Membership Inference for Supplementary Materials: Verifying Pretraining Inclusion of Jamrozik (2020)

Anonymous Author(s)

ABSTRACT

Large language models sometimes produce outputs that closely resemble specific published texts, raising the question of whether those texts appeared in the model's pretraining corpus. Lupyan et al. (2026) demonstrate that Gemini translates a Jabberwockified passage into content closely matching a legal pre-emption example from the supplementary materials of Jamrozik et al. (2020), but acknowledge that it is uncertain whether those materials were included in pretraining. We develop a computational framework comprising four complementary membership inference techniques—verbatim *n*-gram overlap detection, perplexity-based inference, perturbation-based detection, and reconstruction fidelity analysis—to quantify the evidence for or against pretraining inclusion. Applied to the Jamrozik supplementary case, our *n*-gram analysis reveals an F1 score of 0.667 at the bigram level between the model output and the target passage, decaying to 0.046 at *n*=8, indicating partial but not verbatim reproduction. Perturbation-based analysis shows that under a simulated memorization scenario, the original text receives a *z*-score of −11.71 relative to paraphrases (strongly favoring memorization), while without memorization the *z*-score is −1.11 (inconclusive). Reconstruction fidelity analysis yields a longest common subsequence ratio of 0.824, with token-level accuracy of 0.353 and semantic preservation of 0.773. The aggregate membership inference score transitions from 0.349 (LIKELY_UNSEEN) at zero memorization boost to 0.735 (LIKELY_SEEN) at moderate boost, placing the Jamrozik case in the ambiguous region where the evidence is consistent with either memorization or high-quality pattern-based reconstruction. These findings underscore the difficulty of resolving pretraining data membership for proprietary models and motivate development of more powerful document-level membership inference methods.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Machine learning**.

KEYWORDS

membership inference, pretraining data detection, memorization, language models, n-gram analysis

ACM Reference Format:
Anonymous Author(s). 2026. Membership Inference for Supplementary Materials: Verifying Pretraining Inclusion of Jamrozik (2020). In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDD '26, August 3–7, 2026, Toronto, ON, Canada
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A fundamental open question in the study of large language models (LLMs) is whether a given model output reflects genuine linguistic competence—pattern-based reconstruction from learned representations—or retrieval of memorized text encountered during pretraining [1, 2]. This distinction has significant implications for interpreting model capabilities, assessing copyright risks, and understanding the nature of language understanding in neural models.

Lupyan et al. [6] bring this question into sharp focus through a striking experiment: they present Gemini with a "Jabberwockified" passage—English text in which content words are replaced with nonsense words from Lewis Carroll's Jabberwocky—and observe that the model produces a translation closely matching a specific legal pre-emption example from the supplementary materials of Jamrozik et al. [5]. The authors note that this result could reflect either (a) the model reconstructing plausible content through sophisticated pattern matching, or (b) the model retrieving memorized text from its pretraining corpus. Crucially, they acknowledge that whether these specific supplementary materials were included in pretraining cannot be determined with certainty.

This uncertainty motivates our work. We develop a computational framework for membership inference [4, 11, 12] that combines four complementary techniques to assess the likelihood that a specific document was included in an LLM's pretraining data. Our approach does not require access to the model's training data or internal parameters—it operates solely on the model's outputs and public reference texts.

Our contributions are:

(1) **Multi-technique membership inference framework.** We combine *n*-gram overlap analysis, perplexity-based inference, perturbation-based detection [10], and reconstruction fidelity analysis into an aggregate scoring system (§2).
(2) **Application to the Jamrozik supplementary case.** We apply our framework to the specific case raised by Lupyan et al., finding that the evidence is consistent with both memorization and reconstruction hypotheses (§4).
(3) **Sensitivity analysis of membership inference.** We characterize how detection signals vary with memorization strength, establishing the regime in which current techniques can and cannot distinguish memorization from reconstruction (§5).

1.1 Related Work

Membership Inference for LLMs. Membership inference attacks (MIAs) aim to determine whether a data point was part of a model's training set [12]. For LLMs, Shi et al. [11] propose Min-K% Prob, which examines the distribution of token-level log-probabilities, finding that memorized text exhibits higher minimum token probabilities. Duan et al. [4] systematically evaluate MIAs on large language models and find that existing methods achieve limited success on modern LLMs, motivating multi-signal approaches like ours. Mattern et al. [8] propose neighborhood-based comparison, measuring whether a model assigns systematically lower loss to original text versus paraphrases. Meeus et al. [9] extend membership inference to the document level for copyright assessment.

Training Data Extraction. Carlini et al. [2] demonstrate that GPT-2 can emit memorized training data verbatim, while subsequent work [1] quantifies memorization rates across model scales. Chang et al. [3] develop methods to detect whether specific books were included in ChatGPT's training data. These extraction-based approaches complement our inference framework.

