Investigating Privacy Concerns and Mitigations for Large Language Models in Healthcare

Jenny Chim

December 2024

1 Introduction

This report describes work undertaken during the internship project on "Investigating Privacy Concerns and Mitigations for Healthcare Language and Foundation Models" at NHS England from July to December 2024.

1.1 Motivation

Natural language processing (NLP) has long been applied to textual data in healthcare, supporting applications including analysis, decision-making, and personalised care provision. Recent advances have introduced Large Language Models (LLMs), deep learning models typically based on the transformer architecture [47] trained on vast corpora through self-supervision, which surpass their predecessors in both scale and capability. These models hold transformative potential in healthcare, with use cases including, inter alia, new modes of interacting with electronic medical records [18] and clinical text generation [46], which can help alleviate the well-known documentation burden among healthcare professionals [3, 48].

However, the adoption of LLMs warrants caution. Along with other limitations such as their generating plausible-sounding but factually incorrect information [25], amplifying biases [19], and introducing information clutter [35], LLMs can also bring privacy risks [44]. *This project examines the privacy implications of healthcare LLMs, focussing on data memorisation and leakages in AI-assisted information sharing.*

1.2 Work Packages

This project is split into three work packages (WP), with outputs (O) and main learnings (L) summarised below.

- **WP1 (Memorisation)** involves reviewing relevant literature on how memorisation manifests and is captured in language data.
  - **O1:** Code for memorisation-related metrics.

- **WP2 (Mitigation)** involves running experiments to assess memorisation of instruction-tuning data, using approaches mapped in WP1.
  - **O2a:** Code to compile instruction-tuning dataset.
  - **O2b:** Code to finetune model on instruction-tuning.
  - **O2c:** Code to test memorisation, including prefix-suffix, using an LLM-attacker, likelihood-based methods.
  - **L2:** Current membership inference attack methods are ineffective at distinguishing between actual clinical notes seen in training and notes not seen in training.
- **WP3 (Contextual Privacy)** involves creating a framework to test privacy leakages in clinical documentation workflows, focussing on an ambient AI note generation scenario. *Note: this section is currently redacted due to publication guidelines, but will be available in future report versions.*

The repository containing code for this project is at: `https://github.com/nhsengland/pvt_p71_privLMextended`.

2 Memorisation

This section surveys perspectives on data memorisation related to privacy.

Overview Neural NLP models demonstrate significant data memorisation during training. Research has found that LLMs can reproduce training data both verbatim [8, 10] and approximately [23, 11]. Text reproduction increases with model size [38], training data duplication [26], presence of outliers in deduplicated training sets [7], and use of probabilistic decoding [30]. Membership inference attacks (MIA) can determine whether specific data was used in training [43], although at the scale of LLMs, existing attacks for earlier pretrained language models (e.g. BERT [14]) seem to be ineffective [15], possibly due to the common practice of training LLMs on a huge corpus over one epoch rather than multi-epoch training on a (relatively) small corpus, which was a commonly taken approach for training earlier models.

What makes data more susceptible to memorisation? Factors contributing to memorisation and mitigation effectiveness are an ongoing area of research. The likelihood of a data point being memorised is impacted by its *exposure frequency* and *predictability*.

- *Exposure*: Text sequences repeatedly exposed to a model are more likely to be memorised, for example when entire documents or substrings are found across multiple training data points. Deduplication helps mitigate this type of memorisation [24, 26].
- *Predictability*: Text sequences comprising patterns, such as ordered lists or numbered table cells, are inherently more predictable and more susceptible to memorisation [40]. Outliers are also more likely to be memorised [24].

The position of a data point in the training set does not seem to impact memorisation [5].

How does memorisation manifest? Once a model memorised data, it can reproduce the information in literal and non-literal ways [12, 30]. *Literal* or verbatim reproduction includes the model generating direct identifiers seen in training, such as names, phone numbers, and account numbers. *Non-literal* reproduction can occur in the same level of detail as the original sequence, i.e. as *paraphrases*, for example using different words to express that a specific patient has received a diagnosis. Non-literal reproduction can also occur more abstractively at an *idea* level, for example generating a short story that follows the high-level plot of an existing long novel instead of writing an original story from scratch, generating patent ideas similar to existing patents, or generating a short description of a patient's hospital course having seen parts of their records during training.

