Empirical Analysis of Feature-Level Data Minimization Trade-offs in Machine Learning

GPT-5.2, Gemini 3 Pro and Copilot

Independent Researcher
Republic of Korea

Abstract

Data minimization is legally mandated in major data protection frameworks, yet practitioners still lack decision-oriented evidence on how staged, feature-level minimization changes model utility and error profiles across model families. We conduct a controlled experimental study using the UCI Adult dataset, comparing four minimization strategies defined by the authors with LLM-assisted decision support : a human-designed heuristic baseline and three strategies whose feature-removal order is instantiated with large language model (LLM) decision support and validated by the authors. We evaluate Logistic Regression, Random Forest, and Gradient Boosting under four progressive minimization levels.

Across models, aggressive minimization can trigger a sharp collapse of recall while accuracy remains comparatively stable. At the strictest level, the most aggressive setting reduces recall to near zero (0.0–0.039) with accuracy remaining above 0.74, yielding extreme utility loss (ΔF1 down to -0.68). In contrast, the utility-aware Robust strategy shows substantially smaller degradation at the same level (ΔF1 between -0.09 and -0.11 across models).

These results provide a practical, implementation-light evaluation framework for compliance-oriented feature minimization, and show that the primary cost of minimization can manifest as strategy-dependent recall collapse rather than a uniform accuracy drop.

1 Introduction

Recent advances in artificial intelligence (AI) have significantly increased the scale and complexity of data-driven decision-making systems. As AI models increasingly rely on large-scale personal data, privacy protection and regulatory compliance have become central considerations in AI system design. In particular, data minimization, a core principle in modern privacy regulations, requires that only data strictly necessary for a given purpose be collected and processed.

A wide range of privacy-preserving techniques have been proposed to address these concerns, including anonymization, differential privacy, and secure computation. While such mechanisms can provide formal privacy guarantees, they often introduce additional system complexity, performance overhead, or noise-induced utility loss, making them difficult to deploy in practice. By contrast, data minimization occupies a unique position: it is not only a legally mandated requirement under regulations such as the GDPR, but also one of the earliest and most implementation-light design choices available to practitioners, directly affecting data collection, storage, and model inputs.

Preprint. Under review.

Despite its legal importance, data minimization is often treated as a qualitative guideline rather than a quantitatively analyzable design choice for model design and deployment. In practice, reducing input features may degrade predictive performance, yet the extent, structure, and failure modes of this degradation remain insufficiently characterized. Moreover, feature removal can alter model error behavior in non-obvious ways, potentially impairing critical decision functions even when aggregate performance metrics such as accuracy appear stable. Existing research has largely focused on privacy-preserving learning mechanisms that modify training procedures or model outputs, while comparatively less attention has been paid to feature-level data minimization as an independent design dimension. As a result, practitioners lack decision-oriented evidence on how staged, compliance-driven feature removal impacts model utility and error characteristics across different model families.

In this study, we examine the impact of feature-level data minimization on model performance in AI-based classification tasks. Using the UCI Adult dataset as a controlled benchmark, we evaluate how progressive feature removal affects predictive performance and error profiles across multiple machine learning models. To reflect realistic design choices under regulatory and operational constraints, we consider multiple data minimization strategies, including a human-designed heuristic baseline and strategies defined by the authors, with feature-removal orders informed by large language models (LLMs) as decision aids.

The main contributions of this work are summarized as follows:

- We propose a structured experimental framework that operationalizes legally mandated data minimization as progressive feature reduction, enabling controlled and reproducible evaluation across strategies and model architectures.
- We provide empirical evidence that feature-level data minimization can induce severe, strategy-dependent recall collapse that is not captured by accuracy alone, highlighting a critical risk for compliance-driven deployments.
- We demonstrate how large language models can support systematic, transparent feature-removal decisions as AI co-scientists, without acting as model optimizers or replacing human responsibility in experimental design.
- We bridge legal requirements, technical feasibility, and deployment-oriented evaluation by offering an implementation-light methodology that aligns privacy compliance with practical machine learning workflows.

2 Related Work

2.1 Data Minimization in Machine Learning

Data minimization has emerged as a core principle in both regulatory frameworks and machine learning system design, most notably formalized in the GDPR [5]. Recent research has moved beyond legal interpretation toward operational definitions of data minimization in learning systems. Staab et al. provide a comprehensive systematization of knowledge, organizing data minimization methods across the machine learning pipeline and clarifying their technical and conceptual boundaries [10]. Building on this foundation, Staab et al. further demonstrate how vertical data minimization can be applied in practice by selectively removing features while maintaining task utility [11]. Complementarily, Ganesh et al. frame data minimization as a first-class design principle in machine learning, emphasizing its interaction with model utility and privacy risk in real-world deployments[7].

