – Brazil)	110+
pt_PT (Portuguese – Portugal)	45
pl_PL (Polish – Poland)	150+
hi_IN (Hindi – India)	150+
fi_FI (Finnish – Finland)	150
ar_AE (Arabic – UAE)	150
nl_NL (Dutch – Netherlands)	105
nl_BE (Dutch – Belgium)	45
no_NO (Norwegian – Norway)	150

Table 2: Benchmark Sample Count by Locale

Locale	Phase I	Phase II	Phase III	Δ (1→2)	Δ (2→3)	Δ (1→3)
sv_SE	0.396	0.614	0.703	55.05%	14.50%	77.53%
vi_VN	0.468	0.539	0.571	15.17%	5.94%	22.01%
zh_CN	0.594	0.632	0.680	6.40%	7.60%	14.48%
zh_SG	0.590	0.742	0.758	25.76%	2.16%	28.48%
nl_NL	0.582	0.597	0.625	2.58%	4.69%	7.39%
nl_BE	0.836	0.785	0.834	-6.10%	6.24%	-0.24%
no_NO	0.583	0.664	0.660	13.89%	-0.60%	13.21%
hi_IN	0.486	0.503	0.696	3.50%	38.37%	43.21%
fi_FI	0.573	0.592	0.730	3.32%	23.31%	27.40%
ar_AFB	0.463	0.479	0.567	3.46%	18.37%	22.46%
pt_BR	0.446	0.547	0.659	22.65%	20.48%	47.76%
pt_PT	0.511	0.593	0.615	16.05%	3.71%	20.35%
pl_PL	0.428	0.583	0.602	36.22%	3.26%	40.65%

Table 3: Locale Performance Across Three Phases by F1 Score

Locale	**Approach**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**TP**	**FP**	**TN**	**FN**
hi_IN	NERd	0.328	0.194	0.227	0.209	59	245	159	201
hi_IN	Zero-shot LLM	0.472	0.801	0.534	0.641	310	77	0	22
hi_IN	RECAP	0.534	0.781	0.628	**0.696**	364	102	0	216
fi_FI	NERd	0.347	0.431	0.268	0.330	166	219	192	454
fi_FI	Zero-shot LLM	0.485	0.830	0.538	0.653	534	109	0	458
fi_FI	RECAP	0.574	0.744	0.716	**0.730**	710	244	0	282
ar_AE	NERd	0.420	0.474	0.229	0.309	267	296	622	897
ar_AE	Zero-shot LLM	0.362	0.667	0.442	0.531	886	443	0	1120
ar_AE	RECAP	0.396	0.610	0.537	**0.567**	1078	736	0	928
sv_SE	NERk	0.705	0.599	0.810	0.689	205	137	237	48
sv_SE	Zero-shot LLM	0.344	0.720	0.397	0.512	641	249	0	974
sv_SE	RECAP	0.542	0.714	0.692	**0.703**	1118	447	0	497
vi_VN	NERn	0.737	0.602	0.818	0.694	198	131	292	44
vi_VN	Zero-shot LLM	0.317	0.751	0.354	0.481	796	264	0	1454
vi_VN	RECAP	0.400	0.772	0.453	0.571	1020	302	0	1230
zh_CN	NERj	0.385	0.446	0.228	0.302	120	149	228	406
zh_CN	Zero-shot LLM	0.486	0.811	0.549	0.654	625	146	0	514
zh_CN	RECAP	0.515	0.652	0.709	**0.680**	808	431	0	331
zh_SG	NERj	0.523	0.449	0.172	0.249	31	38	174	149
zh_SG	Zero-shot LLM	0.523	0.799	0.603	0.687	279	70	0	184
zh_SG	RECAP	0.610	0.753	0.762	**0.758**	353	116	0	110
nl_NL	NERj	0.280	0.238	0.294	0.263	160	511	188	385
nl_NL	Zero-shot LLM	0.341	0.878	0.358	0.508	473	66	0	849
nl_NL	RECAP	0.456	0.790	0.517	**0.625**	684	182	3	638
nl_BE	NERj	0.364	0.243	0.342	0.284	50	156	94	96
nl_BE	Zero-shot LLM	0.446	0.840	0.487	0.617	194	37	0	204
nl_BE	RECAP	0.715	0.887	0.786	**0.834**	313	40	0	85
no_NO	NERj	0.436	0.282	0.348	0.312	162	412	390	303
no_NO	Zero-shot LLM	0.433	0.873	0.462	0.604	517	75	0	604
no_NO	RECAP	0.493	0.785	0.570	**0.660**	638	175	0	481
pt_BR	NERj	0.309	0.293	0.210	0.244	136	328	240	513
pt_BR	Zero-shot LLM	0.341	0.886	0.356	0.508	186	24	0	336
pt_BR	RECAP	0.492	0.792	0.560	**0.659**	290	76	0	224
pt_PT	NERj	0.309	0.293	0.210	0.244	136	328	240	513
pt_PT	Zero-shot LLM	0.239	0.878	0.247	0.385	158	22	0	482
pt_PT	RECAP	0.444	0.689	0.555	**0.615**	354	160	0	284
pl_PL	NERb	0.276	0.394	0.195	0.261	133	208	154	550
pl_PL	Zero-shot LLM	0.322	0.744	0.362	0.487	128	44	0	226
pl_PL	RECAP	0.425	0.614	0.579	**0.602**	398	244	0	242
All locales	NER*	0.428	0.395	0.362	0.360	-	-	-	-
All locales	Zero-shot LLM*	0.391	0.795	0.437	0.558	-	-	-	-
All locales	RECAP*	0.492	0.729	0.605	**0.657**	-	-	-	-

Table 4: PII Detection Performance Results by Locale and Approach

Note: The last three rows (marked by *) represent weighted averages across all locales.

