EXPOSURE EMBEDDINGS FOR CORPUS AUDITING AND DOWNSTREAM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Author-conditioned title likelihood offers a behavioral signal for corpus auditing, but its interpretation depends on document identity and the construction of comparison papers. We study this signal using Comma v0.1 and the arXiv component of Common Pile. Starting from metadata-matched paper pairs, an independent identifier audit retains 1,222 pairs whose positive paper is present and negative paper is absent from that component. On these pairs, Comma's title score achieves AUC 0.6523 with matched distractors and 0.6139 with lexically similar distractors, compared with 0.5316 and 0.5154 for GPT-Neo-2.7B. These associations concern component presence, not verified membership in the complete training mixture. Replacing the correct author list by a within-year, one-to-one derangement reduces Comma's AUC by 0.0369 and 0.0337 on the two distractor conditions; the paired intervals exclude zero, while GPT-Neo changes by less than 0.0006. Comma's exploratory within-year AUCs range from 0.642 to 0.663 in 2020–2023, compared with 0.455–0.569 for GPT-Neo. In a separate citation-prediction cohort, holding out query papers gives SPECTER2 AUC 0.90756 and fusion AUC 0.90770; the paired 90% interval for the difference is $[-0.00095, 0.00119]$. The results distinguish behavioral discrimination of an audited corpus component from incremental utility in one downstream task.

1 INTRODUCTION

A model may assign high likelihood to a paper title because it recognizes an author–title association, because the title is plausible, or because similar metadata appeared during training. Behavioral corpus auditing must distinguish these explanations. Membership inference often uses target-text likelihood (Shokri et al., 2017; Carlini et al., 2022; Shi et al., 2024; Duan et al., 2024); an author-conditioned probe instead tests a relationship between metadata fields, but still needs an independently defined corpus outcome.

Common Pile and its Comma models (Kandpal et al., 2025) permit a scan of the arXiv component. OpenAlex (Priem et al., 2022) supplies paper identifiers and matching covariates. This combination exposes a distinction between title occurrence and document identity: a reference list can mention a paper absent from the component, and a heading mismatch can conceal a paper whose identifier is present. We therefore audit identifiers separately from title matching.

We match paper pairs on year, field, and citation-count bin, reducing differences in title length and author count within these strata. A metadata-only identifier audit retains 1,222 complete pairs. On these pairs, Comma v0.1 reaches title-score AUC 0.6523 with matched distractors and 0.6139 with harder distractors, compared with 0.5316 and 0.5154 for GPT-Neo-2.7B. This is a model comparison on fixed questions, not a causal estimate of training exposure.

We also separate corpus-label discrimination from downstream utility. In a distinct citation cohort, a query-held-out evaluation yields text-only AUC 0.90756 and fusion AUC 0.90770. Our contribution is an evidence-linked evaluation in which document identity defines the audited outcome, matched comparisons constrain difficulty, and an independent task tests incremental predictive value. The identifier filtering is post hoc and uses no model scores.

We keep three claims separate throughout: (i) whether a model ranks a title associated with an audited component document, (ii) whether the ranking depends on the correct author–title relation