How is memorisation measured? Methods for estimating memorisation rely on the assumption that models trained to predict text sequences will assign higher likelihoods to, and therefore be more likely to generate, sequences previously seen in training.

Assuming *access to model outputs*, researchers have proposed to identify whether a data point has been used to train a model, i.e. perform membership inference attacks (MIA), by computing the loss of the target sequence directly [50], as well as calibrating it with the loss computed under a reference model, with compression size [10], and with the losses of similar but previously unseen sequences [34]. Others have explored token probabilities; for example, min-k% prob captures the difference in token probability distributions between previously seen and unseen examples under the trained model, by computing the average log likelihood of the k% tokens with the minimum probabilities in the target sequence [42].

When there is only *black-box access* to models, researchers typically estimate memorisation via extractability: intuitively, we know that a model has knowledge of a sequence if we can extract the sequence by interacting with the model, and that given a starting text segment, a model that has seen the full text during training is more likely to follow the segment with content similar to the real continuation. How much 'starting' context to provide and what constitutes sufficient similarity varies. Some require exact matches [6, 4], whereas others require the texts to be similar above a threshold, relying on metrics such as the Levenshtein distance ratio [39] and ROUGE-L [32, 27]. We implement and experiment with several of these introduced approaches in Section 3.

Implications for memorisation of clinical texts From a utility perspective, concerns related to memorisation include model over-fitting, learning spurious correlations rather than actual knowledge, and inflating of benchmark results from memorising test examples leaked into the training corpus (i.e. data contamination [33]). More relevant to the current project, data memorisation has privacy implications, as naively training models on sensitive healthcare data may lead to their regurgitation in deployment. Such risks are compounded by fairness concerns. As mentioned, during training, outliers are more likely to be memorised by models; and in post-hoc mitigation strategies such as machine unlearning, techniques applied to protect the most vulnerable subset can jeopardise the next-most vulnerable subset [7]. Thus, in both model training and post-hoc mitigations, data points constituting the minority group are likely to be the most susceptible to privacy-related risks from memorisation.

Challenges to assessing memorisation. Prior work studying memorisation in the pre-trained language model BERT [14] found that template-based infilling and probing were ineffective at revealing links between patients and conditions in clinical notes [31]. More recently, it was found that existing MIA techniques are ineffective on LLMs, due to large model sizes, the prevalent practice of training on vast datasets over few iterations, and fuzzy boundaries between instances

seen and not seen during training, i.e. member and non-member instances[16]. As will be discussed in Section 3, this challenge with fuzzy boundaries is particularly pronounced in the clinical domain. Clinical texts often exhibit repetitions and redundancies [41], stemming from naturally similar structures (e.g. section headings, standardised language for documentation, automatically populated sections such as test results) and longitudinal record-keeping practices (e.g. patient demographic information and family history, sections carried over from previous notes when the information remains accurate and relevant).

3 Towards Mitigating Data Memorisation

For us to address potential privacy risks from memorisation, we must first identify appropriate metrics that can distinguish between previously seen and unseen data, and preferably, quantify the degree of memorisation. To this end, we fine-tuned models based on LLaMA-3 [45] on a dataset constructed from MIMIC discharge summaries. However, we found that current membership inference attack methods are ineffective even at the basic task of differentiating between records actually present in the training set and those that are absent from it. The remainder of this section describes our experiments, our negative results, possible factors contributing to the metrics' ineffectiveness, and our learnings.

> **Takeaway:** We find that current NLP metrics are ill-suited to capturing memorisation, and *fail to* distinguish between data seen in instruction-tuning from unseen data, if the seen and unseen are from the same distribution (e.g. discharge summaries from the same institution).

3.1 Method

Dataset When we worked on this work package during the internship, there were limited instruction-tuning datasets based on real clinical texts.[1] We therefore gathered pre-existing annotated datasets in clinical NLP that use MIMIC-III or MIMIC-IV discharge summaries as its input, the majority of which hosted on PhysioNet [21]. This allowed us to create a multi-task instruction tuning dataset that contains instances with multiple annotations (e.g. same input summary with human-annotated or human-validated annotations for multiple tasks) that can be used in future projects. Table 1 summarises the tasks and datasets.

> The code to reproduce the dataset construction process is found in the project repository under `data_processing/`.

