Major Research Project: Data Science & Analytics

Literature Review to Support Protecting Personally Identifiable Information (PII) in Abstractive Summaries using Large Language Models (LLMs)

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

Date of Submission: 24 June 2025

Table of Contents

Abstract

Introduction

Access to public records is often hindered by administrative, legal, and technical challenges. Key barriers include the labor-intensive process of reviewing documents against privacy and access laws, difficulties in identifying relevant records using non-standardized content, and the high cost of processing large volumes of documents. This research proposes leveraging Large Language Models (LLMs) and automatic text summarization to address these challenges and improve access to public records.

The current manual processes for identifying, reviewing, and retrieving public records are time-consuming, costly, and often fail to efficiently address requests due to:

- The need to assess compliance with complex legal requirements (e.g., Privacy Act, PIPEDA, FIPPA, MFIPPA).
- Limitations of keyword-based search in identifying responsive documents.
- High resource cost associated with reviewing large volumes of documents or lengthy reports.

This literature review will explore current research in using traditional pre-trained transformer models specialized in named entity recognition (NER) and LLMs such as LLaMA 3 to correctly identify personally identifiable information (PII), ultimately seeking to prevent leaking PII from the fine-tuned model training set, and to finally produce document summaries that are clear and coherent and no do not contain PII.

Literature Review

This literature review covers 28 research articles published from 2016 to 2025 and seeks to understand the effectiveness of utilizing pre-trained models to generate high-quality summaries of a set of documents while also protecting personally identifiable information. Topics covered range from the metrics available to evaluate abstractive summaries, factors that will influence model performance, risks to exposing training data and ability to consistently identify and categorize personally identifiable information. The literature review concludes with some potential strategies to mitigate the risk of leaking training data and exposing PII unessentially, including approaches to fine-tuning the pretrained models to further mitigate the known risks.

As indicated by Rehman et al (2023), at a high level it can be said that pre-trained language models are able to generate high quality extractive and abstractive summaries even without fine tuning. However, issues become apparent when evaluating the produced summaries in detail for consistency, accuracy and relevancy. For example, traditional transformer models like BERT and T5 can produce irrelevant or overly verbose content with repetitious summaries, and LLMs are known to produce hallucinations with factually inaccurate information (Rehman et al., 2021).

Given that neither traditional transformer models nor fine-tuned LLMs offer a perfect solution for summarizations, the best approach will depend on the dataset available, computing considerations and the intended output. As Matkin et al. (2024) have outlined, for some types of structured data extraction, traditional encoder-based models can outperform LLMs. While their research did show that some fine-tuned LLMs such as LLaMA 3.1 were promising, the computing cost and latency were concerns and lagged behind the traditional encoder-decoder transformer-based models. However, in datasets where individual documents are significantly long, these standard transformer models struggle with producing relevant summaries that are not truncated (Koh et al., 2022). For long document summarization tasks, LLMs are better suited. And while researchers such as Obeidat et al. (2025) claim that a tipping point has been reached where recent advances with LLMs now allow these language models to outperform encoder-based applications, particularly in zero-shot and few-shot prompting, further research by Rehman et al. (2025) of two advanced encoder-decoder models (BART and FLAN-T5) and two fine-tuned LLMs (LLaMA 3-8B and Gemma-7B), found that no single model outperformed all others across the various datasets used in their research. Additionally, advanced LLMs still require significant GPUs and memory resources, which may represent an obstacle to deploying the summarization tool at scale.

