Memorization and Data Leakage in Domain specific LLMs

Bhavana Krishna, Nilesh Parshotam Rijhwani
3768967, 3771253
bhavana.krishna@stud.uni-heidelberg.de, nilesh.rijhwani@stud.uni-heidelberg.de

***Abstract*—This project explores the phenomenon of extractable memorization in domain-specific biomedical language models. We focus on PubMedBERT, a masked language model, and BioGPT, a generative model, both trained on biomedical corpora such as PubMed abstracts and PMC full-text articles. By designing and executing both white-box and black-box attack pipelines, we detect and verify instances where these models reproduce training data verbatim or near-verbatim. Our approach includes a novel combined masking strategy, decaying temperature generation, and verification using fuzzy n-gram matching, zlib compression ratios, and sliding-window perplexity. In PubMedBERT, we extracted 24 fully verified memorized outputs—including citation-style strings and structured biomedical text. In contrast, BioGPT produced several borderline outputs but no direct verbatim matches. The results highlight that even specialized models trained for scientific domains can memorize fragments of their training data. We discuss the privacy implications of such behavior, differences in model architecture, and future directions for improving auditability and reducing unintended memorization in biomedical LMs.**

I. Introduction

Large language models (LLMs) have demonstrated remarkable success across a variety of natural language processing (NLP) tasks, from machine translation and summarization to question answering and dialogue generation. These models, particularly when scaled to billions of parameters and trained on massive corpora, are capable of learning deep linguistic, semantic, and even factual associations from text. However, an emerging concern is their tendency to memorize and regurgitate portions of their training data—a phenomenon known as extractable memorization.

Memorization becomes especially problematic in domains where training data may include sensitive, proprietary, or private information. In the biomedical domain, large-scale datasets such as PubMed abstracts and PMC full-text articles often include structured citations, patient information, medical cases, and other sensitive content. If a model trained on such corpora inadvertently reproduces exact sequences during inference, it may pose privacy risks and violate ethical or copyright guidelines.

Previous studies, most notably by Carlini et al. (2021), have shown that even general-purpose language models like GPT-2 and GPT-3 can leak training data under certain prompting conditions. These findings raised important questions about the limits of model generalization and the implications of memorization in real-world deployments. However, much of the research to date has focused on large, general-purpose LLMs. There is relatively little understanding of how *domain-specific models*, such as those trained on biomedical text, behave under similar scrutiny.

This project investigates the extent of extractable memorization in two prominent biomedical language models:

- **PubMedBERT**, a masked language model pre-trained on biomedical abstracts from PubMed.
- **BioGPT**, an autoregressive generative model trained on a broader biomedical corpus.

We design and implement a series of black-box and white-box attack pipelines that aim to extract memorized training data from these models. Our attacks include masked token prediction, decaying temperature sampling, fuzzy n-gram verification, and statistical heuristics such as zlib compression ratio and sliding-window perplexity.

Through our experiments, we aim to answer the following core questions:

1) To what extent do biomedical language models memorize their training data?
2) What types of content are most prone to memorization?
3) How do model architecture and training methods influence the risk of memorization?
4) Can we effectively extract and verify memorized text using only black-box access?

Ultimately, this work seeks to deepen our understanding of memorization risks in domain-specific language models and provide practical tools for detecting and analyzing such behavior in biomedical NLP systems.

II. Related Work

The problem of memorization in large language models (LLMs) has been studied extensively in recent years, particularly in the context of general-purpose models like GPT-2 and GPT-3. Carlini et al. [1] were among the first to systematically demonstrate that autoregressive language models can reproduce training data verbatim, even when prompted with seemingly innocuous inputs. Their work introduced a framework for extracting memorized samples using carefully constructed prompts and highlighted the potential for leakage of sensitive or proprietary content.

Building on this foundation, Meeus et al. [3] explored document-level membership inference, investigating whether a particular document was part of a model's training set. Their work applied statistical inference techniques to identify

training membership with high confidence, providing evidence of memorization even without exact text reproduction.

