OCDR: OVERLAP-AWARE COUNTERFACTUAL DUAL-REPRESENTATION SCORING FOR MEMBERSHIP INFERENCE IN RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) systems can leak whether a target sample is present in the external knowledge base. Existing membership-inference attacks can achieve useful ranking performance while still producing substantially overlapping member and non-member score distributions, leaving an ambiguous decision region. We study whether counterfactual reference evidence can be represented to improve score separability while retaining competitive attack utility. We propose OCDR (Overlap-Aware Counterfactual Dual-Representation), which represents each target using two complementary Out-RAG statistics: the raw deviation from the Out-RAG mean and its variance-normalized counterpart. The two features are standardized on development data, expanded with second-order interactions, and combined using an L2-regularized logistic scorer. On HealthCareMagic with Mistral-7B-Instruct-v0.2, OCDR achieves an AUC of 0.6614 and TPR@5%FPR of 0.260, compared with 0.6270 and 0.175 for DC-MIA. It also reduces KDE overlap from 0.9324 to 0.7743 and 80-bin histogram overlap from 0.895 to 0.570. Additional experiments across generator backbones and datasets further show that ranking performance, low-FPR detection, and score-distribution overlap capture complementary aspects of membership-inference behavior. These results suggest that explicitly modeling and evaluating member/non-member score overlap provides a useful complementary perspective for membership inference in RAG systems.

1 INTRODUCTION

Large language models (LLMs) demonstrate strong capabilities in language understanding and generation, but purely parametric models remain limited in accessing and updating external knowledge. Retrieval-augmented generation (RAG) addresses this limitation by combining language models with external knowledge bases retrieved at inference time (1; 2; 3). While this architecture improves knowledge-intensive generation, it also introduces a distinct privacy risk: sensitive records stored in the retrieval database may be exposed through observable RAG outputs (6; 11).

Membership inference attacks (MIAs) provide a natural framework for studying this risk. Unlike conventional MIAs, which determine whether a sample was used for model training (4; 5), RAG membership inference asks whether a target sample is contained in the external retrieval database. Existing attacks use direct membership-oriented prompts (6), exploit response similarity (7), calibrate sample-dependent response difficulty using reference RAGs (8), suppress retrieval interference (9), or improve attack stealth and query efficiency (10).

Despite these advances, member and non-member scores can remain substantially overlapping. Existing methods mainly improve the underlying membership evidence or attack procedure, while the statistical separability of the *final membership-score distributions* has received comparatively less attention as an explicit design objective. This distinction is important because ranking performance and distributional separation characterize different properties of a membership score.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

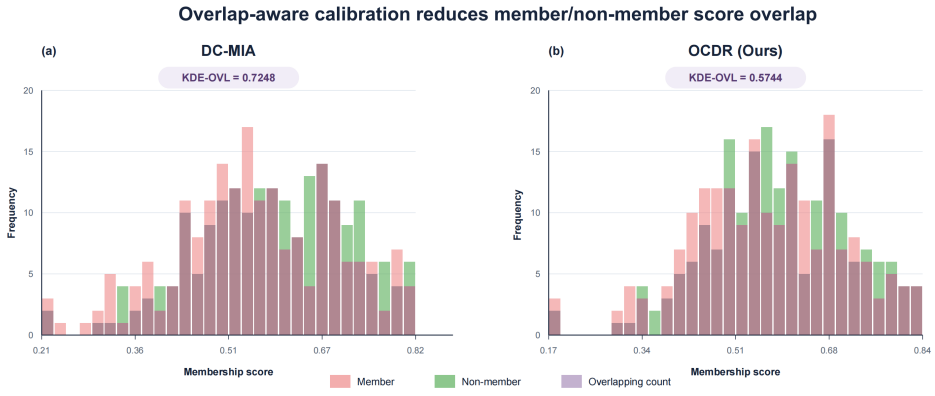

Figure 1: Member and non-member score distributions for DC-MIA and OCDR on HealthCareMagic with Phi-3.5-mini. Each method uses shared histogram bin edges for its member and non-member scores. KDE-OVL decreases from 0.7248 for DC-MIA to 0.5744 for OCDR; lower distributional overlap indicates better score separability.

We address this problem with **OCDR** (Overlap-Aware Counterfactual Dual-Representation), which represents each target relative to matched Out-RAG behavior using two complementary statistics: the absolute deviation from the Out-RAG mean and its variance-normalized counterpart.

Our contributions are threefold:

- We formulate member/non-member score separability as an explicit objective for RAG membership inference, complementing conventional ranking and low-FPR evaluation.
- We propose OCDR, a compact dual-representation scoring method that combines absolute and variance-normalized Out-RAG deviations through second-order nonlinear fusion.
- We evaluate OCDR against representative RAG membership-inference baselines on HealthCareMagic and AG News using both conventional attack metrics and score-overlap measures, and analyze the contribution of its feature representation through ablation studies.

2 Background and Related Work

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) combines parametric language models with external knowledge retrieved at inference time (1; 2). Dense retrievers such as DPR enable efficient semantic search over large document collections (3). By separating external knowledge from model parameters, RAG allows the knowledge base to be updated without retraining the generator. However, this architecture also introduces a distinct privacy surface because the retrieval database may contain sensitive or proprietary records (6; 11).

2.2 Membership Inference

Membership inference attacks (MIAs) aim to determine whether a target sample belongs to a model's training data (4). Subsequent work has shown that attack evaluation should consider not only overall discrimination but also performance at low false-positive rates, where privacy leakage can remain significant (5). In RAG systems, membership inference targets a different source of information: the adversary seeks to determine whether a sample is contained in the external retrieval database rather than in the generator's training set.

2.3 RAG MEMBERSHIP-INFERENCE ATTACKS

Existing RAG membership-inference attacks primarily differ in how they construct and calibrate membership evidence. Early work uses targeted prompts to elicit whether a candidate passage appears in the retrieved context (6). S-MIA uses semantic similarity between the generated response and target content as the membership signal (7). RAG-leaks shows that this signal is affected by sample-dependent response difficulty and proposes DC-MIA, which calibrates target behavior using matched reference RAG systems (8).

More recent work addresses complementary limitations. DCMI uses query perturbation and differential calibration to reduce interference from non-member retrieved documents (9), while the Interrogation Attack improves stealth and query efficiency through target-specific natural-language queries (10). OCDR focuses on a complementary aspect: how counterfactual reference evidence is represented and transformed into a final membership score with improved member/non-member score separability.

3 PROBLEM STATEMENT

3.1 MEMBERSHIP INFERENCE ON RAG

We consider a target RAG system consisting of a knowledge base $\mathcal{K}$, a retriever $\mathcal{R}$, and a generator $\mathcal{M}$. Given a target sample $x = (q, a)$, where q denotes the query and a the corresponding reference answer, the adversary seeks to determine whether x is contained in the external knowledge base. This differs from conventional membership inference, where membership refers to participation in model training. We formulate RAG membership inference as the binary hypothesis test

$$H_0 : x \notin \mathcal{K}, \qquad H_1 : x \in \mathcal{K}. \tag{1}$$

3.2 THREAT MODEL

We consider a black-box adversary with access only to the textual responses generated by the target RAG. For a target sample $x = (q, a)$, the adversary can submit q to the system and observe the resulting response, but has no access to retrieved documents, token-level probabilities, hidden representations, gradients, or other internal states.

