# 大規模言語モデルに対する サンプリングを活用したメンバーシップ推論攻撃

綿 祐貴 [1]　金子 正弘 [2,1]　Youmi Ma[1]　岡崎 直観 [1]
[1] 東京工業大学　[2] MBZUAI
{yuki.wata@nlp.,youmi.ma@nlp.,okazaki@}c.titech.ac.jp
Masahiro.Kaneko@mbzuai.ac.ae

## 概要

本研究は，与えられたテキストがモデルの学習データに含まれていたかを判定するメンバーシップ推論攻撃に取り組む．従来手法は，モデルが計算する尤度を必要としており，適用できるモデルが限られる．そこで，本研究では出力テキストだけから検出する**サンプリングベース・メンバーシップ推論攻撃**を提案する．提案手法は検出対象のテキストを参照テキスト，サンプルされたモデルの複数出力を候補テキストとし，それらの一致度合を計算し，テキストがモデルの学習データに含まれていたか判定する．提案手法は尤度を利用しないにも関わらず，実験では既存手法と肩を並べる性能を発揮し，特に長いテキストを対象とした検出で高い性能を示した．

## 1　はじめに

大規模言語モデル（Large Language Models; LLM）の学習コーパスの規模が拡大するにつれ，GPT-4 [1] や PaLM 2 [2] などの開発者は，組織の競争力を維持するため，学習データの準備や出典などの詳細の公表を控えるようになった．その結果，事前学習に用いた文書が分からないため，LLM が生成したテキストが剽窃にあたるか判断できず，著作権者および利用者の双方にリスクが生じうる [3]．更に，評価用のベンチマークが LLM の学習データに含まれる場合，モデルの性能を適切に評価できない [4, 5, 6]．

本稿が取り組むタスクは，**メンバーシップ推論攻撃**（Membership Inference Attacks; MIA）[7] である．MIA タスクは，検出対象のテキストとモデルが与えられた際に，対象テキストがモデルの学習データに含まれていたかを判定するものである．一般に，モデルはデータに適合するように学習されるため，学習データに含まれるテキストは含まれないテキストよりも高い尤度を示す [8, 9]．MIA の既存研究はこのアイディアに基づくため，モデルの尤度が計算できることを前提としており [8, 10, 11, 12, 13]，尤度を提供しないモデルには適用できない．一方，尤度を出力できるモデルの多くは学習データが公開されており [14, 15, 16]，MIA を使わずとも対象テキストが含まれるかを直接確認できる．そのため，MIA を適用したい対象は学習データが非公開のモデルであるにも関わらず，そのような LLM の多くは尤度を出力しないため [1, 2, 17, 18]，MIA を実際に利用できる場面が限られている．

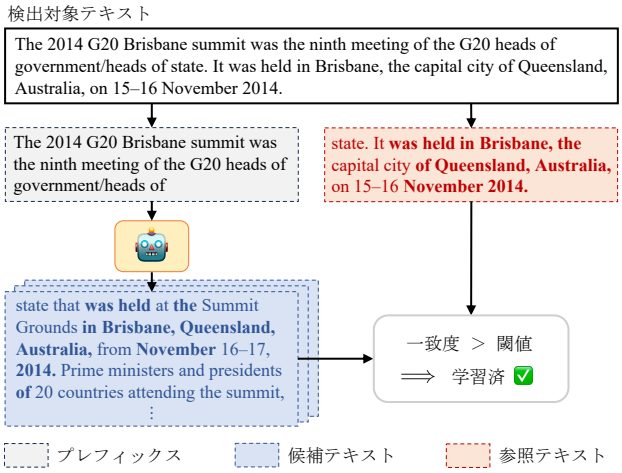


**図 1**　サンプリングベース・メンバーシップ推論攻撃

本稿では，尤度に依存しない MIA として**サンプリングベース・メンバーシップ推論攻撃**（**S**ampling-based **M**embership **I**nference **A**ttacks; **SaMIA**）を提案する．学習データが検出対象を含む場合，LLM は学習した正解テキストをそのまま出力すれば良いため，対象テキストと生成されたテキストは表層の一致率が高い [10, 19]．図 1 は SaMIA の具体的な検出手順を示している．対象テキストの冒頭部分を LLM に与え，その続きをサンプリングにより複数生成する．そして，生成された系列を候補テキス