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

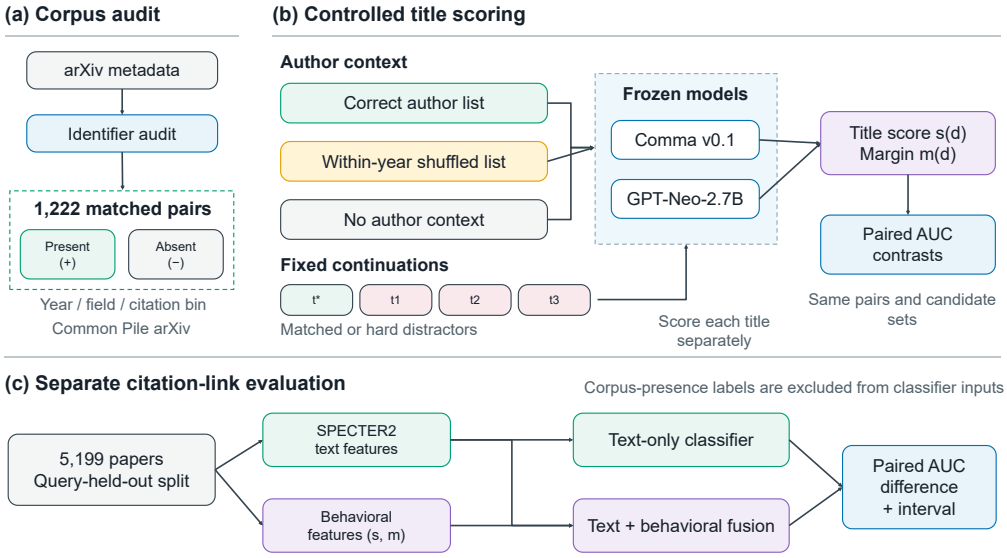


Figure 1: Corpus auditing and downstream evaluation use separate cohorts. (a) An identifier audit retains 1,222 present–absent pairs from the Common Pile arXiv component. (b) Author controls share the same pairs and candidate titles: the true title t^* and three distractors. Each continuation is scored separately by each frozen model. (c) A separate 5,199-paper cohort compares text-only and fused features under a query-held-out split. Component-presence labels are excluded from the classifier inputs.

rather than title plausibility alone, and (iii) whether the resulting behavioral features add predictive value to citation links. Evidence for one claim is not reused as evidence for the others. Figure 1 makes the cohort boundary explicit.

2 Data and Exposure Labels

2.1 Sampling and Metadata Matching

We sample 10,000 OpenAlex works associated with the arXiv source `S4306400194`, published in 2018–2023, using seed 2027. Of these, 9,138 have an arXiv identifier. Heading-based screening initially identifies 1,576 positive candidates.

We form 1,482 positive–negative pairs within exact year, field, and citation-bin strata. Citation bins are 0, 1–2, 3–9, and at least 10. Within each stratum, Hungarian matching minimizes a standardized title-length difference plus 0.5 times standardized author-count difference and 0.25 times standardized log-citation difference. Negative candidates have no title occurrence in the scanned document bodies. All four model–distractor combinations score the same 2,964-paper matched pool.

For matching, title length is the count of alphanumeric words, author count is obtained from the comma-separated author list, and citation count is transformed by $\log(1 + c)$. Each feature is divided by its pooled standard deviation within the stratum before computing the weighted absolute differences. Standardization puts the matching terms on comparable scales; the weights then determine their relative contribution to the assignment. Papers are sorted by identifier before assignment to make the input ordering deterministic. These covariates are used to construct the evaluation population, not as inputs to the language-model probe.

2.2 Independent Identifier Audit

Scanning 22 compressed Common Pile arXiv-papers shards yields 304,048 distinct normalized arXiv identifiers. Among the 1,482 heading-negative papers, 104 have identifiers present in the

component. Among heading positives, one known identifier is absent and three disagree with the matched heading document. Other papers have unresolved identifiers.

We retain a pair only if its positive identifier is present and agrees with the heading document, while its negative identifier is known and absent. This metadata-only rule retains 1,222 pairs, or 2,444 papers. We preserve the originally measured questions and distractors rather than regenerating them after filtering. Present and absent refer to this component, not to the complete training mixture.

The heading screen and identifier audit answer different questions. A heading screen associates a bibliographic title with text in a corpus document; the identifier audit tests whether the associated document is the intended paper. The 104 identifier-present negatives show why lack of a title occurrence is insufficient for assigning absence. Conversely, requiring a positive identifier to agree with the heading document rejects title matches that refer to a different work. Unknown identifiers are excluded rather than treated as negative evidence. The retained labels therefore depend on observable document identity, not on whether either language model answers correctly.

Whole-pair retention also preserves the comparison established by matching. If one member fails the identifier rule, keeping its partner alone would change the balance of the matched covariates. Excluding both instead defines a restricted, balanced population on which the original scores can be compared. Because candidate titles remain fixed, the change in population is separated from a change in question difficulty. The audited estimate describes identifier-resolved pairs selected from this sampling frame; it does not reweight them to the complete arXiv collection.

The retained pair counts are 14, 18, 60, 203, 421, and 506 for 2018 through 2023, respectively (28, 36, 120, 406, 842, and 1,012 papers). Every year has equal positive and negative counts because filtering retains whole pairs.

