機械生成文検出とメンバーシップ推論は相互に転移可能

小池隆斗 [1,2*] Liam Dugan[2*] 金子正弘 [3] Chris Callison-Burch[2] 岡崎直観 [1]
[1] 東京科学大学 [2] ペンシルバニア大学 [3]MBZUAI
ryuto.koike@nlp.comp.isct.ac.jp

概要

本稿では，メンバーシップ推論攻撃 (MIA) と機械文検出の「転移性」，つまり一方のタスクの手法が他方でどれほど有効かを，理論的かつ実験的に明らかにする．理論的には，両タスクの漸近的な最適指標が同一であることを示す．さらに，多くの既存手法がこの最適指標の近似として捉えられ，その近似精度が転移性に寄与するという仮説を提示する．実験的には，幅広いドメインと生成器での MIA と検出手法に対する大規模実験の結果，タスク間の性能に強い順位相関が見られた．特に，特定の検出手法が MIA においても最高水準の性能を達成し，両タスク間での転移性の実用的意義を示した．最後に，MIA と機械文検出のタスク横断的な開発と公平な評価を可能にするための統合評価基盤 MINT[1)] を公開する．

1 はじめに

大規模言語モデル (LLM) は，人間に匹敵する文生成能力と文理解能力を示し，幅広い分野での利用が進んでいる．一方で，LLM が訓練データの記憶によって個人情報や機密情報 [1]，著作物 [2] を漏洩するリスク [3] があり，また LLM を用いたプロパガンダ生成 [4] や学術分野での剽窃 [5] など，文の真正性にも課題が生じている．

これらの問題に対処するため，さまざまな研究が進められている．MIA は，ある文が言語モデルの訓練データに含まれるかどうかを判定するタスク [6] であり，LLM による個人情報や著作物の漏洩の可能性を精査する．一方で，機械文検出は，人間が書いた文と機械が生成した文を識別するタスク [7] であり，誤情報の拡散や学術的不正の抑止に寄与する．

両タスクは目的が異なるものの，いずれも言語モデルの確率分布に基づく指標で判定を行う．MIA では学習データ中の文ほど高い尤度を示し，機械文検出でも機械文が高い尤度を示す傾向があり，尤度やエントロピーが共通のベースライン指標として用いられる．さらに，*Neighborhood Attack* [8]（MIA 手法）と *DetectGPT* [9]（検出手法）は本質的に同一であることが指摘 [10] されており，どちらも文への摂動を通して文群の確率曲線を推定する．図 1 でも，両タスクにおける両手法の予測スコア分布に顕著な類似性が確認できる．しかし，両タスクは独立に研究されており，一方で得られた強力な手法や知見が他方で十分活用されていない可能性がある．

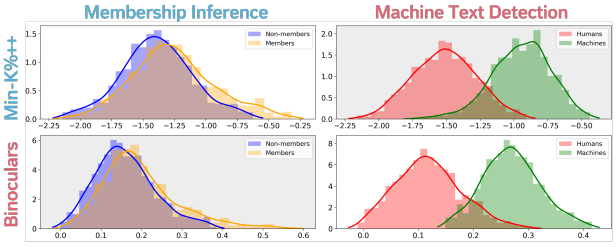

図 1 *Min-K%++*（最先端の MIA 手法）と *Binoculars*（最先端の検出手法）による，両タスクでの予測スコア分布．網掛け部分は cross-task 設定を指す．両手法は別々のタスクで提案されたにもかかわらず，そのスコア分布はきわめて類似しており，高い転移性を示唆する．

これを踏まえ，本稿では，MIA と機械文検出の間の「転移性」，つまり一方のタスクで提案された手法が他方でもどの程度有効に機能するかを，理論的および実験的に検証する．理論的には，両タスクにおいて漸近的に最適な性能を達成する指標が「対象モデル分布と真の母集団分布との尤度比検定」で同一であることを示す（§2）．さらに，多くの既存手法がこの最適指標の近似として統一的に捉えられ，効果的に近似する手法は高い転移性を持ち，両タスクで高い性能を発揮する，という仮説を提示する．