Consistent with reference-RAG-based membership inference, the adversary is assumed to know the general system configuration, including the generator and retriever, and to possess auxiliary data drawn from the same domain distribution. This knowledge permits the construction of locally matched reference RAG systems for counterfactual comparison, while the contents of the target knowledge base remain inaccessible.

3.3 MEMBERSHIP-SCORE FORMULATION

A membership-inference attack assigns each target sample a scalar score $s(x)$ derived from observable RAG behavior. We orient all scores such that larger values favor the membership hypothesis H_1. Given a decision threshold τ, the final prediction is

$$\hat{y} = \mathbb{I}[s(x) > \tau]. \tag{2}$$

Existing attacks differ primarily in how $s(x)$ is constructed. S-MIA directly uses response similarity, whereas DC-MIA calibrates similarity using reference RAG systems. OCDR retains this score-based decision framework but changes how counterfactual reference evidence is represented to construct the final membership score.

3.4 DISTRIBUTION-SEPARABILITY OBJECTIVE

Conventional membership-inference metrics characterize different aspects of attack utility. ROC-AUC measures global ranking performance, while low-FPR TPR evaluates detection performance

under restrictive false-positive operating points. Neither metric, however, directly quantifies the statistical overlap between the final scores assigned to members and non-members.

Let $p_0(s)$ and $p_1(s)$ denote the score distributions induced by non-member and member samples, respectively. We use the overlap coefficient to quantify their shared probability mass, where smaller values correspond to greater distributional separability (18):

$$\mathrm{OVL} = \int \min\{p_0(s), p_1(s)\}\, ds. \tag{3}$$

4 METHODOLOGY

We propose OCDR, an overlap-aware counterfactual dual-representation scoring method for RAG membership inference. OCDR constructs counterfactual reference behavior using matched Out-RAG systems and represents the target response relative to this reference before producing the final membership score.

4.1 OVERVIEW

OCDR transforms the observable behavior of a target RAG system into an overlap-aware membership score through a five-phase pipeline, as shown in Fig. 2. The method does not introduce a new querying interface or modify the underlying membership-inference task. Instead, it progressively constructs counterfactual reference evidence, represents the target-to-reference deviation, and converts this evidence into the final membership score. Fig. 2.

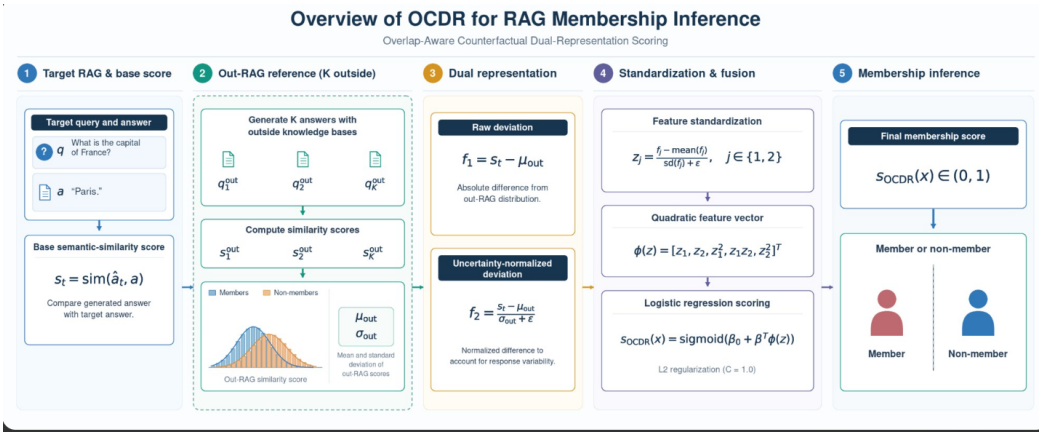

Figure 2: Overview of OCDR for RAG membership inference. Starting from the target-RAG response, OCDR constructs matched Out-RAG references, derives two complementary counterfactual features, standardizes and expands them with second-order terms, and produces the final membership score through regularized logistic regression.

Phase 1: Target-RAG Response and Base-Score Extraction. Given a target sample $x = (q, a)$, the query q is submitted to the target RAG system. The generated response $\hat{a}_t$ is compared with the reference answer a to obtain the base semantic-similarity score s_t.

Phase 2: Out-RAG Counterfactual Reference Construction. OCDR constructs multiple matched Out-RAG systems whose knowledge bases exclude the target sample. Their response similarities form the counterfactual reference set S_{out}.

Phase 3: Dual-Representation Construction. The target-to-reference relationship is represented from two complementary perspectives: the raw deviation from average Out-RAG behavior and the same deviation normalized by reference variability.

Phase 4: Standardization and Quadratic Scoring. The two representations are standardized using development-set statistics and expanded with compact second-order terms before being mapped to a continuous OCDR membership score.

Phase 5: Membership Inference. The resulting score is used for the final member/non-member decision and for evaluating membership-score distributional separability.

4.2 Dual Counterfactual Representation

The target similarity score s_t alone is not sufficient for reliable membership inference because response similarity is also affected by sample-dependent difficulty. A non-member sample may receive a high similarity score when it can be answered from related knowledge, while a member sample may still obtain a relatively low score. OCDR therefore evaluates s_t relative to matched Out-RAG systems whose knowledge bases exclude the target sample.

Let $\{s_1^{\text{out}}, \ldots, s_m^{\text{out}}\}$ denote the Out-RAG similarity scores, with mean μ_{out} and standard deviation σ_{out}. OCDR represents the target-to-reference difference in two complementary forms: the raw deviation preserves its magnitude and direction, while the normalized deviation accounts for the variability of the counterfactual reference. With ϵ for numerical stability, the dual representation is

$$\mathbf{f}(x) = \left[s_t - \mu_{\text{out}},\ \frac{s_t - \mu_{\text{out}}}{\sigma_{\text{out}} + \epsilon} \right]^\top = [f_1, f_2]^\top. \tag{4}$$

4.3 Development-Set Standardization

The two counterfactual features operate on different numerical scales: f_1 preserves the raw target-to-reference deviation, while f_2 normalizes it by sample-specific Out-RAG variability. To prevent scale differences from affecting the learned scorer, OCDR standardizes both features using development-set statistics fixed before held-out evaluation. Let μ_j^{dev} and σ_j^{dev} denote the development-set mean and standard deviation of f_j, respectively. The standardized representation is

$$z_j = \frac{f_j - \mu_j^{\text{dev}}}{\sigma_j^{\text{dev}} + \epsilon}, \qquad j \in \{1, 2\}. \tag{5}$$

4.4 Compact Quadratic Scoring

After standardization, (z_1, z_2) encode the absolute and uncertainty-normalized counterfactual deviations. Because their effects may be nonlinear and interactive, OCDR applies a compact second-order expansion. The linear, squared, and cross terms preserve the original signals while capturing magnitude-dependent effects and interactions:

$$\phi(\mathbf{z}) = \left[z_1,\ z_2,\ z_1^2,\ z_1 z_2,\ z_2^2 \right]^\top. \tag{6}$$

The expanded representation is mapped to a scalar membership score using L2-regularized logistic regression. This design preserves a compact scoring model while allowing a nonlinear decision boundary in the original two-dimensional counterfactual feature space. Let β_0 denote the intercept, $\boldsymbol{\beta}$ the learned coefficient vector, and $\sigma(\cdot)$ the logistic sigmoid. We fix the regularization parameter to $C = 1.0$, with all classifier parameters estimated on development data and frozen before test evaluation. The final OCDR membership score is

$$s_{\text{OCDR}}(x) = \sigma\big(\beta_0 + \boldsymbol{\beta}^\top \phi(\mathbf{z})\big). \tag{7}$$

5 Experimental Setup

We evaluate OCDR on HealthCareMagic and AG News across multiple generator backbones, considering both membership-inference utility and member/non-member score separability.

