International Journal of Engineering Research and Modern Education (IJERME)
International Peer Reviewed - Refereed Research Journal, Website: www.crystalpen.in
Impact Factor: 7.018, ISSN (Online): 2455 - 4200, Volume 10, Issue 2, July - December, 2025

GRADIENT EXPOSURE CONDITIONS AND TEXT RECONSTRUCTION VULNERABILITY IN FEDERATED FOUNDATION MODELS ACROSS PUBLISHED BENCHMARK EXPERIMENTS

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Gradient Exposure Conditions and Text Reconstruction Vulnerability in Federated Foundation Models Across Published Benchmark Experiments", International Journal of Engineering Research and Modern Education, Volume 10, Issue 2, July - December, Page Number 176-183, 2025.

Copy Right: © Crystal Pen Publication, 2025 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI: https://doi.org/10.5281/zenodo.24049639

Abstract:

This analysis examines whether privacy vulnerability rankings remain consistent across text reconstruction procedures, informing international debates on federated learning security. A descriptive secondary analysis uses 72 published benchmark conditions from 2024, covering three attacks, two transformer configurations, three text benchmarks and four batch sizes. Matched comparisons hold the remaining experimental dimensions constant. DAGER exceeds TAG and LAMP in all 24 matched conditions, with average advantages of 61.62 and 49.79 ROUGE 1 score points, respectively. LAMP exceeds TAG in 23 conditions. Increasing batch size from one to eight coincides with lower reconstruction throughout the TAG and LAMP comparisons, whereas DAGER retains ceiling performance on GPT 2 and deteriorates on BERT. Model rankings consequently reverse across attacks. These patterns indicate that aggregation constrains recoverability conditionally rather than providing uniform protection. The contribution is a comparative synthesis showing that observed vulnerability depends on reconstruction capability and model configuration, without establishing causal mechanisms. Organizations should evaluate multiple attacks under matched conditions and distinguish unigram recovery from comprehensive disclosure when interpreting privacy claims.

Key Words: Federated Learning; Gradient Inversion; Text Reconstruction; Privacy Evaluation; Matched Benchmark Comparisons.

1. Introduction:

Keeping training text on a client device does not ensure that its contents remain private. In federated learning, clients transmit model updates that support collaborative training, but those updates can retain information about the records used to compute them. The substantive problem is therefore not only where data are stored, but what an observer can reconstruct from the information exchanged during learning. Kairouz et al., 2021, distinguish decentralized data management from formal privacy protection and identify information leakage as a central challenge. For organizations evaluating collaborative language models, this distinction makes reconstruction assessment consequential: low recovery under one attack may reflect the limitations of that procedure rather than protection inherent in the model. Text reconstruction from gradients is an inverse problem. The client maps private text into a model update, while the attacker searches for inputs compatible with that update. Reconstruction depends on the information retained in the gradient and the procedure used to recover it. Deng et al., 2021, introduce TAG and demonstrate that reconstruction performance varies with transformer characteristics, vocabulary size and input length. Balunović et al., 2022, develop LAMP, combining gradient matching with language model information and alternating continuous and discrete optimization. These approaches establish that gradient visibility can permit textual recovery, while also showing that recovery depends on how the reconstruction search is organized.

Batch size and model architecture introduce competing constraints into this inverse problem. Combining several sequences in an update can make individual contributions harder to separate. However, mixing does not necessarily eliminate recoverable structure. Petrov et al., 2024, demonstrate that DAGER exploits the low rank structure of attention gradients and discrete token embeddings to reconstruct text. Their approach also distinguishes reconstruction procedures for encoder and decoder architectures. Consequently, a larger batch may impede one attack while leaving another effective, and a configuration that appears less vulnerable under one procedure may appear more vulnerable under another. These are conditional relationships between experimental dimensions, rather than evidence that any single dimension determines privacy independently. The closest evidence already establishes substantial differences among TAG, LAMP and DAGER. The present analysis therefore does not claim to discover gradient leakage, introduce a new attack or provide an independent experimental replication. Its contribution is a structured secondary comparison of the published benchmark grid. It makes three features explicit: the consistency of attack rankings across matched conditions, the variation in batch size patterns across models and attacks, and the reversal of model rankings between reconstruction procedures. This focus qualifies interpretations based on pooled performance summaries by examining whether an average advantage persists across the conditions from which it is calculated.

Measurement determines the scope of that contribution. Reconstruction vulnerability is operationalized as the published mean ROUGE 1 score, expressed on a scale from zero to 100. This measure captures unigram overlap between reconstructed and original text. It does not directly establish correct word order, complete document recovery, personal identification or operational harm. Retaining one outcome across matched conditions enables consistent comparison, while preserving these distinctions prevents overlap scores from being interpreted as comprehensive privacy measures. Architecture is likewise treated cautiously because the available encoder and decoder categories are represented by BERT and GPT 2 respectively. Their differences describe observed model configurations rather than an isolated architectural effect. This study aims to determine how reported text reconstruction vulnerability varies across attack methods, batch sizes and model configurations within matched benchmark conditions. The central research question is whether reconstruction rankings and batch size patterns remain consistent when the

other observed dimensions are held constant. The specific objectives are to compare TAG, LAMP and DAGER within identical model, benchmark and batch size combinations; examine reconstruction changes across batch sizes within each attack, model and benchmark combination; and compare BERT with GPT 2 under matched attacks, benchmarks and batch sizes. These objectives require descriptive comparisons rather than causal estimation or an additional moderator.

