2026 Volume: 02 Issue: 07 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-22

THE BLACK-BOX GRAVEYARD: A RESEARCH PROTOCOL FOR QUANTIFYING RESIDUAL ENTROPY OF DELETED TRAINING SAMPLES IN POST-UNLEARNING LARGE LANGUAGE MODELS

***Dr. Akhilesh Kumar**

Chief Technology Officer, Department of Information Technology, Kalra Hospital SRCNC Private Limited, New Delhi, India.

Article Received: 08 June 2026
Article Revised: 28 June 2026
Published on: 18 July 2026

***Corresponding Author: Dr. Akhilesh Kumar**
Chief Technology Officer, Department of Information Technology, Kalra Hospital SRCNC Private Limited, New Delhi, India.

DOI: https://doi-doi.org/101555/ijrpa.3858

ABSTRACT

Background

Large Language Models (LLMs) are increasingly used in healthcare, biomedical research, finance, and other data-intensive sectors. Their dependence on large-scale training datasets raises concerns regarding privacy because sensitive information may remain embedded within model parameters even after data deletion requests. Machine unlearning has emerged as a computational approach for removing the influence of selected training samples without requiring complete model retraining. However, current evaluation methods primarily focus on predictive performance and do not directly quantify the residual information retained after unlearning.

Objective

This research protocol proposes an information-theoretic framework to quantify residual knowledge remaining within post-unlearning LLMs using a novel Residual Entropy Score (RES).

Methods

The proposed study adopts a quantitative computational design using publicly available biomedical datasets and transformer-based language models. Representative machine unlearning algorithms will be evaluated under black-box conditions using utility metrics, privacy attacks, semantic similarity analysis, and information-theoretic measures. The proposed RES combines predictive entropy, embedding divergence, semantic similarity,

reconstruction probability, and membership inference confidence to estimate residual information after unlearning.

Expected Outcomes

The proposed framework is expected to provide a reproducible methodology for comparing machine unlearning algorithms and for quantifying hidden information that may remain encoded after selective deletion. The protocol may support future development of standardized evaluation procedures for privacy-preserving AI.

Conclusion

This manuscript presents a methodological framework rather than completed experimental research. If validated through future empirical studies, the proposed Residual Entropy Score may contribute to trustworthy artificial intelligence by improving the evaluation of machine unlearning in privacy-sensitive applications.

KEYWORDS: Large Language Models, Machine Unlearning, Residual Entropy, Privacy Preservation, Information Theory, Membership Inference, Black-Box Evaluation, Healthcare AI, Trustworthy AI, Data Deletion.

1. INTRODUCTION

Artificial Intelligence (AI) has rapidly evolved over the past decade, with Large Language Models (LLMs) becoming one of the most transformative developments in natural language processing and intelligent decision-support systems. Powered by transformer architectures and trained on massive textual corpora, LLMs are capable of performing a wide range of tasks, including language translation, scientific summarization, clinical documentation, biomedical question answering, software development, and knowledge synthesis. Their remarkable ability to understand context and generate coherent responses has accelerated their adoption across healthcare, finance, education, legal services, and cybersecurity.

In healthcare, LLMs are increasingly integrated into clinical workflows to assist with medical documentation, literature review, disease diagnosis support, patient communication, and biomedical research. These applications promise improvements in operational efficiency, clinical decision-making, and evidence-based practice. However, their performance depends heavily on exposure to vast amounts of training data that may include sensitive information such as electronic health records, clinical narratives, copyrighted publications, and proprietary institutional documents. Although developers generally employ data

anonymization and preprocessing techniques, the possibility that models retain fragments of confidential information remains a significant concern.

The growing deployment of foundation models has intensified discussions surrounding data privacy and regulatory compliance. Modern privacy regulations, including the General Data Protection Regulation (GDPR) and the Digital Personal Data Protection (DPDP) Act, recognize an individual's right to request the deletion of personal information from digital systems. While these regulations clearly define legal obligations for data controllers, implementing selective deletion within deep neural networks presents a substantial technical challenge. Unlike traditional databases, where information is stored explicitly and can be removed directly, LLMs encode knowledge across millions or billions of interconnected parameters. Consequently, deleting a single training sample does not automatically eliminate its statistical influence on model behaviour.