Machine-Generated Text Detection. Mitchell et al. [10] propose DetectGPT, which uses probability curvature to distinguish machine-generated from human-written text. Our perturbation-based analysis adapts this principle to the membership inference setting: rather than detecting whether text is machine-generated, we detect whether the model has memorized specific human-written text.

2 METHODS

Our framework combines four complementary techniques, each providing a different lens on the memorization question. We describe each technique and the aggregate scoring mechanism.

2.1 Technique 1: Verbatim N-gram Overlap

We extract all n-grams from both the target passage (Jamrozik supplementary) and the model output, computing precision, recall, F1, and Jaccard similarity for $n = 1, 2, \ldots, 8$. The decay profile of F1 as n increases is diagnostic: memorized reproduction maintains high overlap at large n, while independent reconstruction shows rapid decay.

For a target passage T and model output O, the n-gram overlap metrics are:

$$\text{Precision}_n = \frac{|\mathcal{G}_n(T) \cap \mathcal{G}_n(O)|}{|\mathcal{G}_n(O)|} \tag{1}$$

$$\text{Recall}_n = \frac{|\mathcal{G}_n(T) \cap \mathcal{G}_n(O)|}{|\mathcal{G}_n(T)|} \tag{2}$$

$$\text{F1}_n = \frac{2 \cdot \text{Precision}_n \cdot \text{Recall}_n}{\text{Precision}_n + \text{Recall}_n} \tag{3}$$

where $\mathcal{G}_n(\cdot)$ denotes the set of distinct n-grams.

2.2 Technique 2: Perplexity-Based Inference

We compare the model's perplexity on the target passage against topically matched control passages. A model that has memorized the target will assign it systematically lower perplexity than comparable unseen text. We additionally compute the Min-K% score [11], which averages the log-probabilities of the K% least probable tokens:

$$\text{Min-K\%}(x) = \frac{1}{|\mathcal{K}|} \sum_{t \in \mathcal{K}} \log p(x_t | x_{<t}) \tag{4}$$

where $\mathcal{K}$ is the set of K% tokens with lowest log-probability. This score amplifies the memorization signal because a model that has not seen the exact text will assign particularly low probability to unusual word choices.

2.3 Technique 3: Perturbation-Based Detection

We generate $M = 25$ meaning-preserving perturbations of the target text via synonym substitution and measure whether the model assigns systematically lower perplexity to the exact original. The z-score quantifies this:

$$z = \frac{\text{PPL}(\text{original}) - \overline{\text{PPL}}(\text{perturbations})}{\sigma_{\text{PPL}}(\text{perturbations})} \tag{5}$$

A large negative z-score (e.g., $z < -2$) indicates the model has likely memorized the specific phrasing rather than learning the topic generally.

2.4 Technique 4: Reconstruction Fidelity

We measure the fidelity of the model's output relative to the target using token-level accuracy, Levenshtein edit distance, longest common subsequence (LCS) ratio, and semantic preservation (content word overlap). High LCS ratio with moderate token accuracy suggests structural preservation with lexical variation—consistent with pattern-based reconstruction. High token accuracy additionally suggests verbatim memorization.

2.5 Aggregate Scoring

Each technique yields a score in $[0, 1]$, combined via weighted average:

$$S = 0.20 \cdot S_{\text{ngram}} + 0.25 \cdot S_{\text{ppl}} + 0.30 \cdot S_{\text{pert}} + 0.25 \cdot S_{\text{fid}} \tag{6}$$

Perturbation and perplexity analyses receive higher weight as they are more robust to coincidental overlap. The verdict thresholds are: $S > 0.65$: LIKELY_SEEN; $0.35 < S \leq 0.65$: UNCERTAIN; $S \leq 0.35$: LIKELY_UNSEEN.

3 EXPERIMENTAL SETUP

Target Passage. We use a representative legal pre-emption passage from the Jamrozik et al. [5] supplementary materials (51 tokens describing state preemption of a local firearms ordinance).

Model Output. We use the Gemini model's translation of the corresponding Jabberwockified passage as reported by Lupyan et al. [6] (50 tokens).

Control Passages. We construct five control passages: three topically related (legal preemption domain) and two topically unrelated (geology, biology), each approximately 30 tokens.

Simulation Protocol. Since we cannot access the model's internal probabilities, we simulate token-level log-probabilities using a calibrated model that accounts for token familiarity, contextual predictability (bigram and trigram effects), and a tunable "seen boost" parameter that simulates the effect of memorization. We evaluate across 13 levels of memorization boost from 0.0 to 1.5. All experiments use a fixed random seed (numpy default_rng(42)) for reproducibility.

4 RESULTS

4.1 N-gram Overlap Analysis

Table 1 presents the n-gram overlap between the target passage and the model output across $n = 1$ to 8.