Locale	Phase	Accuracy	Precision	Recall	F1 Score	TP	FP	TN	FN
sv_SE	I	0.247	0.438	0.362	0.396	584	749	0	1031
sv_SE	II	0.443	0.582	0.650	0.614	1049	753	0	566
sv_SE	III	0.542	0.714	0.692	0.703	1118	447	0	497
vi_VN	I	0.326	0.551	0.407	0.468	915	747	91	1335
vi_VN	II	0.369	0.630	0.471	0.539	1059	622	0	1191
vi_VN	III	0.400	0.772	0.453	0.571	1020	302	0	1230
zh_CN	I	0.481	0.545	0.654	0.594	745	623	199	394
zh_CN	II	0.462	0.582	0.692	0.632	788	566	0	351
zh_CN	III	0.515	0.652	0.709	0.680	808	431	0	331
zh_SG	I	0.487	0.573	0.607	0.590	281	209	90	182
zh_SG	II	0.590	0.723	0.762	0.742	353	135	0	110
zh_SG	III	0.610	0.753	0.762	0.758	353	116	0	110
nl_NL	I	0.434	0.652	0.526	0.582	695	371	69	627
nl_NL	II	0.448	0.665	0.542	0.597	716	361	69	606
nl_NL	III	0.456	0.790	0.517	0.625	684	182	3	638
nl_BE	I	0.718	0.909	0.774	0.836	308	31	0	90
nl_BE	II	0.645	0.871	0.714	0.785	284	42	0	114
nl_BE	III	0.715	0.887	0.786	0.834	313	40	0	85
no_NO	I	0.442	0.677	0.512	0.583	573	273	75	546
no_NO	II	0.497	0.780	0.578	0.664	647	182	0	472
no_NO	III	0.493	0.785	0.570	0.660	638	175	0	481
hi_IN	I	0.347	0.406	0.607	0.486	352	516	44	228
hi_IN	II	0.336	0.426	0.614	0.503	356	479	0	224
hi_IN	III	0.534	0.781	0.628	0.696	364	102	0	216
fi_FI	I	0.416	0.493	0.685	0.573	680	700	42	312
fi_FI	II	0.421	0.515	0.698	0.592	692	652	0	300
fi_FI	III	0.574	0.744	0.716	0.730	710	244	0	282
ar_AE	I	0.316	0.421	0.514	0.463	1032	1421	73	974
ar_AE	II	0.315	0.442	0.522	0.479	1048	1323	0	958
ar_AE	III	0.396	0.610	0.537	0.567	1078	736	0	928
pt_BR	I	0.320	0.453	0.439	0.446	214	258	36	274
pt_BR	II	0.377	0.519	0.579	0.547	294	214	0	214
pt_BR	III	0.492	0.792	0.560	0.659	290	76	0	224
pt_PT	I	0.366	0.536	0.489	0.511	312	270	32	326
pt_PT	II	0.421	0.620	0.568	0.593	352	216	0	268
pt_PT	III	0.444	0.689	0.555	0.615	354	160	0	284
pl_PL	I	0.297	0.416	0.440	0.428	308	432	40	394
pl_PL	II	0.412	0.567	0.601	0.583	364	278	0	242
pl_PL	III	0.425	0.614	0.579	0.602	398	244	0	242
All locales	I*	0.372	0.516	0.522	0.511	-	-	-	-
All locales	II*	0.419	0.584	0.601	0.585	-	-	-	-
All locales	III*	0.492	0.729	0.605	0.657	-	-	-	-

Table 5: PII Detection Performance Results by Locale and Phase
Note: The last three rows (marked by *) represent weighted averages across all locales.

Listing 1: High-level overview of Name-extraction prompt used in experiments

```
{
System Prompt: You are a multilingual PII detection system. Your task is
     to detect and extract personal names from the input text based on
    the specified {locale}. Follow these rules: - Always consider locale-
    specific naming conventions. - Return names exactly as they appear
    in the text (including diacritics, prefixes, and original scripts). -
     If no names are found, return an empty list []. Locale-specific
    examples: - zh_CN / zh_SG: Extract Chinese names in original
    characters. - vi_VN: Vietnamese names follow the order [Family Name]
     [Middle Name] [Given Name]. - nl_NL / nl_BE: Dutch/Flemish names
    may include prefixes (e.g., "van", "de"). User Prompt: LOCALE: {
    locale} TEXT: "{text}"
}
```