トークン分割に起因する分散に基づく トークンベースのメンバーシップ推論の解釈

石原祥太郎 [1]　高橋寛武 [2*]
[1] 株式会社日本経済新聞社　[2] 独立研究者
shotaro.ishihara@nex.nikkei.com

概要

大規模言語モデルの訓練データに対するメンバーシップ推論に注目が集まる中，英語で有効とされる Min-K% Prob のようなトークン絞り込みが，日本語では必ずしも良好な性能を示さないと報告されている．本研究はこの現象の解釈として，トークンベース手法の性能が，サンプル群におけるスコア分散に依存する点に着目する．更に，この分散が言語ごとのトークン分割の分散に起因すると仮定し，既存の実証結果を統一的に理解するための理論的枠組みを提供する．

1　はじめに

大規模言語モデルの実用化が進むにつれ，訓練データのメンバーシップ推論（Membership Inference Attack; MIA）への注目度が高まっている [1]．古典的な手法として「モデルが訓練データに対して低い損失（高い予測値）を示す」ことに注目した LOSS [2] があり，損失をもとに特定のサンプルが訓練セットに含まれるかを予測する．その後より強力な手法として，トークン単位の予測値のうち下位 K% を抽出して用いる Min-K% Prob [3] が登場し，$K = 10$ で高い MIA 性能を示すと報告された．このトークン絞り込みは「訓練データでは予測値が低いトークンはほとんど含まれず，訓練データ以外では一定数含まれる」という仮定に基づき，2 群の違いを強調する．この流れは発展を続け，派生手法も複数提案されている [4, 5]．

既存の取り組みは実証研究が中心で，理論的説明は十分に与えられていない．英語での実証的知見としては，MIA の性能といった大規模言語モデルの暗記が文字列の重複・モデルサイズ・文脈長に関連すると報告されている [6, 7]．これらの知見は日本語で必ずしも再現していないが，その理由を考察するための枠組みは存在しない．特に Min-K% Prob については，K が小さいほど，または文脈長が小さいほど性能が劣化するといった英語と異なる現象が確認されている [8, 9, 10]．

本研究では，MIA における代表的な性能指標である AUC が，サンプル集合に対するスコアの分散に依存するという古典的な知見を再検討する（§3.1）．一般に平均値の差が同一であっても，分散が増大すると AUC は低下することが知られている [11, 12][1)]．更に本研究では，この分散が言語ごとに異なるトークン分割の細かさに起因すると仮定し（§3.2），実験的に検証する（§4）．加えて，本研究で提示した理論的枠組みに基づき，既存研究を再解釈する（§5）．我々の知る限り，トークンベースの MIA に対して，言語間の性能差を含めた理論的説明を与える試みは本研究が初めてである．

2　前提知識

本節では，本研究で扱うトークンベース MIA の基礎的前提と，従来の実証的知見を整理する．

2.1　トークンベース MIA

n 個のトークンの系列から成る文 $x = (t_1, \ldots, t_n)$ の生成確率 $p(x)$ は式 (1) のように計算できる．

$$p(x) = \prod_{i=1}^{n} p(t_i | t_1, \ldots, t_{i-1}) \tag{1}$$

一般に $p(x)$ を直接計算すると非常に小さい値となるため，対数をとった値（対数尤度）を扱う場合が多い．n で割って平均の対数尤度とすると，他との大小比較にも利用しやすくなる．ここで各トークンに対する負の対数尤度を考えると，Loss(x) は式 (2)

＊ 株式会社日本経済新聞社での業務委託

1） 概念的な図示は付録 A に示す．

のように表現できる．

$$\mathrm{Loss}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i, \qquad \ell_i = -\log p(t_i \mid t_{<i}) \tag{2}$$

