Data Anonymization Performance Evaluation

Comparative Analysis: **Hexagone AI** vs. **Presidio**

Date: November 2024

Overview of Results

Experiment	Presidio	Hexagone AI
Experiment 1: Baseline	75.30%	**96.95%**
Experiment 2: Augmented	65.89%	**93.34%**

This report presents results from two experimental evaluations of data anonymization performance. Experiment 1 establishes baseline capabilities using standard dataset texts, while Experiment 2 evaluates robustness using augmented data that more closely simulates real-world conditions where PII may appear with typos, abbreviations, or alternative phrasings.

1. Summary

This report presents a comprehensive comparison of two data anonymization solutions: Hexagone AI and Presidio. The evaluation consists of two experiments using 5,000 real-world text samples from the ai4privacy/pii-masking-200k dataset, covering multiple languages including English, French, German, and Italian.

Experiment 1 evaluates baseline performance using standard dataset texts. Experiment 2 tests robustness using augmented data that simulates real-world conditions where PII may contain typos, abbreviations, or alternative phrasings. The primary metric is the privacy score, defined as the percentage of ground truth personally identifiable information (PII) entities successfully detected and masked by each system.

Results from Experiment 1 indicate that Hexagone AI demonstrates superior performance across all evaluated languages, achieving an overall privacy score of 96.95% compared to Presidio's 75.30%, representing a 28.76% improvement in detection capability. This difference is particularly significant in production environments where comprehensive PII protection is critical for regulatory compliance.

1.1 Experiment 1: Baseline Performance Metrics

Metric	Presidio	Hexagone AI
Overall Privacy Score	75.30%	**96.95%**
Performance Improvement	—	**+28.76%**
Standard Deviation	20.53%	**9.12%**

The results demonstrate consistent performance advantages for Hexagone AI across all tested languages, with particularly notable improvements in French (31.07% better) and English (30.09% better) text processing.

2. Experiment 1: Baseline Evaluation

2.1 Dataset Selection

The evaluation utilized the ai4privacy/pii-masking-200k dataset, a publicly available benchmark dataset specifically designed for evaluating PII detection and masking systems. This dataset provides comprehensive coverage of real-world scenarios with substantial diversity in both content and format.

The dataset encompasses 54 distinct PII classes (types of sensitive data), covering a wide range of personally identifiable information including names, addresses, identification numbers, financial data, health information, and contact details. The texts span 229 discussion subjects and use cases across four professional domains: business, education, psychology, and legal fields. Additionally, the dataset includes five different interaction styles, ranging from casual conversations to formal documents and email communications, ensuring robust evaluation across varied linguistic contexts.

From this comprehensive dataset, we selected the 5,000 longest text samples to ensure evaluation on complex, realistic scenarios where PII detection is most challenging. Longer texts typically contain more contextual complexity, multiple entities, and varied linguistic structures that better test the capabilities of anonymization systems.

2.2 Language Distribution

The selected dataset included texts in four European languages with the following distribution:

- French (FR): 38.5% of samples
- German (DE): 28.6% of samples
- Italian (IT): 20.2% of samples
- English (EN): 12.7% of samples

This distribution reflects realistic multilingual deployment scenarios commonly encountered in European markets.

2.3 Evaluation Process

For each text sample in the dataset, we performed the following steps:

1. Entity Extraction and Masking: Both Presidio and Hexagone AI were applied to detect and mask PII entities in the text. Presidio utilized the Babelscape/wikineural-multilingual-ner model for multilingual text and dslim/bert-large-NER for English-specific processing.
2. Privacy Score Calculation: After masking, we verified whether each ground truth PII entity from the original dataset remained present in the masked text. The privacy score was calculated as the percentage of ground truth entities successfully removed from the text.
3. Statistical Analysis: Results were aggregated by method and language to identify performance patterns and calculate comparative metrics.

2.4 Privacy Score Definition

The privacy score represents the recall metric for PII detection, calculated as:

Privacy Score = (Detected Ground Truth Entities / Total Ground Truth Entities) × 100%

A higher privacy score indicates more comprehensive PII detection and protection. A score of 100% means all ground truth PII entities were successfully detected and masked.