The empirical analysis consolidates 72 experimental condition means from Table 2 of Petrov et al., 2024. The sample covers three attacks, two model configurations, three benchmarks and batch sizes of one, two, four and eight. All observations were published in 2024, within the eligibility period from 2010 to 2024. The dataset is a cross section of experimental conditions rather than an annual panel. Its balanced composition supports comparisons without changes in benchmark representation across attacks or batch sizes. Published confidence margins describe uncertainty around individual condition means, while the absence of batch level observations and covariance information limits inference about matched differences. The results show that DAGER exceeds both alternatives in every matched condition, whereas LAMP exceeds TAG in most conditions but includes one reversal. Larger endpoint batches coincide with lower reconstruction throughout the TAG and LAMP comparisons, while DAGER's deterioration is concentrated in BERT. Model rankings also reverse: BERT has higher reconstruction under TAG and LAMP, whereas GPT 2 has higher or equal reconstruction under DAGER. Together, these findings refine the interpretation of vulnerability as conditional on the reconstruction procedure and model configuration, rather than a stable ranking that can be transferred across attacks. The practical contribution concerns evaluation design. Organizations can obtain a more defensible assessment by comparing several attacks under matched conditions and retaining the settings in which conclusions change. Transferability rests on this comparison principle, not on assuming that the observed scores apply to other model families or deployments. Protocol protections must also remain distinct from reconstruction performance under exposed gradients. Bonawitz et al., 2017, demonstrate that secure aggregation changes the server's access to individual contributions. The present analysis does not evaluate that protection; it clarifies what published reconstruction evidence establishes when gradients are observable.

2. Theory Conceptual Framework and Research Questions Development:

Federated learning creates an informational dependency between clients and a coordinating server. Clients retain their training records but transmit updates that enable collective model development. These updates depend on the underlying records, so keeping data on client devices does not necessarily prevent reconstruction. Kairouz et al., 2021, distinguish decentralized data storage from formal privacy protection and identify information leakage as a central challenge in federated learning. This analysis examines that distinction through published text reconstruction outcomes under different attack methods, batch sizes and model configurations. The conceptual mechanism is an inverse problem. A client computes gradients from private text, while an attacker uses the observed gradients and known model to search for compatible inputs. Reconstruction depends on both the information retained in the update and the attack's ability to extract it. Gradient exposure conditions therefore describe three observed experimental dimensions: attack method, batch size and model architecture. Text reconstruction vulnerability is operationalized as the published mean ROUGE 1 score on a scale from zero to 100. Higher scores indicate greater unigram overlap with the original text, rather than a higher proportion of completely recovered documents.

Attack method represents the procedure used to transform exposed gradients into reconstructed text. TAG searches for inputs whose gradients approximate the observed update. Its effectiveness depends on the optimization procedure and the model characteristics that shape the reconstruction problem. Deng et al., 2021, demonstrate that reconstruction performance varies across transformer configurations, vocabulary sizes and input lengths. These findings establish that attack performance is conditional on the experimental setting rather than a fixed property of gradient access. Language information provides a competing route to reconstruction. LAMP combines gradient matching with an auxiliary language model and alternating continuous and discrete optimization. The language model guides the search toward plausible text, potentially reducing ambiguity among candidate inputs. Balunović et al., 2022, report improved textual reconstruction relative to earlier approaches. Nevertheless, linguistic plausibility and agreement with the exposed gradient remain distinct criteria. The theoretical expectation is that attack design changes how effectively available information becomes recoverable text, while the magnitude and consistency of that difference require comparison under matched benchmark conditions.

RQ1: How does mean ROUGE 1 differ among TAG, LAMP and DAGER when benchmark, model and batch size are held constant?

Batch size changes the number of sequences contributing to an observed gradient. Combining more sequences can make their individual contributions harder to separate, increasing reconstruction ambiguity. This mechanism suggests declining reconstruction as batch size increases. However, aggregation does not necessarily remove the algebraic structure that an attack can exploit. Petrov et al., 2024, show that DAGER uses the low rank structure of attention gradients and discrete token embeddings to recover text. Their results establish that larger batches can remain vulnerable under suitable model and input conditions. The competing expectations are therefore deterioration through greater mixing and persistence through exploitable gradient structure.

This analysis evaluates those expectations across batch sizes of one, two, four and eight sequences. Comparisons remain within each attack, model and benchmark combination because pooling unlike conditions could obscure different reconstruction patterns. Scores near the upper measurement boundary can also remain unchanged across batch sizes, so a uniform negative relationship is not imposed.

RQ2: How does mean ROUGE 1 change across batch sizes of one, two, four and eight within each attack, model and benchmark combination?