5.1 Data and Protocol

HealthCareMagic contains medical questions with free-form answers (6; 12), while AG News follows the benchmark of(13). We evaluate Mistral-7B-Instruct-v0.2 (15), Qwen3-1.7B (17), and Phi-3.5-mini (16). All target RAGs use all-MiniLM-L6-v2 (14) as the dense retriever with retrieval depth $k = 4$.

We use a Dev600→Test400 protocol with balanced splits: Dev600 contains 300 members and 300 non-members, and the disjoint Test400 contains 200 members and 200 non-members. Within each dataset–generator setting, all methods use the same Test400 samples. OCDR uses Dev600 for preprocessing and scorer fitting, while S-MIA uses it for threshold selection. All attacks assume black-box response access.

5.2 Comparison Methods

We compare OCDR with S-MIA and DC-MIA. S-MIA uses response–target semantic similarity as its membership score, while DC-MIA calibrates target behavior using matched reference-RAG responses through its two-stage procedure. OCDR follows the counterfactual reference setting described in Section 4.

5.3 Implementation and Metrics

OCDR uses $m = 8$ matched Out-RAG references per target sample, with all parameters fixed before Test400 evaluation.

We report ROC-AUC, TPR@1%FPR, and TPR@5%FPR for attack utility, together with KDE-OVL and Hist-OVL$_{80}$ for score overlap. Higher AUC and TPR are better, while lower overlap indicates better member/non-member separation.

6 Results and Analysis

6.1 Main Results

We first examine whether OCDR improves membership-inference utility and score-distribution separability across datasets and generator models. Table 1 summarizes the results under the unified Dev600→Test400 protocol.

Across HealthCareMagic, OCDR consistently produces lower score overlap than both baselines across all three generator models and achieves the strongest low-FPR performance. Its AUC behavior is more model-dependent: OCDR achieves the highest AUC with Mistral, whereas DC-MIA obtains higher AUC with Qwen and Phi. On AG News, S-MIA achieves the highest AUC and lowest overlap, while DC-MIA performs best at the reported low-FPR operating points. OCDR nevertheless reduces both overlap measures relative to DC-MIA. These results show that global ranking performance, low-FPR detection, and score-distribution separability need not vary together.

Table 1: Membership-inference performance under the unified Dev600→Test400 protocol. Higher AUC and TPR are better; lower overlap is better.

Dataset	Generator	Method	AUC↑	TPR@1%FPR↑	TPR@5%FPR↑	KDE-OVL↓	Hist-OVL$_{80}$ ↓
HealthCareMagic	Mistral-7B	S-MIA	0.6306	0.0700	0.1600	0.8351	0.6500
		DC-MIA	0.6270	0.0300	0.1750	0.9324	0.8950
		OCDR (Ours)	0.6614	0.0700	0.2600	**0.7743**	**0.5700**
	Qwen3-1.7B	S-MIA	0.6900	0.2450	0.3600	0.6744	0.5850
		DC-MIA	0.7920	0.0000	0.3200	0.6780	0.7400
		OCDR (Ours)	0.7207	0.2600	0.3800	**0.6162**	**0.5500**
	Phi-3.5-mini	S-MIA	0.7004	0.0850	0.2250	0.7033	0.5950
		DC-MIA	0.7870	0.0000	0.3250	0.7248	0.8900
		OCDR (Ours)	0.7591	0.2750	0.4350	**0.5744**	**0.4750**
AG News	Mistral-7B	S-MIA	0.7358	0.1150	0.2700	0.6591	0.5550
		DC-MIA	0.6940	0.1700	0.3150	0.7495	0.6950
		OCDR (Ours)	0.6817	0.1650	0.2650	**0.7201**	**0.6550**

6.2 ROC AND LOW-FPR BEHAVIOR

To examine this distinction more closely, we compare the ROC behavior of the attacks on HealthCareMagic and AG News using the same Mistral-7B generator. Figure **??** reports both the full ROC curves and enlarged low-FPR regions, allowing global ranking behavior to be compared with performance under strict false-positive constraints.

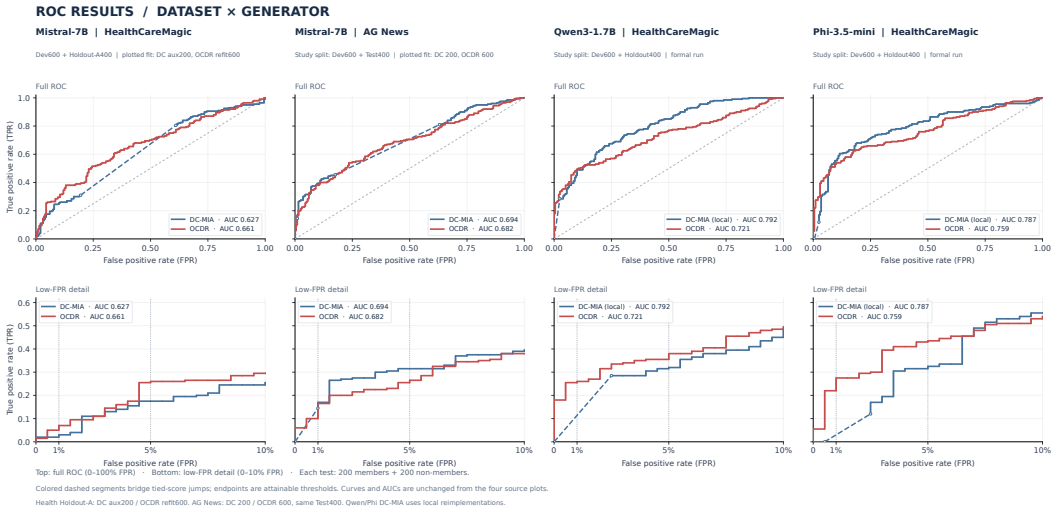

Figure 3: ROC behavior across datasets and generators. The upper row presents the complete ROC curves, while the lower row enlarges the low-FPR region from 0% to 10%. All evaluations use balanced test sets containing 200 members and 200 non-members.

The ROC curves exhibit different behavior across the two datasets. On HealthCareMagic, OCDR provides stronger low-FPR detection while retaining competitive global ranking performance. On AG News, DC-MIA shows stronger performance at the reported low-FPR operating points. Together with Table 1, this contrast further illustrates that global ranking performance, low-FPR detection, and score-distribution separability characterize different aspects of membership-inference behavior.

7 ABLATION AND ROBUSTNESS ANALYSIS

We examine how the OCDR representation affects score overlap and whether similar behavior appears across generator backbones. The representation ablations are refitted and evaluated on the same HealthCareMagic/Mistral AUX200→Test400 split. Cross-generator results are reported descriptively because the available settings do not all use identical development protocols or baseline implementations.