ト，対象テキストの先頭以降の系列を参照テキストとみなし，それらのテキストの単語の一致度合を計算する．一致度合が閾値以上であれば対象テキストが LLM に学習されたと判定する．

SaMIA の有効性を検証するため，事前学習済みモデルに対し，本手法の検出性能を既存手法と比較した．その結果，ROUGE-1 [20] と zlib 圧縮エントロピー [21] を併用した提案手法は尤度を利用していないが，既存手法と肩を並べる性能を発揮し，特に長いテキストを対象とした検出で高い性能を示した．分析では，グラム長，候補テキストの数，検出対象のテキスト長が SaMIA の性能に与える影響を調査した．SaMIA の性能は単語ユニグラムのときに高く，候補テキスト数の増加に伴い改善され，テキストが長いほど効果的であるという知見が得られた．

## 2 サンプリングのみを活用した MIA

### 2.1 MIA タスクの定義

MIA は，モデル $f_\theta$ の学習データセット $\mathcal{D}_{\text{train}}$ に対象テキスト $x$ が含まれるか否かの二値分類タスクである．攻撃者の目標は，適切な攻撃関数 $A_{f_\theta} : \mathcal{X} \to \{0,1\}$ を設計し，テキスト空間 $\mathcal{X}$ 中の事例 $x$ に対し $x \in \mathcal{D}_{\text{train}}$ の真偽を判定することである．

### 2.2 SaMIA

本稿では，LLM からサンプルした出力だけを活用する SaMIA を提案する．本手法はモデル $f_\theta$ が損失 $\mathcal{L}$ やトークン尤度 $P_\theta$ を提供しない前提にあるため，より厳しい利用条件のもとにある．提案手法は損失と尤度を利用しないため，任意の LLM に対して適用できる．具体的な方法は，長さ $n$ の検出対象のテキスト $x = (w_1, w_2, \ldots, w_n)$ を単語数に応じて前半と後半に分割し，前半を LLM に与えるプレフィックス $x_{\text{prefix}} = (w_1, w_2, \ldots, w_{\lfloor n/2 \rfloor})$，後半を参照テキスト $x_{\text{ref}} = (w_{\lfloor n/2 \rfloor+1}, w_{\lfloor n/2 \rfloor+2}, \ldots, w_n)$ として使用する．LLM は $x_{\text{prefix}}$ に続くテキストをサンプリングにより $m$ 個生成し，これらを候補テキスト $x^j_{\text{cand}} (j = 1, \ldots, m)$ として検知に用いる．

LLM は学習した事例をそのままの形で漏洩する可能性がある [10, 19]．そのため SaMIA は，候補テキスト $x^j_{\text{cand}}$ と参照テキスト $x_{\text{ref}}$ の表層的な一致率が高い場合，元のテキスト $x$ が LLM の学習データに漏洩していると考える．テキスト間の一致率の評価指標には，参照テキストの単語の再現率である ROUGE-N [20] を利用する．LLM が生成した候補テキスト $x_{\text{cand}}$ と，参照テキスト $x_{\text{ref}}$ が与えられたとき，ROUGE-N $\in [0,1]$ は式 1 に従って計算される．

$$\text{ROUGE-N}(x_{\text{cand}}, x_{\text{ref}}) = \frac{\sum_{\text{gram}_n \in x_{\text{ref}}} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{gram}_n \in x_{\text{ref}}} \text{Count}(\text{gram}_n)} \tag{1}$$

ここで，$n$ は n-gram の長さを表す．また，数式の分母は参照テキストの n-gram の総数であり，分子は候補テキストと参照テキストに共起する n-gram の総数を表す．例えば ROUGE-1 は，参照テキスト中の単語が LLM に生成されるほど高い値となる．