To address this challenge, researchers have introduced the concept of machine unlearning, which aims to remove the influence of selected training samples without requiring complete retraining of the model. Machine unlearning has emerged as a practical alternative because retraining large-scale language models is computationally expensive, time-consuming, and environmentally demanding. Current approaches include gradient-based optimization, parameter editing, knowledge distillation, influence-function estimation, and other approximate techniques that seek to reduce the impact of deleted information while preserving the model's overall utility.

Despite rapid progress in algorithm development, evaluating the effectiveness of machine unlearning remains an open research problem. Existing studies primarily assess performance using predictive accuracy, benchmark task performance, perplexity, or membership inference attacks. While these metrics provide useful information regarding observable model behaviour, they offer limited insight into whether deleted information continues to persist within latent neural representations. Emerging evidence suggests that models may continue to encode semantic traces of deleted samples even when conventional evaluation metrics indicate successful forgetting. Such hidden information may potentially be recovered through sophisticated prompt engineering, semantic reconstruction, or adversarial querying, creating important privacy implications for organizations deploying LLMs in sensitive domains.

This challenge motivates the central concept explored in the present protocol, referred to as the Black-Box Graveyard. The term describes the possibility that deleted information remains statistically embedded within a model despite appearing inaccessible through routine evaluation. Under this perspective, successful machine unlearning should not merely suppress

direct recall but should demonstrably reduce the residual informational influence of deleted samples. Quantifying this hidden information requires evaluation methods that extend beyond traditional performance metrics and incorporate principles from information theory, semantic analysis, and privacy assessment.

Accordingly, this manuscript proposes a Residual Entropy Score (RES) as a novel information-theoretic framework for estimating residual knowledge retained in post-unlearning LLMs. The proposed metric combines predictive entropy, semantic similarity, embedding divergence, reconstruction probability, and membership inference confidence into a unified quantitative measure. Unlike existing evaluation approaches that rely on isolated indicators, RES is intended to provide a multidimensional assessment of information retention under realistic black-box deployment conditions. The framework is designed to be model-agnostic, reproducible, and applicable across diverse transformer architectures.

The significance of this research extends beyond methodological innovation. A standardized framework for evaluating machine unlearning has the potential to strengthen privacy auditing, support regulatory compliance, and improve public confidence in trustworthy AI systems. In healthcare, where patient confidentiality is paramount, objective evidence that sensitive information has been effectively removed from deployed models could become an essential component of responsible AI governance. Although the present manuscript does not report experimental findings, it establishes a comprehensive research protocol that can guide future empirical investigations into privacy-preserving machine learning.

2. LITERATURE REVIEW

The rapid advancement of transformer-based Large Language Models (LLMs) has significantly influenced artificial intelligence research and its practical applications. Since the introduction of the Transformer architecture by Vaswani et al. (2017), language models have evolved from task-specific systems into foundation models capable of reasoning, summarization, question answering, code generation, and scientific knowledge synthesis. Modern LLMs, including GPT, LLaMA, Mistral, and BioMedLM, have demonstrated remarkable capabilities across domains such as healthcare, finance, education, and legal services. Their widespread adoption is primarily attributed to large-scale pre-training on diverse textual corpora that enable models to capture complex semantic and contextual relationships.

Despite these advances, the extensive data required for training LLMs has raised important concerns regarding privacy and data governance. Training datasets frequently contain

publicly available web documents, scientific publications, institutional reports, and, in some circumstances, sensitive personal information. Studies have demonstrated that language models may memorize portions of their training data and reproduce them when prompted appropriately, creating risks related to confidential information, copyrighted material, and personally identifiable data. Such memorization challenges existing privacy regulations that require organizations to remove personal information upon request.

To address these concerns, researchers have proposed machine unlearning, a paradigm that enables selective removal of the influence of designated training samples without requiring complete retraining of the model. Ginart et al. first formalized the concept of machine unlearning as an efficient alternative to retraining, while subsequent studies introduced practical techniques for removing information from deep learning models. Bourtoule et al. later proposed the SISA (Sharded, Isolated, Sliced, and Aggregated) framework, demonstrating that partitioning training data could substantially reduce the computational cost associated with selective deletion. Since then, numerous approaches have been developed, including gradient ascent, parameter editing, influence-function approximation, knowledge distillation, and low-rank adaptation.

