

TraceMIA: Interpretable Membership Inference Attack via Lexical Patterns in Reasoning Traces

Bo Zhang^{1,2,3†}, Yongtang Wu⁴, Ning Liu⁴, Jun Zhang^{2,3*}, Zhunchen Luo^{2,3},
Yunbo Cao^{2,3}, Xiaoying Bai^{2,3}, Guotong Geng^{2,3}, Zhong Wang^{1*}

¹PLA Rocket Force University of Engineering ²Center of Information Research, PLA AMS

³Discipline and Technology Research Center for Large Model Intelligence Applications

⁴School of Software, Shandong University

mcgrady150318@163.com, dsp863wang@163.com

Abstract

As large language models (LLMs) increasingly rely on web-scale corpora, membership inference attacks (MIAs) have become essential tools for auditing unauthorized data usage. However, traditional token-level MIAs fail to generalize to large reasoning models (LRMs) due to data distribution shifts introduced during fine-tuning. While recent studies address this by encoding reasoning traces into latent semantic spaces, they sacrifice the interpretability strictly required for real-world copyright verification. In this paper, we demonstrate that an LRM’s internal memory states are explicitly leaked through its stylistic word choices. Our analysis reveals a distinct lexical dichotomy: non-member traces tend toward a "guessing-like" tone, whereas member traces exhibit a "remember-like" style. Motivated by this interesting phenomenon, we propose TraceMIA, an interpretable MIA framework that audits training corpora by computing membership scores based on the relative frequencies of discriminative cue words. Extensive empirical evaluations across representative LRMs and diverse datasets demonstrate that TraceMIA achieves robust detection efficacy while providing the transparent, human-readable evidence critical for practical auditing scenarios.

1 Introduction

Large language models (LLMs) have achieved remarkable progress in natural language understanding, code generation, and multi-step reasoning (Guo et al., 2025; Yang et al., 2025). However, because these models are trained on web-scale corpora that encompass copyrighted books, news articles, and proprietary documents, their deployment raises serious concerns regarding the unauthorized use of protected content (Carlini et al., 2021; Karamolegkou et al., 2023).

*Corresponding authors.

†Work performed while an intern at the Center of Information Research, PLA Academy of Military Science.

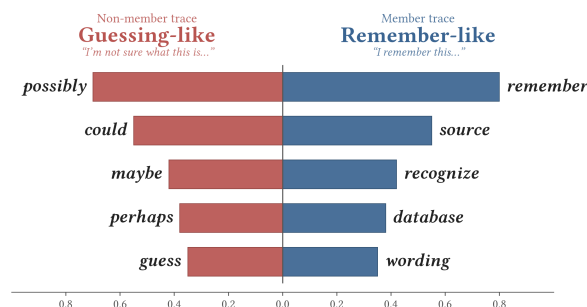


Figure 1: Top-5 discriminative words per class in reasoning traces of Qwen3-8B on WikiMIA-25. Bar lengths show the difference in the fraction of traces containing each word between member and non-member text sequences.

In recent years, membership inference attacks (MIAs) have been widely adopted to audit language model training data (Shokri et al., 2017). These techniques typically determine whether a given text sequence is part of the training corpus by analyzing token-level statistical features extracted from the target model (Carlini et al., 2021; Mattern et al., 2023; Shi et al., 2024). Existing approaches have demonstrated strong performance on established detection benchmarks—such as arXiv-based datasets (Meeus et al., 2024) and WikiMIA (Shi et al., 2024)—when evaluated against pretrained language models (PLMs) like GPT-NeoX (Black et al., 2022) and LLaMA (Touvron et al., 2023).

However, recent work (Hu et al., 2026) demonstrates that prior MIA methods fail to generalize to large reasoning models (LRMs). Detection accuracy drops significantly due to the data distribution shift introduced during the fine-tuning process from PLMs to LRMs, indicating that fragile token-level statistical features alone are insufficient in this new setting. To address this, Hu et al. (2026) propose encoding reasoning traces into a latent semantic space to enhance detection capabilities, albeit at the expense of interpretability. This creates a critical

dilemma: in real-world scenarios, content publishers require transparent, human-readable evidence to verify the unauthorized use of their copyrighted corpora. Rather than mapping traces to opaque latent spaces, we hypothesize that an LRM’s internal memory states are explicitly leaked through its stylistic word choices during reasoning.

To validate this and enhance the transparency of MIAs in LRMs, we conduct a preliminary study on Qwen3-8B. Specifically, we sample 100 member and 100 non-member text sequences from WikiMIA-25 (Yi and Li, 2026) and collect reasoning traces by prompting the model to determine whether each sequence was part of its training corpus. We then analyze the lexical patterns within these traces and visualize the top-5 most discriminative words after χ^2 filtering in Figure 1. Interestingly, vocabulary in the non-member traces tends toward a "guessing-like" style (e.g., *maybe*, *perhaps*, *possibly*), while member traces exhibit a "remember-like" style (e.g., *remember*, *source*, *recognize*). Notably, there is no overlap among the top-5 words across the two classes. Consequently, we argue that such macroscopic lexical patterns inherent in reasoning traces serve as highly discriminative and interpretable indicators that are more robust to fine-tuning shifts.

Motivated by these observations, we propose TraceMIA, an interpretable MIA framework designed to audit the training corpora of LRMs. The framework operates through a transparent, three-stage pipeline. First, utilizing a small labeled reference set, we prompt the target LRM to generate a reasoning trace for each text sequence. Next, we apply statistical filtering to these traces to construct distinct member-indicative and non-member-indicative lexicons. Finally, to evaluate a new query sequence, we compute a membership score by analyzing the relative frequencies of these cue words within its trace; a higher score signifies a greater probability that the sequence is part of the training corpus.

Our main contributions can be summarized as follows:

- We reveal, through a preliminary study, that vocabulary in non-member traces tends toward a "guessing-like" style, whereas member traces exhibit a "remember-like" style. This demonstrates that lexical patterns within reasoning traces serve as highly discriminative indicators for membership inference.

- We propose **TraceMIA**, a novel and interpretable membership inference attack framework. By leveraging inherent lexical patterns in the reasoning traces of LRMs, our method provides transparent evidence and demonstrates high practicality for real-world auditing scenarios.
- Extensive empirical evaluations across three representative LRMs and diverse datasets demonstrate the effectiveness of TraceMIA.

2 Related Work

2.1 MIA with Token-Level Probabilities

Membership inference attacks on language models are commonly grounded in the observation that models tend to assign higher likelihood to training data than to unseen text (Yeom et al., 2018). Early work directly thresholds perplexity as a membership score; subsequent methods improve calibration through likelihood ratio tests against a reference model and Zlib-entropy-based complexity adjustments (Carlini et al., 2021), or via neighbourhood-based perturbation comparisons (Mattern et al., 2023). Min-K% (Shi et al., 2024) takes a complementary perspective, using the average log-probability of the lowest- $k\%$ tokens as a membership signal, on the premise that training members are less likely than non-members to contain outlier tokens. Min-K%++ (Zhang et al., 2025) builds on this by normalizing token-level log-probabilities against the mean and variance of the predicted next-token distribution. Gap-K% (Kwak and Kim, 2026) exploits the fact that training samples exhibit smaller top-1 prediction gaps than unseen text, with sequential smoothing for local correlations. Although these methods demonstrate strong performance when token-level probabilities are accessible, their applicability is substantially limited when only the model’s generated text is available.

2.2 MIA on Generated Text

Under strict black-box constraints where only model-generated text is observable, prior work circumvents the need for token-level probabilities through sampling-based strategies. SaMIA (Kaneko et al., 2025) measures n-gram overlap between sampled and reference suffixes, CDD (Dong et al., 2024) quantifies the peakedness of edit-distance distributions across temperature-perturbed continuations, and SimMIA (Yi and

Li, 2026) performs word-by-word sampling with soft semantic scoring. These methods, however, operate on final outputs and remain blind to the deliberative process now exposed by reasoning models.

Recently, BlackSpectrum (Hu et al., 2026) demonstrates that chain-of-thought traces themselves carry statistically detectable membership signals, but encodes them into a latent semantic space, reducing direct interpretability. Both BlackSpectrum and TraceMIA therefore assume access to textual reasoning traces. TraceMIA instead retains membership-associated stylistic word choices as transparent, human-readable statistical evidence for practical auditing.

3 Method

3.1 Overview

We consider a data-owner auditing scenario where the auditor has strict black-box access to the target LRM’s generated text, including an exposed textual reasoning trace. Models that hide or only summarize their intermediate reasoning fall outside our current threat model. The reference set is constructed at low annotation cost: members from the owner’s corpus and non-members from post-cutoff content. We implement the lexical dichotomy through a three-stage pipeline (Figure 2): **Phase 1** elicits introspective reasoning traces from the target model for a small labeled reference set; **Phase 2** applies χ^2 filtering to construct member-cue (W^+) and non-member-cue (W^-) lexicons; **Phase 3** scores each query passage by the length-normalized cue-frequency difference in its reasoning trace. Phases 1–2 run once offline; Phase 3 requires a single model call per query.

