CAN FACTUAL PROBES DETECT CORPUS EVIDENCE? AN AUDITABLE BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Can factual probes distinguish reference-corpus evidence from answer plausibility? We introduce the Factual Exposure Benchmark (FEB), which records subject–object co-occurrences in an exhaustive scan of 6,407,814 English Wikipedia documents. Its labels describe this reference snapshot, not membership in the evaluated models' training data. The benchmark pairs 878 evidence-absent facts with evidence-present facts matched on entity frequency and length, and includes subject-preserving object corruptions. Across five Pythia checkpoints and GPT-Neo-2.7B, pooled probe features reach mean AUC 0.5583 over five overlapping entity-grouped splits on the matched slice, compared with 0.9230 against corrupted objects. Single-model matched AUCs range from 0.5236 to 0.5676. GPT-Neo continuation-likelihood statistics also yield modest discrimination, with AUCs 0.5488–0.5572. Pythia deduplication twins have nearly identical true-versus-corrupted answer gaps. When labels are instead defined by exact token-window presence in the released Pythia shards, the five checkpoints' scalar answer probabilities yield AUCs 0.501–0.518. A fixed 60-feature classifier yields $0.5545/0.5235$ on all facts but $0.5026/0.4807$ after exact-relation covariate matching. These contrasts show how the negative pool changes the apparent strength of factual auditing signals. They support matched evaluation of reference-corpus evidence, while leaving detection of actual training exposure unresolved.

1 INTRODUCTION

Pretraining corpora are large, partially undocumented, and increasingly contested. Auditors, privacy regulators, and rights holders would like to know whether a specific text or the content of a specific claim was part of the training data of a deployed model. For whole documents, membership inference is a mature field (Shokri et al., 2017; Carlini et al., 2022; Shi et al., 2024; Duan et al., 2024). But the unit that matters in most public disputes is smaller: a single factual claim, e.g. that a person was born in a particular town. Whether such a claim can be audited from model behavior alone determines whether "the model was trained on this" is an answerable question or an unfalsifiable one.

Consider a concrete case. Asked to complete "Pennsylvania Route 696 is located in ____", a model answers *Pennsylvania* correctly. The only question that matters for an audit is *why*: did the model answer correctly because a supporting sentence was in its training corpus, or because the fact is popular, plausible, and reinforced across many documents (Kandpal et al., 2022; Mallen et al., 2023)? A model may know a fact through popularity or inference rather than exposure; *knowing a fact is not the same as having seen its evidence*. Separating the two is the prerequisite for any defensible fact-level exposure claim, and it motivates the matched comparisons below.

Prior probing work examines related but different outcomes. ProPILE (Kim et al., 2023) probes personal-information leakage, while Huang et al. (2022) study email recovery from email contexts and owner names. These settings do not directly establish whether behavioral scores identify corpus evidence for a factual claim. At that granularity, a classifier may exploit entity popularity or answer plausibility unless comparisons control for these differences. Text-level membership inference methods (Min-K% Prob (Shi et al., 2024), MIMIR (Duan et al., 2024), Tag&Tab (Antebi et al., 2025), and time-split variants (Yi & Li, 2026)) operate at document granularity and inherit their labels from corpus construction or release dates, not from an audit of where a *claim* is sup-

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.

ported. Memorization increases with training-data duplication (Carlini et al., 2023), and deduplication reduces extraction risk (Lee et al., 2022; Kandpal et al., 2022), leaving open whether *semantic*, fact-level exposure leaves any detectable trace at all. Our benchmark is designed to make this confound explicit by pairing audit-grade fact labels with frequency-matched negatives and a falsification-oriented evaluation.

The benchmark. FEB defines membership operationally and auditably. From LAMA-TREx (Petroni et al., 2019) we take subject–relation–object facts spanning 41 Wikidata relations. A fact is a *member* iff the subject and object strings co-occur within a fixed window in the English Wikipedia 2023-11-01 snapshot (41 Parquet shards, 6,407,814 documents), located with an Aho–Corasick multi-pattern automaton (Aho & Corasick, 1975); we record the first supporting document, its component, and a verbatim ± 40-character span. A fact is a *true negative* iff the full scan finds zero co-occurrences; 878 true negatives are additionally matched one-to-one to members on entity frequency and length, and we add *corrupted negatives* that keep the subject and replace the object with a same-relation object verified never to co-occur with it. All frequency covariates are computed from the same scan, so every label and every covariate is reproducible from the corpus.

Two falsification diagnostics. Because "plausible knowledge" is such a strong confounder, FEB provides two descriptive controls for examining exposure-detection interpretations. First, an *honest slice*: 878 strictly frequency-matched member–non-member pairs on which pooled multi-model probe features, entity-grouped repeated splits, and white-box statistics can be compared under matched evidence. Second, a *deduplication twin test*: the Pythia suite (Biderman et al., 2023) provides deduped and non-deduped checkpoints trained on the same corpus, so if probe responses track exposure, the probability a model assigns to a gold answer should shift with deduplication *more for members than for non-members*. These controls answer complementary descriptive questions; neither cancels unknown fact-level exposure differences in the actual training corpora.

These controls diagnose different alternative explanations: lexical frequency, label definition, and checkpoint-specific behavior. Their interpretation depends on the corpus construction and measurement protocol; a checkpoint contrast is not a necessary condition for exposure detection. Failure on a diagnostic does not prove that a model never saw a fact. Our benchmark measures performance under explicit reference-corpus and lexical operationalizations rather than providing a universal decision rule for exposure.