SaMIA は，LLM が生成した各候補テキスト $x^j_{\text{cand}}$ と参照テキスト $x_{\text{ref}}$ の間の ROUGE-N を計算し，それら $m$ 個の平均が閾値 $\tau$ を超えるテキスト $x$ は学習データに含まれると判定する（式 2）．

$$A_{f_\theta}(x) = \mathbb{1}\left[\frac{1}{m}\sum_{j=1}^{m} \text{ROUGE-N}(x^j_{\text{cand}}, x_{\text{ref}}) > \tau\right] \tag{2}$$

この検出指標の解釈は直接的で，テキスト $x$ が学習に用いられた場合，LLM は参照テキスト中の n-gram を多く生成するという仮説に基づく．SaMIA の疑似コードを付録 B の Algorithm 1 にまとめる．

### 2.3 サンプルの情報量を利用した改善

既存手法である PPL/zlib では生成テキストの冗長性の特徴を利用して評価するために，zlib が計算する情報量を用いた [10]．未学習データのサンプルは繰り返し生成（例：“I love you. I love you...”）を含む傾向があり，そのようなサンプルの zlib 圧縮後の情報量が小さくなると考えられる．PPL/zlib は，サンプル $x$ のパープレキシティと zlib 圧縮後のビット数 $\text{zlib}(x)$ の比率を検出指標する（式 3）．

$$A_{f_\theta}(x) = \mathbb{1}\left[\frac{\prod_{i=1}^{n} P_\theta(x_i | x_{1:i-1})^{-\frac{1}{n}}}{\text{zlib}(x)} < \tau\right] \tag{3}$$

ここで，$\text{zlib}(x)$ はテキスト $x$ を zlib 圧縮したときのエントロピーをビット数で表す．

$\text{zlib}(x)$ はテキスト $x$ の文字情報のみに依存する指標であるため，SaMIA にも適用できる．また，SaMIA（式 2）は候補テキスト $x^j_{\text{cand}}$ における繰り返し生成の有無は考慮できないため，zlib との併用により性能改善が期待できる（式 4）．

$$A_{f_\theta}(x) = \mathbb{1}\left[\frac{1}{m}\sum_{j=1}^{m} \text{ROUGE-N}(x^j_{\text{cand}}, x_{\text{ref}}) \cdot \text{zlib}(x^j_{\text{cand}}) > \tau\right] \tag{4}$$

**表 1**　提案手法と比較手法の AUC スコア

| | GPT-J-6B | | | | OPT-6.7B | | | | Pythia-6.9B | | | | LLaMA-2-7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 単語数 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | **Avg.** |
| LOSS | 0.64 | 0.62 | 0.67 | 0.69 | 0.61 | 0.57 | 0.62 | 0.64 | 0.64 | 0.61 | 0.65 | 0.68 | **0.55** | 0.50 | 0.56 | 0.59 | 0.62 |
| PPL/zlib | 0.65 | 0.63 | 0.68 | 0.69 | 0.61 | 0.58 | 0.64 | 0.65 | 0.64 | 0.62 | 0.67 | 0.70 | **0.55** | 0.51 | 0.57 | 0.59 | 0.62 |
| Lowercase | 0.59 | 0.57 | 0.58 | 0.60 | 0.58 | 0.57 | 0.57 | 0.59 | 0.59 | 0.55 | 0.57 | 0.55 | 0.49 | 0.50 | 0.49 | 0.59 | 0.56 |
| Min-K% Prob | **0.67** | **0.66** | **0.70** | 0.71 | **0.62** | 0.60 | 0.67 | 0.67 | **0.66** | 0.64 | **0.69** | 0.71 | 0.51 | 0.50 | 0.56 | 0.58 | 0.63 |
| SaMIA | 0.54 | 0.60 | 0.64 | **0.77** | 0.56 | 0.63 | 0.69 | **0.82** | 0.54 | 0.63 | 0.65 | 0.73 | 0.52 | 0.52 | 0.58 | 0.64 | 0.63 |
| SaMIA*zlib | 0.55 | 0.63 | 0.67 | 0.75 | **0.62** | **0.69** | **0.74** | 0.81 | 0.57 | **0.66** | 0.67 | **0.75** | 0.54 | **0.56** | **0.60** | **0.66** | **0.65** |