Model architecture shapes the constraints available during reconstruction. Decoder attention restricts how positions depend on subsequent tokens, whereas encoder attention permits information exchange across the input sequence. These differences can alter how an attack identifies candidate tokens and reconstructs their ordering. However, the available comparison contains one decoder model, GPT 2, and one encoder model, BERT. Architecture and model identity therefore vary together. The analysis compares the two observed configurations without attributing their differences exclusively to architecture.

RQ3: How does mean ROUGE 1 differ between the GPT 2 decoder configuration and the BERT encoder configuration under matched attack methods, batch sizes and benchmarks?

The framework connects the three observed dimensions to reconstruction vulnerability through differences in information separation and computational recoverability. CoLA, SST 2 and Rotten Tomatoes define comparison strata. The dimensions remain separate observed variables rather than indicators of a latent construct or components of a composite exposure index. Their relationships with ROUGE 1 represent descriptive associations within published benchmark conditions. Gradient visibility also defines the framework's boundary. Bonawitz et al., 2017, demonstrate a secure aggregation protocol that allows a server to compute an aggregate without observing individual client contributions. Such protection changes the information available to an attacker. The present questions concern reconstruction from the exposed gradients represented in the dataset and do not evaluate secure aggregation.

The three research questions are exploratory because they are formulated after the published benchmark outcomes became available. They can be resolved through matched descriptive comparisons of the 72 reported conditions. Similarity, reversal and variation across benchmarks remain valid outcomes. The framework does not treat benchmark condition means as independent client observations or interpret their differences as causal effects.

3. Data:

This analysis uses published benchmark results to document data provenance, define the analytical sample and operationalize text reconstruction vulnerability. The resulting dataset supports descriptive comparisons across attack methods, batch sizes and model configurations.

3.1 Data Sources, Population, and Sample Construction:

The empirical source is Table 2 of DAGER: Exact Gradient Inversion for Large Language Models, published by Petrov et al., 2024, in Advances in Neural Information Processing Systems, volume 37. The source reports reconstruction outcomes for TAG, LAMP and DAGER across two models, three text benchmarks and four batch sizes. The proceedings version provides the authoritative numerical record, while the accompanying Excel workbook consolidates the extracted observations. The unit of analysis is a published experimental condition defined by model, benchmark, attack method and batch size. The analytical sample contains 72 conditions published in 2024, within the eligibility period from 2010 to 2024. Publication year identifies the source date rather than the date of experimental execution. The data constitute a cross section of benchmark conditions without a calendar observation frequency. Each outcome is an experimental mean rather than a flow, stock or annual average.

The empirical setting is text reconstruction from exposed gradients in transformer language models. Coverage comprises the GPT 2 decoder configuration and the BERT encoder configuration evaluated on CoLA, SST 2 and Rotten Tomatoes. These are computational benchmarks rather than geographical sampling units. Consequently, the dataset does not represent countries, institutions, industrial sectors or individual users. The conceptual target population consists of text reconstruction conditions involving exposed gradients in transformer models. The accessible sampling frame is narrower: the conditions reported in Table 2 of the source publication. Selection is purposive at the source level and exhaustive within that table. All combinations of the two models, three benchmarks, three attacks and batch sizes of one, two, four and eight are retained. This selection defines a secondary benchmark analysis rather than a representative sample of all gradient inversion experiments.

Table 1: Sample construction

Stage	Distinct Models	Experimental Conditions	Inclusion Rule or Verified Outcome
Accessible published frame	2	72	All conditions in source Table 2
Extracted records	2	72	One record per model, benchmark, attack and batch size
Duplicate exclusions	0 removed	0 removed	No duplicate condition keys
Missing value exclusions	0 removed	0 removed	Required analytical fields complete
Retained descriptive sample	2	72	All extracted conditions retained

The design is balanced across its experimental dimensions. Each benchmark and attack contributes 24 conditions, each model contributes 36 conditions, and each batch size contributes 18 conditions. These counts describe benchmark cells rather than independent clients or reconstructed documents. The workbook does not record an extraction date or a preregistered selection protocol. Public access to the source establishes retrievability but does not establish unrestricted redistribution rights for the underlying benchmark corpora. The analysis uses published aggregate results rather than redistributing those corpora.

3.2 Data Integration, Cleaning, and Quality Assurance:

The dataset combines results from one published table, so no external dataset merge or temporal matching is required. The composite identifier comprises model, benchmark, attack method and batch size. An audit of the workbook confirms 72 unique composite identifiers, with no duplicate conditions or increases in observation counts from integration. Repeated scores remain valid observations when their experimental identifiers differ. Model labels are mapped to the observed architecture categories: GPT 2 to decoder and BERT to encoder. Original benchmark and attack labels retain their substantive meaning. Publication year is constant across all records. No frequency conversion, currency adjustment, inflation correction or calendar alignment is applied. All 72 records contain values for the outcome, experimental dimensions, published confidence interval margins and source location. No deletion, imputation, interpolation or replacement of missing values with zero is required. Zero confidence interval margins are retained as reported values rather than interpreted as missing observations.

