Major Research Project: Data Science & Analytics

Methodology & Experiments for Protecting Personally Identifiable Information (PII) in Abstractive Summaries using Large Language Models (LLMs)

Ryerson
Toronto Metropolitan University

Student: Colin Lacey
Student ID: 501176114
Supervisor: Dr. Tony Hernandez

Date of Submission: 14 July 2025

Table of Contents

Approach to Methodology & Experiments

Aim of Study

The aim of this study is to evaluate the effectiveness of a fine-tuned LLaMA 3 Retrieval-Augmented Generation (RAG) model for automatically generating coherent, relevant, and privacy-preserving summaries of the Town of Whitby's council and committee meeting minutes.

Public access to government records is often restricted by manual and resource-intensive review processes to identify relevant records and then redact content to prevent the exposure of personally identifiable information (PII). Leveraging recent advancements in machine learning and large language models, particularly methods such as Differential Private Data Dropout (DPDD), Named Entity Recognition (NER), Retrieval-Augmented Generation (RAG), this research seeks to evaluate whether these innovative approaches can simultaneously improve access while protecting privacy. As part of the assessment pf privacy protection, this research incorporates Canary Extraction Success Rate (CESR) which is a targeted measure of memorization risk that involves the strategic insertion of identifiable false statements ("canaries") that will allow for the evaluation of privacy leakage risks from the large language model used.

However, as the practical deployment of large language models is constrained by computational resources, this study includes Low-Rank Adaptation (LoRA) as a method to mitigate computational costs and to provide the flexibility to scale the model as required.

Selection of Response Variables

Response variables selected include ROUGE and BERTScore for summarization quality, CESR for privacy leakage assessment, and human evaluator ratings as a final assessment. These response variables will also be compared against document-length categories (short, medium, long) and document groupings by publication year due to the variability observed in the previously submitted Exploratory Data Analysis (EDA) that was included with the Literature Review.

1. Text summarization quality metrics:
    - ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L)
    - BERTScore
2. Privacy and data leakage metrics
    - Canary Extraction Success Rate (CESR)
    - Human-rated privacy assessments (qualitative)

Choice of Factors and Levels:

The experimental factors, informed by the earlier literature review and exploratory analysis, include fine-tuning hyperparameters (LoRA rank and learning rate), retrieval settings (number of documents retrieved, ElasticSearch relevancy threshold), privacy protection methods (NER and DPDD preprocessing), document length categories, and grouping documents by year. Each factor was selected to evaluate distinct dimensions affecting summarization quality, computational efficiency, and ultimately privacy risk.

1. LoRA Rank and Hyperparameters (continuous factors):
    a. Rank size (4, 8, 16)

   b. Learning rate (1e-4, 3e-4, 5e-5)

2. Retrieval settings (continuous or categorical factor):
   a. Number of documents retrieved (1, 3, 5, etc.)
   b. ElasticSearch relevancy threshold (0.3, 0.5, 0.7)

3. Privacy protection methods (categorical factor):
   a. NER combined with DPDD preprocessing
   b. Comparison against baseline (no DPDD)

4. Document length (categorical factor):
   a. Short (under median of 1101 words)
   b. Medium (1101 to mean of 2059 words)
   c. Long (above 2059 words)

5. Time periods (categorical factor):
   a. Before 2018 (less volume, shorter documents)
   b. 2018-2022 (peak volume, longer documents)
   c. After 2022 (recent data, moderate length)

Experimental Design:

The experimental design incorporates a factorial structure varying LoRA rank, retrieval settings, and DPDD preprocessing. Additionally, stratified sampling based on document length categories and publication years (grouped by periods identified during exploratory data analysis) will ensure robust performance across subsets of data. Data will be partitioned into distinct train-

validation-test splits (70%-15%-15%), ensuring that hyperparameter tuning is performed using validation sets, while final model evaluation occurs exclusively on the test set.

Potential limitations include variability introduced by noisy or anomalous data, inherent randomness in model initialization, and complexities arising from interactions among multiple factors.

Factorial Design:

- Conduct experiments in a structured way, varying multiple factors simultaneously (LoRA rank, Retrieval Documents, and DPDD settings).
- Train-validation-test splits for evaluation of model performance (70%-15%-15%)

Performing the Experiment

The experiment will involve the preprocessing steps that were informed from the earlier exploratory data analysis insights, including categorizing documents by length and publication year to manage variability in data volume and structure. Summarization tasks and evaluations will be conducted separately for each identified subgroup to systematically measure model performance and generalizability.

The experiment will be conducted systematically, with explicit attention to replication, randomization, and robustness, alongside detailed human evaluation protocols.

Step 1: Data Preprocessing

- All 1180 PDF documents will undergo initial preprocessing:
  - Text extraction and tokenization.