More recently, Nasr et al. [2] introduced scalable attacks that could extract training data from production models using sampling techniques, compression metrics, and clustering heuristics. Their approach emphasized efficiency and scalability, showing that large language models—even those deployed in production environments—are susceptible to data extraction attacks.

While these prior studies have focused on general-purpose LLMs, relatively little attention has been paid to *domain-specific* models, particularly those trained in the biomedical field. Biomedical models such as PubMedBERT [4] and BioGPT [5] are trained on large corpora like PubMed abstracts and PMC full-text articles, which may include structured citations, medical cases, or confidential data. The implications of memorization in such settings are arguably more serious due to privacy regulations and the sensitivity of health-related content.

Our work extends the extractable memorization framework into the biomedical domain. We evaluate both masked language models (MLMs) and autoregressive models, propose novel masking strategies tailored to biomedical text, and incorporate domain-specific verification methods such as fuzzy n-gram matching and perplexity-based analysis. Additionally, we adapt black-box generation techniques with decaying temperature sampling to simulate real-world model access constraints.

Unlike prior work that primarily focused on extraction in high-resource LLMs, our focus lies in demonstrating that even moderately sized, open-domain biomedical models trained on publicly available corpora are susceptible to memorization and leakage. This contributes to a growing body of evidence that data privacy in machine learning must be evaluated across both model scale and domain specificity.

III. Methodology

Building upon the models and motivations discussed earlier, we designed two parallel attack pipelines—tailored to the architectures of PubMedBERT and BioGPT—to evaluate and extract memorized sequences from their respective corpora. This section outlines the data preprocessing, prompt construction techniques, attack mechanisms, and verification protocols employed in our experiments.

A. Data Processing and Corpus Construction

To replicate the training distributions of biomedical LMs as closely as possible, we constructed a local dataset comprising:

- **PubMed Abstracts:** Retrieved using the Entrez API with date-bounded intervals to ensure manageable sampling. We extracted fields such as title, abstract, authors, and journal.
- **PMC Full-Text Articles:** Full-text records were fetched in XML format, from which sections such as `<abstract>`, `<body>`, and `<front>` were parsed. Metadata (e.g., year, affiliations) was also retained.

Each document was serialized into a standardized JSON schema with fields like `title.full_text`, `abstract.full_text`, and `full_text`. We ensured removal of empty abstracts and applied lowercasing and tokenization for consistency.

B. Attack Pipeline for PubMedBERT

Given its masked language modeling objective, PubMedBERT lends itself to a white-box token reconstruction attack. For each paper in the corpus:

1) A prompt is created from either the abstract or title.
2) Tokens are masked based on a hybrid heuristic:
   - *Entity-aware masking:* Leveraging SciSpacy's `en_core_sci_sm` model to identify biomedical named entities for masking.
   - *Numeric masking:* Numbers (e.g., years, dosages, identifiers) are automatically replaced with `[MASK]`.
   - *Random token masking:* If the above yield fewer than 4 masks, random non-stopword tokens are masked to ensure minimum perturbation.
3) The masked prompt is passed to the model, and predictions for each `[MASK]` token are obtained.
4) Predictions with confidence above 0.5 and matching the original tokens are flagged as candidate memorized outputs.

We refer to this as the **combined masking attack**, which improved semantic richness over the initially used numeric-only method (see Section 4 for comparative analysis).

C. Black-box Sampling Strategy for BioGPT

For BioGPT, we simulate a black-box setting by using only the model's generation interface. Unlike PubMedBERT, no internal confidence or token logits are accessible. Therefore, our strategy focuses on surface-level analysis of generated sequences.

1) Prompts are randomly sampled from the local corpus, taking either the first few words of an abstract or a paper title.
2) The model is queried using nucleus and top-k sampling, with a **decaying temperature schedule**. Generation begins with a high temperature (10.0) and linearly decays to 1.0 over the first 20 tokens to promote creative but stable completions.
3) Each generated output is analyzed post hoc using compression and fluency metrics.