2.2 Privacy Models and Feature-level Protection

Classical privacy-preserving models such as k-anonymity [12] and differential privacy [4] have laid the theoretical groundwork for protecting sensitive information in data analysis. While differential privacy offers strong formal guarantees, its practical deployment often incurs substantial utility loss. Recent studies therefore explore intermediate approaches that optimize the privacy–utility trade-off through feature selection and minimization. Almasi et al. investigate privacy–utility optimization for secure systems, showing that aggressive feature removal can significantly impair predictive performance [1]. These findings motivate empirical studies of feature-level minimization strategies that do not rely on noise injection.

2.3 Quasi-identifiers and Re-identification Risk

Beyond explicitly sensitive attributes, quasi-identifiers (QIDs) pose significant re-identification risk when combined with auxiliary information. Gkoulalas-Divanis and Tassa analyze re-identification driven by QIDs, emphasizing that privacy leakage can persist even after removal of direct identifiers [8]. This issue is particularly relevant for tabular datasets commonly used in privacy and fairness research, including the UCI Adult dataset [3]. Continued reliance on this dataset has motivated revisions and replacements, such as the Retiring Adult benchmark proposed by Ding et al., which addresses both fairness and data quality limitations [2].

2.4 Fairness-related Performance Criteria

Fairness-aware machine learning has highlighted the limitations of aggregate performance metrics in capturing model behavior under data constraints. Foundational work by Hardt et al. introduces equality of opportunity, emphasizing recall parity across protected groups as a core fairness criterion in supervised learning [9]. More recent empirical studies revisit the assumption that removing sensitive attributes necessarily improves fairness; Feng et al. demonstrate that excluding sensitive features can degrade model performance due to proxy effects [6]. These findings underscore the importance of analyzing recall, F1-score, and error asymmetry rather than accuracy alone when evaluating feature removal strategies.

2.5 Positioning of This Work

In contrast to prior work that focuses on theoretical guarantees or individual minimization mechanisms, this study provides a systematic empirical comparison of multiple feature-level data minimization strategies under a unified experimental protocol. By explicitly analyzing recall and F1-score degradation across models and minimization levels, our work connects quasi-identifier removal to strategy-dependent recall collapse, complementing existing research on privacy-aware learning and compliance-oriented data minimization.

3 Experimental Setup

3.1 Dataset

All experiments are conducted using the UCI Adult dataset, a widely adopted benchmark for studies on algorithmic fairness, privacy, and socio-economic prediction. The dataset consists of demographic and employment-related attributes and is commonly used to predict whether an individual's income exceeds a predefined threshold.

Following standard preprocessing practices, instances with missing values are removed, and categorical attributes are converted using one-hot encoding. The resulting dataset contains both numerical and binary features, enabling a controlled evaluation of feature-level data minimization strategies. Although the dataset has known limitations, it is employed in this study as a benchmark to ensure reproducibility and comparability across experiments.

Importantly, the Adult dataset includes a range of demographic and socio-economic attributes that are commonly discussed in privacy and fairness research, such as age, marital status, occupation, and education. While these variables do not constitute direct identifiers, they can function as quasi-identifiers when combined, and many of them also carry strong predictive signals for income classification. This combination makes the dataset particularly suitable for examining how feature-level data minimization reshapes the privacy–utility trade-off.

3.2 Train-Test Split

The dataset is randomly divided into training and test sets using an 80:20 split. Stratified sampling is applied based on the target variable to preserve the original class distribution across splits. This ensures that performance differences across data minimization levels and strategies are not influenced by class imbalance artifacts.

The same train-test split is used for all experimental conditions to guarantee a fair comparison across models, strategies, and minimization levels.

3.3 Data Minimization Levels

In our experimental design, minimization levels and minimization strategies serve distinct roles: levels determine how much data are removed, whereas strategies determine which attributes are prioritized for removal.