実験では，7 つの MIA 手法および 5 つの検出手法を対象に，13 のドメインおよび 10 の生成モデルにわたる大規模評価を実施した（§3）．両タスクにおける各手法の性能順位の相関を分析した結果，一方のタスクで高性能な手法は他方においても高性能と

* 二人の著者は本稿に等しく貢献した．本稿は第一著者がペンシルバニア大学滞在中に実施した．
1) https://github.com/ryuryukke/MINT

なる傾向にあり，強い順位相関（$\rho > 0.6$）が見られた（§3.2）．さらに，検出手法である *Binoculars* [11] が MIA においても最高水準の性能を達成し，両タスク間での転移性の実用的意義が示された．

2　共通の最適検定統計量

2.1　両タスクの定式化

$\mathcal{X}$ をすべてのトークン系列の集合とする．$\mathcal{M}$ を $\mathcal{X}$ 上の分布 $P_{\mathcal{M}}$ を導く言語モデルとし，$\mathcal{D}_{\text{train}} \subset \mathcal{X}$ を $\mathcal{M}$ の学習に用いられた未知の訓練データとする．機械文検出において，$P_{\mathbb{Q}}$ を $\mathcal{X}$ から $\mathcal{M}$ が生成し得る文を除外した集合上の分布とする．MIA において，P_{out} を訓練データ $\mathcal{D}_{\text{train}}$ に含まれない文の集合上の分布，P_{in} を訓練データ内の分布とする．ある文 $x \in \mathcal{X}$ に対し，それぞれのタスクは以下の仮説検定で定式化される．

$$\textbf{機械文検出} - H_0 : x \sim P_{\mathbb{Q}},\quad H_1 : x \sim P_{\mathcal{M}},$$
$$\textbf{MIA} - H_0 : x \sim P_{\text{out}},\quad H_1 : x \sim P_{\text{in}}.$$

両タスクの目的は，それぞれの帰無仮説を最大の統計的検出力で棄却できるような関数 $f(x; \mathcal{M})$ を開発することである．

2.2　尤度比検定による両タスクの統一

定理 2.1（共通最適性）　$\mathcal{X}$ をすべてのトークン系列の集合とする．$\mathcal{M}$ を，訓練データ $\mathcal{D}_{\text{train}} \subset \mathcal{X}$ の尤度を最大化するように学習された言語モデルとし，その確率分布を $P_{\mathcal{M}}$ とする．$P_{\mathbb{Q}}$ を，$\mathcal{X}$ から $\mathcal{M}$ が生成し得る文を除外した集合上の確率分布とする．このとき，ある文 $x \in \mathcal{X}$ に対して検定統計量

$$\Lambda(x) = \frac{P_{\mathcal{M}}(x)}{P_{\mathbb{Q}}(x)}$$

は，標準的な正則条件の下で，機械文検出および MIA の両方において，任意の第一種の過誤において最大の統計的検出力を達成する．また，対応する最大アドバンテージ（ランダムな推測に対する改善度）は以下で抑えられる．

$$\text{adv} \le \sqrt{\frac{D_{\text{KL}}(P_{\mathbb{Q}} \| P_{\mathcal{M}})}{8}}.$$

証明　$\Lambda(x)$ が両タスクにおいて尤度比検定と一致することを示すことで，これを証明する．

***Step 1*：機械文検出**　以下の検定を考える．

$$H_0 : x \sim P_{\mathbb{Q}},\quad H_1 : x \sim P_{\mathcal{M}}.$$

この仮説に対する尤度比検定は

$$\Lambda_{\text{検出}}(x) = \frac{P_{\mathcal{M}}(x)}{P_{\mathbb{Q}}(x)}$$

であり，これは提案した統計量 $\Lambda(x)$ と一致する．ネイマン・ピアソンの補題 [12] により，この検定は任意の第一種の過誤において一様最強である．