Quality checks assess identifier uniqueness, field completeness, numerical bounds and interval consistency. Every reconstruction mean lies between zero and 100. Every interval margin is nonnegative, and every calculated interval contains its corresponding mean. Recalculated bounds agree with the workbook values within numerical precision. No invalid values are identified by these checks. The source PDF takes priority over automatically extracted web text where the two conflict. For the GPT 2 configuration on CoLA under TAG with batch size one, the PDF reports a mean ROUGE 1 score of 7.0, whereas an automated web rendering displays 0.4. The workbook retains 7.0. No trimming, winsorization or outlier exclusion is applied. Scores at the upper boundary of 100 are valid published outcomes and remain unchanged. This preserves the reported distribution,

including ceiling values. Retention of every eligible condition prevents additional selection through cleaning, although it does not remove selection associated with reliance on one publication.

3.3 Variable Measurement, Construction, and Validation:

Text reconstruction vulnerability is measured by the published mean ROUGE 1 score on a scale from zero to 100. The measure captures unigram overlap between reconstructed and original text. It does not directly measure recovery of word order, complete documents or personal identities. The workbook preserves the source scale without normalization or conversion into a composite index. Attack method is a categorical predictor distinguishing TAG, LAMP and DAGER. These categories identify different reconstruction procedures rather than ordered levels of attack strength. TAG uses gradient matching to reconstruct text, as documented by Deng et al., 2021. LAMP incorporates language model information and alternating optimization, as described by Balunović et al., 2022. The analysis uses these labels to compare reported outcomes without imposing an ordinal coding. Batch size records the number of sequences contributing to a batch. Its observed values are one, two, four and eight. Architecture distinguishes the encoder and decoder configurations, while model identity retains the corresponding BERT and GPT 2 labels. Because each architecture is represented by one model, architecture and model identity cannot be separated empirically. Benchmark identifies the corpus within which comparisons are made.

Table 2: Variable dictionary

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Mean ROUGE 1	Outcome	Published mean unigram overlap score	Petrov et al., 2024, Table 2, R1 columns	Score from zero to 100 per condition	Untransformed; higher values indicate greater reconstruction	No missing values	72 conditions; does not measure complete document recovery
Attack method	Predictor	TAG, LAMP or DAGER	Source Table 2, attack labels	Categorical per condition	No ordinal ranking imposed	No missing values	24 conditions per attack
Batch size	Predictor	Number of sequences in the evaluated batch	Source Table 2, batch columns	Sequences per condition	Values of one, two, four and eight	No missing values	18 conditions per value
Model architecture	Predictor	Encoder or decoder mapped from model identity	Source Table 2 and model descriptions	Categorical per condition	No numerical transformation	No missing values	36 conditions per category; inseparable from model identity
Model identity	Matching identifier	BERT or GPT 2	Source Table 2, model labels	Categorical per condition	Original model distinction retained	No missing values	Two configurations
Benchmark	Comparison stratum	CoLA, SST 2 or Rotten Tomatoes	Source Table 2, dataset labels	Categorical per condition	Comparisons retain benchmark identity	No missing values	24 conditions per benchmark
Lower confidence margin	Uncertainty descriptor	Published distance below the mean	Source Table 2, reported interval	Score points per condition	Retained without symmetrization	No missing values	72 conditions
Upper confidence margin	Uncertainty descriptor	Published distance above the mean	Source Table 2, reported interval	Score points per condition	Retained without symmetrization	No missing values	72 conditions
Confidence interval bounds	Calculated descriptors	Mean less lower margin and mean plus upper margin	Workbook formulas using source values	Score points per condition	Original asymmetric margins preserved	Complete inputs required	72 conditions; reflects source uncertainty
Publication year	Provenance field	Year of source publication	Proceedings record	Year per condition	Constant at 2024	No missing values	Does not identify experimental execution dates
Condition identifier	Record identifier	Constructed unique label for each condition	Workbook	Identifier per record	No substantive numerical interpretation	Complete	72 unique labels

Validation concerns the accuracy and consistency of observed fields rather than latent construct reliability. The audit checks category membership, complete experimental combinations, valid score ranges and confidence interval arithmetic. No factor analysis or reliability coefficient is required because the predictors are observed experimental categories and the outcome is a published metric.

Table 3: Descriptive statistics for numerical analytical variables

Variable	Valid Count	Mean	Standard Deviation	Median	Minimum	Maximum
Mean ROUGE 1	72	55.87	36.49	61.60	4.40	100.00
Batch size	72	3.75	2.70	3.00	1.00	8.00
Lower confidence margin	72	1.95	1.55	1.85	0.00	6.90
Upper confidence margin	72	1.91	1.57	1.75	0.00	6.90

Standard deviations in Table 3 describe variation across published condition values. They are distinct from the uncertainty margins reported for individual conditions. The pooled mean summarizes the selected experimental grid and does not estimate average privacy vulnerability across deployed language models.