7.1 REPRESENTATION DESIGN ABLATION

The two counterfactual components are $f_1 = s_t - \mu_{\text{out}}$ and $f_2 = f_1/(\sigma_{\text{out}} + \epsilon)$, with z_1, z_2 denoting their development-standardized values. We evaluate nested prefixes of the quadratic feature expansion:

$$\begin{aligned} D_1 &= [z_1], & D_2 &= [z_1, z_2], \\ D_3 &= [z_1, z_2, z_1^2], & D_4 &= [z_1, z_2, z_1^2, z_1 z_2], \\ D_{\text{final}} &= [z_1, z_2, z_1^2, z_1 z_2, z_2^2]. \end{aligned} \tag{8}$$

All variants use the same L2-regularized logistic classifier and the same balanced Test400 split. We also fit f_1 only and f_2 only. The f_1-only model is equivalent to D_1, while the f_2-only classifier collapses to a constant score, yielding Hist_{80}-OVL$= 1$ and undefined KDE.

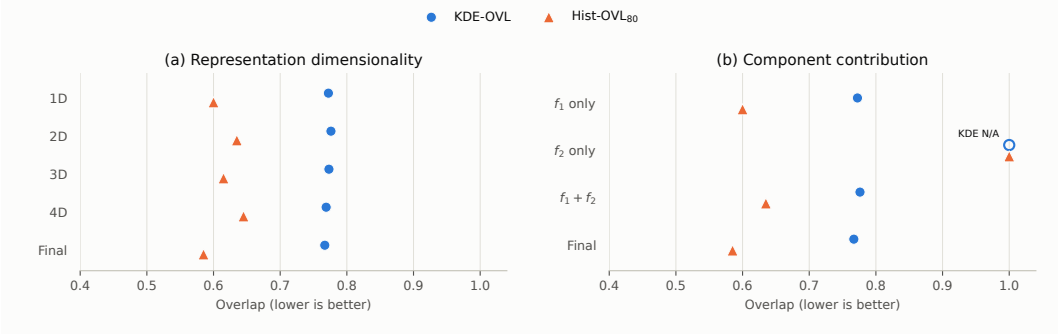

Figure 4: HealthCareMagic/Mistral representation ablation on the AUX200→Test400 split. (a) Nested prefixes of the quadratic OCDR representation. (b) Single components, their linear combination, and the complete scorer. Blue circles denote KDE overlap and orange triangles denote Hist_{80} overlap; lower is better. For f_2 only, KDE is undefined because the fitted scores are constant, while Hist_{80}-OVL is 1.

Table 2: HealthCareMagic/Mistral ablation results on the AUX200→Test400 split. Lower overlap is better. "—" denotes undefined KDE for constant scores.

Variant	Columns	AUC ↑	TPR@1% ↑	TPR@5% ↑	KDE-OVL ↓	Hist_{80}-OVL ↓
D_1 / f_1 only	1	0.659	0.050	0.215	0.772	0.600
D_2 / $f_1 + f_2$ linear	2	0.656	0.045	0.210	0.776	0.635
D_3	3	0.657	0.045	0.210	0.773	0.615
D_4	4	0.661	0.065	0.240	0.769	0.645
Final OCDR	5	0.662	0.070	0.260	0.767	0.585
f_2 only	1	0.500	0.000	0.000	—	1.000

Figure 4 shows that overlap does not decrease monotonically as features are added. Although D_2 has higher overlap than D_1, the complete five-column scorer achieves the lowest observed histogram overlap (0.585) and slightly lower KDE overlap than the other prefixes. Thus, the full representation performs best on this split, but not every added term is individually beneficial.

7.2 Complete Metric View

Table 2 reports attack utility together with overlap for the same ablation. D_1 corresponds to f_1 only, and D_2 to the linear $f_1 + f_2$ model.

The complete scorer has the highest observed TPR@5% FPR (0.260) and the lowest histogram overlap (0.585), while the intermediate variants remain non-monotonic. Thus, the ablation supports the complete feature combination on this split rather than a claim that performance improves with every added term.

7.3 Cross-Generator Comparison

Figure 5 compares DC-MIA and OCDR overlap on HealthCareMagic with Mistral, Qwen, and Phi. Qwen and Phi use matched Dev600-based settings, while the Mistral result uses the AUX200→Test400 configuration and a separate DC-MIA implementation. The comparison is therefore descriptive rather than a fully controlled cross-generator experiment.

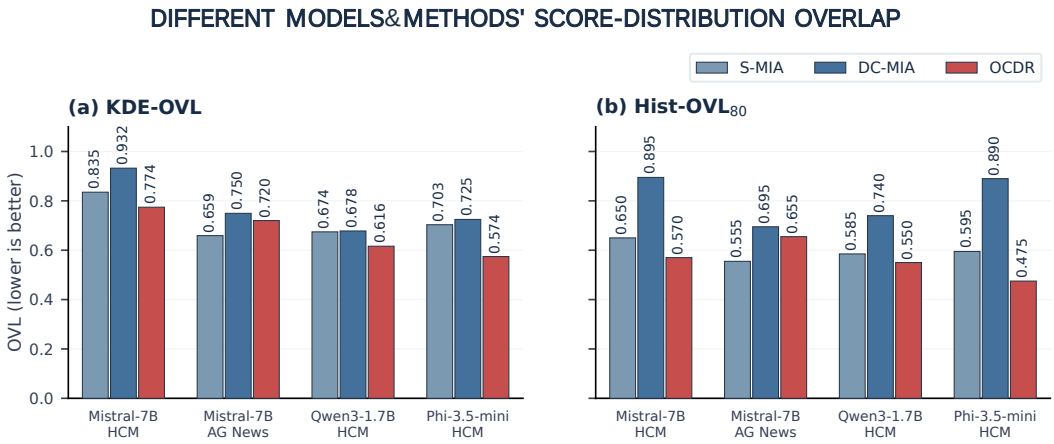

Figure 5: HealthCareMagic overlap for DC-MIA and OCDR across three generator settings. (a) KDE overlap; (b) Hist_{80} overlap. Lower is better. Protocol differences across generators prevent interpreting this as a fully controlled robustness comparison.

OCDR has lower histogram overlap than DC-MIA for Mistral (0.585 vs. 0.895), Qwen (0.550 vs. 0.740), and Phi (0.475 vs. 0.890). KDE overlap is lower for OCDR on Qwen and Phi, but higher on Mistral. The available results therefore show a recurring histogram-overlap reduction, whereas KDE behavior remains generator- or protocol-dependent. Attack utility also varies across settings, reinforcing that AUC, low-FPR TPR, and score overlap capture different aspects of membership-inference behavior.

8 Limitations and Future Work

OCDR requires multiple matched Out-RAG references per target, introducing additional computation compared with single-response attacks. Future work could reduce this cost through reference reuse, adaptive selection, or smaller reference sets.

The current evaluation also covers a limited set of datasets, generators, and protocol configurations. More controlled experiments across generator backbones, retrieval depths, reference counts, and generation settings are needed to assess how consistently the observed overlap behavior generalizes.

9 Discussion and Conclusion

Our results show that membership-inference utility and score-distribution separability capture different aspects of attack behavior. AUC reflects global ranking, low-FPR TPR measures performance at restrictive operating points, and overlap quantifies ambiguity between member and non-member scores. These quantities need not improve together.

OCDR combines raw and variance-normalized counterfactual deviations through a compact quadratic scorer. The ablation results show that adding features does not yield monotonic improvements, although the complete representation provides the strongest observed histogram separation and low-FPR behavior on the evaluated Mistral split.