大規模言語モデルに対する MIA は，あるサンプル x について，モデルの訓練セットに含まれるデータか（member）否か（non-member）を識別する．古典的手法としては，n 個のトークンから成る入力文 x に対する平均損失を指標とし，member に対して損失が小さくなるという仮定に基づく LOSS が広く知られている．更に，より高い性能を示す手法として，Min-K% Prob といったトークンベースの手法が近年注目されている．これは式 (3) のように，n 個のトークンに対する損失列 $(\ell_1, \ldots, \ell_n)$ を昇順に並べた順序統計量 $(\ell_{(1)}, \ldots, \ell_{(n)})$ を用い，下位 K% に相当する $k = \lfloor Kn \rfloor$ 個の損失の平均をスコアとする．

$$S_K(x) = \frac{1}{k}\sum_{i=1}^{k} \ell_{(i)} \tag{3}$$

文中の「特に予測が容易なトークン群」に着目することで, member と non-member の差異を強調し, 性能が向上すると示唆されている．更に Min-K%++ [4] は，全語彙の生成確率の平均・分散を用いて補正したスコアを採用することで性能を改善した．

2.2 実証的知見

既存研究はトークンベース MIA の性能が，次のような要因と強く関連すると報告している [1, 6, 7]．

文字列の重複 訓練セット内の文字列の重複が多いほど，MIA の性能が高くなる．この現象は多くの研究で再現されている．

モデルサイズ モデルのパラメータ数が大きいほど，MIA の性能が高くなる．一方でモデルサイズを大きくしても性能向上が頭打ちになる，あるいは改善が限定的であることも報告されている [3, 4]．

文脈長 与える文字列が長いほど，MIA の性能が高くなる．ただし，文脈長は多くの交絡因子となっており丁寧な議論が必要 [6] で，一定の分量がない設定では性能を正当に測定できないという指摘もある [13]．

これらの知見は主に英語で検証されており [14]，日本語を対象とした MIA 実験では Min-K% Prob が必ずしも高い性能を示さないとの報告がある．日本語の実験 [8, 9] では，Min-K% Prob などのトークンベース MIA の AUC が最大でも 0.60 程度となった．文脈長が大きいほど性能が悪くなるという，英語とは異なる傾向も見られている．小柳ら [10] は，英語では K が小さい場合に良い結果が得られた一方で，日本語では逆に MIA 性能が劣化すると観測した．これらの現象は，語境界が明示されないなどの言語特性に起因する可能性があるが，理論的解釈は十分に与えられていない．本研究は，語境界が明示されない特性がトークン分割数の細かさに繋がり，性能劣化が発生するという解釈を提示する．

3 理論的説明

本節ではトークンベース MIA の手法のうち，特に Min-K% Prob の性能が，サンプル集合におけるスコア分散にどのように依存するかを理論的に整理する．本研究の中心的主張は，トークンベース MIA の性能は平均値の差だけでなく分散に強く制約されており（§3.1），その分散がトークン分割の細かさによって系統的に増大しうる（§3.2）という点にある．

3.1 スコア分布と AUC の近似

訓練セットの集合を $\mathcal{D}_{\mathrm{in}}$，非訓練セットの集合を $\mathcal{D}_{\mathrm{out}}$ とする．文 x に対してスコア $S(x) \in \mathbb{R}$ を返す MIA スコア関数 $S : \mathcal{X} \to \mathbb{R}$ を考える．一般性を失わず，スコアが小さいほど member らしいと判定される設定を仮定する．$\mathcal{D}_{\mathrm{in}}$ および $\mathcal{D}_{\mathrm{out}}$ 上でのスコアを確率変数 S_{in} と S_{out} として表す．

このとき，MIA の二値分類の性能を測る評価指標として一般的な AUC は式 (4) のように表せる．

$$\mathrm{AUC} = \Pr(S_{\mathrm{out}} > S_{\mathrm{in}}) \tag{4}$$

ここで，スコア $S(x)$ が多数のトークン単位損失の平均的統計量であることを踏まえ，スコア分布を式 (5) のように正規分布で近似する．かつ両者は独立とする．このとき差分 $D = S_{\mathrm{out}} - S_{\mathrm{in}}$ は式 (6) に従う．

$$S_{\mathrm{in}} \approx \mathcal{N}(\mu_{\mathrm{in}}, \sigma_{\mathrm{in}}^2), \qquad S_{\mathrm{out}} \approx \mathcal{N}(\mu_{\mathrm{out}}, \sigma_{\mathrm{out}}^2), \tag{5}$$

$$D \approx \mathcal{N}\left(\Delta,\, \sigma_{\mathrm{in}}^2 + \sigma_{\mathrm{out}}^2\right), \quad \Delta = \mu_{\mathrm{out}} - \mu_{\mathrm{in}} \tag{6}$$