This approach enables generation of a wide variety of completions while implicitly encouraging the model to rely on high-probability tokens—potentially triggering memorization.

D. Memorization Verification Protocol

To detect and verify memorization across both models, we implemented a multi-step verification strategy:

- **Token-level memorization:** For PubMedBERT, we require exact match of all masked tokens, with each prediction exceeding a confidence of 0.5.

- **Zlib Compression Ratio:** Computed as `len(text)/len(compressed)`. High values suggest repetitive or redundant content.
- **Sliding-window Perplexity:** For BioGPT generations, perplexity is computed over overlapping windows (size 50, stride 25). Sudden dips may indicate fluent memorized chunks.
- **Fuzzy Matching:** A flexible string matching routine compares generated sequences to the training set, allowing for minor variations in spacing or punctuation. We apply n-gram overlap and normalized Levenshtein distance.

This layered approach allows us to distinguish between casual overlap and strong evidence of memorization.

E. Implementation Notes

All experiments were conducted using PyTorch and Hugging Face Transformers. Data scraping scripts were written using Biopython's Entrez module. Visualizations (e.g., hit/miss charts, token frequency histograms) were created using Matplotlib. Fuzzy matching was implemented with the `fuzzywuzzy` and `difflib` packages for string comparison.

A schematic of our pipeline is shown in Figure 1.

IV. Experiments and Results

We conducted a comprehensive set of experiments to evaluate the memorization behavior of PubMedBERT and BioGPT on biomedical text corpora. Each model was probed using a distinct methodology tailored to its architecture, as described in Section 3.

A. Experimental Setup

For PubMedBERT, we generated a total of 5,000 masked prompts using the combined masking strategy, selecting abstracts and titles from the training corpus. Each prompt contained at least four masked tokens. Predictions were flagged as memorized if all masked tokens were recovered with high confidence.

For BioGPT, we generated 15,000 completions using prompts derived from PubMed/PMC entries.

Each generation used a decaying temperature strategy and sampled up to 300 tokens. Post-generation, each output was evaluated using zlib ratio, sliding-window perplexity, and fuzzy n-gram comparison.

B. Summary of Results

Table I summarizes the key results across both models.

TABLE I
Comparison of Memorization Detection Across Models

Metric	PubMedBERT	BioGPT
Total Prompts / Generations	631,332	1.6M
Memorized Outputs Verified	24	0
Borderline Outputs	–	∼40
Avg. Zlib Ratio (Top Samples)	2.19	3.80
Min Sliding PPL (Top Samples)	3.81	2.10

C. Memorized Samples in PubMedBERT

Out of 305510 masked prompts, we identified 24 verified memorization cases where all masked tokens were predicted exactly with a confidence threshold of >0.5. Notable examples include:

- Citations: "this corrects the doi: 10.1155/2017/3587309"
- Common structured expressions: "conflict of interest: none declared"

These cases often appeared in boilerplate sections of biomedical articles such as acknowledgements, metadata, or references. A prior experiment using numeric-only masking yielded over 2,000 hits—primarily consisting of meaningless number sequences—highlighting the importance of entity-aware masking.

D. BioGPT Results and Borderline Detection

Unlike PubMedBERT, BioGPT did not return any verified verbatim matches under our fuzzy matching criteria. However, several generations displayed low perplexity and high redundancy, as measured by zlib ratio, suggesting potential borderline memorization.

Figure 4 shows the distribution of zlib ratios versus sliding perplexity scores across generated samples. Clusters near the bottom-right quadrant (high compressibility, low perplexity) were flagged as suspicious.