3 Exposure Probes

3.1 Author-conditioned scoring

The prompt is:

```
Which paper has authors {authors(d)}?
Answer:
```

Four continuations are scored separately: the true title and three distractor titles, each preceded by a space. Candidates are not listed together in the prompt. Let $\ell(t)$ be the sum of token log-probabilities for a continuation. The score and margin are

$$s(d) = \frac{\exp \ell(t_d^*)}{\sum_{i=1}^{4} \exp \ell(t_i)}, \qquad m(d) = \ell(t_d^*) - \max_{i \neq *} \ell(t_i). \tag{1}$$

Their pair $\mathbf{e}_d = (s(d), m(d))$ is the behavioral vector. These are summed, not length-normalized, likelihoods. Equal candidate scores yield $s(d) = 0.25$, but this is not a calibrated membership probability or an expected mean for absent papers.

Writing the author prompt as a_d and a tokenized title as $t = (t_1, \ldots, t_L)$ makes the scoring unit explicit:

$$\ell(t \mid a_d) = \sum_{j=1}^{L} \log p_\theta(t_j \mid a_d, t_{<j}). \tag{2}$$

The prompt supplies context; the sum covers continuation tokens under teacher forcing. Prompt and continuation tokenization omit special tokens, and the implementation evaluates log-softmax after casting logits to float32. A longer title introduces more terms, and each title's tokenization depends on the model tokenizer. Matching title lengths constrains this source of variation without making the likelihood length-invariant. We compare the models by their AUCs on shared questions rather than assuming their raw log-likelihoods have the same scale.

The two coordinates summarize different aspects of the same four-way comparison. The margin measures separation from the strongest distractor, whereas the normalized score also depends on the other two distractors. For a fixed true-title likelihood and strongest distractor, raising either

Table 1: Component-presence discrimination on 1,222 identifier-audited pairs. Intervals are matched-pair bootstrap 95% intervals.

Model	Distractors	AUC	95% interval
Comma v0.1	Matched	0.6523	$[0.6291, 0.6753]$
GPT-Neo-2.7B	Matched	0.5316	$[0.5062, 0.5554]$
Comma v0.1	Harder	0.6139	$[0.5939, 0.6332]$
GPT-Neo-2.7B	Harder	0.5154	$[0.4959, 0.5343]$

remaining distractor likelihood lowers $s(d)$ without changing $m(d)$. The vector therefore retains both the nearest alternative and the total competing likelihood. It is a behavioral representation derived from model outputs, not a learned document encoder or an internal hidden-state embedding.

Each title is evaluated as a continuation of the same author prompt. This construction tests conditional title likelihood without exposing alternative titles together as in an instruction-following multiple-choice question. The denominator in Eq. 1 normalizes over the four selected continuations, not over all possible titles. Changing the distractor set can consequently change the score even when the true title and its conditional likelihood stay fixed. Metadata-matched and harder distractors test this dependence while holding the target documents constant.

3.2 Distractors and models

Metadata-matched distractors prioritize the same field and year. If this pool is too small, eligibility broadens to the same field within one year and then to the full pool. Ranking incorporates venue, year, title length, author count, and citation count. Harder distractors use TF–IDF unigram and bigram similarity to prioritize lexically similar eligible titles, with metadata matching as fallback. Comma v0.1 and GPT-Neo-2.7B receive identical candidate sets. They differ in architecture, scale, and training mixture, so their contrast is descriptive rather than an exposure intervention.

3.3 Evaluation units and paired uncertainty

The corpus-label outcome is defined per paper, while matching defines the resampling unit. For n audited pairs, let s_i^+ and s_i^- denote the scores of the present and absent papers. The reported pooled AUC is

$$\widehat{\mathrm{AUC}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\mathbf{1}\{s_i^+ > s_j^-\} + \frac{1}{2} \mathbf{1}\{s_i^+ = s_j^-\} \right]. \tag{3}$$

It compares all present scores with all absent scores, including papers from different matched pairs. This differs from the fraction of pairs in which the present paper outscores its own matched negative. Matching constrains the population; it does not restrict the AUC comparisons to the original partners.