Overall, OCDR provides an overlap-aware perspective on RAG membership inference and shows that score overlap can reveal behavior not captured by AUC or low-FPR TPR alone. This supports overlap-aware evaluation as a useful complement to conventional attack metrics.

AI USE STATEMENT

Generative AI tools were used as research and writing assistants to improve clarity, consistency, LaTeX formatting, and code for tables and figures. AI-assisted suggestions were checked against the underlying experimental records, and all research decisions, interpretations, and conclusions were made by the authors, who take full responsibility for the manuscript.

REFERENCES

[1] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Retrieval Augmented Language Model Pre-Training," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 3929–3938, 2020.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[3] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," *2017 IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership Inference Attacks From First Principles," *2022 IEEE Symposium on Security and Privacy (S&P)*, pp. 1897–1914, 2022.

[6] M. Anderson, G. Amit, and A. Goldsteen, "Is My Data in Your Retrieval Database? Membership Inference Attacks Against Retrieval Augmented Generation," *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP)*, vol. 2, pp. 474–485, 2025.

[7] Y. Li, G. Liu, C. Wang, and Y. Yang, "Generating Is Believing: Membership Inference Attacks against Retrieval-Augmented Generation," *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

[8] G. Wang, J. He, H. Li, M. Zhang, and D. Feng, "RAG-leaks: Difficulty-Calibrated Membership Inference Attacks on Retrieval-Augmented Generation," *Science China Information Sciences*, vol. 68, no. 6, Art. no. 160102, 2025.

[9] X. Gao, X. Meng, Y. Dong, Z. Li, and S. Guo, "DCMI: A Differential Calibration Membership Inference Attack Against Retrieval-Augmented Generation," *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 4184–4198, 2025.

[10] A. Naseh, Y. Peng, A. Suri, H. Chaudhari, A. Oprea, and A. Houmansadr, "Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation," *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1245–1259, 2025.

[11] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, and J. Tang, "The Good and the Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)," *arXiv preprint arXiv:2402.16893*, 2024.

[12] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, and P. Xie, "MedDialog: Large-Scale Medical Dialogue Datasets," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9241–9250, 2020.

[13] X. Zhang, J. Zhao, and Y. LeCun, "Character-Level Convolutional Networks for Text Classification," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 649–657, 2015.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[15] A. Q. Jiang et al., "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[16] M. Abdin et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," *arXiv preprint arXiv:2404.14219*, 2024.

[17] A. Yang et al., "Qwen3 Technical Report," *arXiv preprint arXiv:2505.09388*, 2025.

[18] H. F. Inman and E. L. Bradley, "The Overlapping Coefficient as a Measure of Agreement Between Probability Distributions and Point Estimation of the Overlap of Two Normal Densities," *Communications in Statistics—Theory and Methods*, vol. 18, no. 10, pp. 3851–3874, 1989.

A Experimental Protocol and Configuration

This appendix uses the HealthCareMagic–Mistral Dev600-to-Holdout-A400 export as the primary representation evaluation. Dev600 contains 300 members and 300 non-members; Holdout-A contains 200 of each class. The stored sample identifiers confirm zero development–test overlap. The final OCDR scorer is fitted on Dev600 and applied unchanged to Holdout-A. This statement concerns OCDR fitting, not an assumption that every archived baseline was calibrated using the same development budget. The AUX200 representation ablation in Section 7 is a separate experiment and is not pooled with the primary results.

A.1 Splits and Method-Specific Provenance

Table A1 distinguishes OCDR development data from baseline provenance. Every displayed test comparison uses matching sample identifiers and labels within its setting. Equal test sizes do not imply equal baseline calibration budgets or identical DC-MIA implementations across generators.

Table A1: OCDR development and held-out evaluation data. All test sets contain 200 members and 200 non-members. Baseline calibration provenance is reported separately.

Dataset	Generator	OCDR dev n	Test n	Baseline provenance
HealthCareMagic	Mistral-7B	600	400	Archived DC run: AUX200
HealthCareMagic	Qwen3-1.7B	600	400	Local Dev600 baseline implementation
HealthCareMagic	Phi-3.5-mini	600	400	Local Dev600 baseline implementation
AG News	Mistral-7B	600	400	Complete DC run records AUX200

The AG News summary labels the complete DC-MIA development size as 600, whereas the underlying complete two-stage run records `n_aux=200`. We retain its matched Test400 scores but do not interpret this export as evidence of a Dev600-calibrated complete DC-MIA run. The primary HealthCareMagic–Mistral bundle likewise supplies an OCDR Dev600 refit alongside a baseline run whose metadata records AUX200. The reference count is $m = 8$ and retrieval depth is $k = 4$ in the recorded configurations.

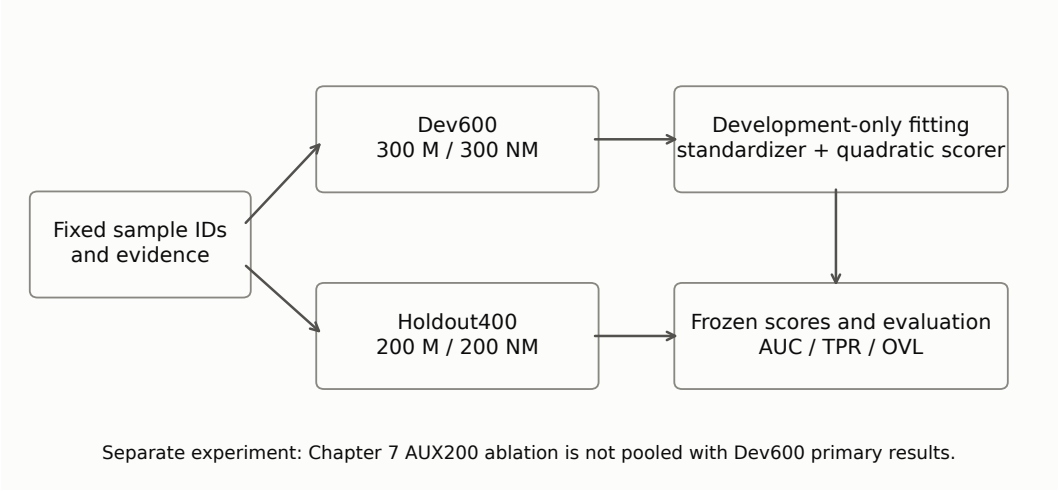

Figure A1: Development-only fitting and frozen held-out evaluation for OCDR. The primary HealthCareMagic scorer uses Dev600 and Holdout-A400. Baseline calibration provenance and the separate AUX200 ablation are not implied to follow this same fitting budget.

A.2 Counterfactual Statistics and Frozen Scorer

The two features are the signed deviation $f_1 = s_t - \mu_{\text{out}}$ and its variability-normalized counterpart $f_2 = f_1/(\sigma_{\text{out}} + \epsilon)$. The first feature is not an absolute-value transformation. For implementations explicitly recording sample standard deviation, the reference dispersion is

$$\sigma_{\text{out}} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(s_i^{\text{out}} - \mu_{\text{out}})^2}. \tag{9}$$

Reproduction of the Mistral Dev600 model uses its stored feature columns and stored standardization parameters rather than silently recomputing them under a different convention. The feature order is $[z_1, z_2, z_1^2, z_1 z_2, z_2^2]$. Table A2 reports the frozen parameters.