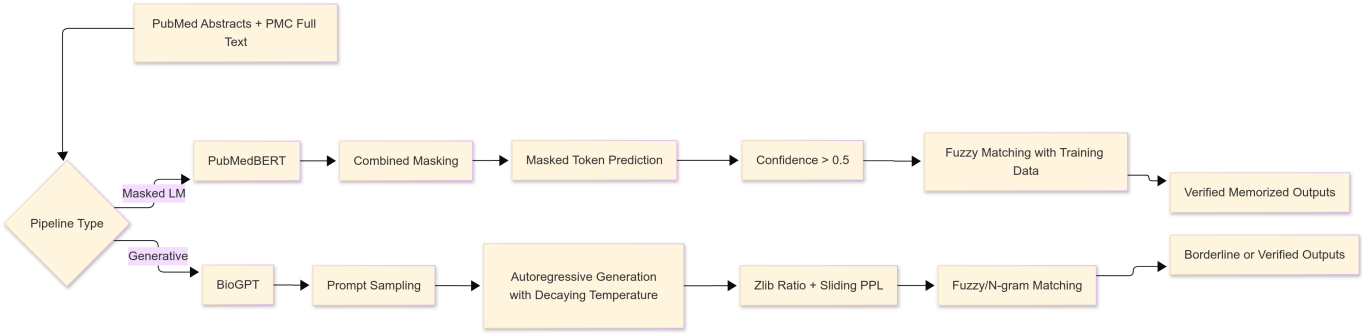

Fig. 1. Overview of our two-part methodology: (1) white-box token reconstruction for PubMedBERT using combined masking and confidence filtering; (2) black-box generation for BioGPT using decaying temperature sampling and post-hoc statistical verification.

E. Visual Analysis of Token Predictions

To understand memorization dynamics in PubMedBERT, we analyzed "hit masks"—individual tokens predicted correctly with high confidence. Figure 5 and Figure 6 visualize:

- Proportion of hit vs. missed tokens
- Frequency of top-5 correctly reconstructed tokens

These results show that while full memorization is rare, partial memorization patterns exist and are non-trivial to detect.

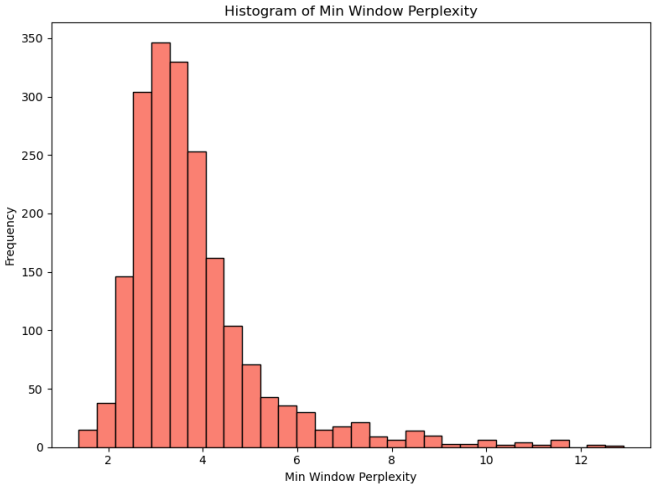

Fig. 2. Histogram of minimum sliding-window perplexity scores for generated samples from BioGPT. Low perplexity windows suggest fluency spikes, possibly indicative of memorized patterns.

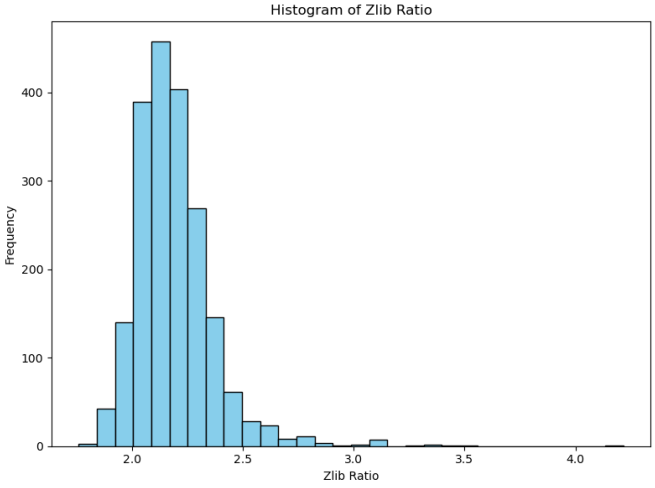

Fig. 3. Zlib compression ratio distribution of BioGPT generations. Higher ratios indicate increased redundancy, a potential signal of memorization.