***Step 2*：MIA**　以下の検定を考える．

$$H_0 : x \sim P_{\text{out}},\quad H_1 : x \sim P_{\text{in}}.$$

言語モデル $\mathcal{M}$ が無限のパラメータを持ち，訓練データ P_{in} の尤度を完全に最大化するという漸近的条件において，$\mathcal{M}$ は訓練サンプルのみを完全に再現する．このとき，$P_{\mathcal{M}}$ は P_{in} に収束する．同時に，$P_{\mathbb{Q}}$ は $\mathcal{M}$ が生成し得る文を除外した分布であり，P_{out} は訓練データを除外した分布である．この漸近的条件下では $\mathcal{M}$ は訓練データのみを生成するため，これら除外される集合は同一となり，$P_{\mathbb{Q}} \approx P_{\text{out}}$ が成立する．したがって，この仮説に対する尤度比検定は

$$\Lambda_{\text{MIA}}(x) = \frac{P_{\text{in}}(x)}{P_{\text{out}}(x)} \approx \frac{P_{\mathcal{M}}(x)}{P_{\mathbb{Q}}(x)}$$

となり，提案した統計量 $\Lambda(x)$ と漸近的に一致する．ゆえに，$\Lambda(x)$ は MIA においても漸近的に最適な検定統計量となる．

***Step 3*：アドバンテージの上界**　統計量 $\Lambda(x)$ を用いる仮説検定において，ベイズ誤り率は

$$\varepsilon^* = \frac{1 - \text{TV}(P_{\mathcal{M}}, P_{\mathbb{Q}})}{2}$$

である．ここで $\text{TV}(\cdot,\cdot)$ は全変動距離を表す．ピンスカーの不等式

$$\text{TV}(P_{\mathcal{M}}, P_{\mathbb{Q}}) \le \sqrt{\frac{1}{2} D_{\text{KL}}(P_{\mathbb{Q}} \| P_{\mathcal{M}})}$$

を適用することで，上述の誤差の境界および対応する最大アドバンテージが得られる．　□

備考　$\mathcal{M}$ が十分なパラメータを持ち，$\mathcal{D}_{\text{train}}$ を用いて漸近的に学習された場合，提案した統計量 $\Lambda(x)$ の MIA における識別性能は最適となる．しかし，現実の言語モデルは多くの場合，訓練データを 1 エポックしか学習しない．したがって，実際には $P_{\mathcal{M}}(x)$ よりも，他の優れた P_{in} の近似手法が高い性能を示す場合がある．本結果は，これらの近似手法に関するさらなる研究を制限するものではなく，機械文検出と MIA がなぜ根本的に関連したタスクなのか，その理論的解釈を与えるものである．

2.3 近似戦略に基づく統一的手法分類

全トークン系列の母集団を $\mathfrak{X}$ としたとき，$\mathcal{M}$ が生成し得る文を除外した集合上の真の分布 $P_{\mathbb{Q}}$ は，一般に未知である．そのため，両タスクの既存手法は，以下の 2 つに大別される戦略を用いて，最適統計量を構成する母集団分布 $P_{\mathbb{Q}}$ の推定に試みてきた．

参照モデルによる近似　この戦略では，母集団分布 $P_{\mathbb{Q}}$ の性質を代表する分布として，既知の参照モデルが持つ分布 $P_{M_{\text{ref}}}$ を利用する．

$$P_{\mathbb{Q}}(x) \approx P_{\mathcal{M}_{\text{ref}}}(x).$$

M_{ref} には，別の言語モデル（MIA: *Reference* [3]），ハフマン符号化による頻度分布（MIA: *Zlib* [3]），外部コーパスに基づくトークン頻度分布（MIA: *DC-PDD* [13]），モデル間の交差エントロピー（検出: *Binoculars* [11]）などが用いられる．