Although these techniques differ in implementation, they share a common objective: minimizing the influence of deleted samples while preserving the predictive capability of retained knowledge. Exact retraining provides the strongest theoretical guarantee of forgetting but remains computationally impractical for billion-parameter language models. Approximate approaches offer substantially lower computational cost but may leave residual information distributed throughout the model's parameter space. Consequently, evaluating the effectiveness of machine unlearning has become an active area of research.

Existing evaluation methodologies primarily focus on utility preservation. Metrics such as classification accuracy, perplexity, BLEU, ROUGE, and downstream benchmark performance are commonly used to determine whether model quality remains stable after unlearning. However, these measures evaluate predictive performance rather than the persistence of deleted information. A model may exhibit excellent benchmark accuracy while continuing to encode semantic traces of deleted samples within latent representations.

To complement utility-based evaluation, researchers have introduced privacy-oriented attack methodologies, including membership inference attacks, model inversion attacks, and prompt extraction techniques. Membership inference determines whether a specific record was included in a model's training data by analyzing prediction behaviour, whereas model inversion seeks to reconstruct characteristics of original training samples. More recently,

adaptive prompt engineering has demonstrated that carefully designed prompts can encourage language models to reveal memorized information even when direct recall appears suppressed. These findings suggest that successful benchmark performance does not necessarily guarantee successful information removal.

Recent investigations have also highlighted the importance of information theory in understanding machine learning behaviour. Shannon's entropy provides a mathematical measure of uncertainty within probability distributions, while mutual information and Kullback–Leibler divergence quantify statistical dependence between variables. These concepts have been widely applied in communication theory, signal processing, and statistical learning but remain underutilized in evaluating machine unlearning. Most existing studies employ entropy only as a measure of predictive uncertainty rather than as an indicator of residual knowledge retained after selective deletion.

In parallel, the concept of trustworthy artificial intelligence has gained increasing attention from governments, regulatory agencies, and international organizations. Frameworks such as the European Union's General Data Protection Regulation (GDPR), the NIST Artificial Intelligence Risk Management Framework, and emerging national AI governance initiatives emphasize transparency, accountability, robustness, fairness, and privacy. However, these frameworks provide limited technical guidance regarding verification of successful machine unlearning in foundation models. Consequently, organizations currently lack standardized methods for demonstrating that deleted information has been effectively removed from deployed AI systems.

Several recent surveys have identified this limitation and emphasized the need for comprehensive benchmarking methodologies capable of evaluating both model utility and privacy preservation. Existing benchmarks often rely on isolated evaluation metrics that capture only a single dimension of model behaviour. Moreover, few studies simultaneously integrate semantic similarity, embedding analysis, reconstruction probability, predictive uncertainty, and adversarial privacy assessment within a unified evaluation framework.

Based on the reviewed literature, three important observations emerge. First, machine unlearning has become a rapidly expanding research area driven by increasing legal and ethical requirements for selective data deletion. Second, existing evaluation methodologies remain fragmented and frequently rely on indirect indicators that do not explicitly quantify residual information retained after unlearning. Third, despite growing interest in trustworthy AI, there is currently no widely accepted information-theoretic metric capable of estimating

hidden knowledge preserved within post-unlearning language models under realistic black-box conditions.

These limitations provide the primary motivation for the present research protocol. The proposed Residual Entropy Score (RES) seeks to address this gap by integrating predictive entropy, semantic similarity, embedding divergence, reconstruction probability, and membership inference confidence into a unified quantitative framework. By combining these complementary indicators, the proposed methodology aims to provide a more comprehensive assessment of machine unlearning than conventional evaluation techniques while supporting reproducible benchmarking and future standardization efforts.

3. RESEARCH GAP, AIM, OBJECTIVES AND RESEARCH QUESTIONS

3.1 Research Gap

The rapid development of machine unlearning has significantly improved the feasibility of selectively removing training samples from Large Language Models (LLMs). However, a critical limitation remains in the evaluation of these approaches. Most existing studies emphasize utility preservation, such as prediction accuracy, perplexity, or benchmark performance, while only a limited number investigate the extent to which deleted information continues to influence model behaviour after unlearning. Privacy-oriented evaluations, including membership inference and model inversion attacks, provide valuable insights but generally examine isolated aspects of information leakage rather than offering a comprehensive quantitative assessment of residual knowledge.