This level-based design operationalizes the data minimization principle as a graded compliance constraint: in practice, organizations often pursue staged reduction of collected attributes to limit unnecessary processing while monitoring utility loss. Although direct identifiers (e.g., resident registration numbers, passport numbers, driver's license numbers, and alien registration numbers) are excluded, even quasi-identifiers and sensitive attributes can enable re-identification when combined with auxiliary information, increasing the risk of identity theft and financial harm in case of leakage. The privacy sensitivity associated with each minimization level is defined in terms of cumulative re-identification risk rather than absolute legal categories. Lower levels retain a broader set of quasi-identifying and socio-economic attributes, resulting in higher potential for attribute linkage and inference. As minimization levels increase, features with higher individual or combinatorial re-identification risk are progressively removed, reducing the amount of personal information exposed to the model. We therefore evaluate progressive feature removal levels to quantify the privacy–utility trade-off under increasingly strict minimization.

To systematically evaluate the impact of data minimization, four levels of feature reduction are defined:

- **Level 0**: No features are removed, serving as the baseline condition.
- **Level 1**: An initial group of features is removed according to a predefined strategy.
- **Level 2**: Features removed at Level 1 plus an additional group.
- **Level 3**: Features removed at Level 1 and Level 2 plus a final group.

Conceptually, Level 1 corresponds to the removal of widely recognized quasi-identifiers, Level 2 extends this by eliminating additional socio-economic attributes with moderate linkage risk, and Level 3 represents a strict minimization setting in which only a minimal, task-essential feature set remains. This progressive design allows us to observe both gradual and extreme effects of data minimization on model performance. The assignment of features to each minimization level is determined by predefined strategy-specific rules rather than ad-hoc or random removal, ensuring consistency across models and experimental runs.

3.4 Data Minimization Strategies

While minimization levels control the overall intensity of feature reduction (i.e., how much information is removed), minimization strategies determine the priority and criteria by which features are removed. Each strategy reflects a distinct decision-making rationale that could plausibly be adopted in practice. We consider four representative strategies, summarized below.

- **Heuristic Strategy (H).** The heuristic strategy serves as a human-designed baseline and reflects common manual practices in privacy-aware data preprocessing. Features are removed based on intuitive judgments about sensitivity and appropriateness, informed by prior experience and conventional norms rather than formal risk or utility metrics. In practice, this leads to early removal of attributes that are widely perceived as sensitive (e.g., explicitly demographic variables), followed by gradual removal of additional socio-economic features at higher minimization levels. This strategy shows experience-driven minimization without systematic optimization.
- **Privacy-first Strategy (P).** The privacy-first strategy prioritizes the mitigation of re-identification and inference risk. Features are ranked according to their potential to function as quasi-identifiers, either individually or in combination with other attributes. Attributes with high linkage or inference risk are removed at earlier minimization levels, even if

they are known to carry predictive utility. Under this strategy, demographic and socio-economic attributes commonly associated with re-identification risk (e.g., age-related or occupation-related variables) are removed before less sensitive, task-oriented features.

- **Conservative Strategy (C).** The conservative strategy enforces an aggressive interpretation of data minimization that approximates maximal compliance. Rather than prioritizing specific attribute types, this strategy removes a large number of features early, retaining only a minimal subset deemed strictly necessary for task execution. As minimization levels increase, most quasi-identifying and utility-relevant attributes are eliminated, resulting in an extremely reduced feature set at the strictest level. This strategy is intended to approximate scenarios where minimization is interpreted primarily as maximal reduction rather than balanced optimization.
- **Robust Strategy (R).** The robust strategy explicitly balances privacy risk against predictive relevance. Features are evaluated along two dimensions: their potential privacy or re-identification risk and their contribution to predictive performance. Attributes with high risk and low utility are removed first, whereas features that exhibit strong predictive value are retained until later minimization levels, even if they carry moderate privacy risk. This strategy indicates a utility-aware approach to data minimization, aiming to preserve stable model behavior under increasingly strict reduction constraints.

Across all strategies, features are conceptually grouped into sensitive, quasi-sensitive, and non-sensitive categories based on their potential privacy risk and re-identification likelihood, and each strategy differs in how aggressively these groups are removed across minimization levels.

The P, C, and R strategies are derived with the assistance of large language models acting as AI co-scientists to support systematic reasoning, while the H strategy serves as a human-only reference baseline.

3.5 Models and Metrics

To analyze the impact of data minimization across models with different representational capacities, three classification models are selected and evaluated in increasing order of complexity: Logistic Regression, Random Forest, and Gradient Boosting.

Logistic Regression is included as a linear baseline model due to its interpretability and widespread use in social-economic prediction tasks. As a parametric model with limited expressive power, it provides a lower-bound reference to understand how feature removal affects simple decision boundaries.

Random Forest represents a non-linear ensemble method that aggregates multiple decision trees trained on bootstrapped samples. Its inherent robustness to feature redundancy and noise makes it suitable for assessing whether ensemble-based models can better tolerate data minimization compared to linear classifiers.