Regardless of approach, it has been well documented that transformer-based models have significantly advanced their abilities to produce text summarization (Jayatilleke et al., 2025). Between the general categories of extractive and abstractive summaries, extractive summaries are characterized by pulling verbatim phrases from the source data. This type of summary tends to maintain accuracy at the cost of awkward phrasing and is prone to repetitious or inclusion of lower value content. Whereas abstractive summaries struggle with factual accuracy and are prone to hallucinations (Balde et al. 2025). The current evaluation metrics available to automate the scoring of the generated summaries, whether extractive or abstractive, are challenging to interpret when provided in isolation. As Balde et al. (2025) have found, while ROUGE (Recall-Orientated Understudy for Gisting Evaluation) is a common evaluation metric for text summaries, the values are often inconsistent with meaningful and truthful rankings produced by human evaluators. In comparison with alternatives to ROUGE, BERTScores are often better correlated with human judgement. Similar findings that ROUGE alone were insufficient as a text summary evaluation metric were documented by Bhattacharya et al. (2019), and El-Kassas et al. (2020) utilized ROUGE, BLEU and METEOR and concluded that these automatic metrics tend to poorly correlate with human assessments of the same summaries. What both El-Kassas et al. (2020) and Balade et al. (2025) noted is that performance is positively impacted by fine-turning the pretrained language models, and that prompt structure significantly affects performance.

On the topic of model performance and fine tuning, Fu et al. (2023) noted that despite the larger model, LLMs were sensitive to small changes that could lead to instability. For example, Fu et al. (2023) observed that smalle changes to the model's hyperparameters could lead to large differences in ROUGE scores. While the research focused on the parameters of learning rate, batch size, optimizer, model size and number of fine-tuning steps completed, they found that changes to learning rate was a major driver of instability. Specifically, a smaller learning rate was associated with more stability but came at the cost of computing resources and Fu et al. recommended that stability could be further increased with more than one fine-tuning run that used multiple random seeds. Gururangag et al. (2020) provide additional recommendations to improve a LLMs performance on domain-specific summaries by first conducting unsupervised pretraining on an unlabeled domain-specific corpora which is then followed by supervised fine-tuning such as instruction-tuning.

Outside of generating text summaries, Muller et al. (2025) are cautious with an LLMs ability to identify sensitive data consistently. What their research found is that an LLMs ability to recognize sensitive data provided in a prompt was highly variable. With respect to the sensitive data provided in a prompt, the risk here is that the model may retain and later regurgitate parts of the information to other users. As such, Muller et al. (2025) conclude that on its own. LLMs are not ready to serve as fully autonomous tools and further investigation of how instruction tuning could help to improve consistency. In a similar vein, Shen et al (2025) investigated a language model ability to identify PII in text of the source data. What they discovered is that while LLMs such as ChatGPT could understand the task definition, these models produced many false positives and lacked precision. Furthermore, the generated responses from the model produced an inconsistent output format as the LLMs prompt sensitivity could significantly alter model behaviour.

While the goal of fine-tuning a pretrained model is to improve the overall performance, it is not without risk. As Carlini et al. (2021) have noted, large language models inherently memorise sensitive content, and fine-tuning may increase the risk of memorization as the models would review the same examples through the multiple epochs. Moral-Gonzalez et al (2025) took this research one step further and were able to show that LLMs can store latent knowledge about entities even without explicitly memorizing the direct data. This latent memory storage not only poses a problem of revealing memorized training data, it also poses a risk of the model unknowingly confusing two similarly named but different individuals or subjects which would limit the factual accuracy.

Given this natural tendency towards memorization, it should be assumed that LLMs could reveal sensitive data, even without malicious prompts (Lukas et al., 2023). In fact, as Pan et al (2020) have proven, LLMs can emit sensitive information verbatim when prompted correctly. Such an outcome presents measurable privacy risks if using a fine-tuned LLM. To help quantify the privacy risk posed by data leakage from LLMs and the impacts of fine-tuning and instruction-tuning on amplifying memorization, Hughes et al (2025) have documented a Canary Extraction Success Rate (CESR) that is designed to insert easily identifiable false statements (aka, the canary) in to the training dataset and will then measure the percent of canaries successfully extracted via prompt engineering from the model post training.