Table A2: Frozen HealthCareMagic–Mistral Dev600 OCDR scorer. Values are rounded for presentation; the figure-generation code uses the full-precision archived parameters.

Quantity	Value
Mean of f_1	0.0131121281
Mean of f_2	0.0295473953
Standard deviation of f_1	0.0786811556
Standard deviation of f_2	4.4709297263
Intercept	-0.0665723613
z_1 coefficient	0.4001462759
z_2 coefficient	0.2209319531
z_1^2 coefficient	0.0132456972
$z_1 z_2$ coefficient	0.0944333063
z_2^2 coefficient	0.0147374279
Regularization	L2, $C = 1.0$

The archived configuration records an initial 84-row training/36-row validation pilot for selecting C, followed by a 600-row development refit. The available record therefore does not support a claim that C was selected by validation across all 600 development rows. The frozen model reconstructs all 400 archived holdout scores with maximum absolute error 1.12×10^{-16}.

B Complete Experimental Results

B.1 Matched Held-Out Metrics

Table B1 reports metrics recalculated from the exported continuous scores. All methods in a setting use the same ordered label vector after sample-ID alignment. The KDE column uses the common numerical procedure in Appendix F.2; it is explicitly distinguished from older reported KDE values.

Table B1: Recomputed matched-test results. OCDR is Dev600-trained in all four settings; baseline development budgets are not uniformly matched. KDE† denotes the appendix's standardized recomputation, not a replacement label for archive-reported KDE. Mistral Health S-MIA is joined from the full-precision reproduction checkpoint after verifying all 400 test IDs and labels.

Dataset	Generator	Method	AUC	TPR@1%	TPR@5%	KDE†	Hist_{80}
HealthCareMagic	Mistral	S-MIA	0.6306	0.070	0.160	0.8330	0.650
		DC-MIA	0.6270	0.030	0.175	0.5869	0.895
		OCDR	0.6614	0.070	0.260	0.7656	0.570
HealthCareMagic	Qwen	S-MIA	0.6900	0.245	0.360	0.6748	0.585
		DC-MIA	0.7920	0.000	0.320	0.6808	0.740
		OCDR	0.7207	0.260	0.380	0.6164	0.550
HealthCareMagic	Phi	S-MIA	0.7004	0.085	0.225	0.7033	0.595
		DC-MIA	0.7870	0.000	0.325	0.7248	0.890
		OCDR	0.7591	0.275	0.435	0.5746	0.475
AG News	Mistral	S-MIA	0.7358	0.115	0.270	0.6605	0.555
		DC-MIA	0.6940	0.170	0.315	0.7499	0.695
		OCDR	0.6817	0.165	0.265	0.7203	0.655

The supplied HealthCareMagic–Mistral summary reports DC-MIA/OCDR KDE values of 0.932350/0.774340, while recomputation on its stored scores with the specified default Gaussian KDE gives 0.586915/0.765618. Its AUC, low-FPR TPR, and Hist_{80} values reproduce. We retain this discrepancy rather than treating the two KDE pairs as interchangeable. In particular, the stored-score KDE comparison does not support lower Mistral KDE overlap for OCDR. Qwen and Phi show lower OCDR KDE point estimates under this common recomputation, but retain higher DC-MIA AUC. Neither point estimates nor this descriptive appendix establish statistical significance.

B.2 Additional Low-FPR Behavior

At 1% FPR there are only two allowable false positives among 200 non-members. TPR values should therefore be interpreted as finite-sample ROC operating-point estimates, not evidence of a universally stable deployment threshold.

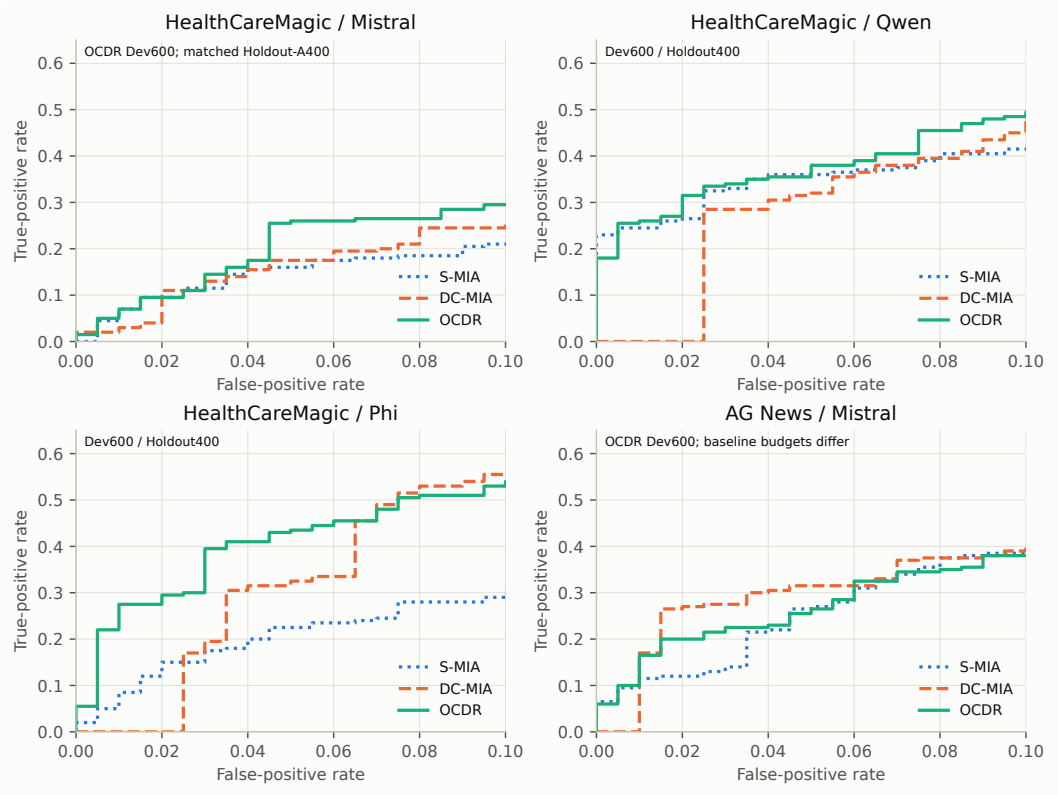

Figure B1: Empirical low-FPR ROC curves for four settings, restricted to $0 \leq \mathrm{FPR} \leq 0.1$. No smoothing or test-time refitting is applied. Mistral Health uses the Dev600 OCDR score and the archived complete DC-MIA score on identical Holdout-A400 records. Qwen/Phi use matched local baseline scores. AG News uses matched Test400 records with differing baseline development provenance. S-MIA is joined from the matched full-precision checkpoint with all 400 IDs and labels verified.

C Detailed Representation Analysis

C.1 Separation of the Main and Ablation Protocols

The complete numerical prefix and component ablation remains in Table 2. It uses the separately archived AUX200 development features and the same 400 test identifiers, whereas the primary model here is the Dev600 refit. Its final AUC/Hist$_{80}$ pair (0.66165/0.585) must not be substituted for the Dev600 pair (0.66140/0.570). The ablation KDE uses a 4096-point observed-range grid with a 10% margin, distinct from the appendix recomputation below. Dimension counts the number of polynomial scorer input columns, not independent counterfactual evidence sources. The constant-score f_2-only ablation is a property of that particular fitted model, not proof that the raw f_2 feature is constant or intrinsically uninformative. An undefined KDE bandwidth is reported as N/A, not as an estimated KDE overlap of 1. If a hollow symbol is displayed at overlap 1 in the main ablation figure, it denotes coincident empirical point masses only.

