Privacy-Preserving Document Summarization via Sensitive-Word Detection and Masked Content Reconstruction

Anonymous ACL submission

Abstract

Identifying and protecting sensitive information in documents is crucial when using cloud-based summarization services. In this study, we introduce a privacy-preserving document summarization framework that first detects potentially sensitive words and then masks them before submitting the document to the AI server. A custom classification model is trained to recognize sensitive terms using a rich feature set. In the proposed pipeline, the sensitive words detected in the original (plain) text are replaced with mask tokens before sending the text to a cloud-based summarization model. The masked text summary is then reconstructed in the user environment by aligning each masked token with its original word using context windows and semantic similarity. Experiments on a labeled dataset of sensitive documents show that using our method, the user can correctly recover 93.87% of masked content in the AI-generated summaries, demonstrating the effectiveness of the proposed masking and reconstruction strategy. Moreover, when masking as much as 50% of sensitive words, the ROUGE-1 quality of multiple summarization models shows an average decrease of only 20% and a drop of about 0.04 in BERTScore compared to the original text summaries.

1 Introduction

In the modern digital landscape, textual data represents a central asset for organizations, yet much of it contains private information (e.g., personal identifiers, medical records) or confidential information (e.g., legal contracts, financial reports). Distinguishing between the two is critical: private data exposes individuals to direct harm, such as identity theft or privacy violations, whereas confidential data exposes organizations to strategic risks, such as loss of competitive advantage or breach of business secrecy. When cloud-based large language models (LLMs) are used for document summarization, both categories are directly relevant: any document submitted to an LLM may contain a combination of private and confidential information, and the leakage of either undermines organizational trust and security.

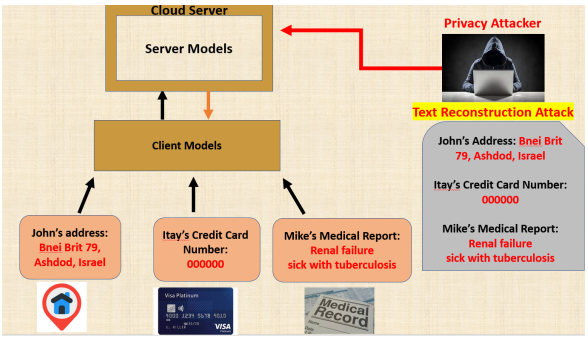


Figure 1: Illustration of Inference Services and Privacy Risks. Despite clients uploading word representations instead of plain text to cloud servers, privacy attackers can still decipher representations into original texts, leaving clients vulnerable to privacy breaches.

Rather than modifying the summarization process itself, we propose a model-agnostic approach that obfuscates sensitive information in the source text prior to its submission to a cloud-based summarization engine. This strategy enables the integration of privacy-preserving mechanisms without altering the architecture or behavior of the underlying LLMs, a principle aligned with recent studies that emphasize input-level redaction as a generalizable, deployment-agnostic solution (Jagielski et al., 2022; Xu et al., 2021; Sheth et al., 2023). It ensures that sensitive content is obfuscated before transmission, regardless of whether the summarizer is extractive, abstractive, or hybrid.

While existing privacy-preserving techniques—such as personally identifiable information (PII) removal, differential privacy, and homomorphic encryption—provide partial mitigation, they often compromise usability. Excessive redaction

may degrade text coherence and summarization quality (Pilán and Others, 2022; Reddy and Knight, 2016), whereas insufficient masking leaves sensitive content exposed (Zhou and Others, 2023; Vats et al., 2023). Differential privacy methods introduce quantifiable noise, but high perturbation levels reduce readability (Igamberdiev and Others, 2022), and homomorphic encryption, although theoretically secure, imposes heavy computational overhead unsuitable for real-time summarization (Gilad-Bachrach and Others, 2016; Chen and Others, 2022).

To bridge this gap, we propose a lightweight and interpretable framework for privacy-preserving text summarization. At its core lies a supervised classifier that estimates the sensitivity of each token based on lexical, contextual, syntactic, and semantic cues. Tokens identified as sensitive are masked prior to transmission to an untrusted third-party summarizer. A local reconstruction module then restores the masked segments in the cloud-generated summary, ensuring that the final output remains both fluent and informative for the end-user. [1]

This work makes the following contributions:

- We introduce a threshold-based masking strategy that explicitly controls the privacy–utility trade-off prior to summarization.
- We present an end-to-end privacy-aware summarization framework that limits exposure of sensitive content to external models while preserving summary quality, with optional local reconstruction.
- We propose an interpretable, context-aware model for detecting sensitive tokens in unstructured text using a rich feature set and continuous sensitivity scores.

2 Related Work

2.1 Privacy-Preserving Techniques in Text Processing

Privacy breaches in text data pose severe risks, particularly in domains such as healthcare, finance, and legal documentation. Regulatory frameworks, such as the General Data Protection Regulation (GDPR), have established legal protections; however, fully preventing privacy violations remains a challenge. As a result, privacy-preserving text processing techniques have gained increasing research attention.

Differential Privacy (DP) Rewriting introduces controlled noise to text, reducing the likelihood of privacy leakage while quantifying privacy risks using an epsilon parameter (Igamberdiev and Others, 2022). For example, in a medical record the phrase "John Smith, 45, diagnosed with diabetes" may be rewritten as "a middle-aged patient diagnosed with diabetes", thereby obscuring the identity while preserving clinical meaning. However, this technique comes with trade-offs—higher privacy levels introduce greater text distortion, reducing readability and usability. Homomorphic encryption offers an alternative, enabling computations on encrypted text (Chen and Others, 2022; Hao and Others, 2022; Gilad-Bachrach and Others, 2016). Despite strong theoretical guarantees, high computational overhead makes homomorphic encryption impractical for real-time text processing.

Machine learning-driven approaches offer a balance between privacy and usability. Semantic perturbation techniques obscure sensitive words by replacing them with functionally similar alternatives while retaining sentence coherence (Zhou and Others, 2023). However, these methods face challenges in high-stakes applications, where even slight distortions in meaning could lead to misinterpretations in legal, medical, or financial contexts. Additionally, many existing privacy-preserving techniques lack formal evaluation against adversarial attacks, raising concerns about their robustness. Beyond semantic perturbation, other obfuscation strategies illustrate alternative ways to navigate this balance. Lexical substitution replaces words to mask details, but word patterns may still reveal author traits (Reddy and Knight, 2016). Secure Binary Embedding (SBE) hashing encrypts summaries so only the sender can retrieve them, but reduces summary utility (Marujo et al., 2015). Syntactic-aware text obfuscation preserves syntax while altering words, though rigid constraints may lead to unnatural phrasing (Hu et al., 2020).