Bootstrap sampling draws pair identifiers with replacement and carries both rows of every selected pair into the resample. We recompute pooled AUC on these rows, then use the empirical 2.5th and 97.5th percentiles over 2,000 draws. Model and prompt-condition differences use shared draws, so a difficult pair affects both systems in the same resample. This paired construction preserves the covariance relevant to a difference in AUC. It estimates uncertainty over the audited pairs with their questions fixed, rather than variation from training new language models or reconstructing the corpus sample.

3.4 Discrimination of component presence

Table 1 reports title-score AUC on the identifier-audited pairs. Intervals use 2,000 matched-pair bootstrap resamples with seed 2027; model differences share the same resamples. Harder distractors reduce Comma's discrimination, demonstrating sensitivity to the candidate comparison. GPT-Neo has a modest matched association, while its harder condition's interval includes chance.

Comma-minus-GPT-Neo AUC differences are 0.1207 (95% interval $[0.1027, 0.1398]$) with matched distractors and 0.0985 ($[0.0844, 0.1124]$) with harder distractors. At 1% false-positive rate,

Comma's true-positive rates are 0.0622 and 0.0409. Thus aggregate ranking discrimination coexists with low sensitivity at a stringent false-positive operating point.

Within-pair comparisons provide a second view of discrimination. Comma's present paper outscores its matched negative in a tie-adjusted fraction of 0.6408 of pairs under either distractor condition. The corresponding 95% pair-bootstrap intervals are $[0.6129, 0.6669]$ for matched distractors and $[0.6137, 0.6678]$ for harder distractors. These estimates compare each paper with its original metadata-matched partner, whereas pooled AUC also compares scores across pairs. The identical rounded win fractions do not mean that the same individual pairs are won in both conditions. Together with the different pooled AUCs, they show why pairwise ordering and population-wide score separation should be distinguished.

The low-false-positive result addresses a different use of the ranking than aggregate AUC. AUC summarizes the ordering of present and absent papers across thresholds; an audit that flags individual papers may instead operate in a narrow part of the ROC curve. At the reported 1% false-positive rate, Comma retrieves approximately 6.2% or 4.1% of component-present papers, depending on distractors. These operating points describe sparse retrieval despite the above-chance aggregate AUC. They cannot be converted into the fraction of correct flags without specifying the prevalence of component presence in the population being audited.

3.5 Author-cue control

To test whether the signal depends on the author–title relation rather than title plausibility alone, we rescored the identical identifier-audited pairs under two controls. The first removes the author list. The second replaces it by the positive-row author list from another pair in the same publication year, using a seed-2027 one-to-one derangement with no fixed points. Both rows of a pair receive the same shuffled authors; titles, labels, choices, pair identifiers, and row order remain fixed.

Figure 2 visualizes all three prompt conditions, and Table 2 reports the stronger shuffled-author comparison exactly. Relative to shuffled authors, correct authors add 0.0369 AUC with matched distractors and 0.0337 with harder distractors; both paired intervals exclude zero. GPT-Neo changes by less than 0.0006, with intervals spanning zero. The shuffled-author AUC is also below the title-only AUC for Comma by 0.0197 ($[-0.0266, -0.0126]$) and 0.0157 ($[-0.0211, -0.0102]$), suggesting that an incorrect author cue actively suppresses the corpus association. By comparison, simply removing authors changes Comma AUC by 0.0171 and 0.0180; GPT-Neo again shows no corresponding change. These results isolate correct metadata matching more directly than author removal, but remain behavioral associations rather than membership proofs.

Author removal and author shuffling are complementary comparisons. Removal changes the information supplied to the prompt, while shuffling keeps an author-list cue but breaks its alignment with the target paper. The one-to-one construction uses each donor list once within its year, and applying the same donor to both rows avoids assigning a different shuffling rule to positives and negatives. It preserves the donor-list multiset at the pair level, while the pairing between authors and titles changes. The resulting AUC contrast measures the effect of that alignment on label discrimination, rather than the probability of retrieving a particular author's paper.