Gradient Boosting is selected as a high-capacity model that iteratively builds decision trees to correct previous errors. While it typically achieves strong baseline performance, its reliance on fine-grained feature interactions makes it more sensitive to aggressive feature removal. Including Gradient Boosting allows us to examine whether data minimization disproportionately impacts models with higher expressive power.

By evaluating these models in a structured complexity order, the experimental design enables a systematic comparison of how data minimization affects classifiers ranging from simple linear models to advanced ensemble methods.

Model performance is evaluated using Accuracy, Precision, Recall, and F1-score. Among these metrics, the F1-score is emphasized as it captures the balance between precision and recall and is particularly informative under class imbalance and asymmetric error costs.

3.6 Evaluation Procedure

For each combination of data minimization strategy and level, models are trained on the reduced feature set and evaluated on the same test set. No model-specific hyperparameter tuning is performed across minimization levels to isolate the effect of feature removal from optimization effects.

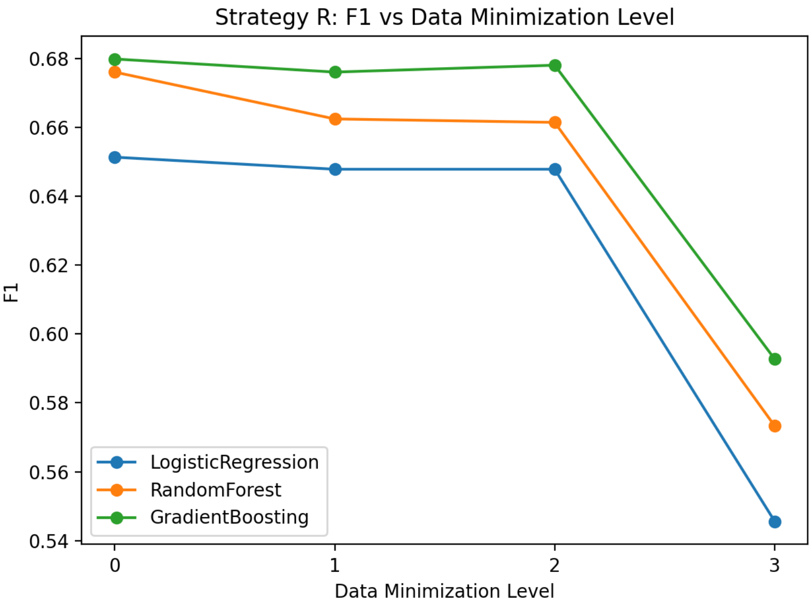

Figure 1: F1-score degradation across data minimization levels under the Robust (R) strategy.

Model performance is assessed using multiple classification metrics, with particular emphasis on the F1-score. In addition to absolute performance values, relative performance degradation with respect to the baseline (Level 0) is analyzed to quantify the cost of data minimization.

This evaluation framework enables comparison of utility degradation across minimization strategies under consistent experimental conditions.

4 Results

4.1 Overall Performance under Data Minimization

Figure 1 presents the general trend of model performance under increasing data minimization levels, measured by the F1-score. Across all evaluated models and strategies, performance consistently decreases as the level of data minimization increases.

The baseline condition (Level 0) achieves the highest F1-score for all models. However, the rate and severity of this degradation differ substantially across minimization strategies, indicating heterogeneous sensitivity to feature reduction.

To illustrate this general trend in a representative and interpretable manner, we focus on the Robust (R) strategy. This strategy preserves a substantial subset of features across minimization levels, allowing performance degradation to be observed progressively rather than collapsing abruptly at early stages. As a result, the figure highlights gradual and model-dependent changes in F1-score, making cross-model comparisons under increasing minimization levels visually clear.

4.2 Comparison of Human Heuristic and LLM-derived Strategies

Different performance patterns emerge when comparing the human heuristic strategy (H) with LLM-derived strategies (P, C, and R). At lower minimization levels, performance differences between strategies remain limited. However, as minimization becomes more aggressive, the divergence between strategies becomes pronounced.

In several configurations, particularly for Logistic Regression and Random Forest, the heuristic strategy maintains higher F1-scores at the same minimization level. In contrast, LLM-derived

Table 1: Model performance and F1-score degradation at the highest data minimization level (Level 3).