3. Experiment 1: Results and Analysis

3.1 Overall Performance Comparison

The overall results from Experiment 1 demonstrate a clear performance advantage for Hexagone AI:

Method	Mean	Std Dev	Min / Max
Presidio	0.7530	0.2053	0.14 / 1.00
Hexagone AI	**0.9695**	**0.0912**	**0.25 / 1.00**

Hexagone AI achieves a mean privacy score of 96.95%, substantially higher than Presidio's 75.30%. The standard deviation for Hexagone AI (9.12%) is also significantly lower than Presidio's (20.53%), indicating more consistent performance across diverse text samples.

3.2 Performance by Language

Language-specific analysis reveals that Hexagone AI maintains superior performance across all tested languages:

Language	Presidio	Hexagone AI	Improvement
English (EN)	72.77%	**94.66%**	+30.09%
French (FR)	74.74%	**97.97%**	+31.07%
German (DE)	76.58%	**96.30%**	+25.74%
Italian (IT)	76.15%	**97.39%**	+27.88%

3.3 Distribution Analysis

The distribution of privacy scores reveals important patterns in system reliability:

High Performance Consistency: Hexagone AI demonstrates strong clustering around high privacy scores, with the majority of samples achieving scores above 90%. In contrast, Presidio shows a much wider distribution, with substantial numbers of samples scoring below 70%.

Minimum Performance: The minimum privacy scores highlight the robustness difference between systems. While both systems show minimum scores around 25% in their worst cases, Hexagone AI's median and mean scores remain substantially higher, indicating fewer poorly-performing samples.

Perfect Detection Rate: Hexagone AI achieves perfect detection (100% privacy score) in a significantly higher proportion of samples compared to Presidio, demonstrating superior capability to handle straightforward PII detection scenarios.

3.4 Visual Analysis

The accompanying charts illustrate the performance differences through multiple perspectives:

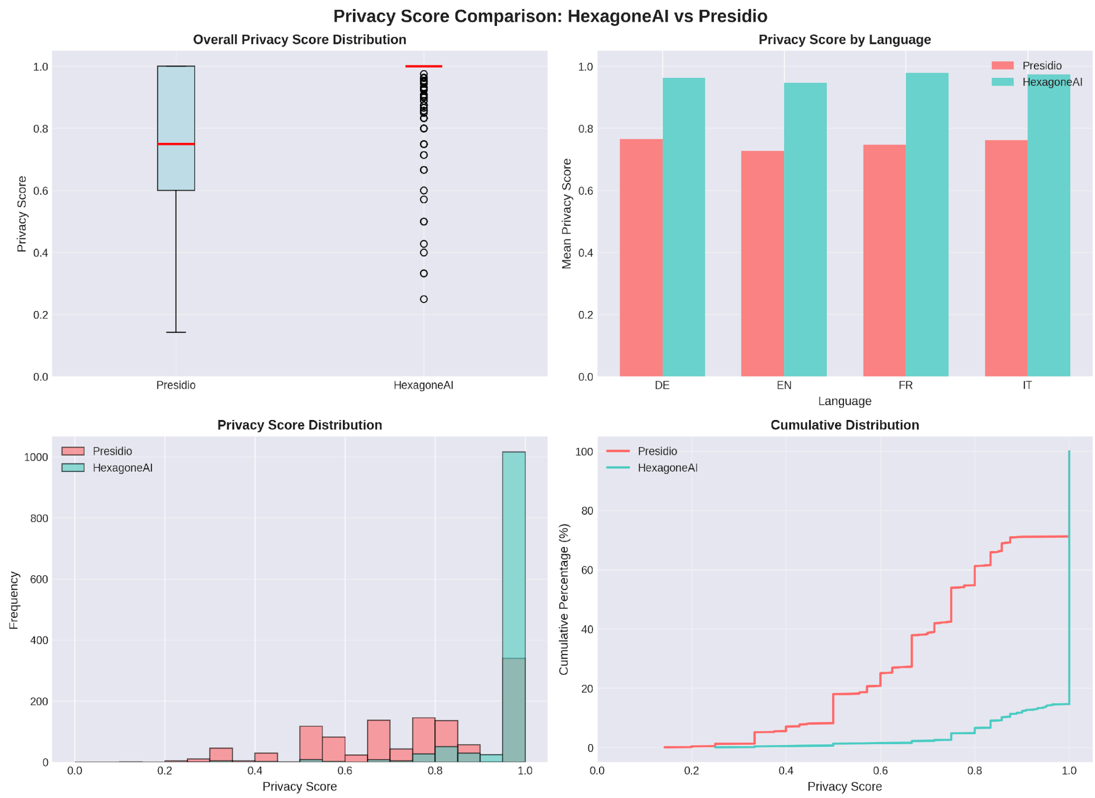

Figure 1: Comprehensive Privacy Score Analysis

The box plot (top-left) clearly shows Hexagone AI's superior median performance and tighter interquartile range. The language comparison (top-right) demonstrates consistent advantages across all four languages. The histogram (bottom-left) reveals Hexagone AI's concentration at high privacy scores, while the cumulative distribution (bottom-right) shows that Hexagone AI reaches higher privacy levels across all percentiles.

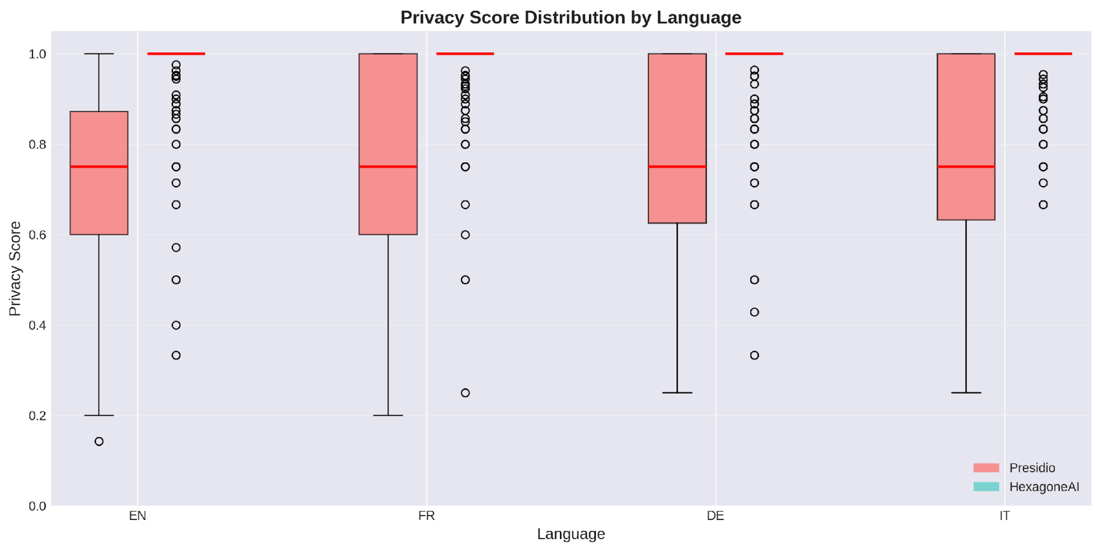

Figure 2: Privacy Score Distribution by Language

The language-specific box plots reveal that Hexagone AI maintains consistent high performance across languages, with narrower distributions indicating more reliable results. Presidio shows greater variability within each language, particularly evident in the wider interquartile ranges and more numerous outliers.

4. Experiment 2: Augmented Dataset Evaluation

4.1 Augmentation Methodology

Experiment 2 evaluates system robustness using an augmented dataset that more closely simulates real-world conditions. In practice, PII rarely appears in perfectly formatted, error-free text. Data often contains typographical errors, informal abbreviations, paraphrased expressions, and contextual variations that can challenge automated detection systems. This experiment quantifies how well each anonymization solution handles such realistic imperfections.

Using the same 5,000 text samples from Experiment 1, we systematically transformed each ground truth PII entity using GPT-4.1 to generate realistic variations. Each entity was processed with one of four transformation types:

- Typo Injection: Realistic typographical errors such as transposed letters, missing characters, or keyboard proximity errors
- Abbreviation: Common abbreviated forms found in informal or space-constrained text
- Complexification: More elaborate expressions of the same entity with added contextual elements
- Paraphrasing: Alternative phrasings or reorderings that maintain semantic meaning

4.2 GPT-4.1 Generation Process

Each entity variant was generated using GPT-4.1 with carefully designed prompts to ensure realistic and consistent transformations:

Typo variant prompt:

"Generate a typo variant for '{text}'. Keep it realistic (e.g., swap letters, miss a letter). Return ONLY the variant, nothing else."