Findings. The original five-Pythia mixed-pool analysis yields AUC 0.849 for probe-plus-frequency features. In the six-model extension, which adds GPT-Neo-2.7B, pooled probe features yield grouped-CV AUC 0.5583 on matched reference-evidence labels and 0.9230 against object corruptions. Single-model matched AUCs range from 0.5236 to 0.5676. The large change with negative construction supports a plausibility confound, but neither matching nor deduplication supplies fact-specific training-membership truth.

Our contributions:

- **An auditable fact-level exposure benchmark.** FEB provides full-corpus scanning with recorded verbatim evidence spans, frequency-matched true negatives, verified counterfactual negatives, and entity-grouped splits, so that every label and covariate is reproducible from the corpus by a single scan (Section 2).
- **A falsification-oriented diagnostic methodology.** Two reusable controls, frequency-matched slices and the deduplication twin comparison, test sensitivity to negative construction and checkpoint choice and can accompany any future exposure-detection claim; we instantiate them with a probe and feature framework covering four black-box probe families, cross-model consistency features, and continuation-likelihood baselines, evaluated with entity-clustered statistics (Section 3).
- **A rigorously documented negative result and its audit implications.** The tested probe and likelihood statistics show modest reference-label discrimination under matching (AUC at most 0.57), compared with AUC above 0.92 against corrupted alternatives. Negative construction therefore materially changes the inference task (Section 4).
- **Public release.** The benchmark (facts, verbatim evidence spans, and matched negatives), the full-corpus scanning pipeline, probe code, and both diagnostics will be released upon publication to support auditable evaluation of future exposure claims.

2 THE FACTUAL EXPOSURE BENCHMARK

2.1 TASK DEFINITION

A fact is a triple $f = (s, r, o)$ with subject s, relation r, and object o from LAMA-TREx (Petroni et al., 2019). Given an audited corpus C with normalization $\phi(\cdot)$ (whitespace canonicalization and lowercasing), define

$$y(f) = 1 \iff \exists\, d \in C \text{ s.t. } \phi(s) \text{ and } \phi(o) \text{ occur within a window } w = 512 \text{ characters in } \phi(d), \tag{1}$$

and let $\mathrm{Ev}(f) = (d^\star, \mathrm{span}^\star)$ record the first supporting document and the verbatim ± 40-character span around the co-occurrence. The auditor's task is to predict $y(f)$ from model behavior alone (black-box) or from internal statistics (white-box), with $\mathrm{Ev}(f)$ used only for evaluation. This differs from document-level MIA in two ways that make it harder: the evidence is a short span inside a 6.4M-document corpus, and the surface patterns are short answer strings rather than long documents, so document-level cues such as near-duplicate structure are unavailable. In our instantiation every object is a single surface word, so the discriminating evidence, when it exists, reduces to a handful of tokens.

2.2 AUDITABLE EVIDENCE VIA FULL-CORPUS SCANNING

We scan the complete English Wikipedia 2023-11-01 snapshot (41 Parquet shards; 6,407,814 documents after filtering). For each fact, the normalized subject and object strings are inserted into an Aho–Corasick automaton (Aho & Corasick, 1975) so that a single pass over the corpus yields all occurrences; a document supports a fact iff an object occurrence falls within $w = 512$ characters of a subject occurrence. The scan also accumulates, for every fact, the number of subject and object occurrences across all documents, giving corpus-derived covariates $\mathrm{freq}(s) = \#s/|C|$ and $\mathrm{freq}(o) = \#o/|C|$. Members are labeled by Eq. 1 with their evidence record $\mathrm{Ev}(f)$; true negatives are facts for which the full scan found zero co-occurrences (our scan covers the entire snapshot, well above a conservative minimum-absence threshold of 5,000 documents). Every label, span, and covariate is recomputable from the snapshot by a single command.

2.3 MATCHED NEGATIVES

Raw non-members are systematically lower-frequency than members; a classifier given unmatched negatives can succeed on frequency alone. We therefore construct two negative families. **(a) Frequency-matched true negatives.** Within each relation, we greedily pair each true negative (the anchoring side) with the member closest in $(\log(1 + 10^6\,\mathrm{freq}(s)),\ \mathrm{len}(s),\ \mathrm{len}(o))$, yielding 878 strictly matched pairs (the *honest slice*; Section 3). **(b) Object-corrupted counterfactual negatives.** For each member we generate corruptions that replace o with one or two other objects of the same relation (sampled from the relation's object pool, stratified by length and first token), keeping the subject fixed, then verify by the full scan that (s, o') never co-occur, a *definitional* non-member whose subject frequency is identical to its member by construction. Corrupted negatives serve as a diagnostic foil: they isolate plausibility from exposure (Section 4.4).

2.4 COMPOSITION AND SPLITS

Table 1 summarizes the benchmark. The primary mixed-pool evaluation uses an 80/20 `GroupShuffleSplit` by subject entity with seed 2027. The honest-slice fixed-feature comparison uses five overlapping 70/30 subject-grouped splits with the same seed. No entity appears in both train and test within a split, so a classifier cannot memorize entity-specific probe values. Per-relation metrics are reported for the 39 relations with at least 20 test rows.