サンプリングによる近似　この戦略では，サンプリングを用いて母集団分布 $P_{\mathbb{Q}}$ を局所的に近似する．具体的には，母集団全体の分布を直接参照する代わりに，対象文 x に対して摂動を加えて得られた近傍文集合に対する $\mathcal{M}$ の尤度の期待値を用いる．

$$P_{\mathbb{Q}}(x) \approx \mathbb{E}_{\tilde{x}\sim\phi(\cdot|x)}[P_{\mathcal{M}}(\tilde{x})].$$

ここで $\phi(\cdot|x)$ は，対象文 x に対して単語置換などを行う摂動モデルである．この戦略は，MIA における *Neighborhood Attack* [8]，検出における *DetectGPT* [9] や *Fast-DetectGPT* [14] で採用されている．[2)]

議論　本分類は，両タスクの多くの手法を網羅するが，*Min-K%* などの尤度比ではない単一統計量に基づく手法は例外となる．これらの転移性は実験的評価に留め，その理論的な統一は今後の課題とする．

3 評価実験

3.1 実験設定

MIA　MIA の評価には，Pile 由来の 5 つの文ドメイン (Wikipedia, Pile CC, PubMed Central, ArXiv, HackerNews) で構成される MIA ベンチマーク MIMIR [15] を用いる[3)]．データは Pile から抽出され，13-gram フィルタリングで訓練（メンバー）とテスト（非メンバー）間のリークが排除される．対象モデルには Pythia シリーズ [16] の異なるパラメータサイズの 5 モデル (160M〜12B) が使用される．

機械文検出　検出の評価には，8 つのドメイン (Wikipedia, News, Abstracts, Recipes, Reddit, Poetry, Books, Reviews) における人間文と機械文で構成される検出ベンチマーク RAID [17] を利用する．対象モデルには，オープンウェイトモデルとして GPT-2-XL [18], MPT-30B-Chat [19], LLaMA-2-70B-Chat [20]，クローズドモデルとして ChatGPT [21] と GPT-4 [22] の計 5 モデルが使用される．

評価指標　両タスク間の転移性を評価するため，両タスクにおける全手法の性能順位を AUROC スコアに基づいて算出し，その順位の一貫性をスピアマンの順位相関係数によって測定する．高い順位相関は，一方のタスクにて優れた手法が，他方のタスクでも高い性能を発揮することを意味する．

対象手法　7 つの最新 MIA 手法 (*Reference* [3], *Zlib* [3], *Neighborhood attack* [8], *Min-K% Prob* [10], *Min-K%++* [23], *ReCaLL* [24], *DC-PDD* [13])，5 つの最新検出手法 (*DetectGPT* [9], *Fast-DetectGPT* [14], *Binoculars* [11], *DetectLLM* [25], *Lastde++* [26])，両タスクに共通するベースライン手法 (*Loss*, *Rank*, *LogRank*, *Entropy*) を対象とする．[4)]

検出設定　順位相関の測定では，両タスクで対象モデルの確率分布にアクセス可能なホワイトボックス設定を採用し，共通条件下で比較を行う．また，実用的な転移性を検証するため，ChatGPT や GPT-4 のクローズドモデルを対象としたブラックボックス設定での検出も実施する[5)]．後者の設定では，代用モデル (Pythia-160M [16], Llama-3-3.2B [27]) を用いて，その平均性能を報告する．

3.2 結果

MIA と機械文検出における顕著な順位相関　図 2 は，MIA と検出の各タスクにおける全手法の順位関係を示す．ここでの順位は，MIA（5 ドメインと 5 対象モデル）と検出（8 ドメインと 3 対象モデル）における平均 AUROC に基づいている．全 15 手法に対するスピアマンの順位相関係数は $\rho = 0.66$ ($p < 0.01$) となり，強い相関が見られた．この結果は，多くの MIA 手法が検出においても優れた性能を示し，その逆も同様であることを示す．また，上位 10 手法に限定すると，さらに強い順位相関 $\rho = 0.78$ ($p < 0.01$) が見られた．これは高い転移性が両タスクでの高い性能に起因するという仮説を裏付ける．