Aside from measuring the degree of data leakage, there are several mitigation strategies that can be implemented to prevent data leakage from occurring. As documented by Abadi et al. (2016), a Differential Privacy strategy can be implemented for deep learning. The Differential

Privacy approach leverages the open-source code available in the “tensorflow_privacy” python package, and is structured to take a random sample that will then “clip” or assign a weight to a response to limit the impact during training of the model. Finally, Differential Privacy will then add noise to the training data in order to mask the private information. The goal being to hide the contribution of the private information while still allowing the model to learn. A key feature of the Differential Privacy approach outlined by Abadi et al. (2016) was the use of the Moment Accountant. The Moment Accountant was designed to balance too little noise added to the training data (privacy not protected) and too much.

More traditional transformer-based models like BERT can also benefit from employing a Differential Privacy approach. As noted by Anil et al (2021), utilizing larger batch sizes in order to reduce the impact of introduced noise, and leveraging adaptive gradient clipping where the clipping bounds are adjusted based on observed behaviours from training data. Anil et al. (2021) also noted that risk to private data could be further mitigated by conducting pretraining on public datasets and then determining optimal hyperparameter setting based on this public data. However, such mitigation measures have their limitations, and Differential Privacy cannot guarantee meaningful privacy protection in all circumstances (Brown et al., 2022). A challenge for privacy protection is that some elements of privacy depend on the context in which it is being reviewed, making it difficult for simplistic assumptions. As Brown et al. (2022) recommend in their research, a stronger approach would be to train a language model solely on data explicitly intended for public dissemination.

Differential Privacy approaches can be further enhanced with elements of data sanitization, such as the Differential Private Data Dropout (DPDD) concept documented in the research article by Pal et al. (2024). At a high level DPDD works by removing memorized training examples from the training data. It accomplishes this by utilizing the formula for Renyi

Differential Privacy (RDP) found in the tensorflow-privacy package. The calculated RDP scores are then used to drop the high-risk data during training. Other strategies to continuously improve a model's performance at preventing sensitive data leakage include those descripted by Jain et al. (2023). In their research, the explored continuing to fine-tune a model based on vague or malicious prompts that produced a sensitive output. While this strategy cannot prevent the initial leak, this method allows for corrects to be made and to have the model generate a more neutral output or a version that has be sanitized of content.

Another promising strategy includes creating a Retrieval-Augmented Generation (RAG) model that will combine retrieval steps such as utilizing elastic search on specific documents to append to the prompt as context (the augment stage) before passing to fine-tuned LLM to generate the text based on the additional context provided. While not a true privacy-protecting mechanism, the additional context provided along with the prompt should mean there is less reliance on memorized information from the training data (Jin et al., 2024).

Finally, in the context of developing a RAG architecture for text summarization, the fine-tuning of the LLM can be improved with the following methods. As Bekman et al. (2025) research showed, the Arctic Long Sequence Training (ALST) is a fine-tuning method that works well with large datasets. ALST works by breaking up large sequences into manageable chunks without hitting memory limits, which is further supported by the use of Ulysses SP to split sequences across multiple GPUs. Low-Rank Adaptation (LoRA) is another fine-tuning method that Rehman et all (2025) has shown to perform well using commercially available GPUs. LoRA achieves a behaviour similar to full fine-tuning but at a fraction of the computing cost by ensuring the base LLM ranks remain unchanged but after training, the low-rank updates are merged.

In summary, the literature review has helped to uncover an emerging approach to exploring how might personal information be protected from being leaked in an abstractive summary through thoughtful planning of the preprocessing of the source data, then leveraging techniques such as DPDD to rank and remove high risk data before passing the training dataset through a RAG architecture where the LLM is fine-tuned through the ALTC or LoRA method and requested documents for summarization as passed through the model as additional context. Considerations for performance evaluation can be measured with the Canary Extraction Success Rate, ROUGE scores and human rating of generated text.