Table 1: **Every FEB label is anchored in a single auditable scan of 6,407,814 Wikipedia documents.** Membership is defined with respect to the English Wikipedia 2023-11-01 snapshot; members carry verbatim evidence spans, and every negative is a verified absence from the same scan (4,053 negatives in total: 878 frequency-matched true negatives and 3,175 object-corrupted counterfactuals).

Quantity	Value
Audited corpus (documents)	6,407,814
Facts (total)	7,878
Members (evidence span found)	3,825
True negatives (verified absent, frequency-matched)	878
Corrupted negatives (verified non-co-occurring)	3,175
Distinct subject entities	4,624
Wikidata relations	41
Honest slice (matched pairs)	878 pairs / 1,756 facts
relations / entities in honest slice	39 / 1,717
Objects that are single surface words	100%

3 Diagnostic Methodology

3.1 Black-box probes

Following the protocol of prior probing work (Petroni et al., 2019; Kim et al., 2023), each fact is rendered into four probe families, all scored as scalar log-probability summaries (vLLM, bf16, fixed seed):

- **A. Direct multiple choice.** The primary relation template (e.g., "s is located in\Answer:") with four options: the gold object plus three same-relation distractors. We record per-option log-probabilities, the gold option's rank, the probability mass on the gold answer p_{correct}, the top-1 margin, and the entropy over options.
- **B. True/false judgment.** The declarative statement ("*s was born in o*. Is this true?") scored by the log-probability ratio of the continuations *True* vs. *False* (tf_ratio).
- **C. Template variance.** The same MCQ under the relation's paraphrase templates (up to five stored per fact; one or two in our instantiation), summarized by the mean and standard deviation of p_{correct} across templates.
- **D. Distractor sensitivity.** A second MCQ whose distractors are matched to the gold object on word count or first token ("hard" distractors); the gap between standard and hard conditions (hard_easy_gap) measures how sharply recall depends on distractor difficulty.

3.2 Features and classifier

The original five-Pythia mixed-pool analysis uses eight probe features (p_{correct}, margin, entropy, max option log-probability, tpl_mean, tpl_std, hard_easy_gap, tf_ratio), and across the five checkpoints we add four cross-model features: the mean, standard deviation, and maximum of p_{correct} and the slope of p_{correct} regressed on log parameter count ($\text{p_mean}, \text{p_std}, \text{p_max}, \text{scale_slope}$), giving 12 probe features in total. Five corpus-derived covariates (entity_freq, answer_freq, answer_len, entity_len, and a year placeholder, constant in this snapshot) complete a 17-dimensional input. The classifier is logistic regression (median imputation, standardization, balanced class weights), chosen for auditability over heavier models; entity-grouped splits and seed 2027 are fixed throughout. We report AUC-ROC, AUPRC, and *TPR@1%FPR*, a low-false-positive operating-point metric, with entity-cluster bootstrap 95% CIs (200 resamples with refit) and entity-block permutation tests (100 permutations).

3.3 Diagnostic 1: the frequency-matched honest slice

The honest slice is the 878-pair subset of Section 2.3 in which members and true negatives are matched on the log entity-frequency key. The six-model extension uses ten features per model:

correct-choice probability, margin, entropy, maximum log-probability, rank, true-response probability, true/false ratio, template mean, hard-choice probability, and hard-choice margin. Concatenation gives 60 features; the earlier five-model comparison has 50. Missing feature values are filled with zero. We average held-out AUC over five entity-grouped shuffle splits, each with 30% test entities and seed 2027. These are repeated splits, not disjoint cross-validation folds. Standardization is fitted to each training split, followed by logistic regression with default class weighting and at most 2,000 iterations. This protocol differs from the mixed-pool classifier above. Matching reduces measured frequency differences but does not eliminate all popularity or plausibility effects; residual discrimination concerns reference-evidence labels, not identified training exposure.

3.4 Diagnostic 2: the deduplication twin test

Let $p_m(f)$ be the probability a model assigns to the gold answer of f in the primary MCQ. For each twin pair (deduped D, original O) define $\Delta p(f) = p_O(f) - p_D(f)$. The original corpus contains every document of the deduped corpus plus its exact duplicates; under a training-exposure interpretation, Δp should be larger for members than for non-members. We test this with a one-sided Mann–Whitney U test and report the twin AUC (classification of members by Δp), both over the full probe set and over strictly paired member–corrupted-negative pairs. The test is one-sided against exposure: it is passed only if members show systematically larger Δp, but shared reference-label errors need not cancel; fact-specific training counts are required for an exposure interpretation (Section 5).

3.5 Continuation-likelihood baselines

We compare probe features with gold-answer continuation likelihood and low-probability token summaries inspired by Min-K% Prob (Shi et al., 2024). These statistics score the answer continuation, rather than the entire evidence document, and are likelihood baselines, not upper bounds on what other detectors can achieve. On GPT-Neo-2.7B we evaluate negative log-likelihood and the lowest-probability 20% and 40% of continuation tokens on the same 1,756 matched facts.

4 Experiments

4.1 Setup

Models. Five Pythia checkpoints (Biderman et al., 2023): 410M (deduped), 2.8B and 6.9B in both deduped and non-deduped variants, all base models, bf16, vLLM inference, seed 2027. The deduped/non-deduped pairs are the twin controls of Section 3.4. **Baselines.** (i) frequency-only LR (covariates only); (ii) probe-only LR; (iii) probe + frequency LR (the main setting); (iv) white-box NLL and Min-K%. The matched-slice extension adds GPT-Neo-2.7B. Original five-Pythia mixed-pool results are identified separately from the six-model grouped-CV comparison.