Strategy	Model	#Features	Accuracy	Recall	F1	ΔF1
H	Logistic Regression	66	0.8103	0.3947	0.5077	-0.1436
H	Random Forest	66	0.8036	0.5651	0.5879	-0.0881
H	Gradient Boosting	66	0.8318	0.4625	0.5769	-0.1029
P	Logistic Regression	20	0.8095	0.3943	0.5064	-0.1449
P	Random Forest	20	0.8214	0.4781	0.5704	-0.1056
P	Gradient Boosting	20	0.8369	0.4567	0.5813	-0.0985
C	Logistic Regression	1	0.7466	0.0393	0.0713	-0.5800
C	Random Forest	1	0.7520	0.0004	0.0009	-0.6751
C	Gradient Boosting	1	0.7519	0.0000	0.0000	-0.6798
R	Logistic Regression	39	0.8202	0.4353	0.5456	-0.1057
R	Random Forest	39	0.8103	0.5143	0.5733	-0.1027
R	Gradient Boosting	39	0.8394	0.4715	0.5927	-0.0871

strategies exhibit steeper performance degradation at higher levels, reflecting differences in the extent and ordering of feature removal.

These results demonstrate that performance under data minimization is strongly influenced by the strategy-specific removal sequence rather than the minimization level alone.

4.3 Impact of Strategy Strictness

A pronounced performance collapse is observed under the Conservative strategy at the highest minimization level. Although accuracy remains relatively stable, recall sharply declines to near-zero values, resulting in severe degradation of the F1-score.

This collapse is consistently observed across all models evaluated, as reflected by extreme negative ΔF1 values at Level 3. In contrast, the Robust strategy exhibits substantially smaller relative performance losses at the same minimization level, avoiding the abrupt collapse seen under aggressive feature removal.

Table 1 summarizes the model performance at the highest minimization level, highlighting the contrasting degradation patterns between strategies with different levels of strictness.

4.4 Model-wise Sensitivity Analysis

Sensitivity to data minimization varies among learning algorithms. Logistic Regression exhibits a gradual decline in performance as features are removed, reflecting its dependence on informative linear feature combinations. Random Forest demonstrates moderate robustness, while Gradient Boosting achieves the highest baseline performance, but shows increased sensitivity under aggressive feature removal.

Across models, similar degradation patterns are consistently observed under the same minimization strategies.

5 Discussion

The results demonstrate that the impact of data minimization on model performance is not uniform, but strongly depends on how features are removed. A central finding of this study is the pronounced collapse in recall under aggressive data minimization, even when overall accuracy remains relatively stable.

This phenomenon can be explained through the role of quasi-identifiers. Features such as age, marital status, and occupation are commonly treated as quasi-identifiers due to their re-identification risk. However, in the Adult dataset, these attributes also carry substantial predictive signal for the positive class. Their early removal therefore disproportionately reduces the model's ability to correctly identify positive instances, leading to recall collapse rather than uniform performance degradation.

Importantly, this shift alters the error profile of the model rather than simply reducing overall capacity. While accuracy remains stable due to the dominance of the negative class, recall sharply deteriorates, resulting in large negative ΔF1 values. This finding suggests that evaluating data minimization solely through aggregate metrics such as accuracy can obscure critical changes in model behavior that are highly relevant for downstream decision-making.

This observation highlights a tension that is not fully addressed in prior work on data minimization and privacy-preserving learning, which has primarily focused on reducing re-identification risk or improving fairness, often evaluating success through aggregate performance metrics such as accuracy or AUC. In contrast, our findings show that aggressive removal of quasi-identifiers can systematically reshape the error profile of models, revealing a trade-off that is not captured by aggregate metrics alone. Our results indicate that minimizing quasi-identifiers without considering their predictive role may inadvertently increase error asymmetry, particularly in applications where false negatives carry high cost.

The comparison between human heuristic and LLM-derived strategies further underscores this point. Although the heuristic strategy relies on intuitive judgments of feature sensitivity, it exhibits pronounced degradation at higher minimization levels. In contrast, the LLM-assisted Robust strategy avoids catastrophic recall collapse by selectively retaining features with high predictive utility while removing high-risk attributes. This suggests that LLMs can function as effective AI co-scientists by contributing structured, semantically informed design choices at the level of experimental configuration, rather than direct model optimization.

From a system design perspective, these findings imply that data minimization should be treated as a design space rather than a binary constraint. Strategy selection plays a critical role in shaping model behavior, and privacy-aware feature removal should be evaluated not only in terms of data reduction but also in terms of its impact on error distribution and task-specific risk.