Furthermore, current evaluation frameworks seldom integrate information-theoretic principles with semantic analysis and adversarial privacy testing. As a result, there is no standardized metric capable of measuring the residual informational influence of deleted samples under realistic black-box deployment conditions, where internal model parameters are inaccessible. This limitation is particularly important in healthcare and other privacy-sensitive domains, where organizations must demonstrate compliance with evolving data protection regulations while maintaining the operational utility of deployed AI systems.

To address these challenges, the present research proposes a novel Residual Entropy Score (RES) that combines predictive uncertainty, semantic similarity, embedding divergence, reconstruction probability, and membership inference confidence into a unified evaluation framework. The proposed methodology is intended to provide a reproducible, model-agnostic approach for assessing the effectiveness of machine unlearning across diverse transformer-based language models.

3.2 Aim of the Study

The primary aim of this research is to develop a comprehensive methodological framework for quantifying residual information retained in post-unlearning Large Language Models through a novel Residual Entropy Score (RES). The framework seeks to establish a standardized approach for evaluating machine unlearning under realistic black-box conditions while balancing privacy preservation and model utility.

3.3 Research Objectives

The proposed study has the following objectives:

1. To develop an information-theoretic framework for measuring residual information retained after machine unlearning.
2. To formulate a composite Residual Entropy Score (RES) integrating multiple indicators of information retention.
3. To compare representative machine unlearning techniques using a standardized evaluation protocol.
4. To investigate the relationship between model utility and privacy preservation following selective data deletion.
5. To establish a reproducible benchmarking methodology that supports future evaluation and standardization of machine unlearning algorithms.

3.4 Research Questions

The proposed research seeks to address the following questions:

RQ1: Can residual information retained within post-unlearning Large Language Models be quantified using a unified information-theoretic framework?

RQ2: How effectively does the proposed Residual Entropy Score differentiate among various machine unlearning algorithms?

RQ3: What relationship exists between utility preservation and residual information after selective data deletion?

RQ4: Can black-box evaluation provide reliable evidence regarding the effectiveness of machine unlearning without requiring access to internal model parameters?

RQ5: How can standardized evaluation methodologies contribute to trustworthy AI and regulatory compliance in privacy-sensitive domains?

3.5 Research Hypotheses

To guide future empirical validation, the following hypotheses are proposed:

H1: Approximate machine unlearning algorithms will retain measurable residual information after selective deletion.

H2: The proposed Residual Entropy Score will provide greater sensitivity in detecting residual information than conventional utility-based evaluation metrics.

H3: Lower Residual Entropy Scores will be associated with reduced privacy leakage while maintaining acceptable downstream model performance.

H4: A multidimensional evaluation framework combining information theory and adversarial testing will provide a more comprehensive assessment of machine unlearning than any single evaluation metric.

Transition to Methodology

The identified research gap and the objectives outlined above form the foundation of the proposed methodological framework. The following section presents the complete experimental design, proposed Residual Entropy Score (RES), evaluation pipeline, validation strategy, and statistical analysis plan intended for future empirical implementation.

Section 3 Summary

- **Research Gap:** Clearly identifies the lack of standardized quantitative evaluation for machine unlearning.
- **Aim:** Introduces the proposed Residual Entropy Score (RES).
- **Objectives:** Five focused objectives aligned with the protocol.
- **Research Questions:** Five questions to guide future experiments.
- **Hypotheses:** Four testable hypotheses for subsequent empirical validation.

4. MATERIALS AND METHODS

4.1 Research Design

This study proposes a quantitative computational research framework for evaluating the effectiveness of machine unlearning in Large Language Models (LLMs). The methodology is designed to measure the residual influence of deleted training samples under realistic black-box deployment conditions, where internal model parameters are inaccessible and evaluation relies solely on observable model outputs.

The proposed framework integrates concepts from information theory, privacy-preserving machine learning, semantic similarity analysis, and adversarial evaluation to estimate residual information retained after machine unlearning. Unlike conventional approaches that primarily focus on model accuracy, the proposed methodology emphasizes both utility preservation and privacy assurance.

The experimental workflow consists of six stages:

1. Dataset preparation
2. Baseline model development
3. Machine unlearning
4. Black-box privacy evaluation
5. Residual Entropy Score (RES) computation
6. Statistical validation

4.2 Benchmark Datasets

The proposed framework recommends the use of publicly available biomedical datasets to ensure reproducibility and avoid ethical concerns associated with identifiable patient information.