Recent studies reveal that even after text sanitization, attackers may still infer sensitive details from the surrounding context. Private information can be reconstructed via contextual reasoning or inference attacks on ostensibly anonymized text (Mireshghallah et al., 2023; Staab et al., 2023). These findings

[1] All code, models, experimental pipelines, and annotated datasets are publicly available via link.

underscore the need for privacy-preserving systems that are robust not only to basic redaction threats but also to intelligent adversarial inference. To mitigate such risks, researchers have explored more adaptive obfuscation strategies. Self-disclosure abstraction automatically identifies and generalizes personal disclosures (e.g., converting "I'm 16F" to "I'm a teenage girl") (Dou et al., 2024). Locally Private Retrieval-Augmented Generation (LPRAG) applies differential privacy selectively to only the sensitive portions of the input, achieving a more favorable privacy-utility trade-off (He et al., 2025).

2.2 Detection of Sensitive Content

Recent studies have expanded the scope of privacy-preserving text processing to include the detection and masking of sensitive words and expressions based on semantic context rather than predefined entity types alone. ExSense, a hybrid model combining rule-based matching with a BERT-BiLSTM-attention classifier, detects both structured PII (Personally Identifiable Information) and contextually sensitive content such as verbs, descriptors, and topic-specific nouns (Guo et al., 2021). Their model achieved an F1 score exceeding 99% on large-scale, unstructured corpora. Similarly, CP-SID, a dual-stage system incorporating ELECTRA-BiLSTM-CRF, captures long-range semantic dependencies in Chinese texts, outperforming conventional NER-based methods in detecting subtle, context-dependent sensitive terms (Ren et al., 2024).

Transformer-based architectures continue to demonstrate superior performance in this domain. Privacy-BERT-LSTM leverages contextual embeddings and attention mechanisms to flag linguistically and semantically sensitive spans—beyond identifiable entities—achieving F1 scores above 85% across multiple domains (Muralitharan and Arumugam, 2024).

Other research has explored graph-based and semantically enriched representations. PRIVAFrame, a knowledge-graph–driven framework, detects sensitive concepts and relations via ontological reasoning rather than token-level classification (Gambarelli and Gangemi, 2022). Similarly, graph convolutional neural networks (GCNs) have been applied to model relationships between words or entities to better identify latent sensitivity, especially in settings where linguistic ambiguity or topic drift can obscure risk signals (Liu et al., 2021; Huo and Jiang, 2023).

2.3 Summarization of Sensitive Information

Privacy-preserving summarization seeks to generate concise yet secure representations of sensitive documents while maintaining informational fidelity. AspirinSum, an aspect-based de-identification framework, aligns extracted summaries with expert-defined content aspects, ensuring that only non-sensitive yet relevant segments of text remain (Li, 2024). However, its dependence on predefined aspects limits adaptability. Privacy risks in summarization models have also been demonstrated, where membership inference attacks can extract training data and expose sensitive information (Tang and Others, 2023). A unified summarization framework integrates query-focused, privacy-aware methods using submodular information measures to balance privacy and utility (Kaushal and Others, 2020). Similarly, EROS, an entity-driven summarization model, ensures that critical privacy-related entities are selectively retained or masked (Singh and Others, 2024). While effective, its reliance on predefined entity filters risks over-sanitization or under-sanitization.

For privacy-preserving summarization, additional evaluation methods have been explored, including measuring privacy leakage via membership inference attacks on summarization outputs (Tang and Others, 2023), applying differential privacy parameters to control obfuscation levels (Igamberdiev and Others, 2022), and testing re-identification or contextual inference risks in generated summaries (Mireshghallah et al., 2023; Staab et al., 2023).

3 Methodology

The methodology adopted in this study is designed to support a modular pipeline for privacy-preserving cloud-based summarization through the identification and masking of sensitive tokens in the original document and contextual reconstruction of masked tokens in the document summary. Section 3.1 presents the induction of the sensitivity classification model, based on probabilistic labeling, feature engineering, and ensemble-based learning. Section 3.2 introduces the proposed masking strategy, which ranks and redacts sensitive tokens based on calibrated sensitivity scores and configurable constraints. Finally, Section 3.3 outlines the summary reconstruction process, which leverages contextual similarity between masked inputs and masked summaries to recover redacted content in the final summary.

3.1 Token Sensitivity Prediction

Given a corpus of annotated documents, where each token was assigned a soft sensitivity label score in the range [0,1], reflecting normalized agreement among participants on whether the token was marked as sensitive, we evaluated a diverse set of features that capture lexical, contextual, syntactic, and semantic properties of each token. These are grouped as follows:

- **Contextual Embeddings (4 features)**:
  - `autoencoded_contextual_embedding` — compressed representation of the embedding via autoencoder, reducing dimensionality while preserving semantic context.
  - `emb_cos_to_doc_mean` — cosine similarity between a token and its document mean embedding; lower values may indicate semantic outliers.
- **Syntactic and Semantic Tags (45 features)**:
  - Part-of-speech (POS) tags encoded as binary features across 17 categories
  - Named entity recognition (NER) tags across 18 entity types
- **Lexical and Positional Features (12 features)**:
  - TF-IDF score of the word within its document.
  - Relative position of the token within the sentence and the full document.
  - `word_length` and number of appearances of the token within the text.
  - Interaction features, such as the product of TF-IDF score and token length, and the ratio between token frequency and its position in text.
  - `surrounding_sentiment` – sentiment polarity of a window of 5–10 tokens around the word.
- **Masked Language Model Predictability (one feature)**:

  For each token, we queried a masked language model (MLM) by masking the word in its context and retrieving the prediction list. The feature value is the *rank* of the true word in the MLM's prediction. If the word is not in the top predictions, it receives a value of −1.

$$\text{MLM\_rank}(w_i) = \begin{cases} r_i, & w_i \in \text{Top-}k\,(\text{MLM}) \\ -1, & \text{otherwise} \end{cases}$$

Our feature selection procedure is described in the Appendix A.1.