3.2 Introspective Reasoning Elicitation

Given a text passage x , we construct an introspective prompt $\mathcal{P}(x)$ that instructs the target model $\mathcal{M}$ to reflect on whether it has previously encountered the passage, considering cues such as writing style, specific phrasing, and recognizable sources, and to emit a binary JSON judgment (full template in Appendix B). We extract the content between the model’s `<think>` tags as the reasoning trace τ_x and preprocess it by lowercasing and retaining alphabetic words of length ≥ 3 . We use two representations of the resulting words: $V(\tau_x)$ denotes the **word sequence** (preserving repetitions), and $\bar{V}(\tau_x)$ denotes its **unique-word set**. The same no-

tation applies to reference samples s via τ_s , $V(\tau_s)$, and $\bar{V}(\tau_s)$. We use τ_x as the unit of analysis rather than the final JSON judgment, since the latter collapses the model’s deliberation into a binary value and discards the richer lexical evidence in the trace; we report the binary self-report as an additional baseline in Appendix G.

3.3 Data-Driven Lexicon Construction

To identify which words carry membership signal, we use a small labeled reference set $\mathcal{D}_{\text{ref}} = \mathcal{D}_m \cup \mathcal{D}_n$, where $\mathcal{D}_m$ and $\mathcal{D}_n$ denote member and non-member samples respectively, each of size K ($|\mathcal{D}_m| = |\mathcal{D}_n| = K$). We elicit reasoning traces $\{\tau_s\}_{s \in \mathcal{D}_{\text{ref}}}$ for all samples with a single inference pass, and compute document-level word frequencies across both classes.

For each candidate word w , let

$$\begin{aligned} a &= |\{s \in \mathcal{D}_m : w \in \bar{V}(\tau_s)\}|, & b &= K - a, \\ c &= |\{s \in \mathcal{D}_n : w \in \bar{V}(\tau_s)\}|, & d &= K - c, \end{aligned}$$

yielding the 2×2 contingency table

	contains w	does not contain w
$\mathcal{D}_m$	a	b
$\mathcal{D}_n$	c	d

where each row aggregates over samples in the corresponding class. Let $\text{df}(w) := a + c$ denote the combined document frequency of w across both classes. We evaluate each word in two steps: (i) discard rare words with $\text{df}(w) < f_{\min}$ to suppress noise from low-count cells, and (ii) apply the χ^2 filtering with Yates’ continuity correction to the remaining table. Let $p_{\chi^2}(w)$ denote the resulting p -value, and define the class-conditional document frequencies $\pi_m(w) = a/K$ and $\pi_n(w) = c/K$. We then partition the surviving words by the direction of the distributional shift:

$$\begin{aligned} W^+ &= \{w \mid \text{df}(w) \geq f_{\min}, \\ &\quad p_{\chi^2}(w) < p_{\text{thresh}}, \\ &\quad \pi_m(w) > \pi_n(w)\}, \\ W^- &= \{w \mid \text{df}(w) \geq f_{\min}, \\ &\quad p_{\chi^2}(w) < p_{\text{thresh}}, \\ &\quad \pi_n(w) > \pi_m(w)\}. \end{aligned} \tag{1}$$

We use the conventional significance level $p_{\text{thresh}} = 0.05$ and minimum document-frequency threshold $f_{\min} = 3$, and set $K = 25$ unless otherwise noted.

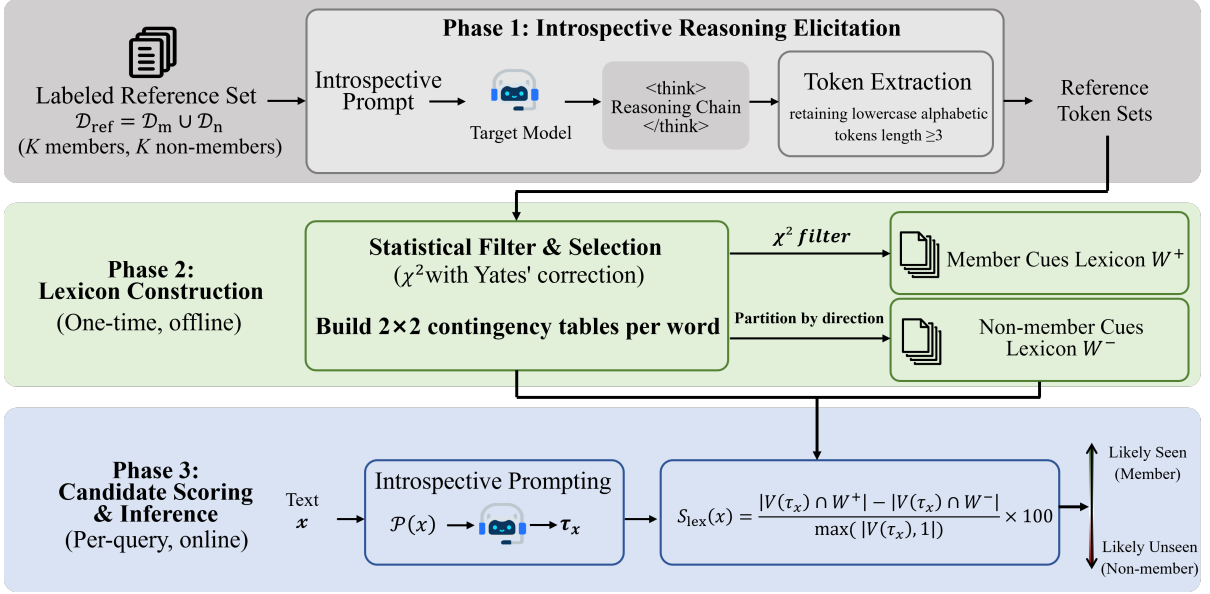


Figure 2: TraceMIA pipeline. Phase 1 elicits introspective reasoning traces from the target model for all labeled reference samples. Phase 2 applies χ^2 filtering to construct W^+ and W^- (see §3.3 for details). Phase 3 scores each query passage by the length-normalized cue frequency difference $S_{\text{lex}}(x)$ in its reasoning trace.

W^+ captures *member cues* reflecting the remember-like style, while W^- captures *non-member cues* characteristic of the guessing-like style. A notable property of this procedure is that high-frequency function words are filtered out automatically: they appear with comparable document frequency across both classes and therefore fail the significance test, obviating manual stop-word curation. The lexicon is constructed once and reused for all subsequent queries.

3.4 Scoring and Inference

For any query passage x , we elicit its reasoning trace τ_x with a single model call and compute the membership score as the length-normalized difference between member-cue and non-member-cue occurrence counts:

$$S_{\text{lex}}(x) = \frac{|V(\tau_x) \cap W^+| - |V(\tau_x) \cap W^-|}{\max(|V(\tau_x)|, 1)} \cdot 100, \quad (2)$$

where $V(\tau_x)$ is treated as a multiset, so each occurrence of a cue word in τ_x contributes once to the numerator. Length normalization removes the effect of variable trace length, the multiplicative factor of 100 rescales scores to a readable range, and the $\max$ in the denominator prevents division by zero in the degenerate case of an empty trace. A high $S_{\text{lex}}(x)$ indicates that the trace is dominated by member cues, exhibiting the remember-like style that signals x was seen in training; a low $S_{\text{lex}}(x)$ in-

dicates the trace is dominated by non-member cues, exhibiting the guessing-like style characteristic of unseen input.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three datasets across three text lengths (32, 64, and 128 tokens): ArxivReasoning and BookReasoning (Hu et al., 2026), consisting of ML/CS paper abstracts and literary excerpts respectively, and WikiMIA-25 (Yi and Li, 2026). Following the standard time-based MIA protocol, all three datasets use passages that predate the evaluated model’s training cutoff as members and post-cutoff passages as non-members. This construction provides benchmark labels but does not by itself exclude residual temporal correlations.

Models. We use three reasoning-capable language models spanning a range of scales: Qwen3-8B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-14B (DS-14B) (Guo et al., 2025), and QwQ-32B (Qwen Team, 2025). These three models are used for the comprehensive baseline comparison and analyses. To further assess model generalizability beyond these Qwen-derived models, we additionally evaluate TraceMIA (fixed) and TraceMIA (lexicon) on GLM-Z1-9B-0414 (GLM-Z1-9B) (Team GLM et al., 2024) and Llama-3.1-Nemotron-Nano-8B-v1 (Nemotron-