3.4 Analytical Framework, Identification, and Reproducibility:

The analytical framework is descriptive. It compares mean reconstruction scores across attack methods, batch sizes and the two observed model configurations while preserving benchmark identity. This interpretation follows the distinction between decentralized training and privacy protection emphasized by Kairouz et al., 2021. The dataset measures reconstruction under specified gradient exposure conditions rather than the privacy guarantees of federated learning as a whole. The comparison structure holds model, benchmark and batch size constant when examining attack differences. It holds attack, model and benchmark constant when examining batch size patterns. Model comparisons match attack, benchmark and batch size. These

comparisons organize the available observations without introducing additional controls, estimated latent variables or causal parameters. The workbook contains aggregate condition means and published confidence intervals rather than batch level outcomes. It does not provide covariance between matched conditions. Consequently, confidence intervals describe uncertainty reported by the source and are not treated as sufficient inputs for significance tests of between condition differences. No regression coefficients, clustering procedure or causal identification strategy is claimed.

The source reports averaging reconstruction outcomes over 100 batches per condition. This replication does not convert the 72 aggregate records into independently observed client records. Architecture comparisons also remain configuration comparisons because model identity changes with architecture. External validity is limited to the represented gradient exposure setting. Secure aggregation changes what the coordinating server can observe, as demonstrated by Bonawitz et al., 2017. Its effectiveness is not measured by this dataset, so the descriptive results cannot establish outcomes under that protocol. Reproducibility rests on the source table identifier, condition identifiers, explicit variable definitions, retained confidence margins and visible interval formulas. The audit confirms complete coverage of the selected experimental grid, unique records and consistent numerical construction. The workbook supplies the consolidated observations and measurement documentation. Its aggregate structure supports transparent descriptive comparison while limiting inference about individual clients, other model families and operational deployments.

4. Methods:

4.1 Research Design and Comparison Strategy:

This analysis uses a descriptive secondary research design to examine reconstruction differences across published experimental conditions. The unit of analysis is one combination of attack method, model configuration, benchmark and batch size. The outcome is the reported mean ROUGE 1 score. The sample construction and variable definitions appear in Tables 1 and 2. The analytical targets are differences between reported condition means within the selected experimental grid. Attack comparisons match model, benchmark and batch size. Batch size comparisons retain the same attack, model and benchmark. Model comparisons match attack, benchmark and batch size. This structure prevents changes in the composition of the observed comparison dimensions from driving the contrasts. The design addresses exploratory research questions rather than prespecified directional hypotheses. It evaluates the consistency, magnitude and exceptions of observed differences without estimating causal effects. Matching aligns the recorded experimental dimensions but does not establish equivalence in unrecorded implementation characteristics. Because BERT represents the encoder category and GPT 2 represents the decoder category, model comparisons cannot isolate architecture from model identity.

4.2 Descriptive Specification and Measurement Alignment:

Let y_{ambs} denote the published mean reconstruction score for attack a, model m, benchmark b and batch size s. The observed dimensions contain three attacks, two models, three benchmarks and four batch sizes. Scores retain their original scale from zero to 100. For each attack, model and batch size combination, the benchmark averaged score is calculated as

$$\bar{y}_{ams} = \frac{1}{3}\sum_{b=1}^{3} y_{ambs} .$$

Here, $\bar{y}_{ams}$ is the arithmetic mean across the three benchmarks, and $y_{ambs}$is the corresponding published condition mean. Equation 1 gives each benchmark equal weight. These averages summarize the selected grid rather than an underlying population of clients or documents. Matched attack differences are calculated as

$$\Delta_{a,c|mbs} = y_{ambs} - y_{cmbs}.$$

Here, $\Delta_{a,c|mbs}$ is the difference between attacks a and c under the same model, benchmark and batch size. Positive values favour attack a. Each attack comparison contains 24 matched pairs. The analysis summarizes their mean difference, minimum, maximum and number of positive differences. Batch endpoint changes are calculated as

$$G_{amb} = y_{amb8} - y_{amb1}.$$

Here, $G_{amb}$is the change from batch size one to eight within a fixed attack, model and benchmark. Negative values indicate lower reconstruction at the larger endpoint. Each attack contributes six endpoint comparisons. Intermediate values at batch sizes two and four are also inspected to distinguish endpoint reductions from consistently nonincreasing sequences. Matched model differences are calculated as

$$D_{abs} = y_{a,GPT2,b,s} - y_{a,BERT,b,s}.$$

Here, D_{abs} is the difference between GPT 2 and BERT under the same attack, benchmark and batch size. Positive values favour GPT 2. Each attack contributes 12 matched model comparisons. All contrasts are expressed in score points, not percentage changes. No normalization, composite index, latent construct or regression specification is introduced. ROUGE 1 remains a measure of unigram overlap rather than complete sequence recovery.

4.3 Computation and Quality Assurance:

The analysis uses all 72 retained conditions without additional sample restrictions. Numerical summaries comprise valid counts, arithmetic means, sample standard deviations, medians and observed ranges. Categorical summaries report the number of conditions within each experimental category. Standard deviations describe dispersion across condition means and remain distinct from the uncertainty reported for individual experiments. Calculations were performed in Python using the standard library modules zipfile and xml.etree. Element Tree to read workbook values, statistics to calculate numerical summaries, and collections to tabulate category counts. Matching used the composite key comprising benchmark, attack, model and batch size. Calculations retained numerical precision until presentation, when reported summaries were rounded to two decimal places. The procedures are deterministic and require no random seed or optimization settings.