For Comma, correct authors outperform both controls, but the controls are not interchangeable: shuffled authors perform below author removal. Thus the correct-minus-shuffled gap combines information provided by the correct cue with the effect of supplying an incorrect cue. Interpreting the entire gap as an isolated amount of memorized knowledge would ignore this difference. Reporting both controls identifies whether removing information and replacing it with conflicting information produce the same behavior; here they do not.

4 Matching and Identifier Sensitivity

Exact year matching makes positive and negative year distributions identical. A deterministic year-only score therefore has AUC 0.5 on these pairs. Matching also fixes field and citation bin, while distractor construction reduces gross differences between candidate titles.

Identifier filtering addresses a separate error: heading occurrence need not agree with document identity. We apply a whole-pair inclusion rule using metadata alone. The scored questions remain

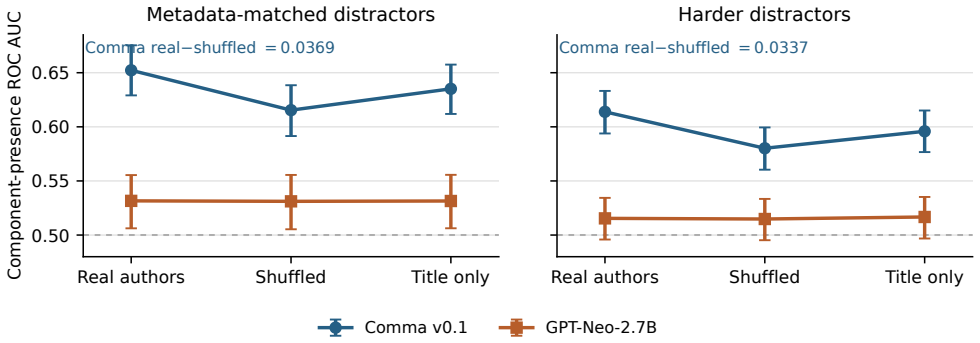


Figure 2: Correct author–title associations, rather than author tokens alone, drive Comma's author-cue effect. ROC AUC is measured on the same 1,222 identifier-audited pairs; error bars are 95% paired-pair bootstrap intervals. Within-year one-to-one author shuffling reduces Comma AUC in both distractor settings, while GPT-Neo remains unchanged within uncertainty.

Table 2: Within-year shuffled-author control on the same 1,222 audited pairs. "Delta" is real minus shuffled AUC; intervals use 2,000 paired-pair bootstrap resamples.

Model	Set	Real	Shuf.	Δ [95% interval]
Comma	Matched	0.6523	0.6154	0.0369 [0.0275, 0.0463]
Comma	Harder	0.6139	0.5802	0.0337 [0.0264, 0.0406]
GPT-Neo	Matched	0.5316	0.5312	0.0004 [−0.0038, 0.0052]
GPT-Neo	Harder	0.5154	0.5148	0.0006 [−0.0027, 0.0041]

unchanged. The resulting AUC is a post-hoc sensitivity estimate conditional on identifier agreement, not a new experiment with regenerated candidates.

As a robustness check, we also report AUC by publication year on the same 1,222 identifier-audited pairs. The 2020–2023 strata contain 60, 203, 421, and 506 pairs, respectively. Comma AUCs are 0.642, 0.663, 0.653, and 0.653; pointwise 95% intervals from 2,000 pair-bootstrap resamples are $[0.544, 0.739]$, $[0.605, 0.720]$, $[0.612, 0.692]$, and $[0.619, 0.686]$. GPT-Neo AUCs are 0.455, 0.569, 0.523, and 0.540; its 2021 and 2023 pointwise intervals also exclude 0.5, so above-chance association is not unique to Comma. No multiplicity adjustment is applied to these exploratory strata. The 2018–2019 strata contain only 14 and 18 pairs; both models' intervals include 0.5. These are within-year associations, not tests of transfer to held-out years.

These controls remove marginal year imbalance and detected identity disagreements. They do not remove all within-year linguistic differences, nor identify which training component caused the behavior.

The year-only result follows from the paired design, not from fitting a weak predictive baseline. Each retained year contributes exactly as many present as absent papers, so any deterministic function of year has identical empirical score distributions in the two classes. The within-year analysis then asks whether the model signal remains after fixing that covariate. Its 2020–2023 estimates for Comma are similar in magnitude, although precision differs with stratum size. Pooling these strata provides more observations, but does not turn them into independent replications across corpora or model-training runs.