Although our accompanying code can create an instruction tuning dataset spanning all described tasks in Table 1, our subsequent experiments involve:

- **A single-task setup** comprising 1.2K records for the task of generating discharge instructions given preceding sections in a discharge summary.
- **A two-task setup** comprising 1.2K records, doubly annotated for classification (30 day out-of-hospital mortality) and for discharge instruction generation (i.e. same as the single-task setup above).

[1] There has since been releases of similar resources addressing this need, such as [49].

NLP Task	Clinical Task	Source
Generation	Brief Hospital Course (BHC) → Discharge Instruction (DI) [1]	IV
	Discharge Summary without DI → DI [1]	IV
	DI with → without unsupported facts [22]	IV
Extraction	Rationale extraction for clinical codes [13]	III
	Medication extraction [20]	IV
	Clinician action item extraction [37]	III
Classification	Social behavioural determinants of health [2]	III
	30 day out-of-hospital mortality [51]	IV
	Patient phenotype/indication [36]	III
QA	Question answering [28]	IV

Table 1: Compiled MIMIC discharge summary-based instruction tuning dataset.

Models We work with models based on LLaMA-3-8B [17]:

(a) The *base model* off-the-shelf.

(b) *FT-LLaMA-3-8B*: (a) fine-tuned on our compiled instruction data.

(c) *Asclepius-3-8B*: (a) fine-tuned on synthetic examples [29].

(d) *FT-Asclepius-3-8B*: (c) fine-tuned on our compiled instruction data.

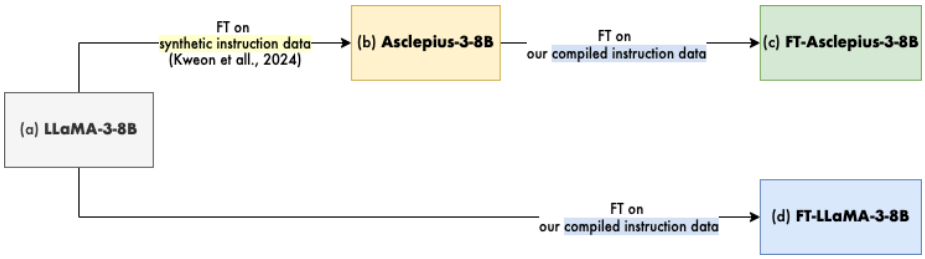

Figure 1: Models used in WP2. We assess the effectiveness of metrics in distinguishing instruction-tuning examples that were seen and unseen by each model.

Metrics We explore methods prior work has used to gauge memorisation. For each note x that we want to see if a model LM has memorised, we compute:

- *LOSS*: The negative log likelihood of x given by the model LM [50].
- *Min-k% Prob*: The average negative log likelihood of the k% most likely tokens in x [42].

- *Prefix-suffix*: Splitting x into two parts – a prefix and a suffix – and asking the LM to continue generating from the prefix, then comparing the similarity between the real suffix and model-generated suffix [9]. We use ROUGE-L [32] as our similarity metric, and always take the first third of x as its prefix and the remaining two-thirds as its suffix [27].

Across these methods, the intuition is that since language models are trained to predict tokens given its context (e.g. complete a text given its prefix), if LM has seen x before, it will assign a higher likelihood to x and generate a completion that is similar to the real suffix, compared to previously unseen notes.

To account for cases where some text sequences are inherently more likely than others (e.g. due to using more 'natural' expressions or appearing more frequently in the training corpus), we also explored *calibrated* versions of these metrics. This involves normalising the likelihood scores with likelihood scores under a reference model, and normalising the ROUGE-L scores with ROUGE-L computed with completions by the reference model.

Experiments We evaluate the metrics' effectiveness at revealing memorisation by comparing their scores on training (seen) and test (unseen) data. Due to resource constraints, we perform our experiments on a subset of the dataset, randomly sampling without replacement 100 notes from the training split as positive instances and 100 notes from the test split as negative instances.

We fine-tune the base LLaMA model (a) for 5 epochs, saving a checkpoint after each epoch. We then apply the metrics on seen and unseen data to see whether there is a significant difference between the two groups. We repeat the process on a model (c) that was previously already fine-tuned from LLaMA on a separate dataset of synthetic notes.