Project Data Source:

This project will be evaluating the data set available from the Town of Whitby public archives of past council and committee meetings. Documents will be downloaded as PDFs from the following location:

- https://www.whitby.ca/en/town-hall/council-meeting.aspx
- Dataset will include all posted minutes from past Town council and committee meeting from January 14, 2008 until June 3, 2025 (approximately 17 years of data).

Preprocessing steps:

Based on the insights gained from the literature review, the Town of Whitby council and committee minutes will be first process to identify and remove NER elements prior to utilizing the tensorflow_privacy package to conduct Differential Private Data Dropout (DPDD) and compute the risk scores before dropping outliers before fine-tuning the LLaMA 3 model with the LoRA method.

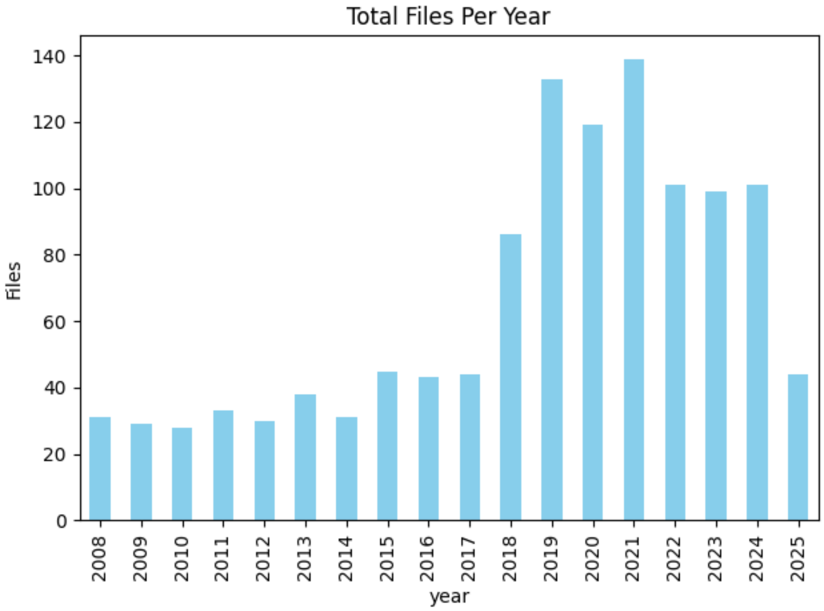
Exploratory Data Analysis

The Town of Whitby council and committee minute dataset from 2008 to 2025 contains:

- 1180 individual PDF files
- 10,941 pages in total

- A mean of 9.27 pages per file
- A median value of 6 pages.
- Total word count of 2,429,861 words across all files.
- A mean value of 2059.20 words per file
- A median value of 1101.5 words per file

Total Files Per Year



year	total_files	total_pages	total_words	mean_pages_per_file	mean_words_per_file	median_pages	median_words
2008	31	259	33049	8.354839	1066.096774	5.0	548.0
2009	29	256	46526	8.827586	1604.344828	4.0	612.0
2010	28	241	43582	8.607143	1556.500000	4.5	543.5
2011	33	284	22759	8.606061	689.666667	4.0	123.0
2012	30	274	22402	9.133333	746.733333	7.5	274.0
2013	38	355	82369	9.342105	2167.605263	5.5	1167.0
2014	31	276	66800	8.903226	2154.838710	4.0	592.0
2015	45	412	98255	9.155556	2183.444444	5.0	1014.0
2016	43	408	94507	9.488372	2197.837209	4.0	691.0
2017	44	419	98046	9.522727	2228.318182	4.5	984.5
2018	86	713	169073	8.290698	1965.965116	6.0	1314.0
2019	133	1243	306481	9.345865	2304.368421	6.0	1333.0

2020	119	1035	253481	8.697479	2130.092437	6.0	1412.0
2021	139	1355	339889	9.748201	2445.244604	6.0	1325.0
2022	101	1137	271668	11.257426	2689.782178	6.0	1200.0
2023	99	928	198228	9.373737	2002.303030	6.0	1044.0
2024	101	893	191491	8.841584	1895.950495	5.0	912.0
2025	44	399	83219	9.068182	1891.340909	5.0	922.0