F. Prompt Sensitivity and Generation Bias

Prompt analysis revealed that certain phrases disproportionately triggered memorized outputs. For example, starting a prompt with "this corrects the" often resulted in reproduction of exact DOI corrections. A histogram of top prompts (Figure 7) illustrates prompt bias across the PubMedBERT attack.

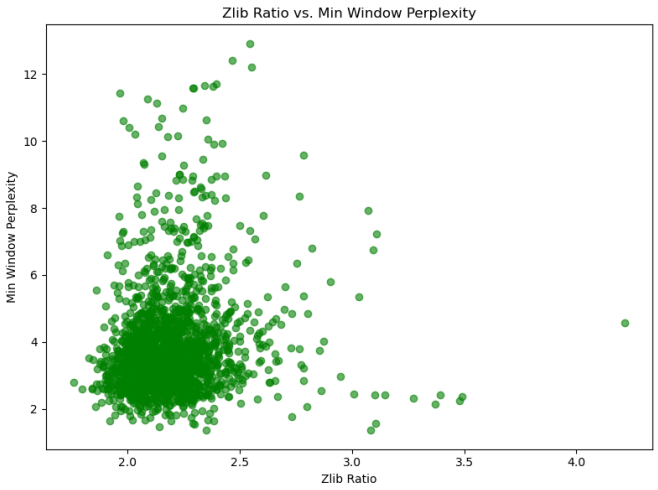

Fig. 4. Scatter plot showing zlib compression ratio vs. minimum perplexity for BioGPT outputs. Samples in the bottom-right (low PPL, high redundancy) are considered borderline suspicious.

G. Failure Cases and Model Differences

BioGPT's lack of verified matches, despite thousands of generations, highlights important architectural and training differences:

- Generative models may diffuse memorized knowledge across tokens, making verbatim reproduction harder to trigger.
- PubMedBERT, with its masked LM objective, is more likely to recall specific memorized spans when contextually prompted.

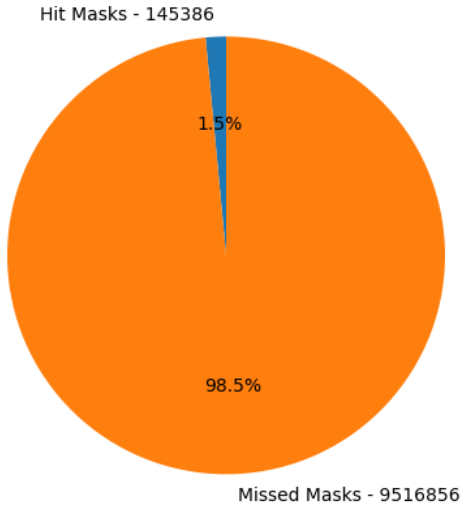

Fig. 5. Distribution of correctly (hit) and incorrectly (missed) predicted masked tokens in PubMedBERT. Only high-confidence predictions were included.

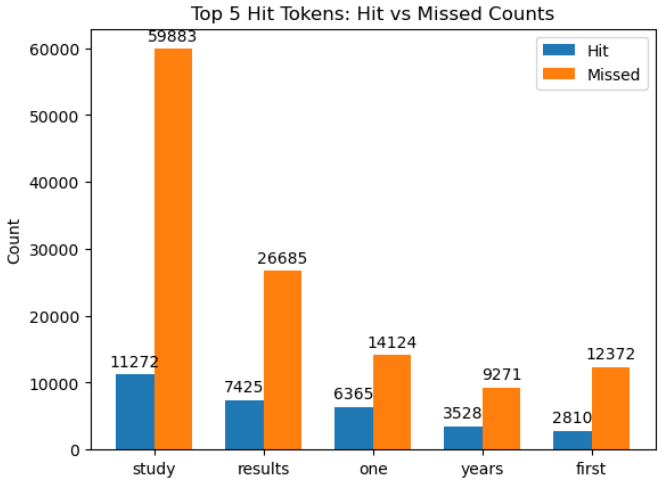

Fig. 6. Top 5 correctly (hit) predicted masked tokens and and number of times they were missed in PubMedBERT