Length	Method	ArxivReasoning						BookReasoning						WikiMIA-25					
		Qwen3-8B		DS-14B		QwQ-32B		Qwen3-8B		DS-14B		QwQ-32B		Qwen3-8B		DS-14B		QwQ-32B	
		AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
32	CDD	0.4972	0.0448	0.5000	0.0500	0.5033	0.0563	0.4889	0.0280	0.4998	0.0496	0.5006	0.0511	0.5013	0.0525	0.5000	0.0500	0.5020	0.0538
	SaMIA	0.6611	0.1877	0.6032	0.0761	0.5846	0.0765	0.6871	0.2509	0.5447	0.1167	0.5284	0.1122	0.5859	0.1585	0.6033	0.1040	0.6211	0.1077
	SimMIA	0.5267	0.0596	0.4980	0.0660	0.6020	0.1280	0.6287	0.1093	0.5920	0.0850	0.6210	0.1000	0.7080	0.1835	0.6997	0.1720	0.5650	0.0900
	BlackSpectrum	0.5790	0.0680	0.6070	0.1270	0.5460	0.0510	0.4580	0.0180	0.5170	0.0520	0.5000	0.0260	0.5650	0.0630	0.5830	0.0760	0.7230	0.1440
	TraceMIA (fixed)	0.7764	0.2931	0.7515	0.2866	0.8539	0.4858	0.4752	0.0523	0.4578	0.0346	0.4841	0.0532	0.9255	0.5722	0.7842	0.1956	0.9754	0.8774
	TraceMIA (lexicon)	0.7739	0.2744	0.8270	0.3490	0.8940	0.5292	0.6972	0.1725	0.7596	0.3323	0.8593	0.4870	0.9546	0.7000	0.9045	0.4670	0.9873	0.9741
64	CDD	0.4932	0.0369	0.5007	0.0512	0.4941	0.0386	0.4973	0.0448	0.5002	0.0504	0.5109	0.0710	0.4994	0.0489	0.5011	0.0522	0.4983	0.0467
	SaMIA	0.6620	0.1601	0.6795	0.1773	0.6489	0.1483	0.7417	0.3243	0.6533	0.1946	0.5980	0.1262	0.5168	0.0743	0.5610	0.0869	0.5818	0.0974
	SimMIA	0.6121	0.0662	0.5970	0.0950	0.6990	0.1680	0.6518	0.1062	0.6030	0.0910	0.6370	0.1120	0.6121	0.1152	0.6369	0.0954	0.4730	0.0580
	BlackSpectrum	0.5950	0.0890	0.5850	0.1080	0.5610	0.0540	0.5590	0.0530	0.5000	0.0640	0.5810	0.0470	0.5800	0.0650	0.5470	0.0630	0.7130	0.1580
	TraceMIA (fixed)	0.7832	0.2866	0.7416	0.2282	0.8054	0.4182	0.4621	0.0384	0.4015	0.0224	0.4997	0.0267	0.9412	0.6600	0.7887	0.2308	0.9859	0.9675
	TraceMIA (lexicon)	0.8412	0.3935	0.8423	0.4431	0.9200	0.6619	0.7694	0.3727	0.8179	0.5029	0.8562	0.5332	0.9722	0.8508	0.8699	0.3988	0.9876	0.9939
128	CDD	0.4873	0.0252	0.4987	0.0475	0.4967	0.0437	0.4987	0.0474	0.4998	0.0496	0.4956	0.0414	0.5000	0.0500	0.5019	0.0536	0.4953	0.0409
	SaMIA	0.6639	0.2021	0.7951	0.3912	0.7589	0.3467	0.7750	0.3870	0.7747	0.4121	0.7194	0.3239	0.5243	0.0670	0.5870	0.1265	0.6167	0.1395
	SimMIA	0.6633	0.1298	0.7890	0.3520	0.7610	0.2810	0.6313	0.1489	0.6120	0.1280	0.6490	0.1450	0.6189	0.0939	0.6277	0.0613	0.4610	0.0330
	BlackSpectrum	0.5680	0.1020	0.5730	0.0880	0.5900	0.0680	0.6340	0.1450	0.4520	0.0410	0.6670	0.0950	0.6810	0.1020	0.5430	0.1210	0.7550	0.1780
	TraceMIA (fixed)	0.7757	0.3577	0.7110	0.2402	0.8120	0.4418	0.4836	0.0476	0.3779	0.0104	0.5718	0.0349	0.9473	0.7405	0.7718	0.1631	0.9857	0.9540
	TraceMIA (lexicon)	0.8675	0.4925	0.8515	0.4597	0.9225	0.6841	0.7758	0.3970	0.8318	0.5232	0.8707	0.5588	0.9709	0.8286	0.8897	0.3401	0.9914	0.9929

Table 1: Performance comparison across three datasets, three reasoning models, and three input lengths. Reported metrics are AUC-ROC and TPR@5%FPR. TraceMIA (lexicon) is evaluated with $K = 25$ reference samples per class. The best result per column is bolded; rows corresponding to our methods are shaded. Per-cell mean and standard deviation across five reference splits are provided in Appendix A.

8B) (Bercovich et al., 2025) under the same datasets, input lengths, and reference-set protocol (Table 5).

Baselines. We compare against four black-box MIA baselines: **SaMIA** (Kaneko et al., 2025) (n-gram overlap of sampled continuations), **CDD** (Dong et al., 2024) (peakedness of edit-distance distributions under temperature sampling), **SimMIA** (Yi and Li, 2026) (word-by-word sampling with soft semantic scoring), and **BlackSpectrum** (Hu et al., 2026) (latent-encoder over reasoning traces). We additionally introduce **TraceMIA (fixed)**, a variant that requires no labeled reference set and scores each trace against a predefined, hand-crafted lexicon of certainty/uncertainty cues, isolating the contribution of data-driven lexicon learning (full implementation in Appendix D). For clarity, we refer to the main method as **TraceMIA (lexicon)** when disambiguation is needed; in §1–§3.4, **TraceMIA** refers to this lexicon-based variant.

Metrics. We report two standard MIA metrics: *Area Under the ROC Curve (AUC-ROC)* for overall ranking quality, and *True Positive Rate at 5% False Positive Rate (TPR@5%FPR)* for high-precision performance.

Implementation details. We sample reasoning traces from each target model with temperature $T = 0.6$ and nucleus sampling at $p = 0.95$, following the official inference configuration recommended for reasoning-mode generation, and cap the generation length at 1024 tokens with extended-thinking mode enabled. To simulate the data-owner

setting, where reference and query passages come from the owner’s corpus, we partition each benchmark into a disjoint reference set ($K = 25$ per class) for lexicon construction and a held-out evaluation set for all reported metrics.

4.2 Experimental Analysis and Findings

To evaluate the effectiveness of TraceMIA, we answer the following questions.

Q1: How effective is TraceMIA at recovering membership signal from reasoning traces? Table 1 shows that TraceMIA (lexicon) attains the highest AUC in nearly all configurations, with the hand-crafted-lexicon variant TraceMIA (fixed) collapsing below chance on BookReasoning. Among prior baselines, CDD remains near chance (AUC ≈ 0.50) across all settings, suggesting its contrastive scoring mechanism fails to capture membership signal under this setup. More tellingly, BlackSpectrum—the only prior method that also operates on reasoning traces—trails our trace lexicon by 0.14–0.39 AUC and falls below even the simpler SaMIA baseline in most comparisons. Operating on the same modality, the gap suggests that the encoding choice is consequential: latent-space projection may discard membership information that lexical filtering recovers.

The gap between the two TraceMIA variants is largest on BookReasoning, where (fixed) frequently drops below chance (AUC < 0.50) while (lexicon) sustains AUC of 0.6972–0.8593 at length=32. The +0.22 to +0.38 AUC gap isolates the contribution of data-driven lexicon construction on out-of-domain text: hand-crafted cer-

Dataset	Experiment	AUC		TPR@5%FPR		$ W^+ $	$ W^- $
		Mean	Std	Mean	Std		
WikiMIA-25	Inference	0.9674	0.0006	0.7981	0.0126	27	11
	Annotation	0.9546	0.0106	0.7000	0.0807	31.6 ± 5.3	12.8 ± 4.5
ArxivReasoning	Inference	0.8272	0.0068	0.3482	0.0209	7	13
	Annotation	0.7739	0.0240	0.2744	0.0380	16.0 ± 3.9	3.6 ± 2.1

Table 2: Stability analysis of TraceMIA (lexicon) on Qwen3-8B (length=32). The **Inference** experiment fixes the lexicon and repeats test inference 5 times; the **Annotation** experiment samples 5 independent reference splits with fixed test inference. Lexicon sizes: point estimates (Inference), mean $\pm$ std (Annotation).

tainty cues do not generalize beyond encyclopedic and technical domains. Two additional patterns emerge from Table 1. First, performance improves monotonically on Arxiv and near-monotonically on BookReasoning, while WikiMIA remains near-ceiling across all lengths—consistent with longer passages providing more surface area for recognition cues to accumulate. Second, model scale is not always the dominant factor: QwQ-32B is consistently strongest, but on WikiMIA-25 DS-14B (0.9045) underperforms the smaller Qwen3-8B (0.9546), suggesting that reasoning style can modulate the recoverable lexical signal independently of scale. We further isolate the contribution of the lexical trace by comparing TraceMIA against the model’s binary JSON judgment emitted under the same introspective prompt (Appendix G). While the direct self-report already exceeds prior black-box baselines, TraceMIA improves over it by +0.07 to +0.24 AUC across all six tested settings, indicating that the lexical trace provides a stronger basis for membership inference than the model’s binary self-report. Per-length visualizations of the learned lexicons, score distributions, and ROC curves on WikiMIA-25 are provided in Appendix C. **TraceMIA (lexicon) surpasses all prior baselines on AUC in every configuration, and the reference-free (fixed) variant alone already outperforms prior methods in 17 of 18 configurations on encyclopedic and technical text.**

Q2: Is the extracted signal stable across runs and reference splits?