4.2 Naive mixed-pool evaluation manufactures a strong signal

Table 2 shows the standard evaluation: members versus all 4,053 negatives in one pool, entity-grouped 80/20 split (6,332/1,546). Probe features dominate: pooled probe+frequency features reach AUC 0.8490 (95% CI $[0.8303, 0.8697]$), AUPRC 0.8309, and the block permutation test confirms the gain over the frequency-only baseline is significant (observed ΔAUC 0.2951, $p < 0.01$, 100 permutations). Per-relation AUCs in the original five-model analysis range from 0.6508 (P106) to 0.9883 (P1001), with 23 of 39 relations above 0.9; per-relation details are in Appendix C. A reader of this table alone might overinterpret black-box probe separability as evidence of training exposure; the matched audit and the training-window scan in Section 4.4 are required for that claim.

Figure 1 previews this label sensitivity: the easy corruption control is high, whereas matched reference evidence and exact-relation shard matching are near chance.

Two details already caution against that reading. First, the metric that matters for audits barely moves: even the full model reaches TPR@1%FPR of only 0.1123 (CI $[0.0828, 0.2764]$), and five relations have TPR@1%FPR $= 0$. Second, the deduped and non-deduped twins score within 0.007

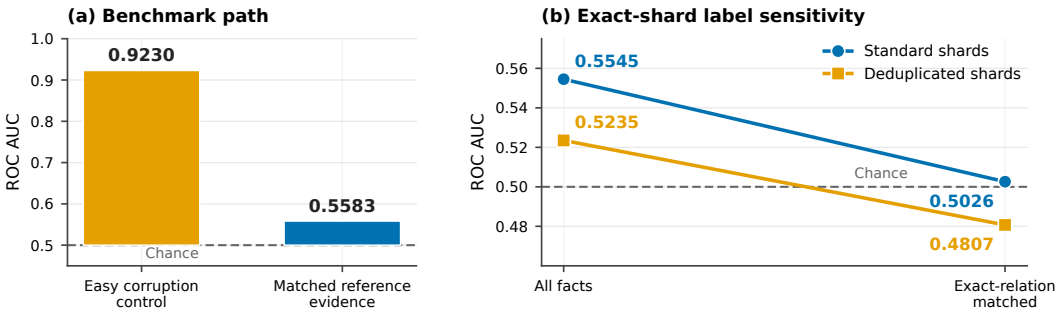


Figure 1: **Aligned labels contract the apparent signal.** (a) Six-model pooled ROC AUC for the easy object-corruption control ($n = 6{,}078$) and frequency-matched reference evidence (878 pairs; 1,756 facts), averaged over five entity-grouped splits: 0.9230 versus 0.5583. (b) Pooled 60-feature ROC AUC for exact token-window shard labels on all 1,756 facts versus exact-relation/covariate-matched subsets (274/292 pairs): standard shards change from 0.5545 to 0.5026, and deduplicated shards from 0.5235 to 0.4807. Dashed lines mark Chance; panel (b) uses a near-chance y-range. These descriptive AUCs do not estimate training-set membership.

Table 2: **The naive mixed-pool evaluation manufactures a strong apparent signal (AUC 0.8490) that its audit-grade operating point does not support.** Member vs. all negatives; entity-grouped 80/20 split; LR. The probe+frequency AUC carries an entity-cluster bootstrap 95% CI of $[0.8303, 0.8697]$, and its TPR@1%FPR a CI of $[0.0828, 0.2764]$. Even here, TPR@1%FPR never exceeds 0.12.

Setting	AUC-ROC	AUPRC	TPR@1%FPR
Frequency-only	0.5539	0.5390	0.0281
Probe-only	0.8418	0.8133	0.0963
Probe + frequency	**0.8490**	**0.8309**	**0.1123**

AUC of each other at both scales (Appendix D), even though only the non-deduped corpus contains duplicated documents.

4.3 Matched evidence yields modest discrimination

The six-model pooled probe features reach a mean AUC of 0.5583 over five entity-grouped splits on the 878 matched pairs. Single-model results range from 0.5236 for Pythia-410M to 0.5676 for GPT-Neo-2.7B. The Pythia-2.8B deduped and original models both yield 0.5508; the corresponding 6.9B models yield 0.5581 and 0.5440. Thus, adding a second model family does not reproduce the high discrimination observed against corrupted objects.

GPT-Neo continuation NLL yields AUC 0.5572, compared with 0.5490 and 0.5488 for the 20% and 40% low-probability token summaries. These are modest associations with reference-snapshot labels. They do not establish either successful training-membership inference or its impossibility. Table 3 retains the five-Pythia comparison and reports the six-model extension separately; its mixed-pool and matched columns use different evaluation protocols.