C.2 Feature-Space Geometry of the Dev600 Model

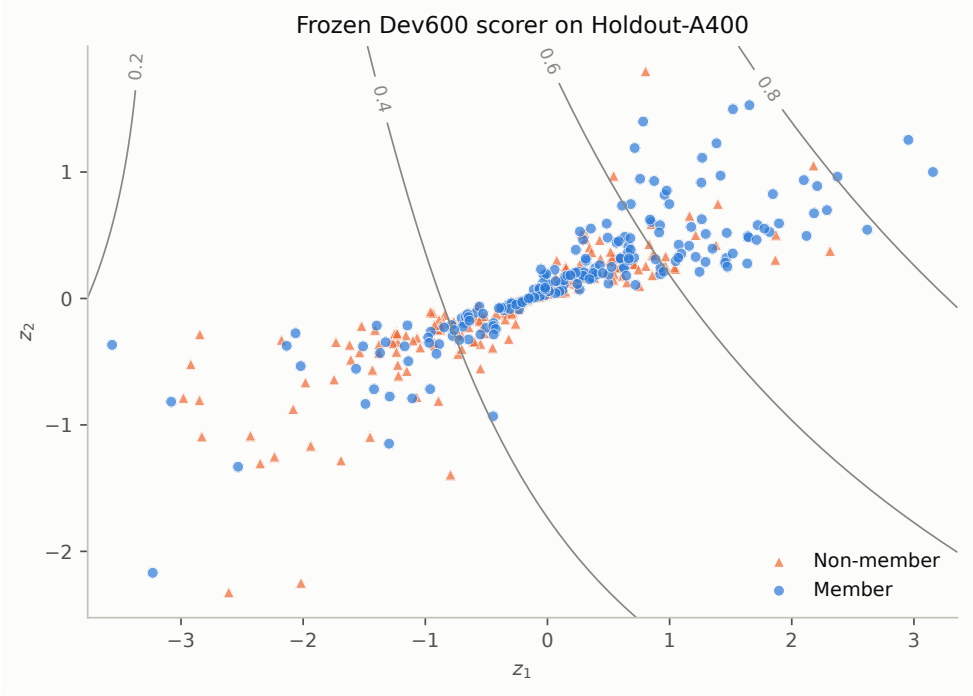

Figure C1: All 400 Holdout-A samples in the standardized feature space of the frozen HealthCareMagic–Mistral Dev600 scorer. Circles denote members and triangles non-members. Gray curves are score contours at 0.2, 0.4, 0.6, and 0.8 computed from the archived coefficients, without fitting to holdout labels.

The two features are related views of the same signed target-to-reference deviation. Their visible dependence and class intermingling do not imply perfect separation. The contours illustrate the nonlinear score surface; they are a visualization of the frozen classifier rather than an additional independent evaluation.

D Additional Overlap Analysis

All analyses in this section keep the primary Mistral Health Holdout-A400 labels and S-MIA/DC-MIA/OCDR scores fixed. They change only the estimator used to summarize member/non-member overlap. S-MIA is taken from the matched full-precision replication checkpoint, not inferred from the OCDR target similarity.

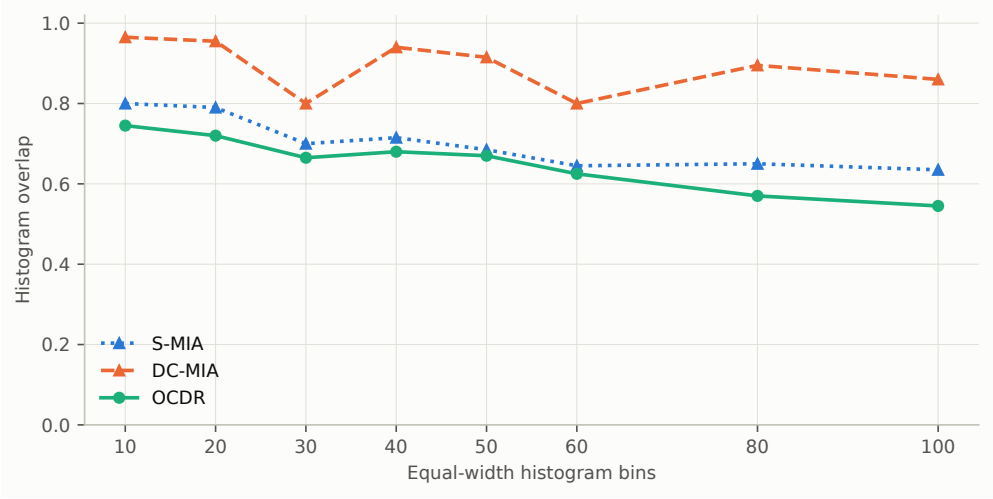

Figure D1: Hist-OVL at 10, 20, 30, 40, 50, 60, 80, and 100 equal-width bins on fixed Holdout-A400 scores. Members and non-members share edges within each method. No bin count is selected to optimize test performance.

Table D1: Numerical histogram-resolution sensitivity; lower is better.

Method	10	20	30	40	50	60	80	100
S-MIA	.800	.790	.700	.715	.685	.645	.650	.635
DC-MIA	.965	.955	.800	.940	.915	.800	.895	.860
OCDR	.745	.720	.665	.680	.670	.625	.570	.545

D.1 Histogram-Resolution Sensitivity

OCDR has lower empirical histogram overlap at every listed resolution. The coefficients are not monotonic in the number of bins, especially for DC-MIA; this illustrates why a histogram estimate should be accompanied by its partition definition rather than treated as an estimator-independent constant.

D.2 Primary Score Distributions

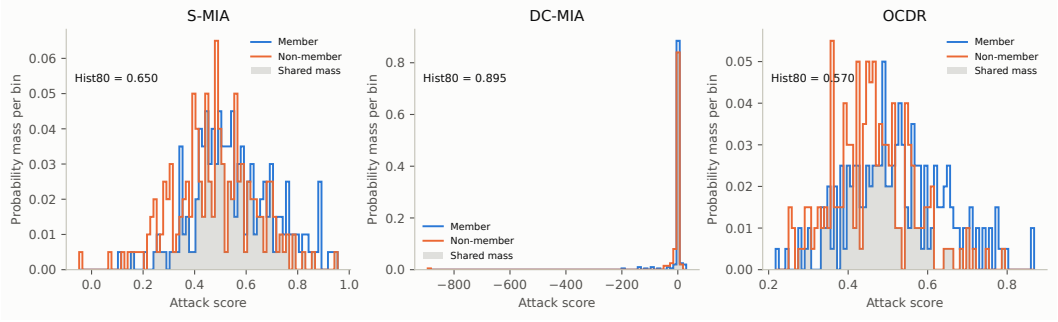

Figure D2: Primary member/non-member score histograms with 80 shared equal-width bins per method. Gray shading denotes the shared probability mass. Each panel uses its own score axis; DC-MIA scores and OCDR probabilities are not forced onto an artificial common scale.

The distributions contain the same scores used for the primary metrics. Their different score scales, concentrations, and tails help explain why smoothed KDE and empirical histogram overlap can produce different rankings. These plots are not independent experimental replications.