TraceMIA involves two independent sources of randomness: generation stochasticity in trace elicitation, and reference set selection in lexicon construction. We disentangle them through two orthogonal experiments on Qwen3-8B (length=32), summarized in Table 2. The **Inference** experiment fixes a single reference lexicon and repeats test infer-

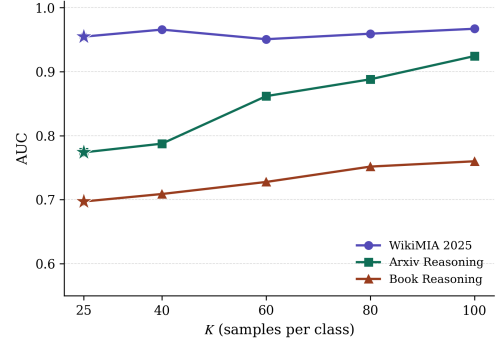


Figure 3: Effect of reference set size K on AUC of TraceMIA (lexicon) (Qwen3-8B, length=32). The horizontal axis denotes K (samples per class). The starred point at $K = 25$ corresponds to the setting used in all main experiments.

ence 5 times; AUC standard deviation is 0.0006 on WikiMIA-25 and 0.0068 on ArxivReasoning, showing that generation stochasticity has negligible impact on downstream scoring. The **Annotation** experiment fixes test inference and samples 5 independent reference splits, each yielding a new lexicon; AUC std remains low (0.0106 on WikiMIA, 0.0240 on Arxiv). The two experiments target different questions, and their mean AUCs are not directly comparable: the Inference experiment reports a single-lexicon point estimate, while the Annotation experiment characterizes expected performance under random reference selection. The higher Arxiv variance under the Annotation experiment tracks its smaller learned lexicon ($|W^+| = 16.0 \pm 3.9$ vs. 31.6 ± 5.3 on WikiMIA): fewer discriminative words make scoring more sensitive to reference selection. **Reference set selection dominates inference stochasticity as a source of variance, but both remain small enough at our default $K = 25$ to ensure a stable membership signal across independent runs.**

Q3: How sensitive is TraceMIA to reference set size K ?

We vary K from 25 to 100 on Qwen3-8B (length=32). As shown in Figure 3, WikiMIA-25 maintains near-ceiling performance, ArxivReasoning improves from 0.774 to 0.924, and BookReasoning improves from 0.697 to 0.760. In a separate Book-only sweep from $K = 20$ to $K = 200$ (not shown), AUC increases monotonically from 0.649 to 0.856 and TPR@5%FPR from 0.116 to 0.336. This indicates that literary-domain performance benefits from additional reference samples, albeit with a larger reference budget. **All main experi-**

Model	Method	AUC	TPR@5%FPR	$ W^+ $	$ W^- $
Input-Lexicon (model-indep.)		0.5811	0.1176	1	2
Qwen3-8B	Trace-Lexicon	0.9722	0.8671	20	20
	Δ	+0.3911	+0.7495		
DS-14B	Trace-Lexicon	0.9085	0.4459	11	15
	Δ	+0.3274	+0.3283		

Table 3: Input text vs. reasoning trace as the source of lexical signal (WikiMIA-25, length=64; fixed split with $K = 25$ per class). Δ denotes the gain of Trace-Lexicon over Input-Lexicon.

Model	Dataset	Variant	AUC	TPR@5%FPR	Trace coef
Qwen3-8B	Arxiv	Input only	0.8324 $\pm$ 0.0198	0.3474	–
		Input + Trace	0.9002 $\pm$ 0.0145	0.5577	+1.621
	Book	Input only	0.6923 $\pm$ 0.0089	0.1499	–
		Input + Trace	0.7377 $\pm$ 0.0060	0.2139	+0.594
DS-14B	Arxiv	Input only	0.8324 $\pm$ 0.0198	0.3474	–
		Input + Trace	0.9126 $\pm$ 0.0148	0.6092	+1.834
	Book	Input only	0.6923 $\pm$ 0.0089	0.1499	–
		Input + Trace	0.7774 $\pm$ 0.0098	0.3293	+0.901

Table 4: Controlled feature ablation on ArxivReasoning and BookReasoning (length=32, 5-fold CV). **Input only**: logistic regression on surface input features (text length, mean word length, type-token ratio, TF-IDF topics). These features are model-independent, so the Input only rows are identical across Qwen3-8B and DS-14B by construction. **Input + Trace**: same classifier augmented with S_{lex} . Trace coef: regression weight on S_{lex} in the joint model.

ments retain $K = 25$; the extended sweep is an additional analysis rather than a change to the main protocol.

Q4: Does the membership signal reside in the reasoning trace or in the input text?

To test whether S_{lex} merely re-encodes surface properties of the query passage, we train a logistic regression classifier on surface input features—text length, mean word length, type-token ratio, and TF-IDF topics—under 5-fold cross-validation, and compare it against the same classifier augmented with S_{lex} . We focus this ablation on Arxiv and Book, where input-only baselines leave substantial headroom and the marginal contribution of S_{lex} is most diagnostic. Table 4 shows that adding S_{lex} yields consistent AUC improvements: +0.07 to +0.08 on Arxiv and +0.05 to +0.09 on Book. The learned coefficient on S_{lex} is uniformly large and positive (+0.594 to +1.834); were it merely re-encoding information already present in the input features, it would be largely collinear with them and L_2 regularization would substantially shrink its coefficient, splitting the weight across the redundant features. As a sanity check, applying the same χ^2 pipeline directly to raw input text yields

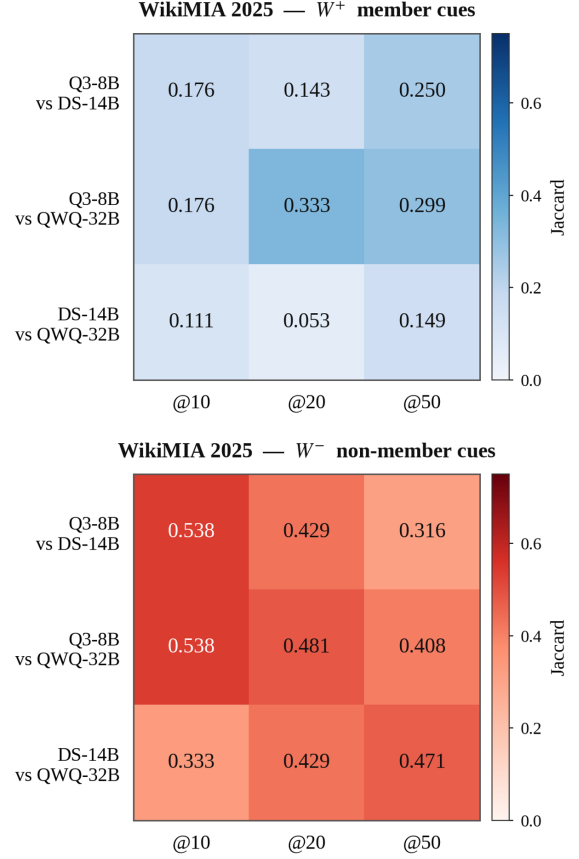


Figure 4: Lexical overlap among the three main-evaluation models (Jaccard similarity) at top-10/20/50 on WikiMIA-25. W^- (non-member cues, bottom) exhibits substantially higher overlap than W^+ (member cues, top), peaking at 0.538 at top-10.

only 1–2 cue words and a low AUC of 0.58 (Table 3), providing evidence that the signal is not explained solely by surface lexical properties of the passage itself. **The trace lexicon constitutes a source of membership signal distinct from, and complementary with, surface input features.**

Q5: Do the learned lexicons reflect shared membership-related patterns across models?

We test whether the learned lexicons reflect a shared membership-related pattern or model-specific idiosyncrasies by independently learning lexicons from three models and measuring their overlap. As shown in Figure 4, non-member cues exhibit substantially higher cross-model consistency than member cues at top-10: W^- Jaccard peaks at 0.538, anchored by a shared core of temporally-bounded and epistemically distancing terms (Table 6). W^+ overlap remains low (top-10 Jaccard ≤ 0.18), indicating that member-side cues are expressed through more model-idiosyncratic vocabulary. Despite their shared Qwen lineage,

Length	Method	ArxivReasoning				BookReasoning				WikiMIA-25			
		GLM-Z1-9B		Nemotron-8B		GLM-Z1-9B		Nemotron-8B		GLM-Z1-9B		Nemotron-8B	
		AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
32	TraceMIA (fixed)	0.7500	0.2947	0.7371	0.2366	0.4636	0.0382	0.4931	0.0431	0.9135	0.5395	0.7345	0.2136
	TraceMIA (lexicon)	0.7307	0.1981	0.7643	0.2440	0.7223	0.1759	0.6324	0.1157	0.9599	0.7411	0.8077	0.3240
64	TraceMIA (fixed)	0.7196	0.1997	0.7105	0.2535	0.4569	0.0395	0.4372	0.0394	0.9085	0.5412	0.7359	0.1786
	TraceMIA (lexicon)	0.7775	0.2939	0.7739	0.2646	0.7459	0.2518	0.6892	0.1531	0.9715	0.8409	0.7845	0.2216
128	TraceMIA (fixed)	0.7265	0.2556	0.7312	0.2581	0.4478	0.0344	0.4469	0.0305	0.9345	0.6488	0.7913	0.2817
	TraceMIA (lexicon)	0.8214	0.3588	0.7848	0.3023	0.7381	0.2179	0.6475	0.1452	0.9758	0.8798	0.8326	0.2452

Table 5: Generalization performance of TraceMIA on two additional non-Qwen reasoning models across three datasets and input lengths. Results are averaged over five independent reference splits with $K = 25$ samples per class. TPR denotes TPR@5%FPR, and bold indicates the better-performing variant in each setting.