4.4 Diagnosis: a knowledge-plausibility artifact

Where does the mixed-pool AUC come from? Table 3 isolates the culprit: restricting the negatives to object-corrupted counterfactuals gives AUC 0.9230 for the six-model pooled features ($n = 6{,}078$ paired facts). Corrupted negatives keep the subject and relation but replace the object, making the statement *false but fluent-adjacent*: at both 2.8B and 6.9B, the model assigns visibly more probability to the true object than to a corrupted one (the paired member-vs-corrupted gap in p_{correct} is ≈ 0.52, essentially identical in deduped and non-deduped twins; Section 4.5). A classifier trained on the mixed pool therefore can exploit knowledge plausibility rather than "was the evidence in the corpus?". The white-box columns of Table 3 show the same mechanism quantitatively: NLL drops

Table 3: **Matched reference-evidence discrimination is modest.** "Mixed pool" = all 7,878 facts (LR settings: entity-grouped split; white-box settings: ranking AUC over the full set). "Honest slice" = 1,756 frequency-matched facts (LR: mean over five overlapping entity-grouped splits; white-box: ranking AUC). The member-vs-corrupted slice (pooled probe features, n=6,078) yields 0.9230 in the six-model extension, as discussed in Section 4.4.

Method	Mixed pool	Honest slice
Frequency-only LR	0.5539	0.5198
Probe LR (pooled, 5 Pythia models)	0.8418	0.5448
Probe LR (pooled, 6 models)	–	0.5583
GPT-Neo continuation NLL	–	0.5572
GPT-Neo low-probability tokens, 20%	–	0.5490
GPT-Neo low-probability tokens, 40%	–	0.5488
NLL, Pythia-2.8B-deduped (white-box)	0.7751	0.5636
NLL, Pythia-6.9B-deduped (white-box)	0.7918	0.5657
Min-K% k=20%, Pythia-2.8B-deduped (white-box)	0.6238	0.5401
Min-K% k=20%, Pythia-6.9B-deduped (white-box)	0.6494	0.5478

Table 4: Whole-shard lexical scan over the 1,756 matched benchmark facts. Counts are token-variant co-occurrences and are not interpreted as training membership.

Corpus	Facts with hits	Window hits	Sequence hits	Total subject occurrences
Pythia standard	1,468	96,539	76,568	2,317,819
Pythia deduped	1,451	89,878	70,808	2,594,705

from ≈ 0.78–0.79 on the mixed pool to ≈ 0.56 on the honest slice, because NLL is also dominated by plausibility. Feature importance corroborates this reading: on the mixed pool the top features by permutation importance are cross-model agreement statistics ($\mathrm{p_mean}$, AUC drop 0.0394; margin 0.0279; $\mathrm{scale_slope}$ 0.0218; $\mathrm{p_max}$ 0.0211; $\mathrm{hard_easy_gap}$ 0.0166), while the frequency covariates that would expose popularity leakage contribute nothing ($\mathrm{entity_freq}$ 0.0001; $\mathrm{answer_freq}$, $\mathrm{answer_len}$ 0.0). In other words, the classifier's evidence is exactly the signal that cannot distinguish exposure from mere knowledge.

4.5 Deduplication twins show similar plausibility gaps

The twin comparison measures response differences between original and deduped checkpoints. In the original full-pool analysis, $\Delta p = p_{\mathrm{orig}} - p_{\mathrm{dedup}}$ has reference-label AUCs 0.4942 and 0.4957 at 2.8B and 6.9B (Table 6). The one-sided rank tests do not detect larger differences for reference-evidence positives.

The paired true-versus-corrupted comparison contains 3,039 pairs per scale. Its mean answer-probability gap is 0.5198 for deduped versus 0.5190 for original 2.8B, and 0.5400 versus 0.5392 at 6.9B. The corresponding twin AUCs are 0.4994 and 0.4960. These results show similar plausibility gaps across checkpoint variants. Because Wikipedia reference labels do not measure how deduplication changes each fact's actual training count, the comparison cannot identify the causal effect of factual exposure.

Exact token-window scan of the Pythia shards. As an orthogonal corpus check, we scanned every fixed-length sequence in the released Pythia standard and deduplicated token shards with the tokenizer used by the checkpoints. The scan searched surface and lower-case subject/object token variants within a 128-token window, reset at sequence boundaries, and recorded both window hits and sequence-level hits. This is a finite lexical co-occurrence audit, not a semantic or membership oracle. Of the 1,756 benchmark facts, 1,468 had at least one sequence-level hit in standard Pythia and 1,451 in deduped Pythia; the per-fact window counts correlate at $r = 0.9788$, while 1,337 facts change count between variants.

We next align these lexical labels with the probe scores themselves. Across all five checkpoints, AUC for predicting exact sequence-level presence is only 0.501–0.518 (Table 8 in Appendix B),

Table 5: Fixed-feature classifiers. Reference and "all" entries are mean AUCs over five overlapping entity-grouped splits on all 1,756 facts. "Matched" entries are pooled out-of-fold AUCs on 274/292 relation- and covariate-matched standard/deduplicated pairs. Single models use 10 features; pooled uses 60.

Features	Reference	Std. all	Std. matched	Dedup. all	Dedup. matched
Pythia-410M dedup.	0.5236	0.5170	0.5199	0.5090	0.5271
Pythia-2.8B dedup.	0.5508	0.5157	0.4093	0.5176	0.4559
Pythia-2.8B original	0.5508	0.5309	0.4450	0.5221	0.4593
Pythia-6.9B dedup.	0.5581	0.5232	0.4840	0.5111	0.4921
Pythia-6.9B original	0.5440	0.5039	0.5222	0.5013	0.5028
GPT-Neo-2.7B	0.5676	0.4923	0.4805	0.5005	0.4791
Pooled six models	0.5583	0.5545	0.5026	0.5235	0.4807

Table 6: **Checkpoint differences weakly discriminate reference-evidence labels.** $\Delta p = p_{\text{original}} - p_{\text{deduped}}$ per fact; members should show larger Δp if duplication-driven exposure moved probe responses. One-sided MWU tests the alternative "members have larger Δp" ($n = 7{,}878$ per pair).