ここで式 (4) の AUC は，式 (7) のように近似できる．Φ は標準正規分布の累積分布関数である．

$$\mathrm{AUC} = \Pr(D > 0) \approx \Phi\left(\frac{\Delta}{\sqrt{\sigma_{\mathrm{in}}^2 + \sigma_{\mathrm{out}}^2}}\right) \tag{7}$$

式 (7) から，トークンベース MIA の性能は平均差 Δ だけでなく，スコア分布の分散（$\sigma_{\mathrm{in}}^2 + \sigma_{\mathrm{out}}^2$）に制約されると分かる．この性質は，言語やモデルに依

らず成立する一般的な特徴である [11]．式 (3) で示される Min-K% Prob のスコア $S_K(x)$ では，一般に K が小さい場合，少数のトークンの値がスコア全体を支配するため分散は大きくなりやすい．

3.2 トークン数とスコア分散の増大

次に，文の意味内容とは独立に，トークン分割の細かさが Min-K% Prob のスコア $S_K(x)$ の分散（$\mathrm{Var}(S_K) \approx \sigma_{\text{in}}^2 + \sigma_{\text{out}}^2$）を増大させる仕組みを考える．文 x は文字列として与えられ，トークナイザによって n 個のトークンに分割される．同程度の内容を持つ文であっても，トークン分割の結果として得られる n は言語ごとに異なる．特に日本語のように語境界が明示されない言語では，n のばらつきが大きくなる傾向がある．

$S_K(x)$ では，$k = \lfloor Kn \rfloor$ が文ごとに変化するため，トークン分割数のばらつきは次の二つの要素を通じて，$S_K(x)$ の分散を増大させる．

- スコアに含まれるトークン数 k の変動
- 下位 k に含まれるトークン集合の不安定化

形式的には，全分散の公式より $S_K(x)$ の分散は式 (8) のように分解できる．

$$\mathrm{Var}(S_K) = \mathbb{E}[\mathrm{Var}(S_K \mid n)] + \mathrm{Var}(\mathbb{E}[S_K \mid n]) \quad (8)$$

第 1 項はトークン損失自体のばらつきに由来する分散であり，第 2 項はトークン数 n の変動に起因する分散である．後者の項は，トークン分割が不安定であるほど増大する．したがって，トークン分割が細かく，かつ文ごとのばらつきが大きい言語では，$\mathrm{Var}(S_K)$ が系統的に大きくなり，式 (7) の分母が増大することで，AUC が低下しやすくなる．

4 実証的検証

前節の理論的説明では，トークンベース MIA の性能がスコアの分散に依存し，特に Min-K% Prob のスコア $S_K(x)$ の場合はトークン分割の細かさが分散を増大させる可能性を示唆した．本節では実証的検証として，日本語と英語でトークン分割数が異なり（§4.1），英語においてトークン分割数が大きいほどスコア分散が大きい（§4.2）と確認する．

4.1 言語によるトークン数分散の違い

英語と日本語では，同程度の意味内容・文長を持つ文であっても，n の分布が大きく異なることが経験的に知られている．英語では語境界が空白により明示されており，大規模言語モデルのトークナイザとして一般的なサブワード [15] を適用しても 1 語あたりのトークン数は比較的安定している．一方で日本語では語境界が文字列上に明示されないため，漢字・仮名・記号の混在や表記揺れに応じて，同一の文長・意味量であってもトークン分割が細かくなる傾向がある．その結果，日本語文では英語文と比較して n が大きくなりやすく，かつ文ごとのばらつきも大きくなる．

表 1 WikiMIA と日本語訳でのトークン数の分散

言語	#words: 128	#words: 256
英語	299 + 497 = 796	773 + 1073 = 1847
日本語	1298 + 1569 = 2868	4575 + 3301 = 7877