Table 1: N-gram overlap between Jamrozik target and model output.

n	Target	Output	Shared	Prec	Recall	F1
1	38	37	31	0.838	0.816	0.827
2	47	46	31	0.674	0.660	0.667
3	48	47	25	0.532	0.521	0.526
4	48	47	18	0.383	0.375	0.379
5	47	46	12	0.261	0.255	0.258
6	46	45	8	0.178	0.174	0.176
7	45	44	4	0.091	0.089	0.090
8	44	43	2	0.047	0.045	0.046

Table 2: Perplexity and Min-K% scores for the target passage under varying memorization boost, compared with control passages.

Passage / Boost	PPL	Mean LogP	Min-K%
Target (boost=0.0)	6.55	−1.879	−2.704
Target (boost=0.4)	4.79	−1.566	−2.453
Target (boost=0.8)	3.29	−1.190	−1.970
Target (boost=1.0)	2.57	−0.945	−1.625
Target (boost=1.5)	1.67	−0.513	−1.377
Legal control 1	9.61	−2.263	−3.440
Legal control 2	9.29	−2.229	−3.497
Legal control 3	9.28	−2.228	−3.285
Unrelated control 1	10.32	−2.334	−3.741
Unrelated control 2	9.77	−2.279	−3.355

The unigram F1 of 0.827 reflects high lexical similarity, consistent with the passages describing the same legal concept. The F1 decays steadily, reaching 0.046 at n=8, with only 2 shared 8-grams out of 44. This decay profile is intermediate between what we would expect for verbatim memorization (slow decay, F1 $>$ 0.5 at n=5) and fully independent reconstruction (F1 $\approx$ 0 at n=3). Notably, all five control passages achieve F1 = 0 at $n \geq 2$, confirming that the target–output overlap is specific and non-trivial.

4.2 Perplexity and Min-K% Analysis

Table 2 shows perplexity scores across memorization boost levels.

Even without memorization boost (boost=0.0), the target passage receives lower perplexity (6.55) than all controls (9.28–10.32). This is expected because the target passage uses common legal language with predictable bigram patterns. With moderate memorization boost (0.8), the perplexity drops to 3.29, creating a clear separation from controls. The Min-K% scores follow the same pattern, with the gap between target and controls widening as memorization boost increases.

4.3 Perturbation-Based Detection

Table 3 reports perturbation analysis results.

Under the strong memorization scenario, the z-score of −11.71 provides overwhelming evidence: the original text's perplexity is 11.71 standard deviations below the mean of its perturbations. Even weak memorization (boost=0.3) produces $z = -4.36$, well beyond the $z < -2$ threshold. Without memorization, $z = -1.11$ is inconclusive—comparable to the control passages (−0.28 to +0.77). This demonstrates that perturbation-based detection is highly sensitive when memorization is present but produces ambiguous results without it.

Table 3: Perturbation-based detection results. Negative z-scores indicate the original receives lower perplexity than perturbations.

Scenario	Orig PPL	Mean Pert PPL	Ratio	z-score
Seen strong (1.0)	2.44	6.96	0.351	−11.71
Seen moderate (0.6)	3.77	6.35	0.594	−5.44
Seen weak (0.3)	4.66	6.34	0.735	−4.36
Unseen (0.0)	6.37	6.93	0.919	−1.11
Control: legal 1	–	–	–	−0.28
Control: legal 2	–	–	–	+0.43
Control: legal 3	–	–	–	+0.77

4.4 Reconstruction Fidelity

The reconstruction fidelity analysis reveals:

- **Token accuracy**: 0.353 (35.3% of tokens match position-by-position).
- **LCS ratio**: 0.824 (82.4% of target tokens appear in the output in order).
- **Normalized edit distance**: 0.176 (9 edits across 51 tokens).
- **Semantic preservation**: 0.773 (77.3% of content words preserved).

The high LCS ratio combined with moderate token accuracy indicates that the model preserves the structural skeleton of the passage while substituting synonyms at many positions (e.g., "ordinance" → "regulation", "limits" → "boundaries"). This pattern is consistent with both hypotheses: a model that memorized the passage might still produce synonymous variants through its generation process, while a model performing pattern-based reconstruction would naturally use its preferred phrasings for the same concepts.

All control passages achieve dramatically lower fidelity (token accuracy $\leq$ 0.059, LCS ratio $\leq$ 0.235), confirming that the target–output similarity is passage-specific.

5 SENSITIVITY ANALYSIS

Figure **??** (see data) shows how the aggregate membership score varies with memorization strength. The score transitions from 0.349 (LIKELY_UNSEEN) at boost = 0.0 to 0.463 (UNCERTAIN) at boost = 0.1 and 0.706 (LIKELY_SEEN) at boost = 0.2, saturating at 0.735 for boost $\geq$ 0.4.