# 3　実験

## 3.1　比較手法

SaMIA との比較に用いる既存手法を説明する．**LOSS** [8] は最もシンプルな MIA であり，サンプル $x$ の損失（負の対数尤度）$\mathscr{L}$ が閾値 $\tau$ より小さい場合に学習データに含まれると判定する（式 5）．

$$A_{f_\theta}(x) = \mathbb{1}[\mathscr{L}(f_\theta, x) < \tau] \tag{5}$$

**PPL/zlib** [10] では，テキスト $x$ の zlib 圧縮エントロピーを活用し，$x$ のパープレキシティとの比率を検出指標とする（式 3）．**Lowercase** [10] は LOSS を拡張した手法の 1 つで，対象テキスト $x$ を小文字化した $x_{\text{lower}}$ の損失との差を検出指標とする（式 6）．

$$A_{f_\theta}(x) = \mathbb{1}[\mathscr{L}(f_\theta, x) - \mathscr{L}(f_\theta, x_{\text{lower}}) < \tau] \tag{6}$$

LOSS はテキスト $x$ の全トークンを用いて検出する一方，**Min-K% Prob** [13] は $x$ の中で尤度の低い $k$% トークン Min-K%($x$) のみを検出に用いる（式 7）．

$$A_{f_\theta}(x) = \mathbb{1}\left[\frac{1}{E}\sum_{x_i \in \text{Min-K\%}(x)} \log P_\theta(x_i|x_{1:i-1}) > \tau\right] \tag{7}$$

ここで，$E = |\text{Min-K\%}(x)|$ は選出されたトークンの総数を表す．Min-K% Prob のハイパーパラメータである $k$ の値は原著論文で推奨された $k = 20$ とする．

## 3.2　実験設定

**データセット**　既存研究 [13] に倣い，本稿ではベンチマークに WikiMIA[1] を用いる．MIA の性能はテキスト長に依存するため [13]，異なる単語数[2]（32, 64, 128, 256）の WikiMIA に対して検出性能を評価する．WikiMIA は，Wikipedia から収集されたイベントページにより構成される．2023 年以降のイベントを未学習データ，2017 年以前のイベントを学習済データとする．イベントページは特定の時期に関連する情報であるため，未学習データは事前学習されていない新しい情報であることが保証される．また，Wikipedia は事前学習データの一般的な情報源であり，学習済データは事前学習に使われたとする．

**モデル**　提案手法および既存手法の性能を，GPT-J-6B [16]，OPT-6.7B [14]，Pythia-6.9B [15]，LLaMA-2-7B [22] を用いて評価する．これら 4 つの LLM は，事前学習データのカットオフ日が 2022 年 9 月以前であるため，WikiMIA の使用要件を満たす．

**評価指標**　既存研究 [12, 13] に倣い，検出方法の有効性を真陽性率（True Positive Rate; TPR）と偽陽性率（False Positive Rate; FPR）を用いて評価する．各閾値 $\tau$ における TPR と FPR を用いて ROC 曲線を描画し，ROC 曲線下の領域面積である AUC と低 FPR における TPR（TPR@10 %FPR）を評価指標とする．

**実装の詳細**　SaMIA（式 2）と SaMIA*zlib（式 4）には ROUGE-1 を用い，サンプル数は $m = 10$ と設定した．事前学習済みモデルは HuggingFace[3] で公開されているものを使用した．トークン生成時のパラメータは全て，`temperature=1.0`，`max_length=1024`，`top_k=50`，`top_p=1.0` に統一した．

## 3.3　実験結果

AUC の評価結果を表 1 に示す．表の右端の **Avg.** には行の平均値を掲載している．SaMIA*zlib の AUC は，256 単語の WikiMIA において全ての比較手法を上回ったが，32 単語においては性能が低かった．また，zlib は SaMIA を改善しており，サンプルの冗長性の考慮が有効であることを確認できる．更に，**Avg.** では SaMIA*zlib の性能が一番高いことから，提案手法は総合的に最も性能が優れていると言える．TPR@10%FPR の結果は付録 A の表 2 に記載しており，AUC と似た傾向を示している．これらの

1）https://huggingface.co/datasets/swj0419/WikiMIA
2）WikiMIA の各事例は，スペース区切りで指定の単語数となるよう切られている．
3）https://huggingface.co/