ここでは，Min-K% Prob の論文 [3] で提案された WikiMIA データセットを分析する．WikiMIA データセットは英語の Wikipedia から構築され，大規模言語モデルの MIA の主要なベンチマークとして認知されている．モデルの事前学習以前の 2017 年以前に発生したイベントの記事から訓練セットの集合を，事前学習後の 2023 年のイベントの記事から非訓練セットの集合を構築している．記事から長さの異なる 4 種類（{32, 64, 128, 256}単語）のテキストが用意されている．

言語ごとの差異を確認するため，128，256 単語の英語テキストを日本語訳し，英語と日本語でのトークン分割数を調べた．日本語訳には Gemini 3 Flash[2] を使用した[3]．各テキストは，LLaMA-7b [16] のトークナイザ[4] で分割した．表 1 に示す通り，英語に比べ日本語は訓練セットでも非訓練セットでもトークン数の分散が大きくなり，その和（$\sigma_{\text{in}}^2 + \sigma_{\text{out}}^2$）も約 4 倍大きい結果となった．

4.2 トークン数とスコア分散の増大

次に全 4 種類の長さを持つ 82 のサンプル対象に，各サンプルの Min-K% Prob および Min-K%++ のスコアと，トークン数の関係性を分析した．具体的には，各単語数の集合をトークン数の大小で二分し，訓練セットと非訓練セットに該当するサンプルのスコア分散の和（$\sigma_{\text{in}}^2 + \sigma_{\text{out}}^2$）を求めた．スコアを算出するモデルは LLaMA-7b で，実装には Fast-MIA [17] を用い，$K = 20$ に設定した．

表 2 に示す通り，Min-K% Prob では全ての単語数において，トークン数が大きい集合の方がスコア分

2) https://deepmind.google/models/gemini/flash/
3) 翻訳に用いたプロンプトは付録 B に示す．
4) https://huggingface.co/huggyllama/llama-7b

表 2　各設定におけるスコア分散（$\sigma_{\text{in}}^2 + \sigma_{\text{out}}^2$）．トークン数が多い群ほど，スコア分散が大きくなる傾向がある．

手法	#words	スコア分散	トークン数で二分した集合での分散		トークン数の分散	
			大	小	大	小
MIN-K% PROB	32	1.38 + 1.35 = 2.73	1.67 + 2.33 = **4.01**	1.09 + 0.72 = 1.82	**135.0**	18.14
	64	0.98 + 1.01 = 2.00	0.92 + 1.15 = **2.07**	0.54 + 1.11 = 1.65	**186.6**	42.49
	128	0.57 + 0.62 = 1.20	0.51 + 0.66 = **1.18**	0.40 + 0.35 = 0.75	**153.5**	106.5
	256	0.40 + 0.41 = 0.81	0.32 + 0.67 = **0.99**	0.28 + 0.51 = 0.79	**420.5**	260.6
MIN-K%++	32	0.25 + 0.81 = 1.07	0.57 + 0.67 = 1.25	0.30 + 1.19 = **1.50**	**135.0**	18.14
	64	0.12 + 0.20 = 0.33	0.35 + 0.23 = **0.58**	0.09 + 0.19 = 0.29	**186.6**	42.49
	128	0.08 + 0.07 = 0.16	0.14 + 0.17 = **0.31**	0.07 + 0.06 = 0.13	**153.5**	106.5
	256	0.05 + 0.05 = 0.10	0.06 + 0.12 = **0.19**	0.05 + 0.03 = 0.09	**420.5**	260.6

散が大きかった．MIN-K%++ でも単語数 32 の場合を除いて同様の結果となった．MIN-K%++ では，確率の補正の結果としてスコア分散が全体的に小さくなっている．

5　既存研究の再解釈

本節では，§3 で導出した式 (7) の枠組みに基づき，既存研究のトークンベース MIA 手法や実証的知見を再解釈する．

5.1　トークンベース MIA の性能の言語差

英語では MIN-K% PROB の K が小さいほど良い結果が得られているが，日本語では必ずしも再現しない理由として，スコア分散の影響が考えられる．K を小さくするとスコアの算出に用いられるトークン数も減少し，一般に分散は拡大する．この分散の増大はトークン分割が細かい日本語でより顕著となり，特徴的なトークンに絞り込む利点を打ち消し，時に上回る可能性がある．