3.2 Sensitive Tokens Masking

Algorithm 1 Threshold-Based Sensitive Word Masking

1: **Input:** Original text $T = [w_1, w_2, \ldots, w_n]$, trained sensitivity model M.
Self user-defined parameters: masking threshold $\theta \in [0, 1]$, max masking ratio ρ
2: **Output:** Masked text T_m with up to $\rho \cdot n$ masked tokens
3: Extract feature vector $\vec{f}_i$ for each token $w_i \in T$
4: Compute sensitivity score $s_i = M(\vec{f}_i)$ for each token
5: Sort tokens by s_i in descending order
6: Initialize masked count $m \leftarrow 0$, maximum allowed masks $m_{\max} = \rho \cdot n$
7: **for** each token w_i in sorted order **do**
8: **if** $s_i \geq \theta$ **and** $m < m_{\max}$ **then**
9: Replace w_i with `[MASK]`
10: Increment $m \leftarrow m + 1$
11: **end if**
12: **end for**
13: **Return:** Masked Text T_m

The sensitive word masking algorithm receives as input the original document T, and initializes the trained sensitivity classifier and the summarization module to be used downstream. The masking threshold θ, as well as the maximum allowed proportion of masked words ρ, are also defined as configurable parameters (lines 1–2). The algorithm then iterates through each token w in the document and extracts its associated features (e.g., syntactic tags, contextual embeddings, positional data). These features are passed through the classifier, which returns a sensitivity score $s_w \in [0, 1]$ indicating the probability that token w is sensitive in context (lines 3–5).

All tokens are then ranked by their sensitivity scores, and those with $s_w \geq \theta$ are selected as candidates for masking. However, to maintain summary quality,

the algorithm enforces a strict cap on redaction extent by masking only the top-ranked tokens until a predefined budget $\rho \cdot |T|$ is reached (lines 6–8). If the masking threshold θ is set too high, the masking budget ρ may not be fully utilized, resulting in limited redaction and reduced privacy coverage. Conversely, lower thresholds increase masking but may degrade summary quality. The framework therefore allows users to adjust θ based on their preferred privacy–utility balance, effectively choosing how much summary quality to trade for stronger privacy.

These selected tokens are replaced in the original document with a special token (e.g., `[MASK]`), resulting in a redacted version of the input (lines 9–10). This masked version can then be passed to an untrusted summarization server, which generates a summary without access to the original sensitive content (line 11).

3.3 Contextual Reconstruction of Masked Words

UnmaskingAlgorithm Description. The algorithm receives as input the masked original text T_m, the list of masked tokens L_m, and the masked summary S_m (lines 1–2). It initializes two dictionaries, D_{S_m} and D_{T_m}, which will store context windows extracted around each mask location (line 3).

The algorithm begins by iterating over all masked tokens appearing in the summary S_m (lines 4–7). For every `[MASK_i]`, it computes a local context window around the mask position by taking a fixed number of neighboring tokens to the left and right, with boundaries clipped to the valid token range (line 5). The resulting window is then stored in D_{S_m} (line 6). Next, the algorithm performs the same context-extraction procedure for each masked token in the masked original text T_m (lines 8–11). For each `[MASK_j]`, a context window is computed using the same boundary rules and stored in D_{T_m} (line 10).

To identify the appropriate replacement for each summary mask, the algorithm performs a pairwise comparison between every context in D_{S_m} and every context in D_{T_m} (lines 12–13). For each context pair, it computes a lexical similarity score based on the longest common subsequence (LCS) between the two context windows (line 14). This lexical score is then averaged with an independently computed semantic similarity score to form a total similarity measure (line 15). The total similarity values for all context pairs are stored in a similarity matrix (line 16).

Algorithm 2 Unmasking Algorithm

1: **Input:** Masked original text T_m , masking words list L_m, masked summary S_m
2: **Output:** Reconstructed summary with replaced `[MASK]` tokens
3: Initialize two dictionaries: D_{S_m} and D_{T_m}
4: **for each** `[MASK_i]` in S_m **do**
5: Compute context window:

$$\text{start} = \max(0, \text{index} - \text{window})$$

$$\text{end} = \min(\text{len}(\text{tokens}), \text{index} + \text{window} + 1)$$

6: Store in D_{S_m}
7: **end for**
8: **for each** `[MASK_j]` in T_m **do**
9: Compute context window:

$$\text{start} = \max(0, \text{index} - \text{window})$$

$$\text{end} = \min(\text{len}(\text{tokens}), \text{index} + \text{window} + 1)$$

10: Store in D_{T_m}
11: **end for**
12: **for each** key k_1 in D_{S_m} **do**
13: **for each** key k_2 in D_{T_m} **do**
14: **Compute Lexical Similarity:**

$$\text{lexical similarity} = \frac{2 \times \text{len}(\text{LCS}(D_{S_m}[k_1], D_{T_m}[k_2]))}{\text{len}(D_{S_m}[k_1]) + \text{len}(D_{T_m}[k_2])}$$

15: **Aggregate Similarities:**

$$\text{total similarity} = \frac{\text{lexical similarity} + \text{semantic similarity}}{2}$$

16: Store total similarity in the similarity matrix
17: **end for**
18: **end for**
19: **for each** `[MASK_i]` in S_m **do**
20: Find the highest similarity match from D_{T_m}
21: Replace `[MASK_i]` with the corresponding word from L_m
22: **end for**
23: **Return** Reconstructed summary

Finally, for each masked token in the summary S_m (lines 18–22), the algorithm selects the entry from D_{T_m} with the highest similarity score (line

20). The summary mask is then replaced with the corresponding original token retrieved from L_m (line 21). The fully reconstructed summary is returned as the output (line 23).

To illustrate the end-to-end functionality of our proposed framework, we provide a concrete example in the Appendix A.2.

4 Design of Experiments

To train a robust model for sensitive word detection, we first curated a gold-standard dataset composed of 10 long documents sampled from the WikiLeaks corpus (WikiLeaks, 2010). Four human participants were recruited according to expertise criteria to ensure both technical competence and domain awareness: two participants were recent graduates with a B.Sc. in Software Engineering, one was a cybersecurity and information security professional, and the fourth was a graduate in Data Engineering. This selection aimed to balance academic knowledge of computational methods with practical expertise in handling sensitive data contexts. Participation was voluntary, with the option to withdraw at any stage without penalty. Informed consent was obtained from all participants, who were explicitly informed that the collected annotations would be used exclusively for academic research purposes.

participants were presented with documents covering diverse domains, including national security, politics, sports, and current news events (see Table 1). They were instructed to label any token they deemed *sensitive*, under the assumption that the document contained potentially confidential information not intended for public disclosure. The definition of sensitivity was intentionally left open, allowing participants to rely on personal interpretation. A word could thus be considered sensitive in one context but not in another, depending on perceived confidentiality risks. Tokens not labeled as sensitive were implicitly treated as non-sensitive, forming the ground truth for model training.The full annotation guidelines provided to participants are detailed Appendix A.3.