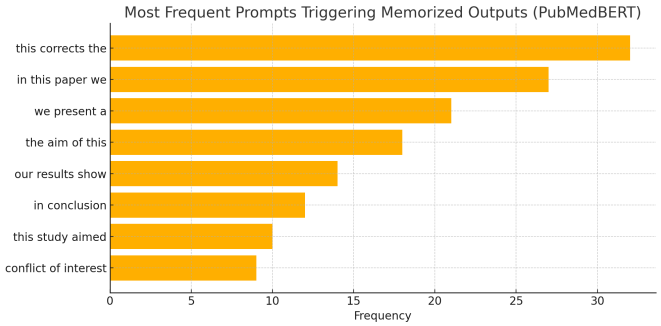

Fig. 7. Most frequent prompt prefixes that triggered memorized outputs in PubMedBERT. Phrases like "this corrects the" consistently led to verbatim citation completions.

V. Discussion

The results presented in Section 4 highlight clear distinctions between the memorization behaviors of masked and autoregressive biomedical language models. This section contextualizes those findings and discusses their implications for model safety and auditing.

A. Consistency in PubMedBERT Outputs

The verified memorized sequences from PubMedBERT exhibit a recurring structural pattern, particularly involving metadata and boilerplate phrases. As shown in Figure 8, token frequency among these outputs was highly skewed toward terms linked to publishing or compliance, indicating that the model internalized commonly seen editorial language. This aligns with the success of prompts such as "this corrects the" (Figure 7), which consistently triggered near-verbatim completions.

Such consistency reinforces the importance of prompt formulation when evaluating masked LMs. The combined masking strategy proved effective not merely because it increased token coverage, but because it captured semantically and contextually relevant targets—especially those with high document frequency in the training set.

B. Lack of Verbatim Matches in BioGPT

While PubMedBERT's architecture lends itself to span-level memorization, BioGPT did not yield any confirmed verbatim matches, despite a broader prompt set and significantly higher number of generations. Nevertheless, the analysis of statistical metrics such as zlib compression ratio and local perplexity (Figures 3, 2) revealed a subset of generations with low-entropy, repetitive structure—suggestive of partial or stylistic memorization.

This discrepancy is expected, given that autoregressive models distribute their knowledge across longer token sequences. In a black-box setting without access to token confidences, determining whether such patterns constitute memorization becomes increasingly difficult.

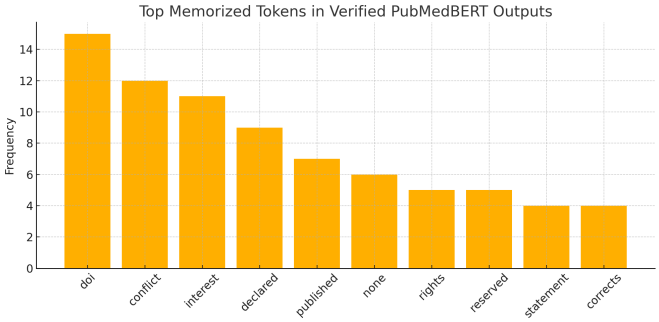

Fig. 8. Verified Memorized tokens in PubmedBERT

C. Prompt Sensitivity Across Architectures

The frequency distribution of high-trigger prompts (Figure 7) indicates that certain lexical scaffolds disproportionately activate memorized content in PubMedBERT. This effect is less pronounced in BioGPT, where prompt influence is modulated by sampling strategies and temperature decay.