5.2　文字列の重複

訓練セット内の文字列の重複は本研究の観点から見ると，平均差を拡大する効果だけでなく，スコア分散を縮小するという側面を持つ．文字列が重複しているほどサンプル群のスコアの順序統計量が安定化し，結果として σ_{in}^2 が低下して性能が向上する．サブワードトークナイザでは頻度に基づきトークン分割を決定するため，訓練セット内で重複している文字列ではトークン分割数が小さくなる．このことも，トークン数を通じたスコア分散の増大を妨げる方向に機能する．

5.3　モデルサイズ

モデルサイズの増加は，member 側の平均スコアを低下させ，平均差を拡大する効果を持つ．一方でモデルサイズの増加に伴って予測の鋭敏化や文脈依存性の増大が起こり，スコア分散が拡大する可能性がある．結果として平均差の拡大と分散の増大が拮抗し，性能の改善が限定的になり得ると理解できる．

5.4　文脈長

一般に入力文の文脈長が長くなるにつれて MIA の性能が改善する現象が報告されているが，日本語の MIN-K% PROB においては，逆の傾向が観測されている．表 2 に示す通り，一般には文脈長が増えるほどスコア分散が低下し，このことが性能改善に寄与する．しかしトークン分割が細かい日本語では，表 1 で確認したように英語に比べ大きなトークン数の分散が生じ，スコア分散も大きくなると見込まれる．結果として文脈長の増加に伴うスコア分散の減少が限定的になっている可能性が考えられる．

6　結論と限界

本研究では，トークンベース MIA の性能が平均差だけでなくスコア分散にも制約されることを理論的に整理し，この分散が言語ごとに異なるトークン分割の細かさとばらつきで誘発されるという仮説を提示・検証した．加えて本枠組みに基づき，英語で有効とされる MIN-K% PROB のようなトークン絞り込みが，日本語では必ずしも良好な性能を示さないといった既存の実証的知見を再解釈した．

主要な限界として，言語間の差異の実証が間接的である点が挙げられる．本研究では英語データセットとその日本語訳を分析したが，理想的には日本語コーパス上で MIA の実験を直接実施するのが望ましい．しかしながら，現時点ではこの要件に見合う日本語の公開データセットは存在しない．そのため，本研究では理論的な説明を与えつつ，間接的な分析を通じて実証的な裏付けを実施した．

参考文献

[1] Shotaro Ishihara. Training data extraction from pre-trained language models: A survey. In Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta, editors, **Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)**, pp. 260–275, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

[3] 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.

[4] 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.

[5] Roya Arkhmammadova, Hosein Madadi Tamar, and M Emre Gursoy. Win-k: Improved membership inference attacks on small language models. **arXiv [cs.AI]**, August 2025.

[6] Bowen Chen, Namgi Han, and Yusuke Miyao. A statistical and multi-perspective revisiting of the membership inference attack in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 22854–22874, Vienna, Austria, July 2025. Association for Computational Linguistics.

[7] 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.

[8] Shotaro Ishihara and Hiromu Takahashi. Quantifying memorization and detecting training data of pre-trained language models using Japanese newspaper. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, **Proceedings of the 17th International Natural Language Generation Conference**, pp. 165–179, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[9] Hiromu Takahashi and Shotaro Ishihara. Quantifying memorization in continual pre-training with Japanese general or industry-specific corpora. In Robin Jia, Eric Wallace, Yangsibo Huang, Tiago Pimentel, Pratyush Maini, Verna Dankers, Johnny Wei, and Pietro Lesci, editors, **Proceedings of the First Workshop on Large Language Model Memorization (L2M2)**, pp. 95–105, Vienna, Austria, August 2025. Association for Computational Linguistics.

[10] 小柳響子, 佐藤美唯, 梶浦照乃, 倉光君郎. LLM の事前学習データ検知法の日英比較. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 4Xin298–4Xin298, 2024.

[11] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern recognition**, Vol. 30, No. 7, pp. 1145–1159, July 1997.

[12] Tom Fawcett. An introduction to roc analysis. **Pattern Recognition Letters**, Vol. 27, No. 8, pp. 861–874, 2006. ROC Analysis in Pattern Recognition.