Overall, this study demonstrates that data minimization is not a free operation: the way data is minimized fundamentally reshapes model behavior. Recognizing and managing this trade-off is essential for deploying privacy-aware AI systems that remain reliable under regulatory constraints.

6 Limitations

This study has several limitations that should be considered when interpreting the results.

First, all experiments are conducted using the UCI Adult dataset, which is known to reflect historical and demographic biases inherent to its census-based origin. Consequently, the absolute performance values reported in this study should not be interpreted as indicators of real-world deployment outcomes or contemporary socio-economic conditions.

Second, the Adult dataset was not originally designed for privacy-focused machine learning research. As a result, certain privacy risks commonly observed in modern datasets—such as high-dimensional behavioral signals, temporal dependencies, or cross-platform data linkability—are not captured. This limits the external validity of the findings when generalizing to more complex, real-world data environments.

Despite these limitations, the Adult dataset is intentionally employed as a controlled benchmark. Its widespread adoption in studies on fairness, privacy, and algorithmic accountability enables reproducible experimentation and facilitates meaningful comparison with prior work. By using a well-understood dataset, this study isolates the effects of data minimization strategies from confounding factors related to data collection or preprocessing.

Third, feature sets across data minimization strategies are not matched by cardinality. While this design reflects realistic applications of data minimization—where strategies inherently differ in aggressiveness and prioritization—it prevents direct comparison under identical feature budgets. Future work could explore controlled experimental settings in which strategies are evaluated under fixed feature cardinality to further disentangle the impact of feature selection order from feature quantity.

Finally, this study focuses exclusively on feature-level data minimization and does not incorporate complementary privacy-preserving mechanisms such as differential privacy, secure computation, or federated learning. Extending the proposed framework to integrate such techniques, as well as evaluating it on additional datasets from diverse domains, remains an important direction for future research. Future work will explore QID-aware data minimization strategies that explicitly account for both re-identification risk and error sensitivity, enabling more nuanced trade-offs between privacy protection and model reliability.

7 Conclusion

This study demonstrates that data minimization is not a uniform or risk-free operation, but one that fundamentally reshapes model utility and error profiles depending on how features are removed. Our results show that aggressive minimization can trigger severe recall collapse even when overall accuracy remains comparatively stable, whereas selective and risk-aware strategies substantially mitigate this effect.

By systematically comparing a human-designed heuristic with LLM-assisted minimization strategies across multiple model families, we show that the primary cost of feature-level minimization can manifest as strategy-dependent recall collapse rather than a uniform accuracy drop. These findings provide decision-oriented, implementation-light evidence for compliance-driven feature minimization, highlighting that data minimization should be treated as a strategy design problem rather than a binary constraint when deploying privacy-aware AI systems.

Acknowledgments and Disclosure of Funding

The authors declare no external funding and no competing interests.

A Appendix / Supplemental Material

This appendix provides additional experimental details, figures, and tables that complement the main results presented in the paper. All materials are included to support reproducibility and to provide finer-grained insights into the impact of data minimization strategies.

A.1 Complete Trade-off Figures

Figures 2 and 3 present the complete set of performance trade-off curves across all data minimization strategies. While the main paper focuses on aggregated trends, these figures allow for strategy-wise inspection of model behavior.

A.2 Delta Performance Tables

Table 2 reports the relative change in F1-score at Level 3 compared to the baseline (Level 0) for all strategies and models. Negative values indicate performance degradation caused by feature removal.

Table 2: Relative F1-score change at Level 3 compared to Level 0 (ΔF1 = F1$_{L3}$ – F1$_{L0}$).

Strategy	Model	Features Remaining	ΔF1
H	Logistic Regression	66	-0.1436
H	Random Forest	66	-0.0881
H	Gradient Boosting	66	-0.1029
P	Logistic Regression	20	-0.1449
P	Random Forest	20	-0.1056
P	Gradient Boosting	20	-0.0985
C	Logistic Regression	1	-0.5800
C	Random Forest	1	-0.6751
C	Gradient Boosting	1	-0.6798
R	Logistic Regression	39	-0.1057
R	Random Forest	39	-0.1027
R	Gradient Boosting	39	-0.0871