Dataset	W^+ (shared)	W^- (shared)
WikiMIA-25	<i>certain, remember, future, hypothetical, fiction, wording</i>	<i>tional, scenario, cutoff</i>
arXiv Reasoning	<i>bert, neural, net-works, deep, nlp, tems, terms</i>	<i>robotics, science, imagenet</i>

Table 6: Representative words appearing in the top-20 of all three models on each dataset.

Dataset	Cluster	Qwen3-8B	DS-14B	QwQ-32B
WikiMIA-25	W^+ : recall_source	4.8%	4.1%	3.3%
	W^+ : certainty_attribution	1.3%	0.3%	1.1%
	W^- : uncertainty_hedge	1.5%	1.1%	1.1%
	W^- : temporal_cutoff	3.9%	3.0%	4.6%
	W^- : fictional_hypothetical	3.9%	3.0%	3.5%
ArxivReasoning	W^+ : recall_source	2.9%	2.6%	1.9%
	W^+ : certainty_attribution	0.7%	1.2%	1.0%
	W^- : uncertainty_hedge	1.1%	1.9%	1.5%
	W^- : temporal_cutoff	1.3%	1.1%	1.1%
	W^- : fictional_hypothetical	0.2%	0.6%	0.7%

Table 7: Semantic cluster distribution of learned lexicons across models (percentage of total lexicon words falling into each cluster). Residual mass corresponds to low-frequency or unclustered words.

the three models differ in scale and post-training procedure; their W^- overlap therefore identifies a recurring non-member pattern within the main evaluation. Table 7 further shows that the distribution across five hand-curated semantic clusters (see Appendix F) is broadly stable across models. This semantic-level stability strengthens the interpretation: the shared signal is not a coincidence of specific word choices but reflects a systematic difference in how reasoning models verbalize the remember-like recognition versus the guessing-like uncertainty.

The shared W^- words are primarily epistemic or hypothetical terms (e.g., *future*, *hypothetical*, and *scenario*) rather than specific years or dates, although temporal correlations inherent in cutoff-

Model	Length	0%	10%	20%	30%	Non-member
Qwen3-8B	32	1.645	0.926	0.540	0.371	-3.898
	64	4.363	2.897	1.974	1.667	-4.098
	128	8.472	6.509	5.783	5.689	+1.093
DS-14B	32	0.686	0.361	0.048	-0.169	-3.226
	64	1.716	1.236	0.980	0.792	-2.179
	128	-0.196	-0.293	-0.402	-0.452	-3.389

Table 8: Mean S_{lex} for member passages under increasing character-level perturbation on WikiMIA-25. The Non-member column reports the mean score for unperturbed non-member passages.

based MIA benchmarks cannot be fully excluded.

The three main-evaluation models share a recurring non-member lexical pattern, while member-side cues remain more model-specific.

Q6: Does TraceMIA track verbatim memorization or mere topical familiarity?

We perturb each alphabetic token in member passages with probabilities of 10%, 20%, and 30% using a random insertion, deletion, replacement, or swap, altering the exact surface form while largely preserving the topic. As shown in Table 8, S_{lex} decreases monotonically across all model-length configurations as the perturbation rate increases. Nevertheless, the mean member score remains higher than the corresponding non-member mean even at 30% perturbation. This pattern suggests that the membership signal is sensitive to surface-form recognition rather than topical familiarity alone.

Q7: Does TraceMIA generalize beyond Qwen-derived reasoning models?

To assess model generalizability beyond the Qwen-derived models used in our main evaluation, we additionally evaluate TraceMIA (fixed) and TraceMIA (lexicon) on GLM-Z1-9B and Nemotron-8B using the same datasets, input lengths, and reference-set protocol.

As shown in Table 5, TraceMIA (lexicon) achieves higher AUC than TraceMIA (fixed) in

17 of the 18 settings, with the only exception being GLM-Z1-9B on ArxivReasoning at length 32. On WikiMIA-25, its AUC ranges from 0.9599 to 0.9758 on GLM-Z1-9B and from 0.7845 to 0.8326 on Nemotron-8B. On BookReasoning, the learned lexicon achieves AUCs of 0.6324–0.7459, whereas the fixed variant remains below chance. These results provide further evidence that TraceMIA’s data-driven lexical signal generalizes beyond the Qwen-derived models used in the main evaluation.

5 Conclusion

We propose TraceMIA, a black-box membership inference method that recovers a discriminative lexical signal from the exposed introspective reasoning traces of reasoning models via χ^2 filtering over a small labeled reference set. Under a data-owner setting where the reference set is low-cost to construct, TraceMIA outperforms existing baselines in nearly all settings across three reasoning models, three datasets, and three text lengths, while yielding interpretable lexicons that expose the remember-like style of member traces and the guessing-like style of non-member traces, providing transparent, human-readable statistical evidence for auditing. Supplementary experiments on two additional non-Qwen models provide further evidence of generalizability, while TraceMIA’s applicability remains limited to models that expose textual reasoning traces.

Limitations

TraceMIA requires a labeled reference set and access to textual reasoning traces, and is therefore inapplicable when such traces are hidden or summarized. Its performance may decline in domains with dispersed membership vocabulary, which require larger reference sets. Providers may weaken the attack by filtering introspective queries, rewriting traces, or fine-tuning models to homogenize their lexical style; targeted vocabulary-suppression defenses remain untested. Finally, temporal correlations and post-training calibration effects cannot be fully excluded. TraceMIA scores should therefore be treated as statistical auditing evidence rather than definitive proof of training inclusion.

Ethics Statement

TraceMIA is intended for data-owner auditing of content from the auditor’s own corpus; applying it to third-party data falls outside this intended use.

As membership inference is inherently dual-use, providers may restrict introspective queries or access to reasoning traces. Moreover, because cue-word counts can be manipulated, results should be corroborated with repeated samples and independent evidence rather than treated as standalone legal proof. All experiments use publicly released benchmarks and reasoning models and involve no PII or closed-source systems. Code, reference splits, and elicited traces will be released with the final paper.

Acknowledgments

We sincerely thank the Discipline and Technology Research Center for Large Model Intelligence Applications for their invaluable support throughout this research. The center provided essential computational resources, research infrastructure, and an intellectually stimulating environment that made this work possible. A generative AI assistant was used to support language and organizational editing during the preparation of the camera-ready manuscript. All technical claims, numerical results, and citations were verified by the authors.

References

- Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 114 others. 2025. [Llama-nemotron: Efficient reasoning models](#). *Preprint*, arXiv:2505.00949.
- Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, and 1 others. 2022. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5—Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of*

- the Association for Computational Linguistics: ACL 2024, pages 12039–12050.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Ruihan Hu, Yu-Ming Shang, Wei Luo, Ye Tao, and Xi Zhang. 2026. When reasoning leaks membership: Membership inference attack on black-box large reasoning models. In *Proceedings of the ACM Web Conference 2026*, pages 3323–3334.
- Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. 2025. Sampling-based pseudo-likelihood for membership inference attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8894–8907.
- Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412.
- Minseo Kwak and Jaehyung Kim. 2026. Gap-k%: Measuring top-1 prediction gap for detecting pretraining data. *arXiv preprint arXiv:2601.19936*.
- Justus Mattern, Fatemehsadat Miresghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343.
- Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385.
- Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pages 51826–51843.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.
- Jiatong Yi and Yanyang Li. 2026. Membership inference on llms in the wild. *arXiv preprint arXiv:2601.11314*.
- Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, volume 2025, pages 64845–64862.

Appendix

A Complete Results for the Main Experiments 10

B Prompt Template 11

C Lexicon Visualization on WikiMIA-25 ... 11

C.1 Length = 32 11

C.2 Length = 64 13

C.3 Length = 128 14

D TraceMIA (fixed) Details 14

E Prompt Ablation Study 15

F Semantic Cluster Definitions 16

G Trace vs. Self-Report 16

A Complete Results for the Main Experiments

The complete per-cell mean and standard deviation of AUC and TPR@5%FPR are reported for our main method, TraceMIA (lexicon), in Table 9, and for the reference-free baseline, TraceMIA (fixed),

in Table 10, computed over five independent reference splits.

B Prompt Template

Table 11 presents the system and user prompts implementing $\mathcal{P}(x)$ from Section 3.2, used uniformly across all MIA queries. The prompt is deliberately introspective: the system message frames the model as an agent capable of reflecting on its own training data, while the user message enumerates four recognition-oriented cues—provenance, stylistic familiarity, recognizable phrases, and likely source—to direct the model toward a recognition-oriented reasoning trajectory. The model is queried with `enable_thinking=True` to elicit an extended reasoning trace enclosed in `<think>...</think>` tags; only the content within these tags is used for lexicon construction and scoring, while the final JSON response is discarded.

The JSON instruction is retained purely as a termination signal: requiring a structured final answer prompts the model to conclude its reasoning before emitting the closing `</think>` tag, yielding cleaner trace boundaries. The binary member field plays no role in the scoring pipeline.