2) 本稿で検証する各手法を数式の再構成を通じて上述の近似戦略へと対応付けた結果については，付録 A を参照．
3) 機械文検出との公正な比較のため，文ドメインに限定．
4) 各手法の詳細は付録 A を参照．
5) 訓練データが非公開のモデルでは MIA の正解データが得られないため，MIA の評価はホワイトボックス設定に限定．

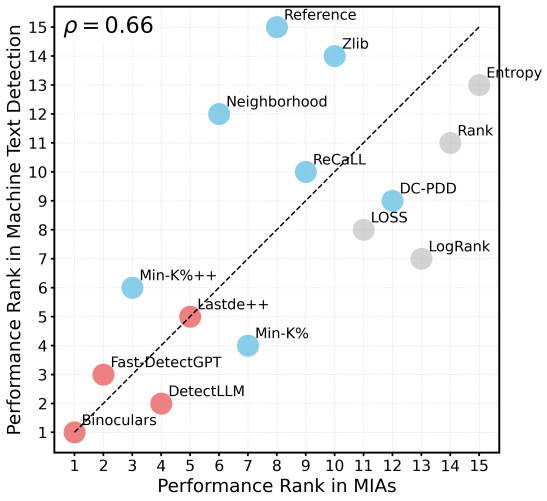

図 2　MIA と機械文検出における手法間の順位の関係．青色と赤色のプロットは，それぞれ MIA 手法と検出手法を示す．灰色のプロットは共通するベースラインを示す．破線は両タスクで順位が等しいことを表す．

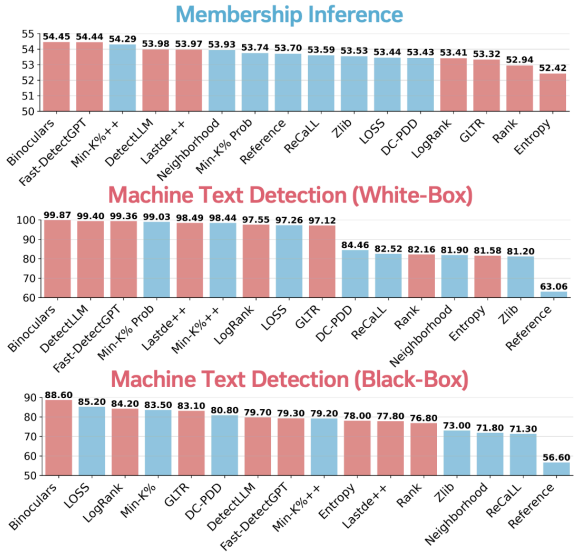

図 3　両タスクにおける MIA 手法と検出手法の平均 AUROC．**上段**：MIA ベンチマークにおける結果（5 ドメイン，5 モデル）．**中段・下段**：検出ベンチマークにおけるホワイトボックス（8 ドメイン，3 モデル），ブラックボックス設定（8 ドメイン，2 モデル）での結果．

検出手法による MIA での最高水準性能　図 3 上段・中段に，MIA と検出の各タスクにおける各手法の平均 AUROC を示す．驚くべきことに，検出手法 Binoculars が，両タスクで最高水準の平均性能を達成した．この結果は，既存の MIA 研究が検出分野の強力な手法を十分に反映できておらず，その評価におけるバイアスを示唆する．一方で，検出においても MIA 手法が高い性能を示しており，タスク横断した手法開発と公平な評価の重要性を提示する．

4　分析

より実用的な設定における転移性　ChatGPT や GPT-4 といったクローズドモデルの生成文を対象に，ブラックボックス設定下での MIA 手法の検出性能を評価した．図 3 下段に，各手法の平均 AUROC を示す．*Binoculars* は依然として他手法を大差で上回っているが，他の MIA 手法も，強力な検出器に匹敵する性能を達成した．これらの結果は，実用的な設定においても，MIA から検出への高い転移性が維持されることを裏付ける．