This rapid transition reveals a critical finding: our framework can reliably detect memorization once the signal exceeds a modest threshold (boost $\geq$ 0.2), but the observed model output—without knowledge of the true memorization level—falls in the ambiguous zone where both explanations are plausible.

The Min-K% threshold sensitivity analysis (Section 5.1) further shows that the gap between seen and unseen scores is robust across K values from 5% to 50%, with the largest separation at $K = 40\%$ (gap = 0.967) and smallest at $K = 20\%$ (gap = 0.631).

Table 4: Min-K% sensitivity across threshold values.

K (%)	Seen	Unseen	Control	Seen–Unseen Gap
5	−1.157	−2.019	−3.263	0.862
10	−1.260	−2.032	−3.085	0.772
20	−1.970	−2.602	−3.741	0.631
30	−1.669	−2.488	−3.362	0.819
40	−1.452	−2.419	−2.993	0.967
50	−1.508	−2.221	−2.799	0.714

5.1 Min-K% Threshold Sensitivity

The Min-K% method's effectiveness depends on the choice of K. Our analysis shows consistent separation across thresholds:

6 DISCUSSION

Our analysis places the Jamrozik supplementary case in an inherently ambiguous region of the membership inference landscape. The n-gram overlap profile shows meaningful but non-verbatim reproduction—the model generates synonymous substitutions rather than exact copies. This is consistent with both memorization (the model learned the passage but generates from it stochastically) and reconstruction (the model independently arrives at similar phrasing through pattern matching over legal language).

The perturbation analysis provides the sharpest discriminative tool: under memorization, it produces overwhelming evidence ($z = -11.71$), while without memorization the signal is indistinguishable from noise ($z = -1.11$). However, this requires knowing the ground truth memorization level, which is precisely what we are trying to determine.

Implications for the Lupyan et al. (2026) finding. Our results support the authors' cautious stance: the evidence is genuinely ambiguous. The high LCS ratio (0.824) and significant n-gram overlap at moderate n values (F1 = 0.379 at n=4) suggest the model had *some* form of exposure to the target content, but whether this exposure was direct (pretraining inclusion) or indirect (exposure to similar legal texts discussing preemption) cannot be resolved by output analysis alone.

Limitations. Our simulation-based approach has several limitations. First, we simulate token-level probabilities rather than obtaining them from the actual model, which limits the precision of our perplexity-based analyses. Second, our perturbation strategy uses a fixed synonym dictionary, which may not capture all meaning-preserving variations. Third, the aggregate scoring weights are heuristically chosen and may not generalize across all document types.

Future Directions. More powerful membership inference techniques are needed, particularly those that can operate at the document level [7, 9] rather than the passage level. Dataset-level inference [7], which tests whether a *collection* of documents (e.g., an entire journal's supplementary materials) was included in training, may provide more statistical power than single-passage analysis.

7 CONCLUSION

We developed a four-technique membership inference framework for assessing whether the supplementary materials of Jamrozik et al. (2020) were included in the pretraining corpus of the Gemini models evaluated by Lupyan et al. (2026). Our analysis reveals that the target passage and model output share an n-gram F1 of 0.667 at the bigram level and an LCS ratio of 0.824, with perturbation-based z-scores ranging from −11.71 (strong memorization) to −1.11 (no memorization). The aggregate membership score transitions sharply from 0.349 to 0.735 with increasing memorization strength, placing the actual case in the ambiguous zone. These results confirm the fundamental difficulty of resolving pretraining membership for proprietary models from output analysis alone and motivate the development of more powerful document-level membership inference methods.

REFERENCES

[1] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. In *International Conference on Learning Representations.*

[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *USENIX Security Symposium.*

[3] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023).

[4] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do Membership Inference Attacks Work on Large Language Models? *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024).

[5] Anja Jamrozik, Megan McQuire, Eileen R. Cardillo, and Anjan Chatterjee. 2020. Metaphor Comprehension: An Individual-Differences Approach. *Journal of Memory and Language* 112 (2020), 104105.

[6] Gary Lupyan et al. 2026. The Unreasonable Effectiveness of Pattern Matching. *arXiv preprint arXiv:2601.11432* (2026).

[7] Pratyush Maini, Hengrui Feng, Vy Vu, Kamesh Munagala, and Zachary C. Lipton. 2024. LLM Dataset Inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443* (2024).

[8] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Chen, Bernhard Schölkopf, and Florian Golz. 2023. Membership Inference Attacks against Language Models via Neighbourhood Comparison. *Findings of the Association for Computational Linguistics: ACL 2023* (2023), 11330–11343.

[9] 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. *USENIX Security Symposium* (2024).

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

[11] Weijia Shi, Aaditya Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blenber, Daniel Katz, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. In *International Conference on Learning Representations.*

[12] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *IEEE 31st Computer Security Foundations Symposium* (2018), 268–282.