Quality assurance checked the uniqueness of composite keys, completeness of required fields, expected category coverage and numerical consistency. Reconstruction scores and interval endpoints were checked against the permissible scale. Confidence margins were checked for nonnegative values, and calculated endpoints were compared with the corresponding mean

less the lower margin and mean plus the upper margin. No records required deletion or imputation. Valid extreme scores, including values of 100, were retained. No trimming, winsorization, interpolation or transformation was applied. Where automatically extracted web text conflicted with the published table, the source PDF supplied the retained value. This rule preserves the published numerical record rather than treating a transcription discrepancy as an empirical outlier. The analysis examined whether aggregate rankings persisted across every matched condition. It counted reversals and ties rather than treating an average advantage as universal superiority. Likewise, inspection of all four batch sizes distinguished monotonic patterns from endpoint changes. These procedures assess descriptive consistency within the available sample.

4.4 Uncertainty, Interpretation and Replicability:

The source experiments report reconstruction means over 100 batches and provide confidence margins using their stated uncertainty procedures, as documented by Petrov et al., 2024. The secondary analysis retains those margins, including asymmetric values, and reconstructs their endpoints without imposing symmetry. The workbook contains aggregate condition means rather than batch level observations. It does not provide covariance between matched outcomes. Consequently, published intervals are reported as uncertainty descriptors for individual conditions and are not converted into confidence intervals or significance tests for matched differences. The analysis does not assume that the 72 condition means are independent participant observations. Equal weighting defines the descriptive target explicitly. Each condition contributes equally to its relevant summary, and every matched comparison uses the same recorded comparison dimensions. Reported differences therefore describe the selected experimental grid. They do not estimate prevalence of disclosure across deployed systems or establish a causal mechanism. The consolidated workbook documents condition identifiers, source locations, measurement definitions, published interval margins and calculated bounds. Equations 1 to 4 specify the contrasts underlying the reported findings. Together, explicit matching, complete retention of eligible conditions and transparent treatment of uncertainty permit replication of the descriptive comparisons. The methodological contribution is the consistent organization and interpretation of published evidence rather than a new estimator or identification strategy.

5. Findings:

This analysis examines text reconstruction vulnerability across attack methods, batch sizes and model configurations. The findings address three research questions through descriptive summaries and matched comparisons of published benchmark conditions. The evidence identifies consistent differences alongside exceptions that qualify their interpretation. The analytical sample contains 72 unique experimental conditions published in 2024, covering three attacks, two models, three benchmarks and four batch sizes. All required analytical fields are complete, and no conditions were excluded. Each attack contributes 24 observations, each model contributes 36 observations, and each batch size contributes 18 observations. The observations are published condition means rather than individual clients or reconstructed documents. Mean ROUGE 1 across all conditions is 55.87, with a standard deviation of 36.49, a median of 61.60 and values ranging from 4.40 to 100.00. Fifteen conditions attain the upper boundary of 100.00. These ceiling observations affect interpretation because unchanged scores cannot distinguish improvements beyond the measurement boundary. ROUGE 1 measures unigram overlap rather than complete document recovery. The original experiments and reporting conventions are documented by Petrov et al., 2024.

Table 4: Reconstruction scores by attack method

Attack Method	Conditions	Mean ROUGE 1	Standard Deviation	Median	Minimum	Maximum
TAG	24	31.39	27.99	16.70	4.40	78.90
LAMP	24	43.22	29.14	40.75	6.30	89.60
DAGER	24	93.00	15.68	100.00	37.10	100.00

DAGER produces the highest average reconstruction score and exceeds both alternatives in every matched comparison. Holding model, benchmark and batch size constant, its mean advantage is 61.62 score points over TAG and 49.79 points over LAMP. The corresponding differences range from 11.70 to 95.60 points and from 9.80 to 93.20 points. The consistently positive differences establish that DAGER's overall ranking does not result from averaging across unlike experimental compositions.

Table 5: Matched differences between attack methods

Comparison	Matched Pairs	Mean Difference in Score Points	Minimum Difference	Maximum Difference	Positive Differences
DAGER relative to TAG	24	61.62	11.70	95.60	24
DAGER relative to LAMP	24	49.79	9.80	93.20	24
LAMP relative to TAG	24	11.83	Negative 2.80	66.30	23

LAMP exceeds TAG in 23 of 24 matched conditions, with an average advantage of 11.83 points. The exception occurs for BERT on Rotten Tomatoes at batch size two, where LAMP scores 46.40 and TAG scores 49.20. This exception qualifies a universal ranking between the two methods without overturning their average difference. The broader pattern is consistent with the reconstruction benefits associated with language model information described by Balunović et al., 2022, while retaining variation across experimental settings. RQ1 is therefore resolved descriptively: DAGER exceeds both alternatives throughout the matched sample, whereas LAMP generally exceeds TAG but does not dominate it in every condition. Batch size patterns differ across attacks and models. Table 6 averages the three benchmark scores within each attack, model and batch size combination. Equal benchmark representation ensures that changes across columns reflect the same benchmark composition.