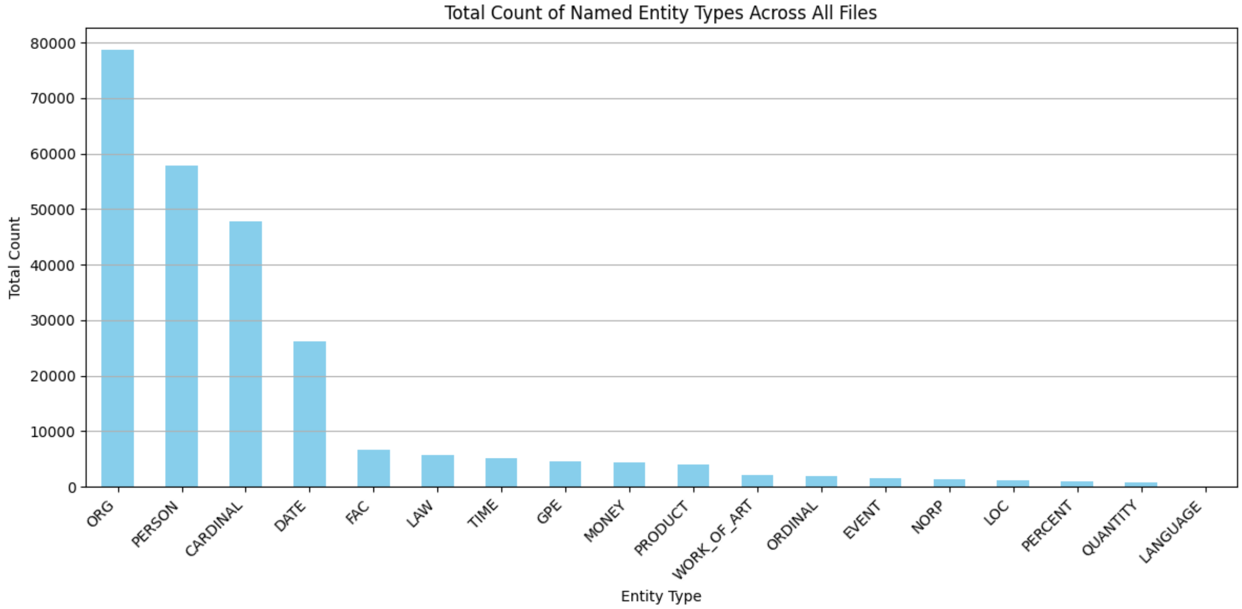
File Categories

- Accessibility Advisory Committee
- Active Transportation and Safe Roads Advisory Committee
- Animal Services Appeal Committee
- Audit Committee
- Brooklin Downtown Development Steering Committee
- Committee of Adjustment
- Committee of the Whole
- Compliance Audit Committee
- Downtown Whitby Development Steering Committee
- Heritage Whitby Advisory Committee
- Joint Accessibility Advisory and Diversity and Inclusion Advisory Committees
- Municipal Licensing and Standards Committee
- Property Standards Appeal Committee
- Public Meetings
- Regular Council Meetings
- Special Council Meetings
- Whitby Diversity and Inclusion Advisory Committee
- Whitby Sustainability Advisory Committee

Table: Named Entity Recognition (NER): First 5 rows

ORG	DATE	PERSON	TIME	GPE	LAW	CARDINAL	FAC	NORP	EVENT	WORK_OF_ART	PERCENT	PRODUCT	MONEY	ORDINAL	LOC	QUA	LANG
60	31	49	4	3	5	34	11	1	5	2	1	3	1	0	0	0	0
62	13	35	9	3	3	32	12	0	0	0	0	2	0	1	0	0	0
59	16	43	7	5	6	29	5	0	0	1	2	2	0	0	2	0	0
9	2	18	3	0	3	7	0	0	0	0	0	0	0	0	0	0	0
32	11	26	3	1	2	17	1	0	1	0	0	2	1	0	1	0	0