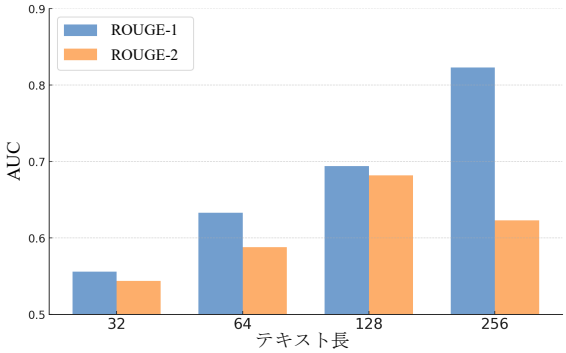


**図 2**　ROUGE-1 vs. ROUGE-2（モデル：OPT-6.7B）

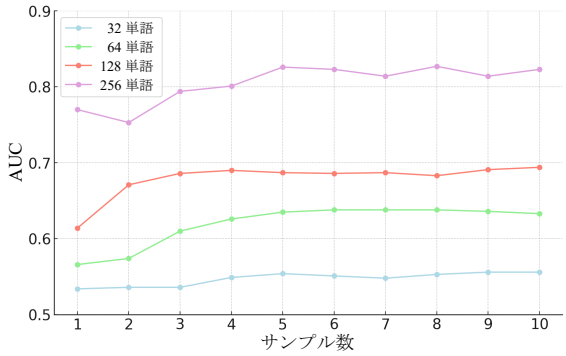


**図 3**　サンプル数 $m$ と SaMIA の性能（モデル：OPT-6.7B）

結果から，SaMIA は尤度や損失を用いずとも，それらを用いた検出手法を上回る，もしくは匹敵する性能を達成できることが分かる．

## 4　分析

SaMIA の性能に影響を与える要因としてグラム長，サンプル数，テキスト長の観点から分析する．

**グラム長**　主要な結果である表 1 には，ROUGE-1 を検出指標に用いた SaMIA の性能を掲載した．単語バイグラムの方が検出しやすくなるかを検証するため，ROUGE-2 を指標に用いた SaMIA（式 2，N = 2）の性能を調べた．図 2 より，ROUGE-2 による検出は ROUGE-1 よりも劣る結果となり，単語ユニグラムを用いた方が良いという知見が得られた．

**サンプル数**　表 1 では，LLM が生成した 10 サンプルを用いた SaMIA の性能を報告した．ここで，サンプル数 $m$ を変化させた際の SaMIA の性能への影響を調査する．直感的には，多くのサンプルがより堅牢な比較を提供すると考えられる．図 3 の結果は，この仮説の正当性を裏付けており，サンプル数 $m$ を増やすことで性能が改善することを示している．しかし，サンプル数が 5 を超えると，全てのテキスト長において検出性能は横ばいとなるため，推論コストを考慮しても $m = 5$ が最適である．