5 Downstream Link Prediction

A score that discriminates corpus labels need not improve a task defined by semantic relationships. We test whether the two-dimensional behavioral vector adds information beyond scholarly text embeddings in citation-link prediction. This experiment uses a separate 5,199-paper OpenAlex cohort and its previously measured behavioral vectors, rather than the 1,222 identifier-audited pairs. Corpus-presence labels are not inputs to the citation classifier.

Table 3: Citation-link prediction with held-out query papers. The primary test set contains 192 queries and 3,956 pairs. Intervals are query-cluster bootstrap 95% intervals. Fusion combines behavioral and text pair features.

Pair features	AUC	95% interval
Behavioral	0.55495	$[0.52547, 0.58206]$
SPECTER2	0.90756	$[0.89100, 0.92327]$
Fusion	0.90770	$[0.89118, 0.92351]$

5.1 Task setup

Candidate papers were published through 2020, and query papers in 2021 or later. Among 956 eligible queries, the fixed evaluation pool contains 20,000 query–candidate pairs: 5,000 observed references and 15,000 uniformly sampled uncited candidates. All compared representations are available for both endpoints. An uncited pair is a sampled negative for this classification task, not evidence of semantic unrelatedness.

We hold out query papers using an 80/20 grouped split. The primary split (seed 2027) contains 764 training queries and 192 test queries, with 3,956 test pairs. No query contributes pairs to both partitions. Candidate papers can recur across partitions, so the evaluation measures generalization to held-out queries within a shared candidate pool. We repeat the grouped split 20 times using seeds 2027–2046, keeping the sampled pairs fixed.

5.2 Representations and classifier

The behavioral vector is $\mathbf{e}_d = (s(d), m(d))$, measured with Comma v0.1 using the earlier uniform-distractor protocol. The text representation uses SPECTER2 (Singh et al., 2023). For either representation, a pair is encoded by cosine similarity, Euclidean distance, and dot product. Fusion concatenates these two sets of pair features into six dimensions; it does not concatenate document embeddings before computing similarities. Each classifier is a balanced logistic regression, with standardization fit on training pairs only and a maximum of 3,000 iterations.

The fusion comparison holds the prediction problem fixed. It changes the available pair features while retaining the same sampled links, query partition, and classifier family. Cosine similarity describes directional agreement, Euclidean distance describes separation, and the dot product also reflects vector magnitude. Applying the same feature map to both representations gives the classifier an explicit opportunity to exploit the behavioral coordinates alongside the text representation. The test is whether these additional features improve held-out ranking, not whether the two representations encode identical information.

The split unit matters because multiple labeled pairs share a query. With a random split of pairs, links from one query could occur in both training and test. Grouping by query instead evaluates new query papers against the existing candidate pool. The temporal restriction establishes which papers can serve as candidates before this split is made. Candidate reuse is part of the task definition, so the evaluation is not a test of generalization to a wholly unseen citation graph.

5.3 Results: little incremental value for held-out queries

Table 3 gives the primary split. Behavioral features yield AUC 0.55495, while SPECTER2 pair features yield 0.90756. Fusion changes AUC by $+0.00013$. Query-cluster bootstrap intervals preserve dependence among links from the same held-out paper. The paired fusion-minus-text 90% interval is $[-0.00095, 0.00119]$, and the 95% interval is $[-0.00117, 0.00137]$.

Across 20 grouped splits, the mean AUC difference is $+0.00049$, with a central 90% range of $[-0.00026, 0.00131]$. This range describes split sensitivity, not a population confidence interval. The primary paired 90% interval lies within a post-hoc tolerance of ± 0.01 AUC, supporting little incremental value at that tolerance. It does not establish absence of training contamination or lack of utility on other tasks.

The paired interval is computed for the difference in AUC, not inferred from whether the two marginal intervals overlap. Both systems score the same held-out links. A cluster draw carries all of a query's links into both evaluations, preserving the shared sample and the dependence among links. The narrow difference interval can therefore coexist with wider marginal intervals for SPECTER2 and fusion: uncertainty shared by the two evaluations need not translate into uncertainty of the same magnitude in their contrast.