Suggested benchmark datasets include:

Dataset	**Application**
PubMedQA	Biomedical question answering
BioASQ	Biomedical semantic understanding
MedMCQA	Medical reasoning
MIMIC-III (de-identified)	Clinical language modelling
Synthetic Electronic Health Records	Privacy evaluation

The complete training dataset is represented as **D**, where:

[
D = {x_1,x_2,x_3,\ldots,x_n}
]

A subset of records designated for deletion is represented by:

[
D_u \subset D
]

The remaining records constitute:

[
D_r = D-D_u
]

An independent evaluation dataset ($\mathbf{D_e}$) will be reserved for measuring downstream model utility.

4.3 Large Language Models

The proposed methodology is model-independent and can be applied to multiple transformer-based architectures, including:

- Llama-3
- Mistral
- Falcon
- GPT-style decoder models
- BioMedLM
- ClinicalBERT

All models should be evaluated using identical prompting strategies, decoding parameters, and benchmark datasets to ensure fair comparison across machine unlearning techniques.

4.4 Machine Unlearning Algorithms

The protocol recommends evaluating representative machine unlearning approaches that capture the major paradigms currently reported in the literature.

Exact Retraining

Exact retraining serves as the theoretical benchmark by retraining the model after excluding deleted samples. Although computationally intensive, it provides the strongest theoretical guarantee of complete forgetting.

Gradient-Ascent Unlearning

Gradient-ascent methods reduce the influence of deleted samples by reversing optimization updates associated with those records. These approaches are computationally efficient but may introduce instability if excessive parameter updates are applied.

Parameter Editing

Parameter-editing techniques modify localized neural representations associated with deleted information while minimizing disruption to unrelated knowledge. Representative methods include knowledge editing frameworks such as ROME and MEMIT.

Knowledge Distillation

Knowledge distillation transfers retained knowledge from a teacher model to a student model after excluding deleted samples. This approach aims to preserve downstream performance while reducing residual information.

Influence Function-Based Unlearning

Influence-function methods estimate the contribution of individual training samples to learned parameters and selectively reverse their influence. Although mathematically principled, these methods may become computationally demanding for large foundation models.

4.5 Proposed Residual Entropy Score (RES)

The principal contribution of this protocol is the development of a Residual Entropy Score (RES) to estimate hidden information retained after machine unlearning.

The proposed score integrates five complementary indicators:

- **Predictive Entropy (H):** Measures uncertainty in generated token probabilities.
- **Semantic Similarity (S):** Quantifies conceptual similarity between deleted samples and generated outputs using embedding-based cosine similarity.
- **Embedding Divergence (E):** Measures separation between latent representations before and after unlearning.
- **Reconstruction Probability (R):** Estimates the likelihood that deleted information can be regenerated through adversarial prompting.
- **Membership Inference Confidence (M):** Evaluates whether deleted samples remain distinguishable from unseen data.

The composite score is defined as:

[
RES=\alpha H+\beta S+\gamma E+\delta R+\epsilon M
]

subject to:

[
\alpha+\beta+\gamma+\delta+\epsilon=1
]

where α–ε represent weighting coefficients to be determined during future empirical validation.

Lower RES values are expected to indicate more effective removal of residual information.

4.6 Black-Box Evaluation Framework

Because many commercial LLMs provide only application programming interface (API) access, the proposed evaluation assumes black-box conditions.

Model behaviour will be assessed through structured interaction using:

- Standard prompts
- Context-expanded prompts
- Multi-turn dialogue
- Prompt inversion
- Semantic paraphrasing
- Membership inference attacks

This evaluation strategy reflects realistic deployment environments in which model internals remain inaccessible.

4.7 Evaluation Metrics

The proposed framework evaluates both model utility and privacy preservation.

Utility Metrics

- Accuracy
- BLEU
- ROUGE
- BERTScore
- Perplexity

Privacy Metrics

- Membership inference accuracy
- Reconstruction success rate
- Semantic similarity
- Prompt leakage score
- Residual Entropy Score

Together, these measures provide a multidimensional assessment of machine unlearning performance.