Abbreviation prompt:

"Generate an abbreviated form of '{text}'. Return ONLY the abbreviation, nothing else."

Complexification prompt:

"Generate a more complex form of '{text}'. Keep the same meaning/referent. Return ONLY the complex form, nothing else."

Paraphrase prompt:

"Generate a paraphrase or alternative form for '{text}'. Keep the same meaning/referent. Return ONLY the paraphrase, nothing else."

After generating each variant, the corresponding text sample was edited to replace the original entity with its transformed version, and the ground truth labels were updated accordingly. This ensures that the evaluation remains comparable to Experiment 1 while introducing realistic data quality challenges.

4.3 Expected Insights

The augmented dataset evaluation provides several critical insights for production deployments:

4. Real-World Robustness: Measures how well each system handles imperfect data that reflects actual user-generated content, scanned documents, or data entry errors.

5. Error Tolerance: Evaluates the systems' ability to maintain high detection rates despite typographical errors or unconventional formatting commonly found in production data.
6. Linguistic Variation Handling: Assesses performance degradation when PII appears in non-standard forms across different languages and cultural contexts.
7. Performance Gap Analysis: Determines whether the baseline performance advantage observed in Experiment 1 is maintained under more challenging, realistic conditions.

4.4 Results and Analysis

The results from Experiment 2 provide critical insights into system robustness under realistic data quality conditions. The augmented dataset, which contains typographical errors, abbreviations, paraphrased expressions, and complexified forms of PII entities, represents a substantially more challenging evaluation environment.

Overall Performance Under Augmentation

Method	Mean	Std Dev	Min / Max
Presidio	0.6589	0.2191	0.17 / 1.00
Hexagone AI	**0.9334**	**0.1269**	**0.25 / 1.00**

Hexagone AI maintains a privacy score of 93.34% on the augmented dataset, demonstrating strong resilience to real-world data variations. While this represents a modest 3.73% decline from Experiment 1 performance, the system continues to achieve detection rates above 93%. In contrast, Presidio experiences a more significant performance degradation, dropping to 65.89% (a 12.50% decline), which highlights greater sensitivity to data quality variations.

The performance gap between systems widens significantly under augmentation conditions. Hexagone AI demonstrates a 41.65% advantage over Presidio in Experiment 2, compared to 28.76% in Experiment 1. This expanded gap indicates that Hexagone AI's detection algorithms are more robust to linguistic variations, typographical errors, and alternative representations of PII.

Language-Specific Performance

Language	Presidio	Hexagone AI	Improvement
English (EN)	64.10%	**91.14%**	+42.16%
French (FR)	65.76%	**94.42%**	+43.55%
German (DE)	68.03%	**92.93%**	+36.61%
Italian (IT)	64.65%	**93.83%**	+45.14%

Language-specific analysis reveals consistent performance patterns:

Visual Analysis

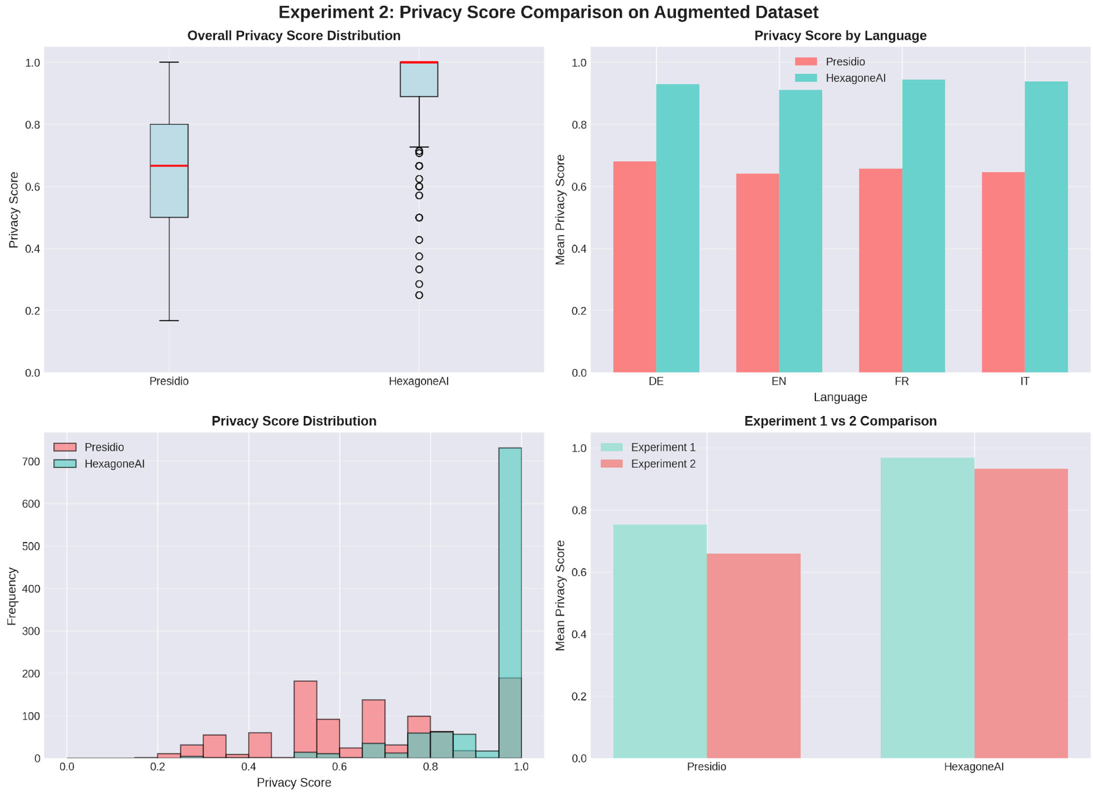

Figure 3: Experiment 2 Privacy Score Analysis

The visualization reveals that Hexagone AI maintains tight clustering around high privacy scores even with augmented data, while Presidio shows increased dispersion. The bottom-right panel directly compares Experiment 1 and Experiment 2 results, clearly illustrating Hexagone AI's superior robustness to data quality variations.