To validate that the chosen training corpus size is sufficient for stable and reliable learning, we evaluated model performance under progressively increasing training-set sizes. In each iteration, one document was held out as the test set, while the remaining documents formed the candidate pool for training. For a given test document, we trained

Table 1: Comparison of the training and test datasets

Statistic	**WikiLeaks (Train)**	**CNN/Daily (Test)**
Number of documents	10	10
Total tokens	3,495	4,092
Average tokens / doc	349.5	409.2
Median tokens / doc	334.5	332.5
Std. deviation	253.9	244.0
Minimum tokens / doc	84	125
Maximum tokens / doc	972	907

the model multiple times using training subsets of varying sizes (from 1 to 9 documents), computed the mean testing accuracy for each subset size, and then averaged these results across all possible test-document selections (nested leave-one-out scheme). As shown in Figure 2, testing accuracy consistently improves as more training documents are included, with performance stabilizing around seven to eight documents. This convergence indicates that using ten documents for training provides a sufficiently large and representative dataset, yielding stable accuracy and diminishing returns from adding additional documents.

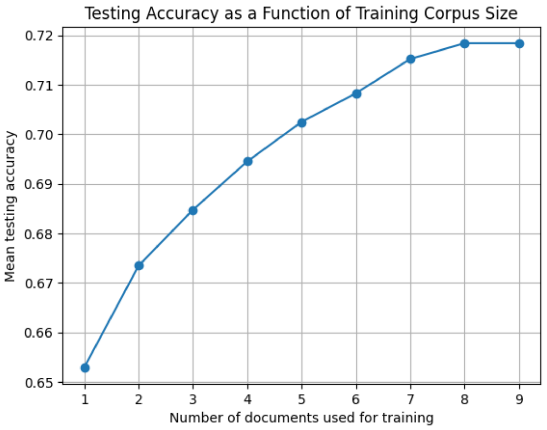


Figure 2: Learning stability by increasing training-set size (1–9 documents). The curve shows a clear and consistent increase in mean testing accuracy as more documents are included in the training set, followed by stabilization around seven to documents

Importantly, participants worked independently without consultation, ensuring that labeling reflected individual perspectives and minimizing bias due to joint discussion. This process yielded *soft labels* per token, ranging from 0 (no participants

flagged the token) to 1 (all participants agreed). Intermediate values (e.g., 0.25, 0.5, 0.75) captured partial consensus and were later used as proportional weights during model training.

Evaluation was conducted using the following metrics:

1. **Reconstruction Accuracy:** For each document, we computed the proportion of masked tokens in the summary that were correctly reconstructed by Unmasking Algorithm 2.

$$\text{Accuracy}_{\text{recon}} = \frac{1}{D} \sum_{d=1}^{D} \frac{|T_{\text{correct}}^{(d)}|}{|T_{\text{masked}}^{(d)}|}, \quad (1)$$

   where $T_{\text{masked}}^{(d)}$ denotes the set of masked tokens in document d, $T_{\text{correct}}^{(d)}$ the subset correctly reconstructed, and D the number of evaluated documents.

2. **Summary Quality:**

   ROUGE Score: Measures n-gram overlap between the generated summary and reference summaries, capturing lexical similarity and content coverage.

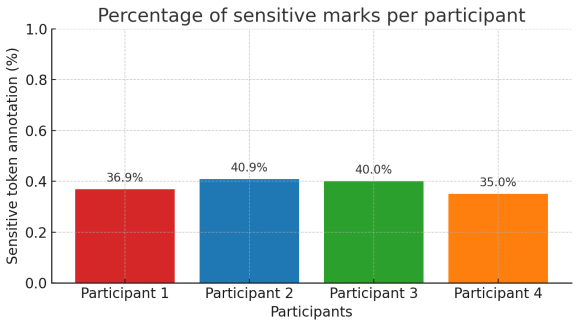
   BERTScore: Evaluates semantic similarity between generated and reference texts using contextual embeddings from a pretrained language model.

5 Results

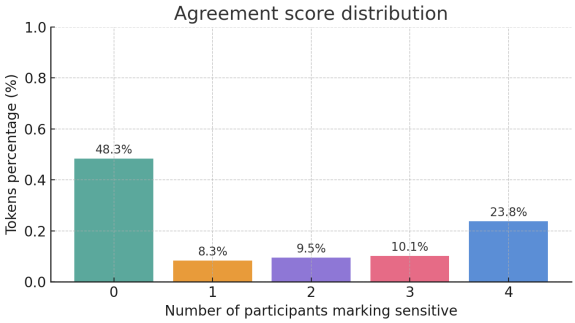
To quantify annotation consistency, we computed inter-participant agreement using both pairwise Cohen's κ and Fleiss' κ. The results of both metrics were highly consistent: the mean pairwise Cohen's κ and Fleiss' κ were 0.671. According to the commonly used Landis and Koch scale (Landis and Koch, 1977), these values correspond to *substantial agreement* (0.61–0.80). This level of agreement reflects that participants frequently converged on certain tokens, while still leaving room for subjective and context-dependent differences in sensitivity judgments.

Comparative results across the sensitive token detection models are presented in Table 2.

We selected the ensemble combination of all four classifiers (`Extra Trees`, `Random Forest`, `LightGBM`, and `XGBoost`) as our final sensitivity scoring model, based on its superior performance across all evaluation metrics. This model is used to assign a probability score to each token, estimating the likelihood that the token contains sensitive information.



(a) Percentage of sensitive marks per participant.



(b) Distribution of agreement scores across participants.

Figure 3: Annotation statistics. (a) shows the proportion of tokens each participant labeled as sensitive. (b) shows the distribution of agreement levels, i.e., how many participants agreed on labeling a token as sensitive.

Table 2: Performance comparison of tree-based models and their ensembles for sensitive token detection (mean ± std).