C Lexicon Visualization on WikiMIA-25

This appendix provides per-length visualizations of the lexicon construction and inference behavior of TraceMIA on WikiMIA-25, evaluated on Qwen3-8B with $K = 25$ reference samples per class. All visualizations are based on a single reference split; the reported metrics are therefore single-split point estimates and may differ from the 5-split averages in Table 9. For each text length ($\ell \in \{32, 64, 128\}$ tokens), we report three artifacts: the discriminative vocabulary W^+/W^- ranked by the signed χ^2 statistic, the empirical distribution of S_{lex} over member and non-member samples, and the corresponding ROC curve. Lexicon construction follows the χ^2 filtering procedure of §3.3 with thresholds $p_{\text{thresh}} = 0.05$ and $f_{\text{min}} = 3$.

C.1 Length = 32

At length = 32, the χ^2 filter recovers 38 discriminative words ($|W^+|=29$, $|W^-|=9$; Figure 5). The member lexicon W^+ is dominated by source-attributive and certainty-oriented terms (*wording*, *sources*, *historical*, *factual*, *considering*, *Wikipedia*), consistent with the model invoking verifiable provenance when reasoning over

training-familiar text. The non-member lexicon W^- comprises temporally-bounded and modal expressions (*future*, *hypothetical*, *cutoff*, *scenario*, *beyond*, *wouldn*) that encode the hedging behavior characteristic of unseen passages. Words at the axis extremes exhibit the largest distributional shift between classes and contribute most strongly to the membership score.

The induced score distribution (Figure 6) shows clear separation between the two classes. The member distribution is right-skewed with median 5.58 and is supported almost entirely above zero, whereas the non-member distribution concentrates around the origin with median 0.00; overlap is confined to a narrow central band. The corresponding ROC curve (Figure 7) attains $\text{AUC} = 0.9608$ with $\text{TPR} = 0.744$ at 5% FPR. The steep rise in the low-FPR regime indicates that TraceMIA recovers a high-precision operating point despite the brevity of the input and without access to token-level probabilities.

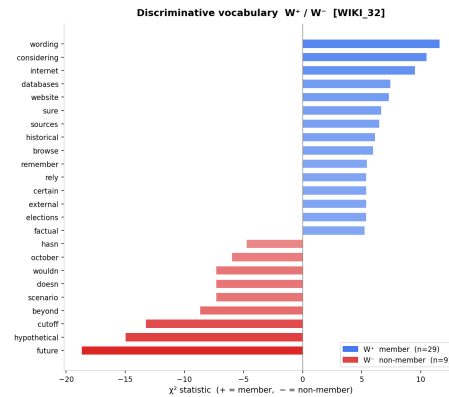


Figure 5: Discriminative vocabulary on WikiMIA-25 at length = 32 ($|W^+|=29$, $|W^-|=9$), ordered by the signed Yates-corrected χ^2 statistic.

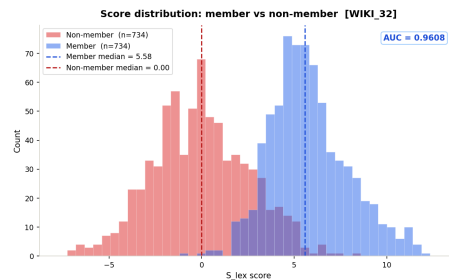


Figure 6: S_{lex} score distribution on WikiMIA-25 at length = 32. Member median = 5.58; non-member median = 0.00; $\text{AUC} = 0.9608$.

Dataset	Model	Length=32		Length=64		Length=128	
		AUC	TPR	AUC	TPR	AUC	TPR
WikiMIA-25	Qwen3-8B	0.9546 $\pm$ 0.0106	0.7000 $\pm$ 0.0807	0.9722 $\pm$ 0.0059	0.8508 $\pm$ 0.0449	0.9709 $\pm$ 0.0040	0.8286 $\pm$ 0.0548
	DS-14B	0.9045 $\pm$ 0.0125	0.4670 $\pm$ 0.0600	0.8699 $\pm$ 0.0449	0.3988 $\pm$ 0.0614	0.8897 $\pm$ 0.0214	0.3401 $\pm$ 0.0614
	QwQ-32B	0.9873 $\pm$ 0.0017	0.9741 $\pm$ 0.0067	0.9876 $\pm$ 0.0010	0.9939 $\pm$ 0.0032	0.9914 $\pm$ 0.0018	0.9929 $\pm$ 0.0027
ArxivReasoning	Qwen3-8B	0.7739 $\pm$ 0.0240	0.2744 $\pm$ 0.0380	0.8412 $\pm$ 0.0340	0.3935 $\pm$ 0.1200	0.8675 $\pm$ 0.0206	0.4925 $\pm$ 0.0670
	DS-14B	0.8270 $\pm$ 0.0359	0.3490 $\pm$ 0.0727	0.8423 $\pm$ 0.0299	0.4431 $\pm$ 0.0810	0.8515 $\pm$ 0.0190	0.4597 $\pm$ 0.0548
	QwQ-32B	0.8940 $\pm$ 0.0394	0.5292 $\pm$ 0.1387	0.9200 $\pm$ 0.0208	0.6619 $\pm$ 0.0642	0.9225 $\pm$ 0.0175	0.6841 $\pm$ 0.0451
BookReasoning	Qwen3-8B	0.6972 $\pm$ 0.0430	0.1725 $\pm$ 0.0377	0.7694 $\pm$ 0.0403	0.3727 $\pm$ 0.0551	0.7758 $\pm$ 0.0080	0.3970 $\pm$ 0.0511
	DS-14B	0.7596 $\pm$ 0.0247	0.3323 $\pm$ 0.0680	0.8179 $\pm$ 0.0185	0.5029 $\pm$ 0.0347	0.8318 $\pm$ 0.0105	0.5232 $\pm$ 0.0532
	QwQ-32B	0.8593 $\pm$ 0.0293	0.4870 $\pm$ 0.0288	0.8562 $\pm$ 0.0185	0.5332 $\pm$ 0.0277	0.8707 $\pm$ 0.0210	0.5588 $\pm$ 0.0365

Table 9: Full results of the TraceMIA (lexicon) variant (mean $\pm$ std over 5 independent annotation splits). TPR denotes TPR@5%FPR.

Dataset	Model	Length=32		Length=64		Length=128	
		AUC	TPR	AUC	TPR	AUC	TPR
WikiMIA-25	Qwen3-8B	0.9255 $\pm$ 0.0013	0.5722 $\pm$ 0.0129	0.9412 $\pm$ 0.0011	0.6600 $\pm$ 0.0096	0.9473 $\pm$ 0.0009	0.7405 $\pm$ 0.0067
	DS-14B	0.7842 $\pm$ 0.0024	0.1956 $\pm$ 0.0032	0.7887 $\pm$ 0.0010	0.2308 $\pm$ 0.0023	0.7718 $\pm$ 0.0029	0.1631 $\pm$ 0.0110
	QwQ-32B	0.9754 $\pm$ 0.0008	0.8774 $\pm$ 0.0037	0.9859 $\pm$ 0.0005	0.9675 $\pm$ 0.0032	0.9857 $\pm$ 0.0010	0.9540 $\pm$ 0.0019
ArxivReasoning	Qwen3-8B	0.7764 $\pm$ 0.0022	0.2931 $\pm$ 0.0019	0.7832 $\pm$ 0.0010	0.2866 $\pm$ 0.0062	0.7757 $\pm$ 0.0015	0.3577 $\pm$ 0.0020
	DS-14B	0.7515 $\pm$ 0.0015	0.2866 $\pm$ 0.0050	0.7416 $\pm$ 0.0021	0.2282 $\pm$ 0.0069	0.7110 $\pm$ 0.0010	0.2402 $\pm$ 0.0040
	QwQ-32B	0.8539 $\pm$ 0.0011	0.4858 $\pm$ 0.0043	0.8054 $\pm$ 0.0005	0.4182 $\pm$ 0.0068	0.8120 $\pm$ 0.0030	0.4418 $\pm$ 0.0045
BookReasoning	Qwen3-8B	0.4752 $\pm$ 0.0013	0.0523 $\pm$ 0.0005	0.4621 $\pm$ 0.0005	0.0384 $\pm$ 0.0005	0.4836 $\pm$ 0.0003	0.0476 $\pm$ 0.0005
	DS-14B	0.4578 $\pm$ 0.0005	0.0346 $\pm$ 0.0003	0.4015 $\pm$ 0.0009	0.0224 $\pm$ 0.0007	0.3779 $\pm$ 0.0004	0.0104 $\pm$ 0.0002
	QwQ-32B	0.4841 $\pm$ 0.0006	0.0532 $\pm$ 0.0009	0.4997 $\pm$ 0.0013	0.0267 $\pm$ 0.0006	0.5718 $\pm$ 0.0007	0.0349 $\pm$ 0.0002

Table 10: Full results of the TraceMIA (fixed) variant (mean $\pm$ std over five runs with randomly sampled reference sets). TPR denotes TPR@5%FPR.

System prompt

You are an AI assistant with the ability to introspect about your own training data.