4.8 Statistical Analysis

The protocol recommends summarizing continuous variables using means, standard deviations, medians, and interquartile ranges.

Comparisons among machine unlearning algorithms should be performed using repeated-measures ANOVA when statistical assumptions are satisfied. Non-parametric alternatives, such as the Friedman test, may be employed where appropriate.

Effect sizes, 95% confidence intervals, and bootstrap resampling are recommended to improve robustness and reproducibility. Statistical significance should be interpreted using a threshold of $p < 0.05$ during future empirical evaluation.

4.9 Ethical Considerations

This protocol does not involve human participants or identifiable patient information. Future empirical implementation should use publicly available, de-identified, or synthetic datasets unless appropriate institutional ethical approval has been obtained. All future studies should comply with applicable data protection regulations and responsible AI principles.

5. PROPOSED VALIDATION FRAMEWORK AND DISCUSSION

5.1 Proposed Validation Framework

A robust evaluation framework is essential to determine whether machine unlearning algorithms effectively remove the influence of deleted training samples while preserving the overall performance of Large Language Models (LLMs). Existing studies frequently rely on downstream task accuracy or isolated privacy attacks to assess forgetting. However, these approaches provide only a partial understanding of post-unlearning behaviour because they do not directly estimate the residual information retained within latent model representations. The proposed validation framework addresses this limitation by integrating utility preservation, adversarial privacy assessment, semantic analysis, and information-theoretic evaluation into a unified benchmarking strategy.

The proposed validation process consists of six sequential stages:

1. **Baseline Model Evaluation:**

A baseline LLM trained on the complete dataset will be evaluated to establish reference values for predictive performance, semantic quality, and privacy-related metrics.

2. **Selective Data Deletion:**

A predefined subset of training samples containing privacy-sensitive information will be designated for removal using representative machine unlearning algorithms.

3. Machine Unlearning:

Selected algorithms—including exact retraining, gradient-ascent unlearning, parameter editing, knowledge distillation, and influence-function methods—will be applied independently under standardized computational conditions.

4. Black-Box Privacy Assessment:

The post-unlearning models will be evaluated through structured interactions using standard prompts, adversarial prompt engineering, semantic paraphrasing, prompt inversion, and membership inference attacks to identify any remaining traces of deleted information.

5. Residual Entropy Estimation:

The proposed **Residual Entropy Score (RES)** will be calculated by integrating predictive entropy, semantic similarity, embedding divergence, reconstruction probability, and membership inference confidence into a single composite metric.

6. Comparative Statistical Evaluation:

Results obtained from different unlearning algorithms will be compared using appropriate statistical procedures to examine differences in privacy preservation and utility retention.

This workflow is designed to facilitate reproducibility and enable meaningful comparison among future machine unlearning approaches.

5.2 Expected Outcomes

Because this manuscript presents a research protocol, no experimental results are reported. Nevertheless, several anticipated outcomes are expected if the proposed methodology is implemented and validated through future empirical studies.

First, the proposed Residual Entropy Score (RES) is expected to provide a more comprehensive quantitative measure of information retention than conventional utility-based evaluation metrics. While existing approaches generally assess model performance through accuracy or perplexity, RES aims to capture hidden residual information that may remain statistically encoded after selective deletion.

Second, different machine unlearning algorithms are expected to demonstrate varying levels of residual entropy despite achieving comparable benchmark performance. Such findings would indicate that similar predictive accuracy does not necessarily imply equivalent privacy protection, highlighting the importance of multidimensional evaluation.

Third, the proposed framework is anticipated to improve detection of latent information leakage under realistic black-box conditions. By integrating adversarial prompting, semantic

similarity analysis, and membership inference attacks, the methodology is expected to provide a more reliable assessment of privacy preservation than isolated attack strategies.

Finally, successful validation of the proposed framework may contribute to the establishment of standardized benchmarking procedures for machine unlearning, supporting future certification of trustworthy AI systems operating in privacy-sensitive environments such as healthcare, finance, and public administration.

5.3 DISCUSSION

The increasing adoption of Large Language Models across healthcare and other knowledge-intensive domains has created an urgent need for reliable mechanisms that ensure compliance with evolving privacy regulations. Although machine unlearning offers a promising solution for selective data deletion, current evaluation practices remain fragmented and frequently emphasize predictive performance at the expense of privacy verification.