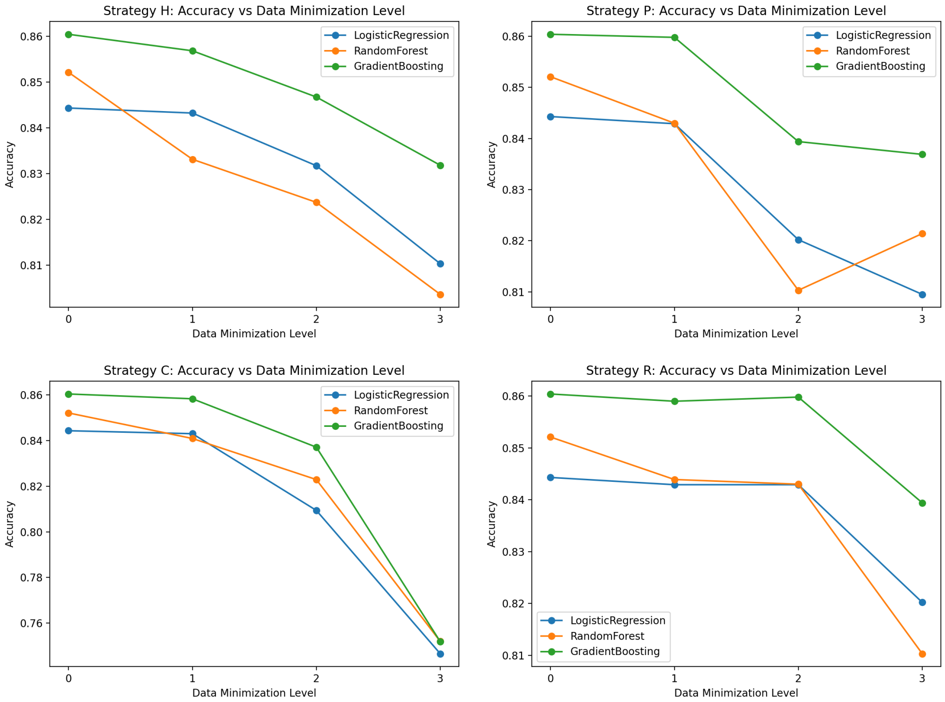

Figure 2: Accuracy trade-offs across data minimization levels for each strategy. Each subplot reports performance for Logistic Regression, Random Forest, and Gradient Boosting models.

A.3 Additional Observations

Several observations can be made from the supplemental results.

First, aggressive data minimization under the Conservative strategy leads to near-collapse of recall, resulting in severe F1-score degradation despite moderate accuracy values. This highlights that accuracy alone may obscure critical failures in minority-class detection.

Second, the Robust strategy consistently achieves the smallest relative F1 degradation across all models, suggesting that selective removal of high-risk features can preserve utility even under strong data minimization.

Finally, differences between models become more pronounced at higher minimization levels, indicating that model complexity interacts non-trivially with feature availability.

A.4 Reproducibility Notes

All experiments are conducted using a fixed train-test split with a fixed random seed. No hyperparameter tuning is performed across minimization levels to isolate the effect of feature removal. The complete experimental pipeline, including feature removal rules and evaluation metrics, is described in the main paper and can be directly reproduced using the provided scripts.