- Identification and removal of personally identifiable information (PII) using Named Entity Recognition (NER).

- Differential Private Data Dropout (DPDD) to quantitatively rank and remove high-risk data points based on calculated Renyi Differential Privacy (RDP) scores.
    - RDP scores will also be generalized to the categories of document types identified in the earlier EDA, to determine if certain types of records pose greater risk than others (mean, standard deviation and maximum RDP per record category)
- Insertion of Canary Statements:
    - Carefully crafted, easily identifiable false statements ("canaries") will be introduced into the training dataset at strategic intervals.
    - Canaries will be designed explicitly not to overlap or conflict with real PII, ensuring they are distinct and unmistakable.
    - The location and frequency of canaries will be systematically documented for accurate subsequent retrieval testing.

Step 2: Experimental Setup and Randomization

- The factorial experimental design will systematically vary:
    - LoRA rank (4, 8, 16) and learning rates (1e-4, 3e-4, 5e-5).
    - Retrieval settings including the number of documents (1, 3, 5) and ElasticSearch relevancy thresholds (0.3, 0.5, 0.7).
    - Privacy protection methodologies (NER with DPDD vs. baseline).

- Document subsets will be explicitly stratified based on length (short, medium, long) and grouping by publication years (pre-2018, 2018-2022, post-2022), ensuring balanced representation across conditions.

- Experimental runs will be randomized within each block (subgroup combination) to eliminate bias due to order effects or learning curves.

Step 3: Replication for Robustness

- To account for variability due to inherent randomness (initialization, training variation, sampling differences):
    - Each experimental condition (factor-level combination) will undergo multiple replications (minimum of three runs per combination).
    - These replications will utilize identical train-validation-test splits to facilitate robust comparisons and ensure variability measured is due to random initialization rather than data differences.
    - The final analysis will aggregate results across replications, providing averages and variances to accurately capture model performance and robustness.

Step 4: Interpretation of Results and Factor Interaction Analysis

- Detailed statistical analysis (i.e., ANOVA) will test for interactions among key factors, especially between privacy protection methods and fine-tuning hyperparameters.
- Interactions will be explored to better understand how varying levels of privacy protection influence overall summarization quality metrics (ROUGE, BERTScore, CESR, precision, and recall).

Step 5: Model Evaluation and Metrics Collection

- Models will be evaluated consistently using train-validation-test splits (70%-15%-15%) to ensure unbiased performance assessment.

- For each experimental condition (each unique combination of LoRA rank, learning rate, retrieval settings, and privacy methods), a consistent sample of generated summaries will be evaluated. Specifically, 50 summaries will be generated per condition, providing a robust basis for calculating summarization quality metrics (ROUGE, BERTScore) and privacy assessments (CESR, Precision, Recall). Additionally, from these 50 summaries, approximately 5 summaries (10%) will undergo human evaluation as further validation.

- Summarization quality and privacy metrics (ROUGE, BERTScore, CESR, Precision, Recall) will be computed for each split separately, then averaged over replications.

- The Canary Extraction Success Rate (CESR) metric will quantify the frequency at which inserted canaries can be retrieved through deliberate prompt engineering, objectively evaluating privacy risk**:**
    - CESR will measure the ability of the trained model to retain and inadvertently reveal inserted canaries will be systematically tested through deliberate prompt engineering.
    - CESR will quantify the percentage of inserted canaries successfully extracted by prompts designed to elicit their retrieval.
    - CESR results will serve as a quantitative metric of the privacy leakage potential of each experimental configuration.

Step 6: Human Evaluation

- Human evaluation will be systematically structured:
    - A trained human evaluator familiar with FIPPA guidelines will assess a representative subset of generated summaries.
    - Evaluation will specifically focus on the presence or absence of basic PII elements defined by FIPPA (name, address, phone number, email, date of birth, social insurance number, and other identifying elements).
    - Evaluations will be recorded categorically (presence or absence) for each summary, with systematic notes on any potential false positives or false negatives.
    - 10% of summaries per experimental condition will undergo human evaluation to validate metrics (CESR, Precision, Recall).

Step 7: Documentation and Transparency

- Complete records of experimental parameters, settings, replications, and human evaluation results will be maintained for reproducibility and auditability.

Stastistical Analysis of Data

Statistical analysis will include ANOVA to identify significant main and interaction effects among experimental factors, document length categories, and temporal groupings. Additional evaluation metrics such as Precision and Recall will be employed specifically to assess the effectiveness and reliability of the privacy-preserving measures (NER + DPDD preprocessing). Metrics including ROUGE, BERTScore, CESR, Precision, and Recall will be statistically analyzed, with descriptive statistics and correlation analyses providing deeper insights into interrelationships and trade-offs between summarization quality and privacy protection.