Table 6: Mean reconstruction scores across batch sizes

Attack Method	Model Configuration	Batch Size 1	Batch Size 2	Batch Size 4	Batch Size 8
TAG	GPT 2 decoder	6.47	7.00	6.70	5.27
TAG	BERT encoder	71.47	63.30	50.33	40.57

Attack Method	Model Configuration	Batch Size 1	Batch Size 2	Batch Size 4	Batch Size 8
LAMP	GPT 2 decoder	55.63	19.80	9.83	7.93
LAMP	BERT encoder	81.03	68.87	56.93	45.70
DAGER	GPT 2 decoder	100.00	100.00	99.77	100.00
DAGER	BERT encoder	100.00	99.13	85.47	59.67

From batch size one to eight, reconstruction decreases in all six benchmark and model combinations for TAG and in all six for LAMP. The mean endpoint reductions are 16.05 points for TAG and 41.52 points for LAMP. However, intermediate scores do not decline uniformly. Four of six TAG sequences and five of six LAMP sequences are nonincreasing across all four batch sizes. The endpoint evidence therefore supports deterioration with larger batches while rejecting a strictly declining pattern in every setting. DAGER exhibits a different configuration specific pattern. For GPT 2, the benchmark averaged score remains at 100.00 at batch sizes one, two and eight, with a small reduction to 99.77 at batch size four. For BERT, the corresponding mean falls from 100.00 to 59.67, a reduction of 40.33 points. At batch size eight, BERT scores range from 37.10 on Rotten Tomatoes to 74.10 on SST 2. Larger batches therefore coincide with substantial deterioration for the BERT configuration but do not produce comparable deterioration for GPT 2. RQ2 is resolved as a conditional pattern: larger endpoint batch sizes coincide with lower reconstruction for TAG and LAMP throughout the matched sample, whereas DAGER's deterioration is concentrated in BERT. Batch aggregation does not provide a uniform reduction in measured vulnerability across attack methods and configurations. Model comparisons reveal a reversal in relative reconstruction performance. BERT exceeds GPT 2 in all 12 matched conditions under TAG and all 12 under LAMP. Its average advantage is 50.06 points under TAG and 39.83 points under LAMP. Under DAGER, GPT 2 exceeds BERT in eight conditions and ties it in four, producing an average advantage of 13.88 points.

Table 7: Matched reconstruction differences between model configurations

Attack Method	Matched Conditions	Higher Scoring Configuration	Mean Advantage in Score Points	Higher Scores	Ties
TAG	12	BERT encoder	50.06	12	0
LAMP	12	BERT encoder	39.83	12	0
DAGER	12	GPT 2 decoder	13.88	8	4

The reversal indicates that the observed vulnerability ranking depends on the reconstruction procedure. It does not support assigning either configuration a universally higher reconstruction risk. This conditional interpretation accords with the dependence of gradient attack performance on model characteristics documented by Deng et al., 2021. Because one model represents each architecture, the comparisons do not isolate architecture from model identity. RQ3 is therefore resolved at the configuration level: BERT has higher reconstruction scores under TAG and LAMP, while GPT 2 has higher or equal scores under DAGER. Published uncertainty margins qualify individual condition means without establishing uncertainty for their differences. For example, DAGER on BERT at batch size eight has a reported interval from 65.50 to 70.10 on CoLA, from 70.80 to 77.40 on SST 2 and from 35.90 to 38.30 on Rotten Tomatoes. The workbook does not contain the covariance needed to construct confidence intervals for matched differences. Consequently, the reported advantages describe observed magnitudes rather than statistically significant effects. The findings establish three conclusions within the selected experimental grid: DAGER consistently produces greater reconstruction than the alternatives, batch size patterns depend on attack and model configuration, and model rankings reverse across attacks. These results concern exposed gradients and do not establish general privacy guarantees for federated learning, a distinction emphasized by Kairouz et al., 2021. They also do not measure secure aggregation, which changes the server's access to individual contributions, as demonstrated by Bonawitz et al., 2017.

6. Discussion:

Privacy vulnerability cannot be assigned to a language model independently of the procedure used to reconstruct its training text. Tables 5 and 7 show that attack performance and model configuration jointly determine the observed reconstruction pattern. DAGER exceeds both alternatives in every matched condition, while the relative ranking of BERT and GPT 2 reverses across attacks. This reversal challenges the interpretation of vulnerability as a stable model characteristic. Deng et al., 2021, established that transformer characteristics influence reconstruction performance; the present comparison sharpens that interpretation by showing that the direction of the model difference itself depends on the attack. The contribution is a comparative synthesis of published conditions rather than a newly discovered attack capability. It demonstrates why conclusions drawn from one reconstruction procedure cannot automatically support a general ranking of model privacy. The batch size results refine the distinction between mixing information and preventing its recovery. Table 6 shows that larger endpoint batches coincide with lower reconstruction under TAG and LAMP, whereas DAGER maintains near complete unigram recovery on GPT 2. The corresponding deterioration on BERT indicates that aggregation difficulty depends on the interaction between the reconstruction procedure and the model configuration. These patterns are consistent with the algebraic account developed by Petrov et al., 2024, in which recoverability depends on exploitable gradient structure rather than batch size alone. They do not independently establish that mechanism because gradient rank and intermediate reconstruction processes are not measured. The analytical implication is nevertheless substantive: batch size should be interpreted as a conditional constraint on reconstruction, rather than a sufficient indicator of protection.