Model	Accuracy	Precision	Recall	F1-score
Extra Trees	0.733 ± 0.027	0.761 ± 0.028	0.842 ± 0.030	0.787 ± 0.025
LightGBM	0.742 ± 0.026	0.799 ± 0.027	0.777 ± 0.029	0.781 ± 0.024
Random Forest	0.753 ± 0.025	0.792 ± 0.026	0.811 ± 0.028	0.793 ± 0.023
XGBoost	0.745 ± 0.025	0.780 ± 0.027	0.819 ± 0.029	0.790 ± 0.024
ExtraTrees + LightGBM	0.750 ± 0.024	0.802 ± 0.025	0.785 ± 0.028	0.793 ± 0.022
ExtraTrees + XGBoost	0.757 ± 0.024	0.794 ± 0.026	0.808 ± 0.027	0.791 ± 0.022
RandomForest + LightGBM	0.758 ± 0.023	0.803 ± 0.025	0.804 ± 0.027	0.796 ± 0.021
RandomForest + XGBoost	0.756 ± 0.023	0.798 ± 0.025	0.811 ± 0.027	0.794 ± 0.021
ExtraTrees + XGBoost + LightGBM	0.762 ± 0.022	0.809 ± 0.024	0.793 ± 0.026	0.798 ± 0.020
RandomForest + ExtraTrees + XGBoost	0.768 ± 0.022	0.812 ± 0.024	0.801 ± 0.026	0.801 ± 0.020
RandomForest + ExtraTrees + LightGBM	0.770 ± 0.022	0.814 ± 0.024	0.803 ± 0.026	0.803 ± 0.019
RandomForest + XGBoost + LightGBM	0.772 ± 0.021	0.816 ± 0.023	0.807 ± 0.025	0.805 ± 0.019
RF + ET + XGB + LGBM	**0.776 ± 0.020**	**0.823 ± 0.022**	**0.840 ± 0.024**	**0.814 ± 0.018**

We evaluated summarization quality using three leading abstractive models: `T5` (Raffel et al., 2020), `BART` (Lewis et al., 2020), and `PEGASUS` (Zhang et al., 2020).

Figures 4 and 7 show that masking approximately 40–50% of sensitive tokens leads to only a moderate degradation in summarization quality, with about a 20% reduction in ROUGE-1 precision and F1 and a decrease of roughly 0.04 in BERTScore precision and F1. A similar trend is observed in the

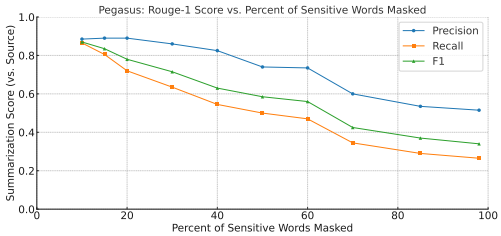


Figure 4: Performance degradation of summarization quality as a function of the proportion of masked sensitive words, evaluated using PEGASUS summarization model.

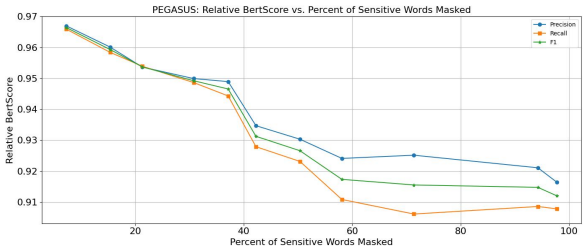


Figure 5: Relative BERTScore (precision, recall, and F1) as a function of the proportion of masked sensitive words, evaluated with the PEGASUS summarization model

BERTScore analysis, where scores remain high and relatively stable up to around 40% masking across all models, followed by a sharper decline beyond this point. Once the masking ratio exceeds 50%, BERTScore values fall below 0.95, indicating a noticeable loss in summary quality. Overall, these results suggest a favorable privacy–utility trade-off, as a substantial portion of sensitive content can be redacted without severely compromising summary informativeness or coherence. Accordingly, we recommend a masking threshold of $\theta = 0.5$, which balances privacy protection with strong summarization performance.

Finally, after running the evaluation on a held-out set of 10 documents from the CNN/DailyMail dataset, the reconstruction module achieved an average accuracy of $\textbf{Accuracy}_{\textbf{recon}} = 93.87\%$ when using the recommended masking threshold of $\theta = 0.5$. This result demonstrates that the vast majority of masked tokens in the generated summaries were correctly recovered, highlighting the effectiveness of the unmasking procedure described in Algorithm 2. Crucially, this high reconstruction accuracy is attainable because the client holds the full set of originally masked tokens.

6 Ablation Study

We conducted a systematic ablation study in which each feature group was removed individually while keeping all other components fixed. The results show a consistent performance decline across all evaluation metrics, confirming that each feature category contributes positively to the model. The most substantial degradation occurred when **lexical and positional features** were excluded (ΔF1 ≈ -0.16, $p < 0.01$), highlighting their critical role in capturing surface-level and positional cues. Removing **syntactic and semantic features** (POS and NER) also led to a significant drop in performance ($p < 0.05$), while excluding the **MLM predictability feature** caused a moderate but significant decline, indicating its complementary value. In contrast, removing **contextual embeddings** resulted in a smaller and non-significant reduction ($p = 0.09$).

7 Discussion

Building on the calibrated sensitivity scores, we evaluated a threshold-based masking algorithm designed to balance privacy with content preservation. By varying the masking threshold and measuring both the proportion of masked sensitive tokens and the total percentage of redacted words, we identified an empirical trade-off curve. The results revealed that moderate masking (around 40-50%) achieves a favorable balance: most participant-labeled sensitive words were redacted, while the overall text remained intelligible for summarization. At this masking level, the summarization model retained approximately 80% of its original precision, indicating that summary quality remained high despite the absence of sensitive content. This sweet spot was robust across documents and suggests a principled basis for threshold selection in real-world applications, with a recommended masking threshold of approximately 50%.

8 Limitations

While the proposed pipeline demonstrates promising results in privacy-preserving document summarization, several limitations warrant attention. First, while our sensitivity classifier is trained at the token level, real-world documents may contain sensitive multiword expressions. Decomposing such expressions into isolated tokens may overlook contextual dependencies or underestimate their overall sensitivity. Second, we have evaluated the sensitivity classifier on human-labeled data drawn from a limited subset of WikiLeaks documents. While this corpus reflects realistic privacy risks, it may not fully capture the diversity of sensitive expressions present in other domains such as healthcare, legal text, or social media.

Finally, the evaluation primarily builds upon automated metrics, including ROUGE-1, BERTScore, and exact-match reconstruction accuracy. While these metrics capture important aspects of content preservation and semantic similarity, they may not fully reflect the human-perceived summary informativeness or the privacy risk. In particular, the current evaluation does not explicitly account for potential attempts to recover sensitive information from masked or summarized outputs. Future work should incorporate human evaluation, richer semantic metrics, and formal privacy analyzes to provide a more comprehensive assessment.

9 Ethical Considerations

The authors acknowledge the use of AI-based tools, including ChatGPT and Grammarly, for language editing and manuscript refinement. The data used in this study consist of publicly available documents and do not involve the collection of new personal information. Potentially sensitive or identifying content was examined through the token-level sensitivity annotation process, and the released data include only sensitivity labels without any participant identifiers or additional personal metadata.

References

X. Chen and Others. 2022. Homomorphic encryption for privacy-preserving nlp models. *IEEE Transactions on Information Security*.