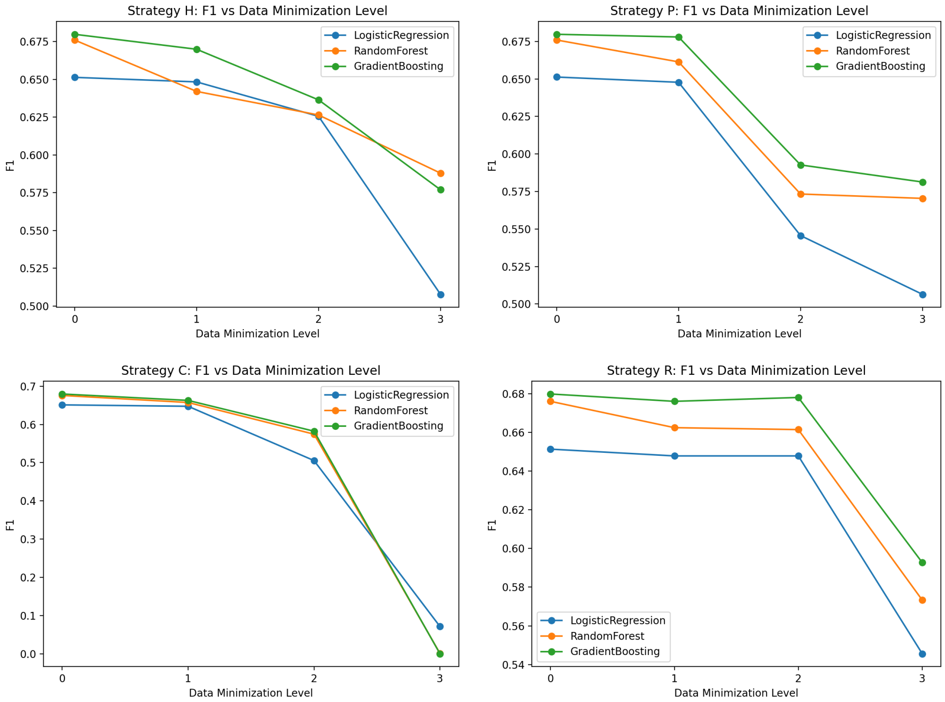

Figure 3: F1-score trade-offs across data minimization levels for each strategy. F1-score reveals substantial differences in recall sensitivity that are less visible when only accuracy is considered.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:The abstract and introduction accurately reflect the scope and contributions of this paper by clearly stating that the goal is an empirical evaluation of feature-level data minimization strategies and their impact on model performance, without overstating generalizability beyond benchmark settings (Sections 1 and 5).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:The paper includes a dedicated Limitations section that explicitly discusses dataset bias, feature-cardinality mismatch across strategies, and the absence of complementary privacy mechanisms, along with their implications and directions for future work (Section 7).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer:[N/A] .

Justification:This paper does not propose theoretical guarantees or formal proofs; its contributions are empirical and experimental in nature.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer:[Yes]

Justification:The paper fully specifies the dataset, preprocessing pipeline, data splits, minimization strategies, models, evaluation metrics, and experimental protocol, enabling independent reproduction of the reported results (Section 4).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All experiments are conducted on the publicly available UCI Adult dataset, and the paper provides sufficient methodological detail to allow faithful reimplementation of the experimental pipeline, even though executable code is not released at submission time.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies data splits, preprocessing, fixed hyperparameter usage, evaluation metrics, and model selection rationale in the Experimental Setup section (Section 4).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[N/A]

   Justification:The experiments are designed as controlled, single-split comparisons to isolate the effect of feature-level data minimization, rather than to estimate statistical variance across repeated trials; therefore, statistical significance measures are not applicable in this setting.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification:All experiments are conducted using standard machine learning models on a moderate-sized public dataset, requiring only commodity CPU resources, making reproduction feasible without specialized hardware.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

Answer:[Yes]

Justification:The study uses publicly available, non-sensitive benchmark data and does not involve human subjects, personal data collection, or deployment-related risks.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer:[Yes]

Justification:The paper discusses how data minimization can support regulatory compliance and privacy protection while also acknowledging potential risks such as performance degradation and fairness implications (Sections 6 and 7).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer:[N/A]

    Justification: The paper does not introduce or release models or datasets that pose a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[Yes]

    Justification: The UCI Adult dataset is properly cited and acknowledged as a publicly available benchmark dataset with established usage in prior research.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A]

    Justification: This paper does not introduce new datasets, models, or code assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[N/A]

    Justification: The study does not involve crowdsourcing or human subject experiments.

    Guidelines:The study does not involve crowdsourcing or human subject experiments.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved, and therefore IRB approval is not required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

[1] Siddiqi Almasi, Mohammed, and Hemmati. Optimization of privacy–utility trade-off for efficient feature selection of secure internet of things. *Journal of Information Security and Applications*, 2024.

[2] Frances Ding, Moritz Hardt, and John Miller. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 2021.

[3] Dheeru Dua and Casey Graff. Uci machine learning repository: Adult data set. `https://archive.ics.uci.edu/ml/datasets/adult`, 2019.

[4] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[5] European Union. Regulation (eu) 2016/679 of the european parliament and of the council. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2016. General Data Protection Regulation.

[6] Ruocheng Feng, Yiming Yang, and Markus Götz. Learning without sensitive attributes: Fairness implications revisited. *Neurocomputing*, 2023.

[7] Pranav Ganesh, Cuong Tran, Reza Shokri, and Ferdinando Fioretto. The data minimization principle in machine learning. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025.

[8] Aris Gkoulalas-Divanis and Tamir Tassa. Assessing re-identification risk via quasi-identifiers. *Designs, Codes and Cryptography*, 2024.

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016.

[10] Robin Staab, Anna Heinemann, Lena Ziegler, and Georgios Kaissis. Data minimization in machine learning: A systematization of knowledge. *ACM Computing Surveys*, 2024.

[11] Robin Staab, Nikola Jovanović, Miloš Balunović, and Martin Vechev. From principle to practice: Vertical data minimization for machine learning. In *IEEE Symposium on Security and Privacy (S&P)*, 2024.

[12] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.