Twin pair	Mean Δp (mem)	Mean Δp (non-mem)	Δp AUC	MWU p
Pythia-2.8B (deduped / original)	-0.00002	$+0.00098$	0.4942	0.814
Pythia-6.9B (deduped / original)	-0.00351	-0.00308	0.4957	0.744

despite the much larger mixed-pool AUCs above. This comparison is descriptive: the exact-label analysis uses a scalar score, whereas the mixed-pool classifier combines features. The lexical labels are limited to the token variants and fixed sequence boundaries defined by our scanner. Entity-cluster bootstrap intervals for these ten estimates all include 0.5 (overall endpoints $[0.466, 0.553]$). None excludes chance, but these intervals do not establish equivalence to chance or the absence of an exposure signal in other features.

To separate feature choice from label choice, we also repeat the six-model classifier on the same 1,756 facts and identical five entity-grouped splits for all three labels (Table 5). The ten features per model, zero imputation, training-only standardization, and default logistic regression match the reference-label analysis; no hyperparameters are selected using these results. The pooled classifier gives AUC 0.5545 for standard-shard presence and 0.5235 for deduplicated-shard presence, compared with 0.5583 for reference labels. Thus the scalar near-chance result does not extend uniformly to a learned feature combination. These exploratory split means are descriptive, not independent replicates or evidence of equivalence. Label prevalence also differs (0.500, 0.836, and 0.826), and lexical presence is not semantic exposure.

The "matched" columns are a predeclared sensitivity analysis for the exact shard labels. After excluding repeated entities, Hungarian matching pairs every minority example within relation on log entity frequency, log answer frequency, and entity/answer length, yielding 274 standard and 292 deduplicated pairs. Five pair-grouped folds are mutually exclusive, so every fact is tested once. The pooled 95% pair-bootstrap intervals are $[0.4565, 0.5495]$ and $[0.4383, 0.5232]$. No feature set has an interval whose lower endpoint exceeds 0.5; two 2.8B configurations generalize in the opposite direction. Maximum absolute post-match SMD is 0.181/0.202, so residual measured imbalance remains. The result strengthens the observed sensitivity to cohort construction but is not an equivalence test or evidence that other detectors must fail.

5 Discussion and Limitations

Negative construction changes the inference problem. Object corruptions preserve the subject but change the answer. A model can distinguish them through factual knowledge without identifying where supporting evidence occurred. The high corrupted-negative AUC therefore does not by itself validate corpus auditing. Frequency matching makes the comparison more demanding, although it does not hold every aspect of plausibility constant.

Reference evidence and training membership. The audited Wikipedia 2023-11-01 snapshot is not the Pile training corpus. Our co-occurrence labels are reproducible reference-corpus outcomes. They do not establish semantic entailment, exact training inclusion, or absence of paraphrased evidence. Similarly, a near-chance twin contrast without fact-specific exposure counts does not prove absence of a duplication effect. The operational contribution is an auditable comparison of probe behavior under different reference-evidence controls.

Limitations. (1) *Corpus and granularity.* FEB audits one language, one domain (Wikipedia), and subject–relation–object facts with single-word objects; multi-token answers, long-tail relations, and other corpora may behave differently. (2) *Snapshot alignment.* Labels are defined with respect to the 2023-11-01 Wikipedia snapshot, while the Pile used by Pythia (Gao et al., 2020) contains an earlier Wikipedia snapshot. Any resulting label noise is shared by both twins. Because deduplication also changes global training dynamics, however, the twin test is a falsification check rather than a clean per-fact exposure intervention. The honest-slice AUC must be read as *detectable-under-our-reference-labels* signal, not as a direct estimate of Pythia training-set membership or as a precise zero. (3) *Models and access.* Results cover base models from 410M to 6.9B with log-probability access; instruction-tuned models, larger scales, and pure-generation (no log-prob) adversaries are untested, and we do not claim impossibility in general; it measures modest discrimination of reference-corpus labels by the tested statistics. (4) *Attack strength.* We evaluate probing-style features, NLL, and Min-K%; reference-model attacks requiring shadow retraining (e.g., LiRA (Carlini et al., 2022)) are infeasible at pretraining scale and were not run. The evaluated models include Pythia and GPT-Neo, both trained on the Pile family of corpora. Replication on models trained on a different documented corpus, and evaluation of multi-token answers, remain open (Section 4).

6 Related Work

Probing memorization and privacy leakage. ProPILE (Kim et al., 2023) probes PII leakage by prompting language models; Huang et al. (2022) study email recovery using email-context and owner-name prompts. Neither evaluates our claim-level corpus-evidence task. Verbatim memorization and extraction are well established (Carlini et al., 2021; 2023); deduplication reduces extraction risk (Lee et al., 2022; Kandpal et al., 2022). We instead audit fact-level evidence and use deduplicated/non-deduplicated checkpoints as a falsification control.

Membership inference on language models. Classical MIA targets classifiers (Shokri et al., 2017; Carlini et al., 2022); LM methods include Min-K% Prob, MIMIR, Tag&Tab, and temporal-split evaluation (Shi et al., 2024; Duan et al., 2024; Antebi et al., 2025; Yi & Li, 2026). These operate at document or passage granularity. FEB instead labels individual facts by exhaustive corpus scanning, a span-addressable analogue of corpus search (Elazar et al., 2024).