D.3 KDE-Bandwidth Sensitivity

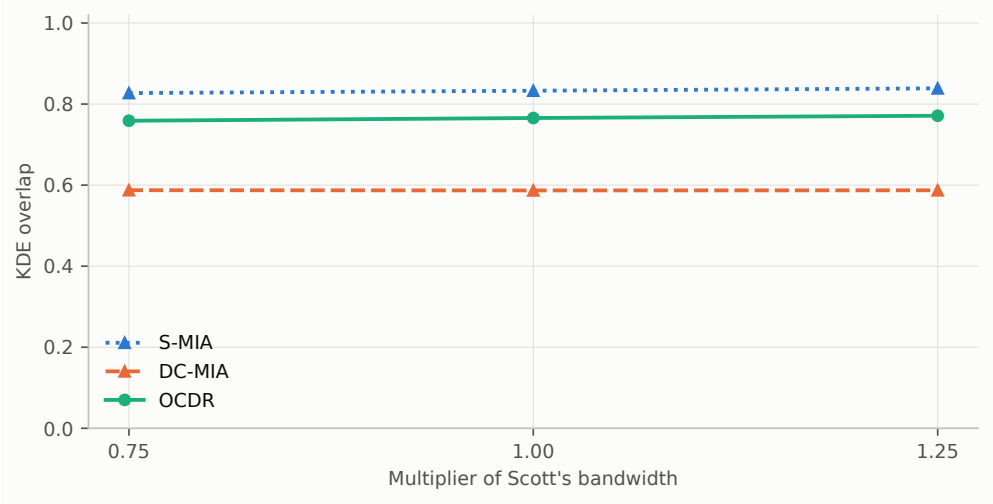

Figure D3: KDE overlap with bandwidth multipliers 0.75, 1.0, and 1.25 relative to Scott's rule. Both class densities receive the same multiplier; held-out scores remain fixed. Integration uses the wide common grid described in Appendix F.2.

Across these multipliers, OCDR has higher estimated KDE overlap than DC-MIA on the primary Mistral scores, despite lower histogram overlap. The reported histogram advantage should consequently not be generalized to every overlap estimator. Bandwidth sensitivity is a diagnostic, not a procedure for tuning attack performance on the test set.

Table D2: KDE-bandwidth sensitivity under the appendix recomputation.

Method	0.75	1.00	1.25
S-MIA	0.8271	0.8330	0.8388
DC-MIA	0.5875	0.5869	0.5872
OCDR	0.7589	0.7656	0.7712

E Randomness and Configuration Sensitivity

E.1 Available Evidence and Unrun Experiments

The supplied archives do not contain a completed series of independent generation runs with different random seeds or a completed reference-count sweep. We therefore do not report a seed-sensitivity curve, uncertainty across independent generations, or a claim that $m = 8$ is optimal. Resampling 400 fixed scores would quantify test-sample uncertainty, not generation-seed robustness.

Table E1: Status of additional robustness analyses in the supplied artifacts.

Analysis	Available evidence	Supported interpretation
Histogram bins	Fixed-score recomputation	Estimator resolution sensitivity
KDE bandwidth	Fixed-score recomputation	Estimator smoothing sensitivity
Generation seeds	No completed multi-seed series	Not evaluated
Reference count m	Formal results use $m = 8$	No optimality claim
AG development size	AUX200 and Dev600 OCDR	Confounded by $C = 3$ versus $C = 1$

E.2 Reference-Cost and Development-Size Limitations

Changing m requires recomputing reference statistics, fitting preprocessing and the scorer on the corresponding development evidence, and evaluating the frozen model on the same held-out IDs. Subsampling saved reference scores would be a new offline ablation, not an independent end-to-end generation run; none is presented here. The AG News AUX200/Dev600 results also change regularization, so their difference does not isolate the causal effect of development-set size.

F Implementation and Reproducibility

F.1 Randomness Control

The Qwen/Phi formal implementation records seed 0 and sets Python, NumPy, PyTorch, and CUDA seeds to that value. Its model generation uses `do_sample=False`, so these exports must not be described as repeated stochastic decoding runs. The Mistral evidence filenames identify seed 0, but the supplied Dev600 refit metadata does not fully document separate fitting RNG states; those states are not inferred from the filenames.

Table F1: Verified randomness controls; an undocumented state is not guessed.

Component	Qwen/Phi implementation	Mistral Dev600 export
Split/evidence seed	0	Seed-0 evidence identifiers
Python/NumPy seed	0 / 0	Not separately documented
PyTorch/CUDA seed	0 / 0	Not separately documented
Generation sampling	Disabled	Not established by refit metadata
Logistic RNG state	Not separately specified	Not separately documented

F.2 Metric Implementation

AUC is calculated from continuous scores with equal scores receiving average ranks. The empirical ROC aggregates tied thresholds; TPR at nominal FPR α is the maximum observed TPR among ROC points with $\mathrm{FPR} \leq \alpha + 10^{-12}$. This is a descriptive ROC operating point, not a separately frozen development-calibrated deployment threshold.

Hist-OVL uses the common score minimum and maximum within each method to form B equal-width bins. Each class's counts are divided by its own sample count before summing the minimum probability mass across bins. The main primary histogram estimate uses $B = 80$.

For the standardized appendix KDE, each class uses SciPy Gaussian KDE with Scott's bandwidth. A shared 32,768-point grid covers the full observed score range plus a margin of eight times the larger class-specific kernel standard deviation at each end. We integrate the minimum densities with the trapezoidal rule. In the bandwidth experiment both class bandwidths are multiplied by the same factor and the integration interval is recalculated accordingly. This wide-grid approximation is explicitly different from the finite 4096-point grids of other exports. The delivered CSVs retain the recomputed numbers without substituting them for archive-reported estimates.

F.3 Artifact Provenance and Reproduction Procedure

The primary Mistral scores are read from `holdoutA400_all_features_all_scores.csv`, paired with `development600_all_17_features.csv` and `ours_top2_training_and_coefficients.json` in the supplied Dev600/Holdout400 bundle. Complete DC-MIA is the `dc_mia_staged` column, not its LLR-only component; final OCDR is `original_top2_refit600_score`, not the pilot, V4-17D, or RCDR score. Qwen/Phi use their formal `holdout_scores.csv`. Mistral Health S-MIA uses the `score` field of `mistral-7b-healthcaremagic-formal2000-seed0_seed_0_smia.jsonl`, filtered to the same 400 IDs; its AUC is 0.63055 and Hist_{80} is 0.650. AG News aligns the final Dev600 OCDR scores with complete DC-MIA and S-MIA records by sample ID and checks labels before comparing them.

1. Read the archived development/test IDs and verify disjointness and class counts.
2. Preserve the corresponding target/reference evidence and final method-specific score columns.
3. Apply the stored development-only feature means and standard deviations.
4. Construct $[z_1, z_2, z_1^2, z_1 z_2, z_2^2]$ in the recorded order.
5. Evaluate the frozen logistic parameters without using holdout labels for fitting or parameter selection.
6. Align compared methods on identical test IDs and verify all labels.
7. Recompute continuous-score AUC, empirical low-FPR TPR, and both overlap estimators from the same ordered rows.
8. Record estimator definitions and retain differences from older reported values instead of silently overwriting them.

The delivered reproducibility directory contains the matched score vectors, full-precision frozen parameters, feature-space coordinates, estimator-sensitivity tables, and plotting script. Generation-seed robustness and reference-count robustness remain unperformed experiments rather than missing plotting steps.