> Details and code for metrics, experiments, and analyses can be found in the project repository under `memorisation/`.

3.2 Results

None of the examined metrics were successful at distinguishing between previously seen and unseen examples in the first three epochs. This applies to models directly fine-tuned from base LMs (Fig. 2) and those fine-tuned from an already tuned LM (Fig. 3), with and without calibrating scores from reference models.

The prefix-suffix method is ineffective throughout, whereas likelihood-based methods begin to differentiate between seen and unseen data by the fourth epoch. Two-sided permutation tests (n=10,000, α=.05) performed between scores for seen and unseen examples at each epoch for each model corroborate these trends. However, given the size of the dataset used in our experiments (a magnitude or two smaller than common instruction tuning datasets),[2] by this point the model is already at risk of over-fitting. We argue that if the metrics

[2]With notable exceptions such as [52].

are unable to identify seen examples in this simplified setting, it is even more unlikely that they will be effective in realistic and more challenging applications involving larger datasets and fewer fine-tuning epochs.

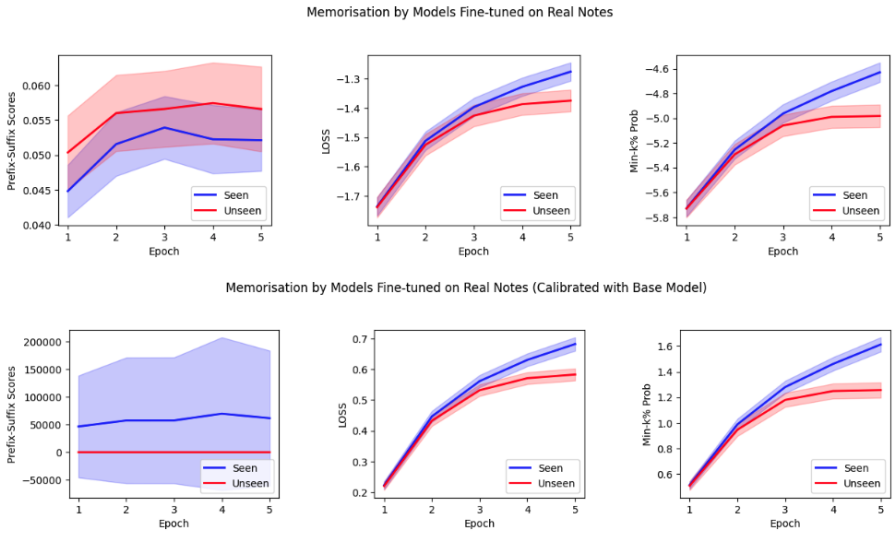

Figure 2: Metric scores on notes seen and unseen by a model fine-tuned from LLaMA, over 5 epochs of instruction tuning. Top row shows raw scores, bottom row shows calibrated values using base LLaMA as the reference model.

In addition, when we ran the metrics on LMs which previously underwent a round of fine-tuning on synthetic data (d), we found consistent significant differences between scores on previously-seen synthetic data and current real data, across all epochs, but did not find differences between seen and unseen real data in the first few epochs. What this indicates is that, while the metrics are picking up certain differences in data distributions, instead of revealing whether an LM has seen a document before, they reflect the textual (e.g. stylistic) differences between synthetic notes and real notes, and thus are not very useful for assessing memorisation.

3.3 Discussion

Our results are in line with recent work on membership inference attacks that find existing methods ineffective at detecting pretraining data from LLMs, particularly when applied to data from the same domain [16]. This is exactly our case, as our dataset involves notes of the same type from the same institution in a constrained time range.

Current metrics are ill-suited for estimating memorisation – and by extension privacy risks introduced by memorisation – in the clinical domain. For one, clinical texts are long and contain natural repetitions [41], including formatted