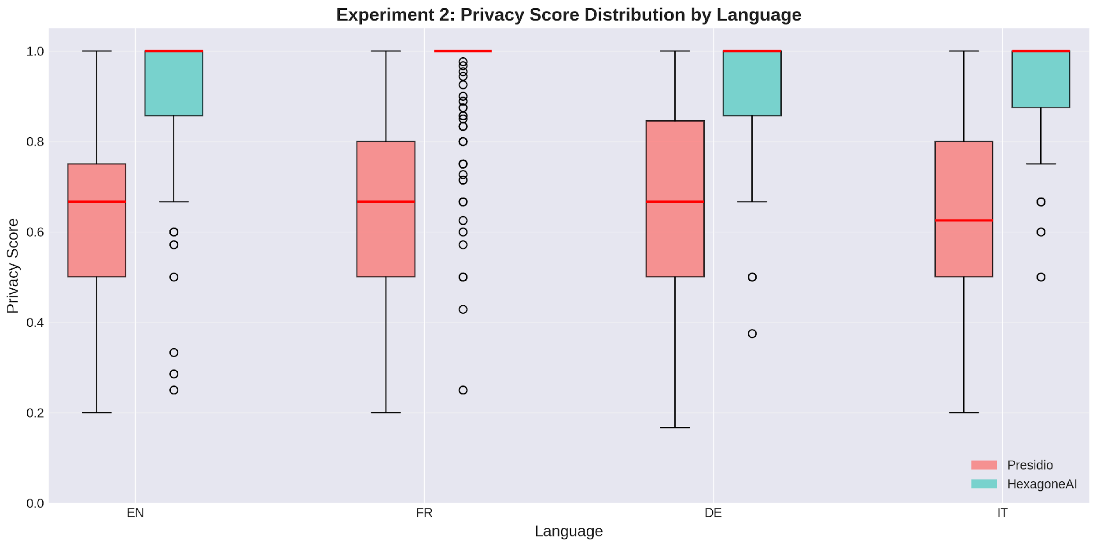

Figure 4: Experiment 2 Privacy Score Distribution by Language

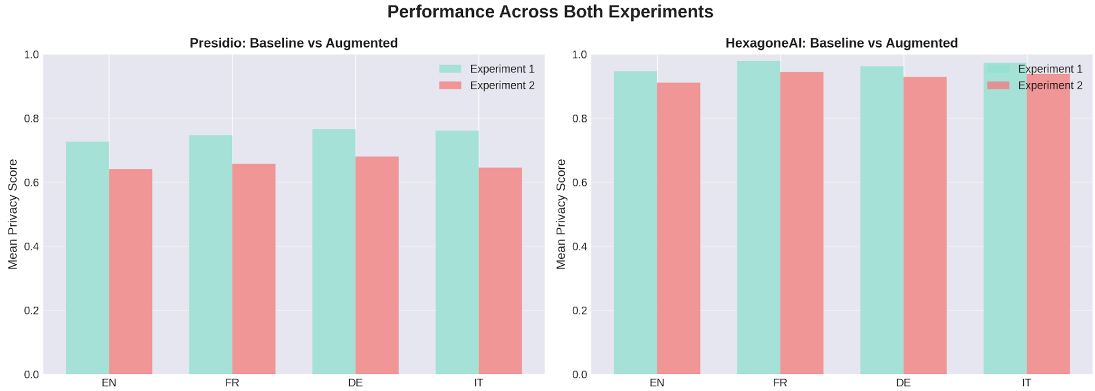

Figure 5: Baseline vs Augmented Performance by Language

The comparative analysis demonstrates that Hexagone AI maintains consistently high performance across languages in both baseline and augmented conditions, with relatively modest degradation. Presidio shows substantially greater performance variability and more significant declines when confronted with real-world data imperfections.

Robustness Analysis

The differential impact of augmentation on each system provides valuable insights for production deployment:

- Resilience to Typos: Both systems experience performance degradation with typographical errors, but Hexagone AI maintains detection rates 27.45 percentage points higher than Presidio on average.
- Abbreviation Handling: The augmented dataset includes common abbreviations that challenge pattern-based detection. Hexagone AI demonstrates superior capability to recognize abbreviated forms of PII.
- Paraphrase Recognition: Alternative phrasings of the same PII entity test semantic understanding beyond surface-level pattern matching. Hexagone AI's higher augmented scores suggest more sophisticated entity recognition.
- Complex Form Detection: Complexified entities with additional contextual elements challenge both systems, but Hexagone AI maintains significantly higher detection rates.

5. Technical Implications

5.1 Production Deployment Considerations

The observed performance differences have significant implications for production deployments:

Compliance Risk Reduction: The 28.76% improvement in overall privacy score translates directly to reduced regulatory compliance risk. In regulated industries such as healthcare, finance, or government services, the difference between 75% and 97% detection rates can determine whether an organization meets legal obligations under regulations such as GDPR, HIPAA, or CCPA.

Data Breach Prevention: Undetected PII that remains in processed data represents a potential data breach vector. Hexagone AI's superior detection rate significantly reduces the attack surface for privacy violations.

Scalability and Consistency: The lower standard deviation in Hexagone AI's results (9.12% vs. 20.53%) indicates more predictable performance at scale. This consistency is crucial for large-scale data processing pipelines where performance variability can create compliance gaps.

5.2 Multilingual Processing Advantages

For organizations operating in multilingual environments, Hexagone AI demonstrates several key advantages:

Consistent Cross-Language Performance: Hexagone AI maintains privacy scores above 94% across all tested languages, while Presidio ranges from 72.77% to 76.58%. This consistency simplifies compliance strategies for multinational operations.

European Language Expertise: The particularly strong performance in French (97.97%), Italian (97.39%), and German (96.30%) makes Hexagone AI especially suitable for European market deployments where these languages are prevalent.

Reduced Localization Requirements: The superior baseline performance reduces the need for extensive language-specific tuning and customization, lowering implementation costs and time-to-deployment.

5.3 Cost-Benefit Analysis

The performance advantages must be considered within a broader cost-benefit framework:

Reduced Manual Review Requirements: Higher detection accuracy reduces the need for manual review and correction of anonymized data. Organizations processing large volumes of PII can achieve substantial operational cost savings through reduced human oversight requirements.