The proposed Residual Entropy Score (RES) addresses this challenge by introducing an information-theoretic perspective into machine unlearning evaluation. Rather than treating successful forgetting as a binary outcome, the framework conceptualizes information retention as a continuous phenomenon that can be estimated through complementary statistical indicators. This multidimensional perspective aligns with the growing recognition that latent neural representations may continue to encode deleted information even when direct textual recall is no longer observable.

A notable strength of the proposed methodology is its applicability under black-box deployment conditions, reflecting the operational reality of many commercial LLMs. Because internal parameters are generally inaccessible in production environments, practical evaluation methods must rely on observable model behaviour. The proposed integration of semantic similarity analysis, adversarial prompting, and membership inference attacks therefore provides a realistic and reproducible strategy for assessing privacy preservation.

The protocol also supports the broader objectives of trustworthy artificial intelligence by promoting transparency, reproducibility, and standardized evaluation practices. A unified benchmarking methodology could assist researchers, healthcare organizations, and regulatory authorities in comparing machine unlearning techniques using common criteria rather than relying on isolated performance indicators.

Nevertheless, the proposed framework has several limitations. As a methodological protocol, it has not yet undergone empirical validation, and the weighting coefficients of the Residual Entropy Score remain to be optimized using experimental data. Furthermore, the

effectiveness of the proposed framework may vary depending on model architecture, dataset characteristics, and unlearning strategy. Future studies should therefore evaluate the methodology across multiple foundation models, biomedical datasets, and multilingual environments to determine its robustness and generalizability.

Despite these limitations, the proposed framework provides a structured foundation for future investigations into privacy-preserving machine learning. If validated empirically, the Residual Entropy Score may contribute to the development of standardized certification procedures for machine unlearning and strengthen confidence in the responsible deployment of foundation models across privacy-sensitive application domains.

5.4 Significance of the Proposed Framework

The proposed methodology offers several practical and scientific contributions:

- Introduces a novel information-theoretic metric for evaluating machine unlearning.
- Provides a reproducible benchmarking framework suitable for black-box LLMs.
- Integrates utility and privacy assessment into a unified evaluation strategy.
- Supports future regulatory auditing and trustworthy AI certification.
- Establishes a methodological foundation for subsequent empirical research.

Collectively, these contributions position the proposed framework as a step toward more rigorous and standardized evaluation of machine unlearning in foundation models.

6. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The rapid evolution of Large Language Models (LLMs) has transformed artificial intelligence by enabling sophisticated natural language understanding, reasoning, and knowledge generation across diverse application domains. Despite these advancements, increasing concerns regarding privacy, data ownership, and regulatory compliance have highlighted the need for reliable methods capable of removing sensitive information from trained models. Machine unlearning has emerged as a promising solution by enabling selective removal of designated training samples without requiring computationally expensive retraining. However, the absence of standardized methodologies for evaluating the effectiveness of machine unlearning remains a significant limitation in current research.

This research protocol proposes a novel Residual Entropy Score (RES) as a unified information-theoretic framework for quantifying residual information retained within post-unlearning Large Language Models. Unlike conventional evaluation methods that primarily

focus on predictive performance or isolated privacy attacks, the proposed methodology integrates predictive entropy, semantic similarity, embedding divergence, reconstruction probability, and membership inference confidence into a multidimensional assessment framework. By combining these complementary indicators, the protocol aims to provide a more comprehensive understanding of residual information that may remain encoded after selective data deletion.

The proposed framework emphasizes evaluation under realistic black-box deployment conditions, reflecting practical scenarios in which model internals are inaccessible to end users. This characteristic enhances the applicability of the methodology to commercial foundation models and aligns with emerging requirements for transparent and trustworthy artificial intelligence. The protocol further promotes reproducibility through standardized datasets, clearly defined evaluation procedures, and a structured statistical analysis plan.

Although the present manuscript does not report experimental findings, it establishes a scientifically grounded methodological foundation for future empirical investigations. If validated through comprehensive benchmarking, the proposed Residual Entropy Score may contribute to the development of standardized evaluation procedures for machine unlearning, strengthen privacy auditing practices, and support regulatory compliance in healthcare and other privacy-sensitive sectors.

6.2 Future Scope

The proposed framework offers several opportunities for future research and methodological enhancement.