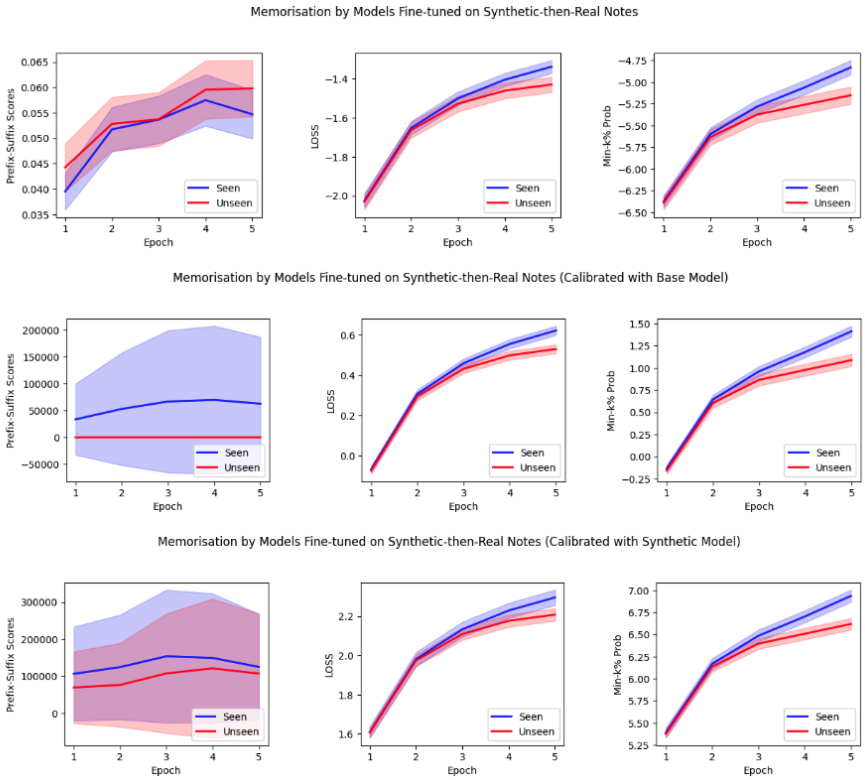

Figure 3: Metric scores on notes seen and unseen by a LLaMA-based model that already underwent a round of instruction tuning on synthetic notes, over 5 epochs of instruction tuning. Top to bottom: raw scores, scores calibrated with base LLaMA, scores calibrated with finetuned-LLaMA.

texts (e.g. section headings), meaningful repetitions of important information over time (e.g. routine test results, ongoing information about disease progression), and redundancies (e.g. repeated sections from automatically populated fields).

These characteristics introduce ambiguity to data group membership, since they can inflate the similarity between seen and unseen examples. For example:

- Isolated encounters may be documented in similar language despite them describing completely different patients, by nature of adhering to standardised clinical practices. An LM that has been trained on records from patient A may assign high likelihood to records from patient B even if it has genuinely never seen patient B's data before.

- A finetuned LM which takes the prefix of a note and completes the rest of the note in a way that is similar to the original continuation might not have

actually memorised the original note. Instead, it may have just learned to generate notes that adhere to standard documentation practices.

On the other hand, relying on current (document-level) metrics can underestimate memorisation. Even if a model fails to faithfully reconstruct a full text or assign significantly higher likelihood to it, so long as it generates/can be used to recover *fragmented* information (e.g. a short unique identifier nested in a long text, or seemingly benign information that can be pieced together over record sequences to compromise identity), it already poses privacy risks.

All in all, these shortcomings motivate developing metrics that (1) work over longer sequences over time, (2) consider notions of similarity beyond exact string matching, (3) model different granularities of information, and (4) capture relationships among different pieces of information which differ in importance and sensitivity under varied circumstances.

4 Privacy in Context

This section is currently redacted to abide by publication guidelines. The content will be made public in future report versions.

5 Future Work

We outline several directions for future work.

Memorisation & Mitgation

- Explore memorisation in alignment (e.g. direct preference optimisation (DPO), reinforcement learning with human feedback (RLHF)).
- Improving metrics to assess privacy risks from memorisation that are more nuanced and granular, e.g. fact-based analyses.

Contextual Privacy

- Examining information leakage detection methods' generalisability.
  - We currently use a cosine similarity threshold of 0.4 for embedding similarity, based on manual inspection. Is this still the right threshold for other scenarios? Can we develop a way to automatically calibrate this threshold?
  - We use manually written queries for semantic search. In new settings, how would query generation work? Do more specific queries yield higher leakage recall? How can we ensure that this semantic search method is robust?
  - We currently use a pre-defined taxonomy of information type and subjects. How can we extend our methodology to catching the diverse information types and relationships in real clinical interactions?
- Investigating potential model biases in two-step generate-then-edit pipelines.
- Developing diverse synthetic transcript generation strategies using open weight LLMs to simulate varying information-sharing attitudes, patient non-compliance, and transcription noise reflecting real-life deployment.
- Developing computationally efficient methods, including evaluation of smaller expert models as privacy moderators. From a practical perspective, lightweight models hold particular promise, as they can be more easily deployed within local infrastructure, providing access to in-house guidelines and requirements while enabling sensitive data processing without external dependencies or unnecessary data transmission.