User prompt

Consider the following passage:

{text}

Reflect carefully on whether you have encountered this exact passage before. Think about: where this text might come from, how familiar the writing style feels, whether you recognize specific phrases or facts, and what source it might belong to.

Respond ONLY with a JSON object, no other text:
{ "member": <0 or 1> }

Table 11: System and user prompts used for all MIA queries. The placeholder {text} is replaced by the candidate passage. The trailing JSON instruction serves only as a termination signal; the produced JSON is discarded and all downstream scoring operates on the content of the <think>...</think> block.

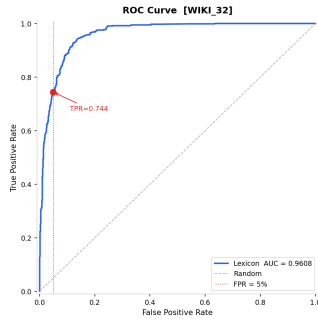


Figure 7: ROC curve on WikiMIA-25 at length = 32. AUC = 0.9608; TPR@5%FPR = 0.744.

C.2 Length = 64

At length = 64, the discriminative lexicon comprises $|W^+|=25$ and $|W^-|=10$ words (Figure 8). Relative to length = 32, W^+ shifts toward explicit memory-retrieval verbs and precision markers (*remember, wording, recall, reliable, straightforward, verbatim, concise*), suggesting that additional context elicits more direct lexical acknowledgement of familiarity. W^- remains anchored on temporal and modal hedges (*future, cutoff, wouldn, hypothetical, scenario, events*), indicating that the core non-member vocabulary is preserved across lengths.

The score distribution (Figure 9) preserves the qualitative pattern observed at length = 32, with member median 5.26 and non-member median 0.00. The right tail of the member distribution is noticeably more elongated than at length = 32:

highly familiar passages now activate a larger fraction of W^+ cues and accumulate higher scores. The marginal AUC gain ($0.9608 \rightarrow 0.9631$) occurs despite a slight reduction in the discriminative vocabulary. The corresponding ROC curve (Figure 10) shows that TPR at 5% FPR improves by 0.042 absolute over length = 32, with the improvement concentrated in the low-FPR regime most relevant to high-precision auditing.

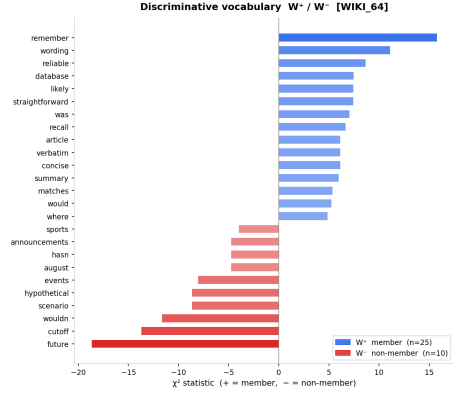


Figure 8: Discriminative vocabulary on WikiMIA-25 at length = 64 ($|W^+|=25$, $|W^-|=10$), ordered by the signed Yates-corrected χ^2 statistic.

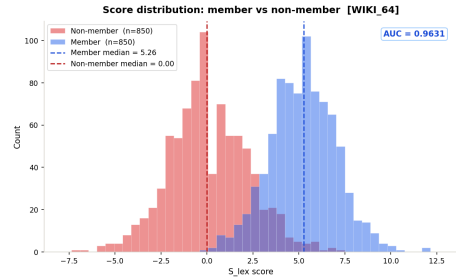


Figure 9: S_{lex} score distribution on WikiMIA-25 at length = 64. Member median = 5.26; non-member median = 0.00; AUC = 0.9631.

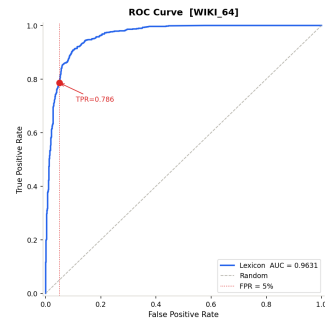


Figure 10: ROC curve on WikiMIA-25 at length = 64. AUC = 0.9631; TPR@5%FPR = 0.786.

C.3 Length = 128

At length = 128, χ^2 filtering yields the largest lexicon across all length conditions ($|W^+|=33$, $|W^-|=15$; Figure 11). W^+ broadens to encompass both explicit recall markers and meta-cognitive certainty terms (*factual, think, concise, given, remember, recall, wording, certain, typical*), indicating that extended reasoning over long familiar passages engages a wider certainty-attribution vocabulary. W^- expands correspondingly, layering epistemically distancing terms (*fictional, relative, knowledge*) on top of the cross-length stable core (*future, hypothetical, scenario, cutoff, yet*).

The score distribution at length = 128 (Figure 12) exhibits the largest inter-class separation across all length conditions, with member median 6.56 and non-member median -1.48 . The negative non-member median is itself informative: the hedging vocabulary is now sufficiently pervasive that S_{lex} is consistently driven below zero on unseen passages, further sharpening the decision boundary. The ROC curve (Figure 13) attains $\text{AUC} = 0.9766$ and $\text{TPR} = 0.845$ at 5% FPR—the strongest operating point across all length conditions—reaching near-unity TPR within the first decile of FPR. Obtained without model internals and with only $K = 25$ reference samples per class, this result supports the practical viability of TraceMIA for high-stakes data auditing on long passages.

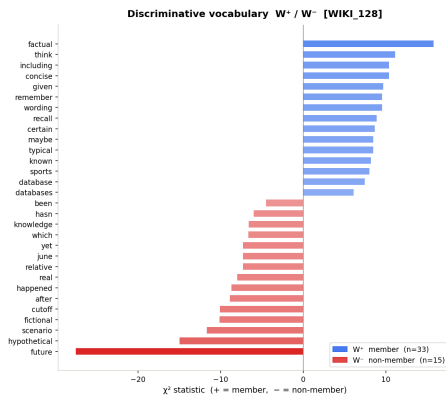


Figure 11: Discriminative vocabulary on WikiMIA-25 at length = 128 ($|W^+|=33$, $|W^-|=15$), ordered by the signed Yates-corrected χ^2 statistic.

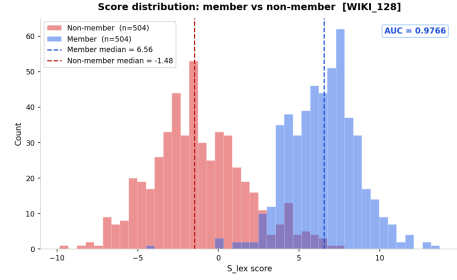


Figure 12: S_{lex} score distribution on WikiMIA-25 at length = 128. Member median = 6.56; non-member median = -1.48 ; $\text{AUC} = 0.9766$.

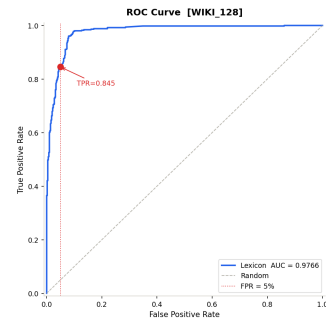


Figure 13: ROC curve on WikiMIA-25 at length = 128. $\text{AUC} = 0.9766$; $\text{TPR}@5\%\text{FPR} = 0.845$.

D TraceMIA (fixed) Details

We provide implementation details of the reference-free variant **TraceMIA (fixed)** introduced in §4.1. It requires no labeled reference set and scores each trace against two predefined hand-crafted cue sets: C (certainty markers) and U (uncertainty markers).

Hand-crafted Cue Sets. C contains: *exact, exactly, specific, clearly, definitely, certainly, directly, original, source, sources, known, present, sentence, paper, papers, article, articles, Wikipedia, database, local, common, model, wording, bert, imagenet*. U contains: *might, maybe, perhaps, possibly, could, seems, appear, appears, appeared, likely, unlikely, chance, hypothetical, future, cutoff, guess, uncertain, probably, roughly, about, around, suggest, suggests, training, data, technical, looking*.

Feature Extraction. Given a reasoning trace τ_x with token sequence $V(\tau_x)$, let $N_\tau = \max(|V(\tau_x)|, 1)$. We compute four lexical cue features—certainty ratio r_C , uncertainty ratio r_U ,

confidence gap g , and an entropy-like quantity H :

$$\begin{aligned} r_C(x) &= |V(\tau_x) \cap C|/N_\tau, \\ r_U(x) &= |V(\tau_x) \cap U|/N_\tau, \\ g(x) &= r_C(x) - r_U(x), \\ H(x) &= -p \log p - (1-p) \log(1-p), \end{aligned}$$

where $p = (|V(\tau_x) \cap U| + 1)/(|V(\tau_x) \cap C| + |V(\tau_x) \cap U| + 2)$.

Score Aggregation. Each feature is min-max normalized to $[0, 1]$ across the evaluation set:

$$\tilde{f}(x) = \frac{f(x) - \min_{x'} f(x')}{\max_{x'} f(x') - \min_{x'} f(x')}.$$

The final score is:

$$S_{\text{fixed}}(x) = 0.35\tilde{r}_C(x) + 0.35\tilde{g}(x) - 0.20\tilde{r}_U(x) - 0.10\tilde{H}(x). \quad (3)$$

The weights reflect the relative reliability of each signal: r_C and g provide positive evidence, while r_U and H contribute negatively with smaller weight. The full procedure is summarized in Algorithm 1.