Ziwei Dou and 1 others. 2024. Learning to abstract self-disclosures in social text. *arXiv preprint arXiv:2402.10462*.

François Frénay and Michel Verleysen. 2013. Is mutual information adequate for feature selection in regression? *Neural Networks*, 37:12–18.

Gianluca Gambarelli and Aldo Gangemi. 2022. Privaframe: Ontology-based detection of sensitive personal information. In *Proceedings of the 21st International Semantic Web Conference (ISWC)*.

R. Gilad-Bachrach and Others. 2016. Computational overhead in homomorphic encryption for nlp. *Journal of Machine Learning Research*.

Xinyu Guo, Ling Liu, and Kai Chen. 2021. Exsense: Automatic sensitive information detection framework based on semantic features. In *Proceedings of the 30th USENIX Security Symposium*.

Y. Hao and Others. 2022. Privacy-aware text processing with homomorphic encryption. *Computational Linguistics Journal*.

Ruotian He and 1 others. 2025. Locally private retrieval-augmented generation for entity-sensitive queries. In *International Conference on Learning Representations (ICLR)*. OpenReview.

Zhifeng Hu, Serhii Havrylov, Ivan Titov, and Shay B. Cohen. 2020. Obfuscation for privacy-preserving syntactic parsing. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 62–72. Association for Computational Linguistics.

Jin Huo and Bo Jiang. 2023. Tsiip: Text sensitive information intelligent perception based on knowledge-enhanced language models. *Journal of Chinese Information Processing*.

A. Igamberdiev and Others. 2022. Differential privacy rewriting in nlp. *Proceedings of ACL*.

Matthew Jagielski, Nicholas Carlini, Florian Tramer, Eric Wallace, Reza Shokri, Adam Roberts, and 1 others. 2022. Auditing privacy in black-box language models. In *NeurIPS 2022*.

V. Kaushal and Others. 2020. A unified framework for generic, query-focused, privacy-preserving and update summarization using submodular information measures. *arXiv preprint arXiv:2010.05631*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Y.-L. Li. 2024. Aspirinsum: An aspect-based utility-preserved de-identification summarization framework. *arXiv preprint arXiv:2406.13947*.

Yuxuan Liu, Wei Zhao, and Huan Liu. 2021. Gcsa: Graph-based contextual sensitive attribute detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Luís Marujo, Wang Ling, Ricardo Ribeiro, Anatole Gershman, Chris Dyer, and Jaime Carbonell. 2015. Privacy-preserving multi-document summarization. *arXiv preprint arXiv:1506.03396*.

Fatemehsadat Mireshghallah and 1 others. 2023. Privacy leakage and reconstruction in language models: Inference attacks via masked context. *arXiv preprint arXiv:2305.06667*.

K. Muralitharan and S. Arumugam. 2024. Privacy-bert-lstm: Context-aware identification of sensitive text using deep language models. *Expert Systems with Applications*.

I. Pilán and Others. 2022. Pii removal and its impact on nlp tasks. *Journal of Computational Linguistics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the Workshop on Stylistic Variation*, pages 17–26. Association for Computational Linguistics.

Yuxuan Ren, Jingbo Li, and Rui Zhang. 2024. Cpsid: A context-aware pii detection framework for unstructured chinese text. *IEEE Access*, 12:43045–43057.

Anish Sheth, Shiva Subramani, and Jordan Boyd-Graber. 2023. Privacy-preserving nlp: A survey. *arXiv preprint arXiv:2304.03447*.

J. Singh and Others. 2024. Eros: Entity-driven controlled policy document summarization. *arXiv preprint arXiv:2403.00141*.

Steffen Staab and 1 others. 2023. Contextual inference attacks on anonymized text: Threats to pseudonymization under gdpr. *arXiv preprint arXiv:2303.01519*.

R. Tang and Others. 2023. Assessing privacy risks in language models: A case study on summarization tasks. *arXiv preprint arXiv:2310.13291*.

Aditya Vats and 1 others. 2023. Recovering from privacy-preserving masking with large language models. *arXiv preprint arXiv:2309.08628*.

WikiLeaks. 2010. Wikileaks. `https://wikileaks.org`. Accessed: 2010.

Canwen Xu, Eric Wallace, Shiva Subramani, Yu Sun, Yang Feng, and Jordan Boyd-Graber. 2021. Differentially private language models benefit from public pre-training. In *ACL 2021*.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *International Conference on Machine Learning (ICML)*.

Y. Zhou and Others. 2023. Privacy-preserving text processing using semantic clustering. *Journal of AI Research*.

A Appendix

A.1 Feature Selection Process

Model performance was assessed using 10-fold stratified cross-validation, where folds were stratified at the document level to preserve contextual consistency. In each fold, we selected the top K features using mutual information regression (Frénay and Verleysen, 2013), which captures general (possibly non-linear) dependencies between each feature and the ordinal target values (0 = no agreement, 0< target value< 1 = partial agreement, 1 = full agreement), while preserving the numerical meaning of the labels. The number of selected features was determined empirically by the feature selection process within the chosen ensemble model (the combination of all four classifiers). A feature was retained if it was selected by at least two of the four models. This approach ensured that the final model included only stable and frequently agreed-upon predictors, reflecting consensus across models while maintaining generalizability. Classical features—such as TF-IDF scores, relative token position within the sentence and document, word length, and surrounding sentiment—were standardized using z-score normalization, while embedding-based vectors (`autoencoded_contextual_embedding` and `emb_cos_to_doc_mean`) remained unscaled. The final feature vector was constructed by concatenating the standardized classical features with the embedding-based features, enabling the models to leverage both interpretable lexical/structural cues and dense contextual representations.

To enhance model reliability in downstream threshold-based decisions, we applied isotonic regression to calibrate the output probabilities of the final model, transforming raw scores into well-calibrated sensitivity estimates interpretable as likelihoods. Isotonic regression is a non-parametric calibration technique that fits a monotonic function to the predicted probabilities, preserving the ranking while improving their alignment with observed

frequencies (Zadrozny and Elkan, 2002).

A.2 End-to-End Workflow Example

Given the source text that appears in the figure 6, the sensitivity classifier is applied to each token. The model assigns a sensitivity probability score to each token based on a combination of contextual, syntactic, and semantic features. Tokens that exceed a predefined sensitivity threshold are considered candidates for redaction. However, in accordance with the masking algorithm 1, not all high-scoring tokens are masked: redaction is constrained by two additional criteria—the threshold θ must exceed an empirically determined optimal value, and the total number of masked tokens must not surpass a predefined masking budget proportional to the document length.