***Zlib* が示すタスク間の相違：事前分布**　転移性が限定的な例として，*Zlib* を取り上げる．*Zlib* は文の損失を Zlib 圧縮エントロピーで徐算して補正する手法である．MIA では両クラスが人間文分布に従うが，検出では人間文と機械文の異なる分布に従う．機械文は人間文より圧縮されやすく [28]，検出では損失とエントロピーが同方向に変動し，両者の比で

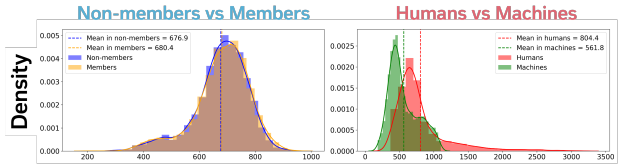

図 4　MIA（非メンバー 対 メンバー）と機械文検出（人間 対 機械）における Zlib 圧縮エントロピーの分布．各データセットから 3,000 文を無作為にサンプリングして算出．

ある *Zlib* スコアがクラス間で相殺され，識別能力が低下する．実際，図 4 では，MIA ではエントロピー分布が重なる一方，検出では分布が乖離している．この性能低下は，両タスク間の相違点の一つである「事前分布の違い」を反映する．

5　おわりに

本稿では，MIA と機械文検出の転移性を理論的および実験的に明らかにし，以下の知見を得た．(1) 両タスクの漸近的な最適指標は同一であり，多くの手法がその近似として位置づけられる．この近似精度が，ある手法の両タスクでの識別性能と転移性を決める要因となりうる．(2) 両タスクの手法間には性能順位に強い相関があり，一方のタスクで優れた手法は他方でも高い性能を発揮する．(3) 特定の検出手法が両タスクで最高水準の性能を達成し，両タスクの転移性には実用的な意義が存在する．これらの知見は，MIA と機械文検出のタスク横断した手法開発と公平な評価の重要性を提示するものである．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究（22501）により得られた．また，本研究は JST 次世代研究者挑戦的研究プログラム JPMJSP2106 の支援を受けた．また実験では，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した．

参考文献

[1] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing Leakage of Personally Identifiable Information in Language Models. In **2023 IEEE Symposium on Security and Privacy (SP)**, pp. 346–363, 2023.

[2] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating Copyright Takedown Methods for Language Models. In **Advances in Neural Information Processing Systems**, Vol. 37, pp. 139114–139150, 2024.

[3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In **30th USENIX Security Symposium**, pp. 2633–2650, 2021.

[4] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is AI-generated propaganda? **PNAS Nexus**, Vol. 3, No. 2, p. pgae034, 02 2024.

[5] The Guardian. Revealed: Thousands of UK university students caught cheating using AI, 2025. Accessed on 2025-07-10.

[6] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In **2022 IEEE Symposium on Security and Privacy (SP)**, pp. 1897–1914, 2022.

[7] Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. A Survey on Detection of LLMs-Generated Content. In **Findings of the Association for Computational Linguistics**, pp. 9786–9805, 2024.

[8] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In **Findings of the Association for Computational Linguistics**, pp. 11330–11343, 2023.

[9] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In **Proceedings of the 40th International Conference on Machine Learning**, pp. 24950–24962, 2023.

[10] 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.

[11] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In **The Forty-first International Conference on Machine Learning**, 2024.

[12] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, Vol. 231, No. 694-706, pp. 289–337, 1933.

[13] Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 5263–5274, 2024.

[14] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In **The Twelfth International Conference on Learning Representations**, 2024.

[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 **Conference on Language Modeling (COLM)**, 2024.

[16] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 2397–2430, 2023.

[17] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 12463–12492, 2024.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. **OpenAI**, 2019. Accessed: 2024-11-15.