Algorithm 1 TraceMIA (fixed)

Require: Target model $\mathcal{M}$; candidate set $\mathcal{X}$; cue sets C, U

Ensure: Scores $\{S_{\text{fixed}}(x)\}_{x \in \mathcal{X}}$

```

1: for each  $x \in \mathcal{X}$  do
2:    $\tau_x \leftarrow \mathcal{M}(\mathcal{P}(x))$ 
3:   Extract  $V(\tau_x)$ ;  $N_\tau \leftarrow \max(|V(\tau_x)|, 1)$ 
4:    $r_C(x) \leftarrow |V(\tau_x) \cap C|/N_\tau$ 
5:    $r_U(x) \leftarrow |V(\tau_x) \cap U|/N_\tau$ 
6:    $g(x) \leftarrow r_C(x) - r_U(x)$ 
7:    $p \leftarrow (|V(\tau_x) \cap U| + 1)/(|V(\tau_x) \cap C| + |V(\tau_x) \cap U| + 2)$ 
8:    $H(x) \leftarrow -p \log p - (1-p) \log(1-p)$ 
9: end for
10: Min-max normalize each feature across  $\mathcal{X}$ 
11: for each  $x \in \mathcal{X}$  do
12:    $S_{\text{fixed}}(x) \leftarrow 0.35\tilde{r}_C(x) + 0.35\tilde{g}(x) - 0.20\tilde{r}_U(x) - 0.10\tilde{H}(x)$ 
13: end for
14: return  $\{S_{\text{fixed}}(x)\}$ 

```

This variant shares the same introspective prompting and trace extraction pipeline as TraceMIA (lexicon), but replaces the data-driven χ^2 -filtered lexicons (W^+, W^-) with fixed cue sets (C, U), isolating the contribution of data-driven lexicon learning.

E Prompt Ablation Study

TraceMIA relies on an introspective prompt that directs the target model to reflect on whether it has previously encountered a given passage. A natural question arises: is the observed lexical membership signal a genuine product of introspective elicitation, or could it emerge under any chain-of-thought prompt? To investigate this, we replace the introspective prompt with a neutral summarization prompt while keeping all other pipeline components identical:

System: You are a helpful AI assistant.

User: Consider the following passage:
{text}

Please summarize the main content of this passage in a few sentences. Think carefully about what topics it covers and what it is trying to convey. Respond ONLY with a JSON object, no other text:
{"summary": "<your summary>"}

Both conditions use `enable_thinking=True`, so the model produces a `<think>...</think>` trace in both cases. Following the Annotation experiment protocol, we sample 5 independent reference splits and report mean $\pm$ std on Qwen3-8B with WikiMIA-25 at length = 32.

Prompt	Method	AUC $\uparrow$	TPR@5%FPR $\uparrow$	$ W^+ $	$ W^- $
Introspective	Lexicon	0.9546 $\pm$ 0.0073	0.7264 $\pm$ 0.0485	29.0	13.8
	Fixed	0.9276 $\pm$ 0.0011	0.6292 $\pm$ 0.0054	—	—
Neutral	Lexicon	0.6002 $\pm$ 0.0595	0.1319 $\pm$ 0.0637	4.6	2.0
	Fixed	0.5192 $\pm$ 0.0013	0.1030 $\pm$ 0.0016	—	—
$\Delta\text{AUC (Introspective} - \text{Neutral)}$		+0.3544 $\pm$ 0.0593			

Table 12: Prompt ablation on WikiMIA-25 (Qwen3-8B, length=32, 5 annotation splits). $|W^+|$ and $|W^-|$ denote mean lexicon sizes per split. This ablation is run independently of the main experiments with a separate set of 5 splits, so the reported standard deviations may differ from those in Table 9.

As shown in Table 12, the neutral prompt yields near-random AUC (0.60) with a degenerate lexicon of only 4.6 member-cue words, compared to AUC 0.9546 and 29.0 cue words under introspective prompting ($\Delta\text{AUC} = +0.3544$). Since both conditions produce chain-of-thought traces of comparable length, the difference cannot be attributed to the presence of extended reasoning per se. Rather, introspective elicitation specifically activates recognition-oriented and uncertainty-hedging vocabulary that discriminates members from non-members; a neutral summarization directive sup-

presses this channel entirely. The weak residual signal under neutral prompting (AUC 0.60 > 0.50) likely reflects incidental lexical overlap between passage content and training-familiar vocabulary, not a structured membership signal. This result indicates that the lexical signal is specifically elicited by introspective prompting rather than emerging from generic chain-of-thought generation, and offers a mechanistic account of why existing black-box methods operating on final outputs or temperature-perturbed continuations fail to exploit the membership-sensitive lexical channel exposed by introspective reasoning.

F Semantic Cluster Definitions

The five semantic clusters used in §4.2 Q5 (Table 7) are defined by hand-curated seed-word lists. A lexicon word is assigned to a cluster iff it appears in the corresponding seed list; words matching no cluster contribute to the residual mass noted in the table caption. The seed lists were constructed by inspecting the top-50 cues per class across the three models and grouping semantically related items; no automated clustering was used.

Member-leaning clusters.

- **recall_source**: *exact, exactly, source, sources, wording, original, directly, passage, sentence, sentences, text, paragraph, document, documents, paper, papers, article, articles, Wikipedia, verbatim, cited, reference, references, quote, excerpt, corpus, dataset, recall, recalling, memorized, memorize, memorization, seen, encountered, recognize, recognized, familiar, familiarity, rely.*
- **certainty_attribution**: *certain, certainly, clearly, definitely, confident, confidence, sure, known, know, knowing, specific, specifically, indeed, confirm, confirms, confirmed, absolutely, undoubtedly, present.*

Non-member-leaning clusters.

- **uncertainty_hedge**: *might, maybe, perhaps, possibly, could, seems, appear, appears, unlikely, chance, guess, uncertain, probably, roughly, suggest, suggests, hypothetical, speculate, speculation, unsure, unclear, ambiguous, vague, approximate, think, thought.*
- **temporal_cutoff**: *future, cutoff, recent, recently, beyond, post, later, newer, current,*

currently, today, modern, latest, upcoming, since, date, knowledge, limit, boundary, exceed, training.

- **fictional_hypothetical**: *fictional, fiction, hypothetical, scenario, imaginary, invented, made, created, fake, constructed, example, illustrative, suppose, assuming, assume, imagine, imagined, event, events, real, realistic.*

Since Table 7 serves as a descriptive cross-model diagnostic rather than a scoring signal, the absolute cluster proportions matter less than the consistency of their pattern across models.

G Trace vs. Self-Report

A natural question is whether the model’s direct binary answer to the introspective prompt—the {“member”: 0/1} JSON field—already suffices for membership inference, obviating the need for trace-based scoring. We compare this **Self-Report** baseline against TraceMIA (lexicon) on the same prompts and traces, at length = 32.

Model	Dataset	Self-Report		TraceMIA (lexicon)	
		AUC	PosRate	AUC	ΔAUC
DS-14B	WikiMIA	0.7164	0.371	0.9045	+0.188
	Arxiv	0.6590	0.326	0.8270	+0.168
	Book	0.6880	0.295	0.7596	+0.072
QwQ-32B	WikiMIA	0.8840	0.415	0.9873	+0.103
	Arxiv	0.6516	0.559	0.8940	+0.242
	Book	0.7494	0.449	0.8593	+0.110

Table 13: Self-Report vs. TraceMIA (lexicon) at length = 32. Self-Report uses the model’s binary JSON answer directly; PosRate is the fraction of samples predicted as members (true base rate = 0.5). Parse-failure samples (up to 19% on QwQ-32B due to truncated reasoning) are excluded from Self-Report but scored normally by TraceMIA.

Three observations emerge from Table 13. **Self-Report provides a non-trivial but limited signal.** The baseline exceeds prior black-box methods (e.g., SaMIA, CDD, SimMIA in Table 1) by a clear margin, indicating that introspective prompting alone elicits a measurable membership signal. TraceMIA (lexicon) nonetheless yields a consistent improvement of +0.07 to +0.24 AUC across all six settings, suggesting that the lexical content of the reasoning trace carries discriminative information not preserved in the final binary judgment.

Self-Report reveals model-specific calibration biases. DS-14B exhibits a conservative bias (PosRate ≈ 0.30, well below the 0.5 base rate), whereas

QwQ-32B shows an overconfident tendency on Arxiv (PosRate = 0.56). These opposing biases induce asymmetric error profiles that degrade Self-Report performance, while TraceMIA’s score remains comparable across both models, suggesting that its lexical score is less directly tied to the calibration of the binary self-report.

Self-Report is undefined on 16–19% of QwQ-32B traces. When extended reasoning exhausts the generation budget, the model fails to emit a closing JSON object, leaving Self-Report inapplicable on these samples; TraceMIA, by contrast, scores them from the partial trace without modification. This robustness margin supports the use of trace-level scoring as a more reliable alternative to direct self-report.

Taken together, these results corroborate the design choice motivated in §3.2: the model’s binary self-report constitutes an informative but lossy summary of its deliberation, whereas the trace’s lexical content offers a stronger and more robust basis for membership inference.