In the current example, five tokens satisfy these conditions: *Nathanyahu*, *Trump*, *met*, *Doha*, *support* and *Gaza*. These tokens are thus selected as sensitive, and subsequently replaced with the special placeholder token `[MASK]` in the redacted version of the input text. This transformation prepares the document for privacy-preserving summarization in the next stage.

Finally, the summarization model received the modified text with masked words and generated an output summary containing masked segments. **The Unmasking Algorithm** 2 is designed to reconstruct masked words in a summary by comparing their contextual environments with those of the original masked text. The method for reconstructing masked words relies solely on the original text being available to the document owner (user). We assume that an attacker does not have access to the original text, and therefore, they would not be able to utilize this reconstruction method effectively. This assumption is critical and should be explicitly noted in the paper to highlight the security of the approach. The process begins by extracting context windows for each masked token in both the summary and the original text.

For the example presented in figure 6, the corresponding context windows are displayed in Tables 4 and 3.

To find the most suitable replacements for the masked tokens, the algorithm computes a similarity ratio for every pair of masked tokens from the summary and the original text. This similarity score is based on the overlap between their context windows, specifically by determining the longest sequence of matching words (known as the longest common subsequence) between the two contexts. The score also considers the total number of words in both contexts to ensure a balanced comparison, favoring replacements that align semantically and structurally well with the surrounding text. ThZe resulting similarity matrix between four masked tokens from the summary (`[MASK]_1`–`[MASK]_4`) and six masked tokens from the original text (`[MASK]_1`–`[MASK]_6`) is shown below, where higher values indicate stronger contextual resemblance:

Original Text:
"Netanyahu and Trump met in Doha last summer. They support the idea of ending the conflict in Gaza."

Token Scores:

Token	Score
Netanyahu	0.9548
Trump	0.9675
met	0.8139
Doha	0.8644
summer	0.7588
support	0.814
idea	0.7873
ending	0.6768
conflict	0.7779
Gaza	0.8825

Masked Text:
Netanyahu, Trump,met Doha, support, Gaza classified as sensitive words.
The user set $\theta = 0.8$, accepting up to a 25-30% drop in ROUGE-1 and a 4-5% drop in BERTScore in order to achieve ~60% redaction of sensitive tokens.

"[MASK] and [MASK] [MASK] in [MASK] last summer. They [MASK] the idea of ending the conflict in [MASK]."

Masked Summary: "[MASK] and [MASK] [MASK] to end the war in [MASK]."

Figure 6: Example of Masking text and Summarization Using Algorithm 2.

Table 3: Masked Context for Text

Masked Token	Context Window
[MASK]$_1$	and
[MASK]$_2$	and in
[MASK]$_3$	and in last
[MASK]$_4$	in last summer. They
[MASK]$_5$	last summer. They the idea of.
[MASK]$_6$	the conflict in .

Table 4: Masked Context for Summary

Masked Token	Context Window
[MASK]$_1$	and
[MASK]$_2$	and the idea
[MASK]$_3$	and the idea of
[MASK]$_4$	the situation in .

$$\text{Similarity} = \begin{bmatrix} 1.000 & 0.611 & 0.490 & 0.116 & 0.144 & 0.169 \\ 0.479 & 0.537 & 0.467 & 0.234 & 0.444 & 0.326 \\ 0.432 & 0.492 & 0.443 & 0.235 & 0.506 & 0.349 \\ 0.260 & 0.405 & 0.368 & 0.290 & 0.288 & 0.643 \end{bmatrix}$$

After calculating the similarity scores, the algorithm selects the best match for each masked token in the summary, the word from the original text with the highest similarity score. The selected word is then used to replace the masked token, ensuring that the reconstructed summary remains coherent and contextually accurate.

The results after applying Algorithm 2 in tables 4 and 3:

- `[MASK]`$_1$ in the masked summary corresponds to `[MASK]`$_1$ in the masked text, resolving to **Nathanyahu**.
- `[MASK]`$_2$ in the masked summary corresponds to `[MASK]`$_2$ in the masked text, resolving to **Trump**.
- `[MASK]`$_3$ in the masked summary corresponds to `[MASK]`$_5$ in the masked text, resolving to **support**.
- `[MASK]`$_4$ in the masked summary corresponds to `[MASK]`$_6$ in the masked text, resolving to **Gaza**.

After replacing the masked tokens with their corresponding matches, the final reconstructed summary is:

Nathanyahu and Trump support to end the conflict in Gaza"

We evaluate the quality of the generated summaries using the ROUGE-1 score, which measures the unigram overlap between the generated summary and a human-written reference summary.

By focusing on the semantic alignment between masked tokens and their potential replacements, the algorithm effectively balances privacy preservation with readability, making it highly suitable for sensitive applications such as medical or legal document summarization.

A.3 Annotation Guidelines

You are given a set of textual documents.

Your task is to read the documents and mark individual tokens that you personally consider sensitive.

A token should be marked only if, in your own judgment, it may reveal or contribute to the exposure of information that feels private or sensitive within the given context.

There is no predefined definition of what makes a token sensitive. Please rely solely on your own intuition and understanding when making decisions.

When making your decisions, imagine that the document contains information that you would not want to be freely circulated or accessed by unintended parties. Mark tokens that, in your view, increase the risk of unwanted exposure or make the document sensitive to share.

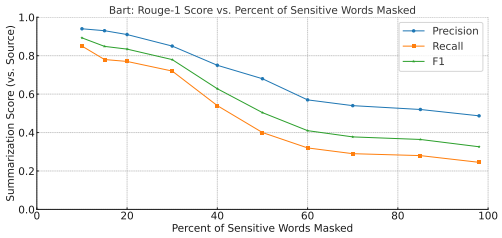
Guidelines

- Consider the local sentence context and, when relevant, the broader document context.
- The same token may or may not be sensitive depending on how it is used.
- Each token occurrence should be evaluated independently.
- There are no correct or incorrect answers.
- Mark sensitivity at the individual token level.

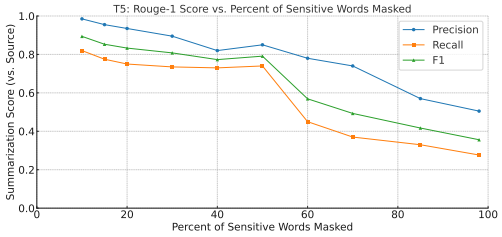
Important Notes

- Do not try to infer what the researchers expect.
- Do not follow any external rules or predefined categories.
- Use your natural judgment only.

A.4 Rouge and Bert Score Results



(a) BART model



(b) T5 model

Figure 7: Performance degradation of summarization quality as a function of the proportion of masked sensitive words, evaluated using Bart and T5 summarization models.