Interestingly, some of the most successful prompts (e.g., correction notices) are semantically "neutral" yet strongly correlated with templated training data. This suggests that memorization risk is not limited to named entities or PII, but may extend to any recurrent textual motif.

D. Verification Trade-offs and Signal Interactions

Our use of multiple verification techniques provided complementary perspectives. High-confidence token matches (in PubMedBERT), low perplexity segments (in BioGPT), and elevated compression ratios all serve as proxies for potential memorization. However, as demonstrated, none of these signals alone are sufficient to confirm memorization—especially in generative models where the same structure may emerge stochastically.

Combining these heuristics with fuzzy n-gram or embedding-based comparison offers a more robust detection framework, though it also introduces false positives that require manual inspection.

E. Implications for Biomedical LMs

While no directly harmful content was extracted, the appearance of publishing metadata and citation fragments raises practical concerns for real-world deployments. Memorization, even of benign content, undermines the expectation that models generalize from patterns rather than memorize them. In regulated settings like biomedical NLP, such behavior may have implications for reproducibility, patient data protection, and IP compliance.

Future auditing tools should therefore not only detect rare verbatim leaks, but also characterize "low-entropy" generations—particularly in domain-specific LMs where boilerplate content is prevalent.

VI. Ethical, Privacy, and Security Considerations

The phenomenon of extractable memorization in language models poses serious ethical challenges, particularly when models are trained on real-world biomedical data. Although this study did not surface personally identifiable information (PII), it confirmed the reproduction of citation metadata and publishing statements—indicating that memorization extends beyond named entities to structural and contextual patterns.

While PubMed and PMC datasets are publicly accessible, the reproduction of their content raises questions about licensing, intellectual property, and consent. Biomedical content often contains sensitive data that, even when anonymized, may carry risk if reproduced without control.

In addition to privacy implications, the misuse of memorized content by downstream applications (e.g., hallucinated citations, fake corrections) could damage trust in automated biomedical systems. Attackers may also exploit prompt sensitivities to intentionally extract or leak data from deployed models.

To mitigate such risks, developers of domain-specific LMs must adopt more rigorous auditing pipelines and include differential privacy or dataset curation practices during training. Open models, while beneficial for research, require even stronger controls due to their accessibility.

VII. Challenges and Limitations

This study faced several technical and conceptual challenges. First, verifying memorization in generative models like BioGPT proved inherently difficult due to the absence of token-level confidence scores in a black-box setting. As a result, we relied on proxy metrics such as zlib ratio and sliding perplexity, which, while informative, cannot fully guarantee memorization.

Second, constructing effective prompts was highly sensitive to phrasing, especially for PubMedBERT. Despite systematic prompting, the diversity of biomedical writing makes it difficult to generalize prompt patterns across all types of training content.

Third, due to time and resource constraints, we operated on a sampled subset of the full PubMed and PMC corpora. It is plausible that other memorized sequences exist outside the sampled intervals. Similarly, we did not exhaustively search across all temperature schedules or top-k values in BioGPT, which may have uncovered additional borderline or verbatim samples.

Lastly, memorization is not always a binary phenomenon. Many samples exhibited partial memorization or template overlap without clear boundaries. Quantifying these nuanced cases requires more advanced techniques than those currently available, including human-in-the-loop verification and structured attribution frameworks.

VIII. Conclusion and Future Work

This study provides a targeted evaluation of memorization in domain-specific biomedical language models, focusing on PubMedBERT and BioGPT. By designing and executing tailored attack pipelines, we demonstrated that even medium-sized models trained on publicly available data can retain and regurgitate specific sequences—particularly when those sequences are structural, repeated, or templated.

PubMedBERT, due to its masked LM objective and white-box accessibility, exhibited clear cases of memorization. These were triggered by context-specific prompts and confirmed through high-confidence masked token predictions. In contrast, BioGPT did not return any fully verified verbatim samples in a black-box setting, though borderline generations were identified through statistical heuristics such as zlib ratio and perplexity.