References

[1] A. Aali, D. Van Veen, Y. Arefeen, J. Hom, C. Bluethgen, E. P. Reis, S. Gatidis, N. Clifford, J. Daws, A. Tehrani, J. Kim, and A. Chaudhari. Mimic-iv-ext-bhc: Labeled clinical notes dataset for hospital course summarization, 2024.

[2] Hassaan Ahsan, Eoin Ohnuki, Avishek Mitra, and Hong Yu. MIMIC-SBDH: A dataset for social and behavioral determinants of health. In *Proceedings of Machine Learning Research*, volume 149, pages 391–413, 2021.

[3] B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W. J. Tuan, C. A. Sinsky, and V. J. Gilchrist. Tethered to the ehr: Primary care physician workload assessment using ehr event log data and time-motion observations. *Annals of Family Medicine*, 15(5):419–426, 2017.

[4] Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Gregory Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[6] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022.

[8] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and others. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805*, 2020.

[10] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

[11] Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation, 2024.

[12] Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15134–15158, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. MDACE: MIMIC documents annotated with code evidence. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186, Minneapolis, Minnesota, 6 2019. Association for Computational Linguistics.

[15] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*, 2024.

[16] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22021–22030, 2024.

[19] Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. Can ai relate: Testing large language model response for mental health support, 2024.

[20] Akshay Goel, Almog Gueta, Omry Gilon, Sofia Erell, and Amir Feder. Medication extraction labels for MIMIC-IV-Note clinical database. *PhysioNet*, 2023.

[21] Ary L Goldberger, Luis A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220, 2000.

[22] Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*, 2024.

[23] Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics.

[24] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1345–1362. USENIX Association, August 2020.

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[26] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.

[27] Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. Alpaca against vicuna: Using llms to uncover memorization of llms. *arXiv preprint arXiv:2403.04801*, 2024.

[28] Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. EHRNoteQA: An LLM benchmark for real-world clinical practice using discharge summaries. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[29] Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. Publicly shareable clinical large language model built on synthetic clinical notes. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[30] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647, 2023.

[31] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. Does BERT pretrained on clinical notes reveal sensitive data? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online, June 2021. Association for Computational Linguistics.

[32] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out*, Barcelona, Spain, 2004.

[33] Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[34] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages

11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics.

[35] Liam G. McCoy, Arjun K. Manrai, and Adam Rodman. Large language models and the degradation of the medical record. *New England Journal of Medicine*, 391(17):1561–1564, 2024.

[36] Ethan Moseley, Leo Anthony Celi, Jianying Wu, and Franck Dernoncourt. Phenotype annotations for patient notes in the MIMIC-III database. *PhysioNet*, 2020.

[37] James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online, August 2021. Association for Computational Linguistics.

[38] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *CoRR*, abs/2311.17035, 2023.

[39] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[40] USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, et al. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*, 2024.

[41] Thomas Searle, Zina Ibrahim, James Teo, and Richard Dobson. Estimating redundancy in clinical text. *Journal of biomedical informatics*, 124:103938, 2021.

[42] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[43] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[44] Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. Identifying and mitigating privacy risks stemming from language models: A survey, 2024.

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[46] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, 2024.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[48] C. P. West, L. N. Dyrbye, and T. D. Shanafelt. Physician burnout: contributors, consequences and solutions. *J. Intern. Med.*, 283(6):516–529, June 2018. Epub 2018 Mar 24.

[49] Zhenbang Wu, Anant Dadu, Michael Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[50] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2017.

[51] WonJin Yoon, Shan Chen, Yanjun Gao, Zhanzhan Zhao, Dmitriy Dligach, Danielle S Bitterman, Majid Afshar, and Timothy Miller. Lcd benchmark: Long clinical document benchmark on mortality prediction for language models. *Under review (Preprint: medRxiv)*, pages 2024–03, 2024.

[52] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.