To ensure robust results, experiments will incorporate randomization in execution order, multiple replications to average out the effects of uncontrollable factors, and blocking techniques (e.g., ensuring identical document samples across experiments) to reduce variability due to nuisance factors.

- ANOVA for significance testing of factor impacts.
- Precision and recall scores for model performance related to exclusion of PII.
- Descriptive statistics (means, standard deviations) of performance metrics.
- Correlation analysis between metrics (ROUGE, BERTScore, CESR, Precision and Recall)

Conclusions and Recommendations

The conclusions of this study will provide key insights that will identify optimal model configurations (LoRA rank, learning rate, retrieval settings) and preprocessing methods (NER and DPDD usage) that achieve the best balance between summarization quality and privacy protection. Detailed analysis of interactions between model hyperparameters and privacy-preserving preprocessing will inform clear, actionable recommendations for practical use in real-world scenarios. Additionally, the evaluation of memorization risk through CESR provides objective measures of privacy leakage, highlighting strengths and limitations inherent to the proposed methodology. Recommendations will include guidelines for deploying similar fine-tuned LLaMA 3 RAG architectures in real-world applications, considering computational efficiency, robustness across varied document lengths and time periods, and best practices for ongoing monitoring of privacy risks through canary-based testing. Finally, potential areas for future research will be identified, including potential approaches to further reduce privacy

leakage, and exploring generalization of findings to broader datasets beyond municipal council minutes.

Github Repository

https://github.com/CDL-DataSci/MRP_AbstractSummary

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, Li. (Oct 24, 2016) “Deep Learning with Differential Privacy” in 23rd ACM Conference on Computer and Communications Security arXiv:1607.00133

2. Alpaydin, E. (2014). “Introduction to Machine Learning, Third Edition”, The MIT Press, Cambridge Massachusetts. IBSN 978-0-262-02818-9.

3. Anil, R., Ghazi, B., Gupta, V., Kumar, R., Manurangsi, P. (Aug 3, 2021). “Large-Scale Differentially Private BERT” arXiv:2108.01624

4. Balde, G., Roy, S., Mainack, M., Ganguly, N. (May 27, 2025). “Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings” in the Findings of the 63rd Annual Meeting of the Association for Computational Linguistics arXiv:2505.21242

5. Bekman, S., Rajbhandari, S., Wyatt, M., Rasley, J., Ruwase, T., Yao, Z., Qiao, A., He, Y. (June 16, 2025). “Arctic Long Sequence Training: Scalable and Efficient Training for Multi-Million Token Sequences” arXiv:2506.13996

6. Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S. (April 2019). “A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments” in Lecture Notes in Computer Science. DOI: 10.1007/978-3-030-15712-8_27

7. Brown, H., Lee, K., Fatemehsadat, M., Shokri, R., Tramèr, F. (February 14, 2022). “What Does it Mean for a Language Model to Preserve Privacy?” in FAccT ‘22 DOI: 10.48550/arXiv.2202.05520

8. Carlini, N., Tramèr, F., Wallance, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Opera, A., Raffel, C. (June 15, 2021). “Extracting Training Data from Large Language Models.” DOI: 10.48550/arXiv.2012.07805

9. Dernoncourt, F., Lee, J. Y., Szolovits, P., Uzuner, Ö. (June 10, 2016). "De-identification of Patient Notes with Recurrent Neural Networks." DOI: 10.48550/arXiv.1606.03475

10. El-Kassas, W. S., Salama, C. R., Rafea, A. A., Mohamed, H. K. (July 2020). “Automatic Text Summarization: A Comprehensive Survey” in Expert Systems with Applications. DOI: http://dx.doi.org/10.1016/j.eswa.2020.113679

11. Fu, Z., Man-Cho So, A., Collier, N. (December 7, 2023). “A Stability Analysis of Fine-Tuning a Pre-Trained Model.” DOI: https://ui.adsabs.harvard.edu/link_gateway/2023arXiv230109820F/doi:10.48550/arXi

v.2301.09820

12. Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltago, I., Downey, D., Smith, N. A. (April 2020). “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks” in ACL 2020. DOI: https://ui.adsabs.harvard.edu/link_gateway/2020arXiv200410964G/doi:10.48550/arXiv.2004.10964

13. Hughes, A., Ma, N., Aletras, N., (May 27, 2025). “How Private are Language Models in Abstractive Summarization?” DOI: https://ui.adsabs.harvard.edu/link_gateway/2024arXiv241212040H/doi:10.48550/arXiv.2412.12040

14. Jayatilleke, N., Weerasinghe, R., Senanayake, N. (February 2025). “Advancements in Natural Language Processing for Automatic Text Summarization” in the International Conference on Computer Systems (ICCS 2024). DOI: https://ui.adsabs.harvard.edu/link_gateway/2025arXiv250219773J/doi:10.48550/arXiv.2502.19773

15. Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Braken-Clarke, D., Xue, E., Yang, Y., Sun, J., Lu, Z. (2024). “Matching Patients to Clinical Trials with Large Language Models” in Nature Communications. DOI: https://doi.org/10.1038/s41467-024-53081-z

16. Koh, H. Y., Ju, J., Liu, M., Pan, S. (July 3, 2022). “An Empriical Survey on Long Document Summarization: Datasets, Models and Metrics” in ACM Computing Systems. DOI: https://ui.adsabs.harvard.edu/link_gateway/2022arXiv220700939K/doi:10.48550/arXiv.2207.00939

17. Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., Wallace, B. C. (April 22, 2021). “Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?” in NAACL Camera Ready Submission. DOI: https://ui.adsabs.harvard.edu/link_gateway/2021arXiv210407762L/doi:10.48550/arXiv.2104.07762

18. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., Zanella-Béguelin, S. (April 23, 2023). “Analyzing Leakage of Personally Identifiable Information in Language Models” in IEEE Symposium on Security and Privacy (S&P) 2023. DOI: https://ui.adsabs.harvard.edu/link_gateway/2023arXiv230200539L/doi:10.48550/arXiv.2302.00539

19. Matkin, N., Smirnov, A., Usanin, M., Ivanov, E., Sobyanin, K., Paklina, S., Parshakov, P. (September 15, 2024). “Comparative Analysis of Encoder-Based NER and Large Language Models for Skill Extraction from Russian Job Vacancies.” DOI: https://ui.adsabs.harvard.edu/link_gateway/2024arXiv240719816M/doi:10.48550/arXiv.2407.19816

20. Miller, J. K., Tang, W. (May 13, 2025). “Evaluating LLM Metrics Through Real-World Capabilities.” DOI: https://ui.adsabs.harvard.edu/link_gateway/2025arXiv250508253M/doi:10.48550/arXiv.2505.08253

21. del Moral-Gonzalez, R., Gomez-Adorno, H., Ramos-Flores, O. (2025). “Comparative Analysis of Generative LLMs for Labeling Entities in Clinical Notes” in Genomics & Informatics. https://genomicsinform.biomedcentral.com/articles/10.1186/s44342-024-00036-x

22. Obeidat, M. S., Al Nanian, S., Kavuluru, R. (April 2025). “Do LLMs Surpass Encoders for Biomedical NER?” in IEEE ICHI 2025. DOI: https://doi.org/10.48550/arXiv.2504.00664

23. Pal, A., Bhargava, R., Hinsz, K., Esterhuizen, J., Bhattacharya, S. (November 8, 2024). “The Empirical Impact of Data Sanitization on Language Models” in Safe Generative AI Workshop at NeurIPS 2024. DOI: https://doi.org/10.48550/arXiv.2411.05978

24. Pan, X., Zhang, M., Ji, S., Yang, M. (July 2020). “Privacy Risks of General-Purpose Language Models” in IEEE Symposium on Security and Privacy 2020. DOI: https://doi.org/10.1109/SP40000.2020.00095

25. Priyanshu, A., Vijay, S., Kumar, A., Naidu, R., Mireshghallah, F. (May 24, 2023). “Are Chatbots Ready for Privacy-Sensitive Applications? An Investigation into Input Regurgitation and Prompt-Induced Sanitization.” DOI: https://doi.org/10.48550/arXiv.2305.15008

26. Rehman, T., Das, S., Sanyal, D. K., Chattopadhyay, S. (February 25, 2023). “An Analysis of Abstractive Text Summarization Using Pre-trained Models” in Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing. DOI: https://doi.org/10.1007/978-981-19-1657-1_21

27. Rehman, T., Sanyal, D. K., Chattopadhyay, S., Bhowmick, P. K., Das, P. P. (2021). “Automatic Generation of Research Highlights from Scientific Abstracts” in EEKE 2021 – Workshop on Extractions and Evaluation of Knowledge Entities from Scientific Documents. https://ceur-ws.org/Vol-3004/paper10.pdf

28. Rehman, T., Ghosh, S., Das, K., Bhattacharjee, S., Sanyal, D. K., Chattopadhyay, S. (March 13, 2025). “Evaluating LLMs and Pre-Trained Models for Text Summarization Across Diverse Datasets.” DOI: https://ui.adsabs.harvard.edu/link_gateway/2025arXiv250219339R/doi:10.48550/arXiv.2502.19339

29. Shen, H., Gu, Z., Hong, H., Han, W. (February 25, 2025). “PII-Bench: Evaluating Query-Aware Privacy Protections Systems.” DOI: https://ui.adsabs.harvard.edu/link_gateway/2025arXiv250218545S/doi:10.48550/arXiv.2502.18545