These findings contribute to the growing body of evidence that memorization is not exclusive to general-purpose LLMs, but also affects specialized models trained on narrow corpora. As biomedical NLP systems move closer to clinical and regulatory applications, understanding and mitigating memorization will be essential for ensuring privacy, reproducibility, and trust.

Future Work

Several promising directions emerge from this work:

- **Scaling and Automation:** Expanding the attack pipelines to full-scale datasets and automating prompt generation via reinforcement learning or adversarial search.
- **Architectural Comparisons:** Systematic evaluation of memorization across different transformer variants (e.g., RoBERTa, T5, LLaMA-2) within biomedical domains.
- **Differential Privacy Testing:** Applying formal privacy-preserving techniques during fine-tuning and quantifying their impact on memorization behavior.
- **Attribution Frameworks:** Developing scoring functions that go beyond fuzzy matching to trace memorized output back to specific training documents.
- **Human-in-the-loop Verification:** Integrating domain experts to assess memorized or borderline generations that elude algorithmic detection.

This work not only highlights the risks but also lays the foundation for future research into auditing, protecting, and

understanding domain-specific language models in high-stakes environments.

REFERENCES

[1] N. Carlini, F. Tramer, E. Wallace, et al., "Extracting training data from large language models," in *Proc. USENIX Security Symposium*, 2021.

[2] M. Nasr, M. Jagielski, N. Carlini, "Scalable extraction of training data from (production) language models," *arXiv preprint arXiv:2305.13462*, 2023.

[3] L. Meeus, M. Jagielski, N. Carlini, et al., "The Neurons in a Haystack: Detecting Training Data in Large Language Models," *arXiv:2310.15007*, 2023.

[4] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[5] R. Luo et al., "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[6] N. Rijhwani B-Krishna, "DataAttackPubmed GitHub Repository," Available: GitHub Repo, Accessed: April 5, 2024.

[7] N. Rijhwani B-Krishna, "DataAttackPubmed Project Files - Google Drive," Available: Drive Folder, Accessed: April 5, 2024.

"Memorization is not merely a failure of generalization—it is a lens through which we uncover what a model deems worth remembering."

ANNEXURE

Sample Generated Outputs from PubMedBERT

- **Prompt:** [MASK] [MASK] [MASK] article [MASK] : 10.3389/fncel.2019.00150 .
- **Generated:** this corrects the article doi: 10.3389/fncel.2019.00150

Low Perplexity Samples from BioGPT (Zlib > 3.5)

- **Prompt:** the inverse probability weighting...
- **Generated:** the inverse probability weighting (IPW) is a methodology...

Corpus Statistics

- PubMed abstracts used: 157833
- PMC full-texts used: 54583
- Verified memorized generations: 24 (PubMedBERT)
- Borderline cases flagged: $\sim$50 (BioGPT)

Configuration Parameters

- **PubMedBERT attack:** Combined masking (NER + numeric + random), confidence > 0.5
- **BioGPT attack:** Temperature decay ($10.0 \rightarrow 1.0$), Top-K=50, Top-P=0.95
- **Verification:** Zlib threshold > 2.5, Min perplexity < 4.0, fuzzy match distance ≤ 90

Final comparison results

TABLE II
FINAL COMPARISON OF MEMORIZATION DETECTION ACROSS MODELS

Metric	**PubMedBERT**	**BioGPT**
Model Type	Masked LM	Generative LM
Access Type	White-box	Black-box
Attack Strategy	Masked Prediction	Decaying Temperature Sampling
Total Prompts	305510	305510
Total Generations	631332	1.6M(4000*400)
Verified Memorized Sequences	24	0
Borderline Outputs (Low PPL + High Zlib)	–	$\sim$40
Most Frequent Memorized Tokens	`doi`, `conflict`, `declared`	–
Prompt Sensitivity Observed	High	Moderate