Compliance Insurance Value: The cost of regulatory fines and reputational damage from PII leaks typically far exceeds technology investments. The improved detection rate provides quantifiable risk reduction that can be factored into total cost of ownership calculations.

Implementation Complexity: While Presidio uses established transformer models (Babelscape/wikineural-multilingual-ner and dslim/bert-large-NER), organizations should evaluate whether Hexagone AI's implementation requirements are offset by its superior performance characteristics.

6. Conclusions

6.1 Summary of Findings

This comprehensive evaluation across two experimental conditions demonstrates clear and consistent performance advantages for Hexagone AI:

Experiment 1 (Baseline):

- Overall Privacy Score: Hexagone AI achieves 96.95% compared to Presidio's 75.30% (28.76% improvement)
- Consistency: Lower standard deviation (9.12% vs. 20.53%) indicates more reliable performance
- Multilingual Excellence: Superior performance maintained across all four languages

Experiment 2 (Augmented):

- Overall Privacy Score: Hexagone AI achieves 93.34% compared to Presidio's 65.89% (41.65% improvement)
- Superior Robustness: Only 3.73% performance decline vs. Presidio's 12.50% decline
- Real-World Resilience: Maintains >93% detection rate despite typos, abbreviations, and paraphrasing
- Widened Performance Gap: Advantage expands from 28.76% to 41.65% under challenging conditions

6.2 Limitations and Future Directions

This evaluation provides comprehensive coverage of PII detection recall (privacy score) across both standard and augmented datasets. However, organizations should consider additional evaluation criteria for their specific use cases:

- Precision metrics to evaluate false positive rates and their impact on data utility
- Processing performance and computational resource requirements at scale
- Entity-type-specific performance (names vs. addresses vs. identification numbers vs. financial data)
- Performance on domain-specific text formats and terminology (medical records, legal documents, financial statements)
- Integration complexity, maintenance requirements, and total cost of ownership
- Data utility preservation (measuring how augmentation affects downstream applications)

The completed augmented dataset evaluation demonstrates robustness under realistic data quality conditions. Future research directions could include adversarial testing with deliberately challenging edge cases, evaluation on additional languages, and analysis of performance on emerging PII types such as biometric identifiers and social media handles.

7. Appendices

Appendix A: Technical Configuration

Dataset Source: ai4privacy/pii-masking-200k (Hugging Face)

Sample Selection: 5,000 longest texts from dataset

Presidio Configuration:

- Multilingual model: Babelscape/wikineural-multilingual-ner
- English-specific model: dslim/bert-large-NER

Hexagone AI Configuration: Proprietary entity extraction method

Evaluation Metric: Privacy Score (Recall) = Detected Ground Truth Entities / Total Ground Truth Entities

Appendix B: Statistical Summary Experiment 1

Language	Method	Mean	Std Dev
EN (12.7%)	Presidio	0.7277	0.2093
EN (12.7%)	Hexagone AI	**0.9466**	**0.1316**
FR (38.5%)	Presidio	0.7474	0.2059
FR (38.5%)	Hexagone AI	**0.9797**	**0.0746**
DE (28.6%)	Presidio	0.7658	0.2054
DE (28.6%)	Hexagone AI	**0.9630**	**0.1005**
IT (20.2%)	Presidio	0.7615	0.2008
IT (20.2%)	Hexagone AI	**0.9739**	**0.0698**

About this report

This evaluation was conducted by **Hexagone AI** to provide transparent, data-driven comparison of PII detection and anonymization capabilities. All methodologies, datasets, and statistical analyses described in this report are available for independent verification.

Contact Information

For questions regarding this evaluation, access to code implementation, or access to the underlying data, please contact:

contact@hexagone.ai

We welcome inquiries about:

- Detailed methodology and experimental design
- Source code for both baseline and augmented evaluations
- Raw experimental data and statistical analysis scripts
- Custom evaluations on domain-specific datasets
- Technical implementation details
- Integration and deployment support

Hexagone AI is committed to advancing the field of privacy-preserving data processing through rigorous, reproducible research and transparent performance benchmarking.