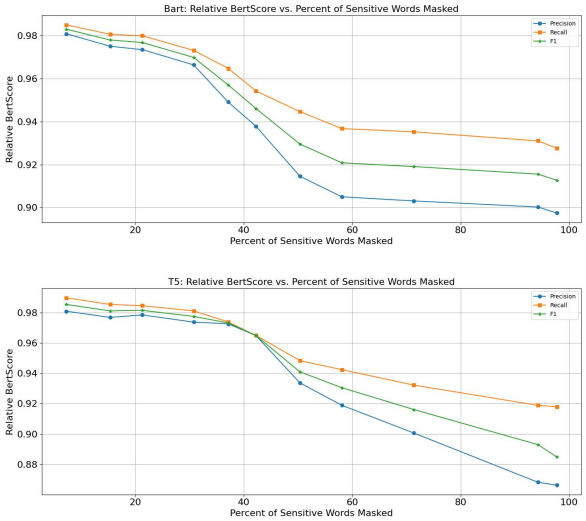


Figure 8: Relative BERTScore (precision, recall, and F1) as a function of the proportion of masked sensitive words across Bart and T5 models.

909

Wikileaks Document Annotation Examples:

"AN ARRANGEMENT FOR THE TRANSFER OF PRISONERS OF WAR, CIVILIAN INTERNEES, AND CIVILIAN DETAINEES BETWEEN THE FORCES OF THE UNITED STATES OF AMERICA, THE UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND, AND AUSTRALIA.This arrangement establishes procedures for the transfer of any Prisoners of War, Civilian Internees, and Civilian Detainees taken during operations against Iraq, from the custody of either US, UK, or Australian forces to the custody of any of the other parties. The Parties undertake as follows: This arrangement will be implemented in accordance with the Geneva Convention Relative to the Treatment of Prisoners of War and the Geneva Convention Relative to the Protection of Civilian Persons in Time of War, as well as customary international law. US, UK, and Australian forces will accept (as Accepting Powers) prisoners of war, civilian internees, and civilian detainees who have fallen into the power of any of the other parties (the Detaining Power), as mutually determined. They will also be responsible for maintaining and safeguarding all such individuals whose custody has been transferred to them. Transfers of prisoners of war, civilian internees, and civilian detainees between Accepting Powers may take place as mutually determined by both the Accepting Power and the Detaining Power. Arrangements to transfer prisoners of war, civilian internees, and civilian detainees who are casualties will be expedited for treatment according to their medical priority. All such transfers will be administered and recorded within the systems established under this arrangement for the transfer of prisoners of war, civilian internees, and civilian detainees. Any prisoners of war, civilian internees, and civilian detainees transferred by a Detaining Power will be returned by the Accepting Power to the Detaining Power without delay upon request by the Detaining Power. The release or repatriation or removal to territories outside Iraq of transferred prisoners of war, civilian internees, and civilian detainees will only be made upon the mutual arrangement of the Detaining Power and the Accepting Power. The Detaining Power will retain full rights of access to any prisoners of war, civilian internees, and civilian detainees transferred from Detaining Power custody while such persons are in the custody of the Accepting Power. The Accepting Power will be responsible for the accurate accountability of all prisoners of war, civilian internees, and civilian detainees transferred to it. Such records will be available for inspection by the Detaining Power upon request. If prisoners of war, civilian internees, or civilian detainees are returned to the Detaining Power, the records (or a true copy of the same) relating to those prisoners of war, civilian internees, and civilian detainees will also be handed over[cite: 13, 14]. The Detaining Powers will assign liaison officers to Accepting Powers in order to facilitate the implementation of this arrangement. The Detaining Power will

be solely responsible for the classification under Articles 4 and 5 of the Geneva Convention Relative to the Treatment of Prisoners of War of potential prisoners of war captured by its forces. Prior to such a determination being made, such detainees will be treated as prisoners of war and afforded all the rights and protections of the Convention even if transferred to the custody of an Accepting Power.Where there is doubt as to which party is the Detaining Power, all Parties will be jointly responsible for and have full access to all persons detained (and any records concerning their treatment) until the Detaining Power has by mutual arrangement been determined. To the extent that jurisdiction may be exercised for criminal offenses, to include pre-capture offenses, allegedly committed by prisoners of war, civilian internees, and civilian detainees prior to a transfer to an Accepting Power, primary jurisdiction will initially rest with the Detaining Power. Detaining Powers will give favorable consideration to any request by an Accepting Power to waive jurisdiction. Primary jurisdiction over breaches of disciplinary regulations and judicial offenses allegedly committed by prisoners of war, civilian internees, and civilian detainees after transfer to an Accepting Power will rest with the Accepting Power. The Detaining Power will reimburse the Accepting Power for the costs involved in maintaining prisoners of war, civilian internees, and civilian detainees transferred pursuant to this arrangement. At the request of one of the Parties, the Parties will consult on the implementation of this arrangement. Done at Camp As Sayliyah, Doha, Qatar on this 25 day of March 2003. For the United States of America: John P. Abizaid LTG, USA Deputy Commander Forward United States Central Command

2.

"Report on Allegations Against a Senior Official, UNMIL 1. Allegations Received: The Investigations Division of the Office of Internal Oversight Services (ID/OIOS) received information concerning a former Senior Official of the United Nations Mission in Liberia (UNMIL). The allegations were related to: An inappropriate relationship with a Local Woman who holds dual American-Liberian citizenship. This Local Woman reportedly has close personal and family links with the former Taylor regime in Liberia. Her family is said to have significant logging interests in Liberia and well-documented close connections with the Taylor regime. Furthermore, the Nobel Peace Prize-nominated NGO \"Global Witness\" has alleged that her uncle has been involved in arms smuggling in the region. The Senior Official invited the Local Woman to functions involving both UNMIL staff and individuals outside the UN, some of which were official in nature. Concerns were raised by several staff members interviewed by ID/OIOS that

the Local Woman was relaying information gathered from the Senior Official and UNMIL to Mr. Taylor and other interested parties. The Local Woman traveled on UNMIL air assets on occasion, despite not being authorized to do so as she was neither a UN staff member nor had an official reason to use them. The Senior Official requested and UNMIL's senior management authorized her use of the UN shuttle. 2. ID/OIOS Investigation Findings: he ID/OIOS investigation found evidence that the Senior Official : Failed to uphold the standards of conduct expected by the United Nations by maintaining a relationship with the Local Woman.\nFailed to carry out his management responsibilities with the best interests of the Organization in mind by authorizing the use of United Nations aviation assets by the Local Woman, a person not authorized to use such assets