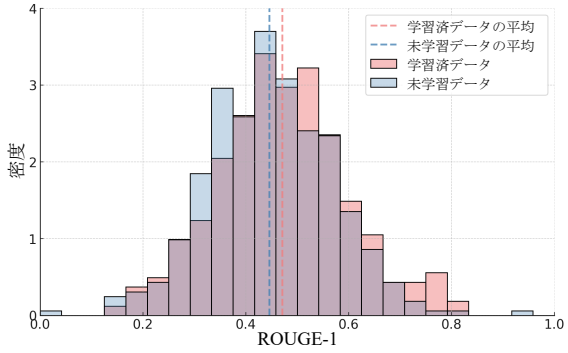


**図 4**　単語数 32 の WikiMIA の ROUGE-1 分布（OPT-6.7B）

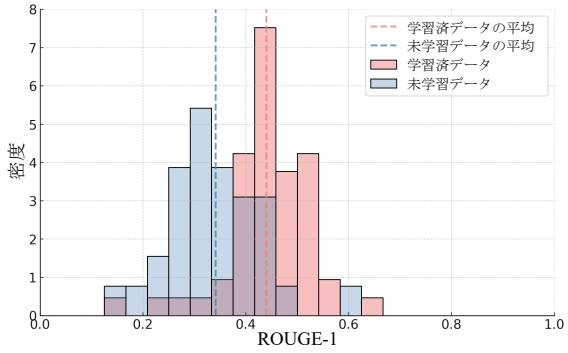


**図 5**　単語数 256 の WikiMIA の ROUGE-1 分布（OPT-6.7B）

**テキスト長**　表 1 によると，SaMIA の性能は対象テキストの長さに依存している．そこで，テキスト長が ROUGE-1 の検出指標（式 2，N = 1）に与える変化をより詳細に分析する．異なる単語数（32，256）における未学習データと学習済データの ROUGE-1 分布を図 4，5 に示す（64，128 単語における分布は付録 A の図 6，7 に載せる）．この図が示すように，長いテキストを対象とした検出では，事前学習の有無が ROUGE-1 に大きな差を生んでいる．この現象は直感的にも正しく，(1) モデルに与えるプレフィックス $x_{\text{prefix}}$ が長いとき，モデルは記憶を特定しやすくなり，(2) 参照テキスト $x_{\text{ref}}$ が長い場合，正解に関する情報が増えて評価が安定すると考えられる．

## 5　おわりに

本稿では，メンバーシップ推論攻撃において尤度に依存せず，サンプリングのみを活用する手法 SaMIA を提案した．本手法を WikiMIA で評価したところ，既存手法と同等の性能を示し，特に長いテキストを対象とした検知では高性能を示した．この結果は，尤度を提供しない LLM に対しても，メンバーシップ推論が可能であることを示唆している．

今後は，短いテキストに対する SaMIA の検知性能の改善を目指したい．

# 謝辞

本研究は JSPS 科研費 19H01118 の助成を受けたものです。

# 参考文献


[1] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.

[2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

[3] Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[4] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. Did chatgpt cheat on your test?, 2023. https://hitz-zentroa.github.io/lm-contamination/blog/.

[5] 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)**, pp. 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Arvind Narayanan and Sayash Kapoor. Gpt-4 and professional benchmarks: the wrong answer to the wrong question, 2023. https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In **2017 IEEE Symposium on Security and Privacy (SP)**, pp. 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society.

[8] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In **31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018**, pp. 268–282. IEEE Computer Society, 2018.

[9] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In **Proceedings of the 36th International Conference on Machine Learning**, pp. 5558–5567, 2019.

[10] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In Michael D. Bailey and Rachel Greenstadt, editors, **30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021**, pp. 2633–2650. USENIX Association, 2021.

[11] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In **Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security**, pp. 3093–3106, 2022.

[12] 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**, pp. 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] 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 **Proceedings of the Neural Information Processing Systems (NeurIPS)**, 2023.

[14] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

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

[16] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[17] James Manyika and Sissie Hsiao. An overview of bard: an early experiment with generative ai, 2023.

[18] Anthropic. Model card and evaluations for claude models, 2023.

[19] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[21] Jean loup Gailly and Mark Adler. zlib compression library, 1995.