Histogram: NER Count Across All Files



Project Approach

The MRP will leverage a research methodology with the following phases:

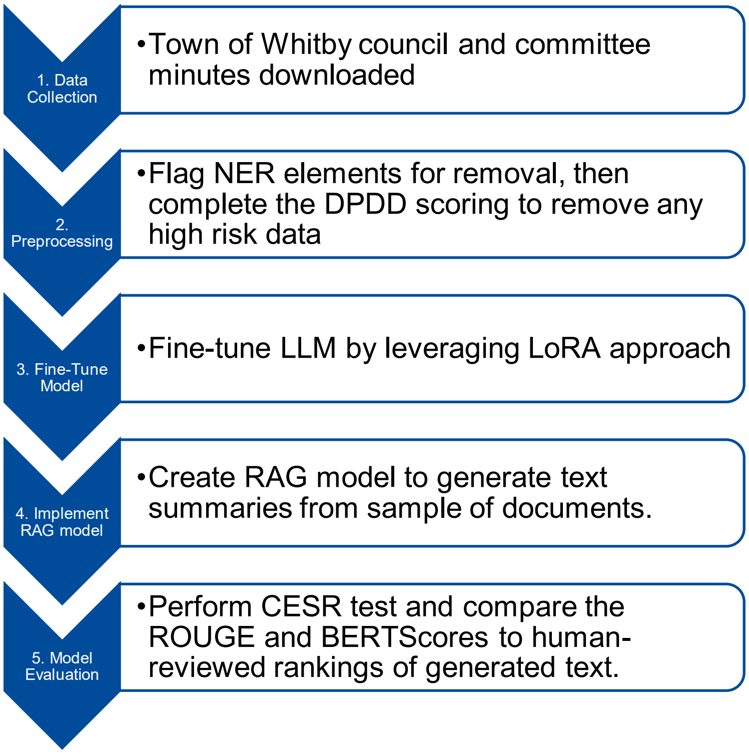
1. Data Collection and Processing:
   - All 1180 PDFs of past council and committee meeting minutes have been downloaded from the public archive on the Town of Whitby's website.
   - The full corpus of documents will be processes using Named Entity Recognition (NER) to flag PII elements to be removed.

   - Differential Private Data Dropout (DPDD) will score and remove sensitive data based on the calculated RDP scores (Carlini et al. (2021) and Moral-Gonzalez et al. (2025)).
   - Introduce “canaries” for post-training evaluation of model (Balde et al., 2025)

2. Fine-Tuning LLaMA 3 using Low-Rank Adaptation (LoRA) to minimize GPU usage while improving performance (Rehman et al., 2025).

3. Develop a Retrieval-Augmented Generation (RAG) model to help minimize hallucinations (Jin et al., 2024):
   - Retrieve relevant and requested documents using ElasticSearch
   - Append retrieved content to prompt to reduce reliance on memorized data

4. Evaluate Model Outputs:
   - Perform Canary Extraction Success Rate tests (Balde et al., 2025)
   - Calculate the ROUGE and BERTScores (El-Kassas et al., 2020)
   - Human ranking of sample of outputs for accuracy, relevance and privacy risk.



1. Data Collection
- Town of Whitby council and committee minutes downloaded

2. Preprocessing
- Flag NER elements for removal, then complete the DPDD scoring to remove any high risk data

3. Fine-Tune Model
- Fine-tune LLM by leveraging LoRA approach

4. Implement RAG model
- Create RAG model to generate text summaries from sample of documents.

5. Model Evaluation
- Perform CESR test and compare the ROUGE and BERTScores to human-reviewed rankings of generated text.

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

3. 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

4. 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

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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/arXiv.2301.09820

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. 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/arX

iv.2505.08253

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

27. 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

28. 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