Factual knowledge in language models. LAMA-TREx popularized factual cloze probes (Petroni et al., 2019); subsequent work locates and edits factual associations (Meng et al., 2022; 2023; Dai et al., 2022; Geva et al., 2021). Recall also tracks popularity (Mallen et al., 2023). FEB tests the distinct question of whether a fact has auditable reference-corpus evidence.

7 Conclusion

FEB makes fact-level reference-corpus evidence auditable. Pooled probes reach AUC 0.5583 on frequency-matched labels but 0.9230 against object corruptions, showing that negative construction can manufacture apparent signal. On matched exact-shard labels, standard/deduplicated AUC is $0.5026/0.4807$. These results support matched corpus audits, but neither prove equivalence to chance nor rule out other membership detectors.

Reproducibility Statement

All experiments use fixed seed 2027; model checkpoints, the Wikipedia snapshot version, scan commands, and filtering rules are recorded in the anonymous supplementary configuration, and every

evidence label is recomputable from the snapshot by a single scan command. All probe scoring uses vLLM in bf16 on a single GPU, so the evaluation pipeline is reproducible without distributed training or shadow retraining. Entity-clustered bootstrap and permutation statistics are specified in Section 3. The exact-shard sensitivity artifact records input/code hashes, pair IDs, mutually exclusive test folds, balance statistics, and 5,000 pair-bootstrap draws. Failed and null results (e.g., relations with TPR@1%FPR $= 0$) are reported in Appendix C. The anonymous supplementary material contains the scan/probe code and compact audited result artifacts; the complete benchmark, including evidence spans, will be released upon publication.

Ethics Statement

This work is defensive: it provides auditors and researchers with a benchmark and two falsification diagnostics for evaluating exposure-detection claims, and it documents that current methods do not support individual-level "the model was trained on this fact" conclusions. Because our central result is that fact-level detection currently fails, we flag two misreadings. First, a null result is not a privacy guarantee: the failure of probes and MIA statistics to detect fact-level exposure does not establish that models never memorize training data. Verbatim memorization is well documented at sufficient exposure counts (Carlini et al., 2023), and the tools released here must not be used to certify that any specific document was or was not in a proprietary training corpus, nor to conclude that training-data privacy risks are absent. Second, the benchmark contains only Wikidata facts and short verbatim spans from a publicly released Wikipedia snapshot; the full-corpus Aho–Corasick scan was run exclusively over that public snapshot and involves no personal or non-public data beyond public encyclopedic content.

AI Use Statement

We used AI assistants for experiment design, code implementation, result analysis, and translation. The authors take responsibility for the final content.

References

Alfred V. Aho and Margaret J. Corasick. Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975. doi: 10.1145/360825.360855.

Sagiv Antebi, Edan Habler, Asaf Shabtai, and Yuval Elovici. Tag&Tab: Pretraining data detection in large language models using keyword-based membership inference attack. *arXiv preprint arXiv:2501.08454*, 2025. URL https://arxiv.org/abs/2501.08454.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Katherine Lee, Daphne Ippolito, Florian Tramèr, Matthew Jagielski, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah Smith, and Jesse Dodge. What's in my big data? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1f7336fd66b6e6e63d1801fdd5930a5a-Abstract-Conference.html.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, 2022. doi: 10.18653/v1/2022.findings-emnlp.148.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10697–10707, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-0911.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022. doi: 10.18653/v1/2022.acl-long.577.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.52202/068431-1262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *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. 2463–2473, 2019. doi: 10.18653/v1/D19-1250.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e32ad85fa27be4a9868d55703f01323e-Abstract-Conference.html.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Jiatong Yi and Yanyang Li. Membership inference on LLMs in the wild. *arXiv preprint arXiv:2601.11314*, 2026. URL https://arxiv.org/abs/2601.11314.

A Matching diagnostics

Table 7 reports standardized mean differences (SMD) between members and negatives for the corpus-derived covariates. Matching anchors on the (smaller) true-negative pool and optimizes $\left|\log(1+10^6\,\mathrm{freq}(s))\right|$ distance first, then lengths; the residual log-domain entity-frequency SMD on the honest slice is $+0.137$, with raw-frequency SMD $+0.006$. Corrupted negatives share their subject with a member, so their entity-frequency match is perfect by construction (residual differences stem from the corrupted object's own frequency, SMD $+0.049$).

Table 7: **Matching leaves negligible residual covariate imbalance.** Standardized mean differences (members minus negatives). TN = true negatives; CN = corrupted negatives; all quantities are corpus-derived from the same scan that produced the labels. All objects in the benchmark are single surface words, so $\mathrm{answer_len}$ SMD is 0 by construction.

Covariate	Honest slice (M vs. TN)	Mixed pool (M vs. TN)	Mixed pool (M vs. CN)
$\mathrm{entity_freq}$	$+0.006$	-0.048	-0.040
$\mathrm{answer_freq}$	-0.044	-0.005	$+0.049$
$\mathrm{answer_len}$	0.000	0.000	0.000
$\mathrm{entity_len}$	-0.177	$+0.096$	$+0.093$

B Exact-shard probe alignment

For completeness, Table 8 reports the per-checkpoint probe-score AUCs for exact sequence-level token-window presence in the released Pythia shards. Positive counts are 1,468 (standard) and 1,451 (deduped) out of 1,756 facts; these lexical labels are not interpreted as training membership.