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

# A　更なる実験結果

提案手法と比較手法の TPR@10%FPR による評価結果を表 2 に示す．表の右端の **Avg.** には行の平均値を掲載している．

**表 2**　提案手法と比較手法の TPR@10%FPR

| | GPT-J-6B | | | | OPT-6.7B | | | | Pythia-6.9B | | | | LLaMA-2-7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 単語数 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | 32 | 64 | 128 | 256 | **Avg.** |
| LOSS | 17.7 | 16.3 | 15.3 | 22.6 | 15.9 | 14.7 | 17.1 | 19.4 | 16.5 | 18.2 | 24.3 | 22.6 | 14.1 | **11.6** | 13.5 | 16.1 | 17.2 |
| PPL/zlib | 18.5 | 15.5 | 18.9 | 32.3 | 17.5 | 13.6 | 16.2 | 22.6 | 17.0 | 16.7 | 18.9 | 25.8 | 14.9 | **11.6** | **15.3** | 22.6 | 18.6 |
| Lowercase | 15.2 | 16.7 | 18.9 | 16.1 | 13.9 | 13.6 | 17.1 | 19.4 | 17.5 | 14.7 | 19.8 | 16.1 | 10.8 | 7.8 | 14.4 | **29.0** | 16.3 |
| Min-K% Prob | **31.0** | **28.2** | **24.5** | 23.5 | **20.9** | 26.4 | 24.5 | 31.4 | **28.9** | **25.4** | **32.4** | 19.6 | 8.0 | 9.9 | 15.1 | 9.8 | 22.5 |
| SaMIA | 10.9 | 16.9 | 12.9 | 23.5 | 12.9 | 16.9 | 27.3 | **49.0** | 14.5 | 18.3 | 18.7 | 27.5 | 11.9 | 8.1 | 10.8 | 11.8 | 18.2 |
| SaMIA*zlib | 12.9 | 20.4 | 22.3 | **35.3** | 17.6 | **27.8** | **38.8** | 47.1 | 17.8 | **25.4** | 20.1 | **33.3** | **17.6** | 10.9 | 14.4 | 15.7 | **23.6** |

WikiMIA の 64，128 単語のデータセットにおける，未学習データと学習済データの ROUGE-1 分布を図 6，7 に示す．

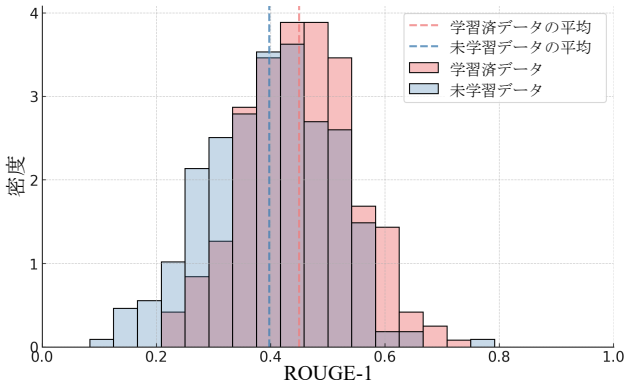


**図 6**　単語数 64 の WikiMIA の ROUGE-1 分布（OPT-6.7B）

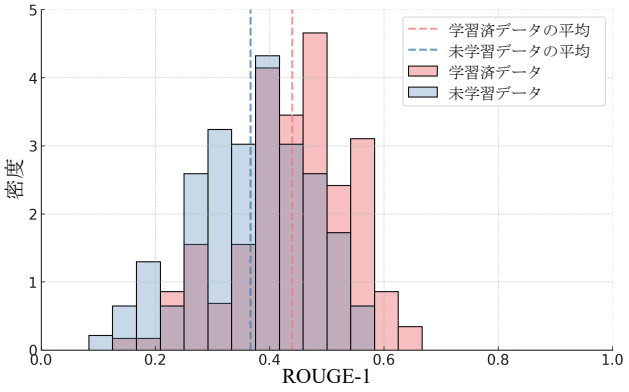


**図 7**　単語数 128 の WikiMIA の ROUGE-1 分布（OPT-6.7B）

# B　SaMIA の詳細

**Algorithm 1** Sampling-based Membership Inference Attacks

**Input:** 検出対象テキスト $x = (w_1, w_2, \ldots, w_n)$，言語モデル $f_\theta$，サンプル数 $m$，グラム長 N，閾値 $\tau$
**Output:** テキスト $x$ はモデル $f_\theta$ の学習済データか否か
$x_{\text{prefix}} = (w_1, w_2, \ldots, w_{\lfloor n/2 \rfloor})$
$x_{\text{ref}} = (w_{\lfloor n/2 \rfloor+1}, w_{\lfloor n/2 \rfloor+2}, \ldots, w_n)$
**for** $j = 1$ **to** $m$ **do**
　　テキスト $x^j_{\text{cand}} \leftarrow f_\theta$ を用いた $x_{\text{prefix}}$ の後続の推論結果
**end for**
$\overline{\text{R}}_m = \frac{1}{m} \sum_{j=1}^{m} \text{ROUGE-N}(x^j_{\text{cand}}, x_{\text{ref}})$
**if** $\overline{\text{R}}_m > \tau$ **then**
　　**return** 学習済である
**else**
　　**return** 未学習である
**end if**