Future empirical investigations should validate the Residual Entropy Score using multiple state-of-the-art Large Language Models, including general-purpose and domain-specific biomedical models. Comparative benchmarking across different model architectures, parameter scales, and training strategies will help determine the robustness and generalizability of the proposed metric.

The methodology may also be extended to multimodal foundation models capable of processing text, medical images, audio, genomic data, and structured clinical records. Such extensions would broaden the applicability of machine unlearning evaluation to emerging healthcare AI systems that integrate multiple data modalities.

Further research should investigate adaptive optimization of the weighting coefficients used in the Residual Entropy Score through data-driven approaches, including Bayesian optimization and meta-learning. Dynamic weighting strategies may improve the sensitivity of the metric across different datasets and application domains.

Another promising direction involves integrating the proposed framework with complementary privacy-preserving technologies, including differential privacy, federated learning, secure multiparty computation, and confidential computing. Combining these approaches may provide stronger theoretical guarantees regarding information removal while preserving model utility.

Longitudinal studies examining continual learning and incremental model updating also represent an important area for future investigation. As foundation models increasingly evolve through continuous retraining, ensuring that previously deleted information remains permanently inaccessible will become an essential challenge.

Finally, collaboration among researchers, healthcare organizations, regulatory agencies, and standards bodies will be necessary to establish internationally accepted benchmarks for machine unlearning evaluation. Such efforts could facilitate certification frameworks supporting trustworthy deployment of artificial intelligence in healthcare and other privacy-sensitive environments.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the contributions of researchers in artificial intelligence, machine learning, information theory, cybersecurity, and privacy-preserving computing whose pioneering work has inspired the development of this methodological framework.

AUTHOR CONTRIBUTIONS

Dr. Akhilesh Kumar conceived the research idea, designed the methodological framework, developed the proposed Residual Entropy Score, prepared the manuscript, and approved the final version for publication.

FUNDING

The author declares that no external funding was received for this research protocol.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

DATA AVAILABILITY STATEMENT

No datasets were generated or analysed during the preparation of this research protocol. Future empirical implementation is expected to utilize publicly available benchmark datasets and de-identified resources in accordance with applicable licensing agreements and institutional policies.

ETHICAL APPROVAL

This manuscript presents a research protocol and methodological framework. It does not involve human participants, patient data, animal experimentation, or identifiable personal information. Consequently, formal ethical approval was not required for the preparation of this manuscript. Future empirical studies should obtain approval from an appropriate Institutional Ethics Committee or Institutional Review Board before data collection.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

REFERENCES

1. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379-423.
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Red Hook (NY): Curran Associates Inc.; 2017. p. 5998-6008.
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019. Minneapolis (MN): Association for Computational Linguistics; 2019. p. 4171-4186.
4. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. 2020;33:1877-1901.
5. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. Stanford (CA): Stanford Center for Research on Foundation Models; 2021.
6. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. In: Proceedings of the 30th USENIX Security Symposium. Berkeley (CA): USENIX Association; 2021. p. 2633-2650.
7. Carlini N, Feldman V, Nasr M, et al. Membership inference attacks from first principles. In: Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos (CA): IEEE; 2022.

8. Ginart A, Guan M, Valiant G, Zou J. Making AI forget you: Data deletion in machine learning. In: Advances in Neural Information Processing Systems. 2019;32.
9. Guo C, Goldstein T, Hannun A, van der Maaten L. Certified data removal from machine learning models. In: Proceedings of the 37th International Conference on Machine Learning. PMLR. 2020.
10. Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Travers A, Zhang B, et al. Machine unlearning. In: Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos (CA): IEEE; 2021.
11. Dwork C, Roth A. The algorithmic foundations of differential privacy. Boston (MA): Now Publishers Inc.; 2014.
12. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. New York (NY): ACM; 2016. p. 308-318.
13. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos (CA): IEEE; 2017. p. 3-18.
14. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York (NY): ACM; 2015. p. 1322-1333.
15. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York (NY): ACM; 2019. p. 220-229.
16. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P, et al. Ethical and social risks of harm from language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York (NY): ACM; 2022. p. 214-229.
17. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv [Preprint]. 2023. Available from: https://arxiv.org/abs/2302.13971
18. European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Off J Eur Union. 2016;L119:1-88.

19. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg (MD): National Institute of Standards and Technology; 2023.
20. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-399.