Table 8: Probe-score AUC for exact sequence-level token-window presence in the released Pythia shards. Positive counts are 1,468 (standard) and 1,451 (deduped) out of 1,756 facts.

Checkpoint	Standard	Deduped
Pythia-410M	0.5121	0.5178
Pythia-2.8B	0.5006	0.5127
Pythia-6.9B	0.5044	0.5115
Pythia-2.8B-deduped	0.5021	0.5147
Pythia-6.9B-deduped	0.5064	0.5160

C Per-relation full tables

Table 9 reports mixed-pool per-relation metrics for the combined probe+frequency classifier (entity-grouped test split; relations with $\geq$ 20 test rows and both classes present). Note that high per-relation AUC co-occurs with very low TPR@1%FPR for many relations, and five relations have TPR@1%FPR $= 0$, consistent with the plausibility-artifact diagnosis of Section 4.4.

Table 9: **High per-relation mixed-pool AUC coexists with very low TPR@1%FPR.** Per-relation mixed-pool results for the probe + frequency classifier (entity-grouped test split).

Rel.	AUC	AUPRC	TPR@1%FPR	Rel.	AUC	AUPRC	TPR@1%FPR
P1001	0.9883	0.9901	0.9375	P27	0.8681	0.8197	0.0000
P101	0.9269	0.9199	0.3158	P276	0.9199	0.9371	0.4762
P103	0.9451	0.9502	0.6429	P279	0.9337	0.9457	0.8095
P106	0.6508	0.6314	0.0455	P30	0.9079	0.8834	0.1500
P108	0.9706	0.9762	0.9286	P31	0.9231	0.9131	0.5385
P127	0.8594	0.8689	0.3750	P36	0.9140	0.8861	0.4615
P1303	0.7769	0.7326	0.0000	P361	0.9375	0.9421	0.5000
P131	0.9428	0.9335	0.5455	P364	0.6667	0.6416	0.1667
P136	0.9321	0.9142	0.2222	P37	0.9389	0.9076	0.3333
P138	0.9150	0.8662	0.0455	P39	0.7335	0.6794	0.0000
P140	0.9414	0.9128	0.3500	P407	0.9077	0.9143	0.5833
P1412	0.8239	0.8143	0.0417	P413	0.6818	0.7180	0.2273
P159	0.8954	0.8585	0.1765	P449	0.6992	0.6667	0.0000
P17	0.7126	0.5724	0.0000	P463	0.9306	0.9457	0.7500
P176	0.9238	0.9251	0.5000	P47	0.9591	0.9600	0.6667
P178	0.9625	0.9679	0.8000	P495	0.9229	0.9061	0.2174
P19	0.7530	0.7228	0.0909	P527	0.9024	0.8738	0.2000
P190	0.8846	0.8894	0.6154	P740	0.7636	0.7358	0.0500
P20	0.9128	0.9166	0.5185	P937	0.7626	0.7992	0.2759
P264	0.8566	0.8883	0.4375				

D PER-MODEL MIXED-POOL RESULTS

Table 10 reports the mixed-pool AUC of probe features for each checkpoint separately. The deduped and non-deduped twins differ by approximately 0.007 AUC at 2.8B and 0.004 at 6.9B, showing similar performance in the original mixed-pool analysis (Section 4.5).

Table 10: **Mixed-pool AUC is nearly identical across deduplication twins at both scales.** Per-model mixed-pool AUC (probe features, same entity-grouped protocol as Table 2). Deduped and non-deduped twins differ by approximately 0.007 and 0.004 at 2.8B and 6.9B, respectively, although only the non-deduped corpus contains duplicated documents (Section 4.5).

Model	Mixed-pool AUC
Pythia-410M (deduped)	0.6429
Pythia-2.8B	0.8416
Pythia-2.8B (deduped)	0.8488
Pythia-6.9B	0.8381
Pythia-6.9B (deduped)	0.8416

E PROBES AND TEMPLATES

Each fact stores: (i) the primary MCQ prompt (relation template + "`\Answer:`"), four options prefixed with a space; (ii) a true/false prompt over the declarative statement with candidate continuations " True"/" False"; (iii) up to five paraphrase-template MCQs (one or two per relation in our instantiation); and (iv) a hard-distractor MCQ whose distractors share the gold object's word count or first token. A worked example (relation P131, "located in"), verbatim after normalization:

> Subject *Pennsylvania Route 696*, object *Pennsylvania*; evidence span (first supporting document): "pennsylvania route 696 (pa 696) is a state highway located in franklin and cumberland counties in pennsylvania . . . ";
> MCQ: "Pennsylvania Route 696 is located in `\Answer:`" with options {*Pennsylvania*, + 3 same-relation distractors};
> TF: "Pennsylvania Route 696 is located in Pennsylvania. Is this true?";
> Hard MCQ: distractors drawn from same-relation objects matching "Pennsylvania" on word count or first token.

All probe scores are scalar summaries of option log-probabilities computed with vLLM in bf16; no full logit tensors are stored. Relation templates and verbs are fixed per relation (e.g., P19 “was born in”, P108 “worked for”, P131 “is located in”), and the same rendered prompts are shared by all five checkpoints so that cross-model features are computed on identical inputs.