[19] MosaicML NLP Team. Introducing MPT-30B: Raising the bar for open-source foundation models, 2023. Accessed: 2023-06-22.

[20] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

[21] OpenAI. Introducing ChatGPT, 2023. Accessed: 2023-05-10.

[22] OpenAI, Josh Achiam, Steven Adler, et al. GPT-4 Technical Report, 2024.

[23] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved Baseline for Pre-Training Data Detection from Large Language Models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[24] Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 8671–8689, 2024.

[25] Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics**, pp. 12395–12412, 2023.

[26] Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free LLM-generated Text Detection by Mining Token Probability Sequences. In **The Thirteenth International Conference on Learning Representations**, 2025.

[27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models, 2024.

[28] Yu Mao, Holger Pirk, and Chun Jason Xue. Lossless Compression of Large Language Model-Generated Text via Next-Token Prediction, 2025.

表 1　MIA および機械文検出における手法の統一的な定式化．$P_{\mathcal{M}}(x)$ は生成確率，$R_{\mathcal{M}}(x)$ は平均対数順位，$\phi(x)$ は任意の摂動関数を示し，$\Phi(x) = x/\sigma_x$ は標準偏差による除算を表す．各式の導出に関する詳細は，付録 A を参照．

手法	タスク	式
Reference [3]	MIA	$-\log(P_{\mathcal{M}}(x)/P_{\mathcal{M}_{\text{ref}}}(x))$
Zlib [3]	MIA	$-\log P_{\mathcal{M}}(x)/\text{Zlib}(x)$
DetectLLM [25]	Detection	$\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log R_{\mathcal{M}}(\tilde{x})]/(\log R_{\mathcal{M}}(x))$
ReCall [24]	MIA	$\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log P_{\mathcal{M}}(\tilde{x})]/(\log P_{\mathcal{M}}(x))$
DC-PDD [13]	MIA	$\mathbb{E}_{\tilde{x}\sim\mathcal{M}}[-\log P_{\mathcal{M}_{\text{ref}}}(\tilde{x})]$
Binoculars [11]	Detection	$(\log P_{\mathcal{M}}(x))/\mathbb{E}_{\tilde{x}\sim\mathcal{M}}[\log P_{\mathcal{M}_{\text{ref}}}(\tilde{x})]$
DetectGPT [9]	Detection	$\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log(P_{\mathcal{M}}(x)/P_{\mathcal{M}}(\tilde{x}))]$
Neighborhood [8]	MIA	$\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log(P_{\mathcal{M}}(x)/P_{\mathcal{M}}(\tilde{x}))]$
Fast-DetectGPT [14]	Detection	$\Phi(\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log(P_{\mathcal{M}}(x)/P_{\mathcal{M}}(\tilde{x}))])$
Min-k% [10]	MIA	$\frac{1}{k}\sum_{i\in\text{min-}k\%}(-\log P_{\mathcal{M}}(x_i))$
Min-k%++ [23]	MIA	$\frac{1}{k}\sum_{i\in\text{min-}k\%}\Phi(-\log P_{\mathcal{M}}(x_i) + \mathbb{E}_{\tilde{x}_i\sim\mathcal{M}}[\log P_{\mathcal{M}}(\tilde{x}_i)])$
Lastde [26]	Detection	$(-\log P_{\mathcal{M}}(x))/\text{StdDev}(\{\text{DE}(x,\tau)\}_{\tau=1}^{\tau'})$
Lastde++ [26]	Detection	$\Phi(\text{Lastde}(x) - \mathbb{E}_{\tilde{x}\sim\phi(x)}[\text{Lastde}(\tilde{x})])$