[13] Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 4165–4182, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[14] Ali Satvaty, Anna Visman, Dan Seidel, Suzan Verberne, and Fatih Turkmen. Memorization is language-sensitive: Analyzing memorization and inference risks of LLMs in a multilingual setting. In Robin Jia, Eric Wallace, Yangsibo Huang, Tiago Pimentel, Pratyush Maini, Verna Dankers, Johnny Wei, and Pietro Lesci, editors, **Proceedings of the First Workshop on Large Language Model Memorization (L2M2)**, pp. 106–126, Vienna, Austria, August 2025. Association for Computational Linguistics.

[15] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. **arXiv [cs.CL]**, February 2023.

[17] Hiromu Takahashi and Shotaro Ishihara. Fast-MIA: Efficient and scalable membership inference for LLMs. **arXiv [cs.CR]**, October 2025.

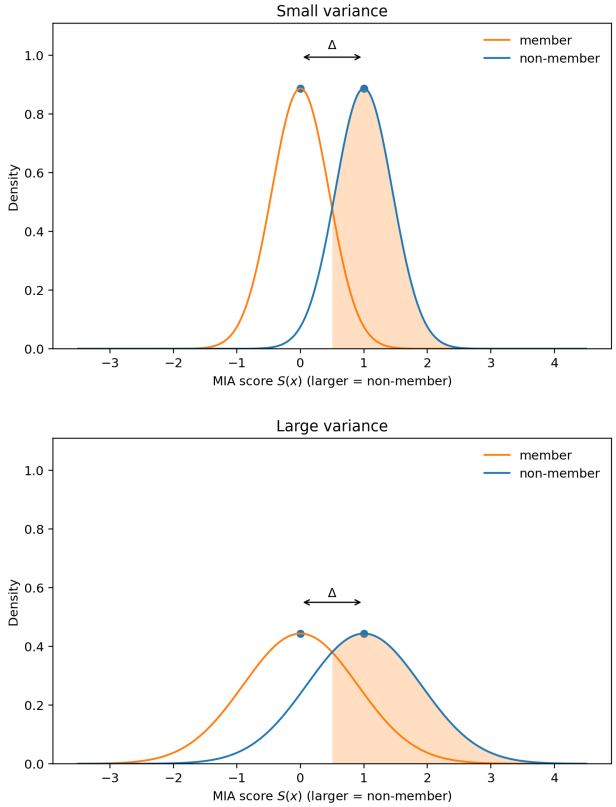

図 1　同一の平均差 Δ を持つ member / non-member のスコア分布で，分散が AUC に与える影響を示す概念図．

A　AUC と分散の関係性の可視化

図 1 に，AUC と分散の関係性を示す．分散が小さい場合（上）には分布の重なりが小さく高い AUC が得られる一方，分散が大きい場合（下）には分布が平坦化して重なりが増え，AUC が低下する．

B　日本語訳に用いたプロンプト

WikiMIA データセットは，次のプロンプトで日本語訳した．

日本語訳に用いたプロンプト

添付した CSV ファイルの input 列の英文は、ある英文を特定の単語数までで区切ったものです。文章としては途中で終わっていることがありますが、そこまででよいので日本語に翻訳してください。日本語訳する際に、絶対に元の英文に含まれていない単語などを補ってはいけません。日本語訳された文章は途中で終わっている不自然な状態で大丈夫です。日本語訳した文章を input_ja 列として保存し、CSV 形式で出力してください。

C　倫理的配慮

本論文は MIA を対象とするが，目的は攻撃の実運用や実行を推奨・助長することではない．トークンベース MIA の手法を科学的に理解することを目的とし，大規模言語モデルのより安全で責任ある開発に資することを目指している．MIA がより信頼性を持つ条件に関する知見が悪用される可能性も否定できないが，これらの知見が防御的にも利用可能である．例えば非英語言語に対するベンチマーク設計や，暗記の緩和の検討に活用できる．

本論文の推敲の補助として生成 AI ツールを使用した．生成されたテキストはすべて著者が確認・検証し，必要に応じて編集されている．