The matched attack comparisons also distinguish consistent superiority from average improvement. Table 5 shows that DAGER's advantage persists across every matched setting, whereas LAMP's average advantage over TAG includes one reversal. Balunović et al., 2022, explain how linguistic information can guide reconstruction toward plausible inputs. The exception observed here indicates that this guidance does not guarantee higher unigram recovery in every experimental condition. It therefore qualifies an interpretation that equates a more informative reconstruction procedure with universal empirical dominance. Table 4 further shows why pooled averages require scrutiny: the attacks differ substantially in both central performance and dispersion. The resulting contribution is a clearer distinction between an attack that performs better on average and one that

dominates throughout the selected comparison grid. The configuration reversal exposes a practical problem in privacy evaluation: a weaker attack can create an appearance of protection that disappears under a different reconstruction method. Table 7 illustrates this risk directly, since GPT 2 has lower reconstruction under TAG and LAMP but higher or equal reconstruction under DAGER. Evaluation should therefore retain matched comparisons across several attacks and report the conditions under which model rankings change. This implication complements the distinction between decentralized data storage and formal privacy protection discussed by Kairouz et al., 2021. It also requires measurement discipline. The ceiling values in Table 4 indicate extensive unigram recovery, but they do not establish complete sequence reconstruction, identity disclosure or operational harm. The evidence refines privacy assessment by linking conclusions to the attack, configuration and outcome actually evaluated.

The international relevance lies in the evaluation problem shared by organizations deploying collaborative language models, rather than in differences between national economies. Tables 5 to 7 indicate that a favourable result under one attack or batch size cannot establish protection across the represented conditions. Organizations can respond by comparing attacks within matched configurations, retaining benchmark specific results and separating reconstruction measurements from claims about protocol security. Bonawitz et al., 2017, demonstrate that secure aggregation changes the server's access to individual contributions, addressing a different layer of the privacy problem. The present evidence does not evaluate that protection. Its broader contribution is to distinguish vulnerability conditional on exposed gradients from protection that changes gradient visibility. This distinction provides a more precise basis for interpreting benchmark evidence across institutional settings without extending the results beyond the observed models and attacks.

7. Conclusion and Implications:

Privacy assessments can mislead when apparent protection depends on the reconstruction procedure used to evaluate it. This analysis shows that recoverability reflects the combination of computational search, information mixing and model structure. Increasing aggregation does not consistently suppress recovery, and the apparent vulnerability ranking of model configurations reverses across reconstruction procedures. The contribution is a comparative interpretation of published evidence that makes these conditional patterns explicit, rather than a new attack or causal model. The findings refine the understanding of privacy vulnerability as a relationship between exposed information and reconstruction capability, rather than a stable property of a model. For decision makers, this means that favourable performance under one evaluation cannot establish general protection. Risk assessments should compare multiple reconstruction procedures under matched conditions, preserve differences across evaluation settings and distinguish textual overlap from complete disclosure. Governance requirements can support this practice by requiring transparent reporting of evaluation coverage and the limits of protection claims. Operational teams should retain conditions that contradict average performance and reassess conclusions when reconstruction capabilities change. Across organizations handling sensitive text, these practices provide a stronger basis for deployment decisions and privacy communication. Their social value lies in reducing misplaced confidence while supporting scrutiny of collaborative learning systems. The evidence covers aggregate results from one publication, two model configurations and three benchmarks. Broader comparisons across independent experimental sources would assess whether the observed rankings persist. Because architecture and model identity coincide, comparisons across several models within each architectural family could separate their contributions. Access to batch level outcomes and covariance information would support uncertainty estimates for matched differences. Measuring gradient structure alongside reconstruction outcomes would permit direct examination of the proposed mechanisms. Extensions using additional languages, sensitive document types and protected communication protocols could establish where the comparative patterns transfer and where different exposure conditions require revised interpretations.

References:

1. Balunović, M., Dimitrov, D. I., Jovanović, N., & Vechev, M. (2022). LAMP: Extracting text from gradients with language model priors. Advances in Neural Information Processing Systems, 35, 7641-7654. https://doi.org/10.52202/068431-0555
2. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175-1191). Association for Computing Machinery. https://doi.org/10.1145/3133956.3133982
3. Deng, J., Wang, Y., Li, J., Wang, C., Shang, C., Liu, H., Rajasekaran, S., & Ding, C. (2021). TAG: Gradient attack on Transformer-based language models. In M. F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 3600-3610). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.305
4. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083
5. Petrov, I., Dimitrov, D. I., Baader, M., Müller, M. N., & Vechev, M. (2024). DAGER: Exact gradient inversion for large language models. Advances in Neural Information Processing Systems, 37, 87801-87830. https://doi.org/10.52202/079017-2787