A　手法の詳細

A.1　ベースライン

Loss: モデル $\mathcal{M}$ の文 x に対する損失．機械文およびメンバーは，平均的に損失が小さいという仮説に基づく．

Entropy: 各ステップにおける次トークンに対する確率の期待値．機械文およびメンバーは，エントロピーが低いという仮説に基づく．

Rank: 各ステップにおける次トークンの確率順位の平均．機械文およびメンバーは，平均順位が高いという仮説に基づく．

LogRank: 各ステップにおける次トークンの対数順位の平均．*Rank* と同様，機械文およびメンバーは平均対数順位が高いという仮説に基づく．

A.2　MIA

Reference: モデル $\mathcal{M}$ と参照モデル $\mathcal{M}_{\text{ref}}$ の間の損失の差．損失を確率を用いて表すと，参照モデルによる近似戦略の一種である．

Zlib: モデル損失と Zlib 圧縮率の比．Zlib 圧縮率は入力文から得られる経験的な部分文字列分布に基づく参照分布であり，同じく損失を確率を用いて表すと，参照モデルによる近似戦略の一種である．

Neighborhood: 文 x の対数確率と，x に摂動を加えて得られた近傍文 $\tilde{\mathbf{x}}$ の対数確率の比の期待値．機械文およびメンバーは近傍文よりも生成確率が高いという仮説に基づく．サンプリングによる近似戦略の一種である．

Min-K%: 文中の確率が最も低い k%のトークンの対数尤度の平均．機械文およびメンバーは，極めて低確率なトークンが含まれにくいという仮説に基づく．

Min-K%++: 文中の確率が最も低い k%のトークンの語彙分布全体で標準化した対数尤度の平均．語彙分布と比較して，特定のトークンの surpirsal を測定する．

ReCaLL: 文 x の対数確率と，非メンバーの一部を文脈として与えたときの条件付き対数確率の比．非メンバーを文脈として与えられたとき，非メンバーはメンバーよりも対数確率が高いという仮説に基づく．サンプリングによる近似戦略の一種である．

DC-PDD: モデル $\mathcal{M}$ と参照コーパス $\mathcal{D}'$ の Unigram 頻度分布の間の交差エントロピー：$-\frac{1}{n}\sum_{i=1}^{n} P_{\mathcal{M}}(x_i)\cdot\log P_{\mathcal{D}'}(x_i)$．これは $\mathbb{E}_{\tilde{x}\sim\mathcal{M}}[-\log P_{\mathcal{M}_{\text{ref}}}(\tilde{x})]$ と等価である．

A.3　機械文検出器

DetectGPT: 文に対する対数確率関数の負の曲率を測定する．文 x の対数確率と，x に摂動を加えて得られた近傍文 $\tilde{\mathbf{x}}$ の対数確率の比の期待値．機械文およびメンバーは近傍文よりも確率が高いという仮説に基づく．サンプリングによる近似戦略の一種である．

Fast-DetectGPT: *DetectGPT* の高コストな摂動を避け，期待値を直接計算する．さらに近傍文群におけるスコア標準化を行う．サンプリングによる近似戦略の一種である．

Binoculars: モデル $\mathcal{M}$ と参照モデル $\mathcal{M}_{ref}$ の間のパープレキシティと交差エントロピーの比．本稿ではこれを $(\log P_{\mathcal{M}}(x))/\mathbb{E}_{\tilde{x}\sim\mathcal{M}}[\log P_{\mathcal{M}_{\text{ref}}}(\tilde{x})]$ として再構成する．参照モデルによる近似戦略の一種である．

DetectLLM: *DetectGPT* の変種であり，対数確率の代わりに対数順位を用いる．本稿ではより高い性能を示す NPR 指標を採用し，$\mathbb{E}_{\tilde{x}\sim\phi(x)}[\log R_{\mathcal{M}}(\tilde{x})]/(\log R_{\mathcal{M}}(x))$ として再構成する．

Lastde++: 時系列解析を用いて，文中のトークン確率の局所的および大局的な変動を測定する．*Lastde++*はこれに対して，*Fast-DetectGPT* と同様の摂動と標準化を適用したものである．確率の分散を用いた指標と考えられる．