The primary bootstrap and repeated splits quantify different sources of variation. The bootstrap conditions on the fitted classifiers and their held-out queries, while the repeated grouped splits change which queries are used for fitting and evaluation. Neither redraws the original 20,000-pair pool. The small mean gain across splits therefore complements the primary paired interval as a sensitivity analysis on this fixed pool. The post-hoc ± 0.01 tolerance gives a scale for interpreting the gain; it is not a preregistered equivalence margin or evidence of exact equality.

6 Discussion and Limitations

Component identity and training exposure. The identifier audit distinguishes papers present in the scanned arXiv component from papers whose known identifiers are absent there. A paper can still occur in another component, in a different version, or as reproduced text. Consequently, the evaluated outcome is component presence, not complete training-mixture membership. Title matches in reference lists cannot by themselves establish document identity.

What matching controls. Pairing fixes publication year, field, and citation-count bin while reducing differences in title length and author count. Lexically similar distractors make title selection harder. Neither procedure eliminates every source of title likelihood: author prominence, subfield vocabulary, title priors, and tokenization may still contribute. The score sums continuation log-probabilities, so title length remains scientifically relevant even after matching. The within-year derangement preserves the marginal author distribution but not all author-level covariates within each matched pair; it tests correct author–title alignment rather than eliminating every metadata confound.

Model comparison. Comma v0.1 and GPT-Neo-2.7B differ in architecture, parameter count, and training mixture. Their paired score contrast measures a model difference on fixed questions; it is not a controlled causal effect of exposure. The modest GPT-Neo matched AUC also counsels against describing the comparison model as a chance-level null.

Scope of downstream evidence. The citation experiment measures a two-dimensional behavioral signature on a separate cohort with uniform distractors. Its query-grouped split removes query overlap between training and test, but candidates remain shared and negatives are uniformly sampled uncited papers. The near-zero fusion gain bounds this representation's incremental predictive value under that protocol. It cannot certify that semantic embeddings or time-split evaluations are uncontaminated.

7 Related Work

Membership inference and pretraining-data detection. Classical membership inference attacks train shadow models to distinguish members from non-members (Shokri et al., 2017; Carlini et al., 2022); for language models, the dominant signals are functions of the likelihood of the target text itself: extraction and memorization analyses (Carlini et al., 2021; 2023), Min-K% statistics (Shi et al., 2024), deduplication-aware risk (Lee et al., 2022; Kandpal et al., 2022), person-level probes (Kim et al., 2023), keyword-based variants (Antebi et al., 2025), and corpus-level contamination searches (Elazar et al., 2024), with recent evaluations finding such attacks fragile on modern LLMs (Duan et al., 2024); all return verdicts about individual strings or corpora. Our work differs in its focus on author-conditioned title comparisons, metadata-matched paper pairs, and identifier-based component labels. We evaluate downstream utility separately from corpus-label discrimination.

Known corpora, scholarly embeddings, and time splits. The Pile with GPT-Neo (Gao et al., 2020; Black et al., 2021), Pythia (Biderman et al., 2023), OLMo (Team OLMo et al., 2025), and the

Common Pile with its Comma models (Kandpal et al., 2025) release pretraining data together with trained models. We use this transparency asymmetrically: the Common Pile's arXiv component defines our labels, Comma v0.1 is the audited model, and GPT-Neo-2.7B is the control. Citation-informed and general embedders (Cohan et al., 2020; Singh et al., 2023; Ostendorff et al., 2022; Chen et al., 2024; Zhang et al., 2025) are standard for scholarly tasks, and citation-link prediction (Liben-Nowell & Kleinberg, 2007; Hu et al., 2020) increasingly uses temporal splits. We use a temporal candidate/query division and hold out query papers to measure the incremental value of behavioral pair features.

8 Conclusion

Author-conditioned title scores discriminate presence in an audited arXiv component after metadata matching and a post-hoc identifier audit. Comma v0.1 reaches AUC 0.6523 with matched distractors and 0.6139 with harder distractors, exceeding GPT-Neo on the same paper pairs. Correct-to-shuffled author gaps of 0.0369 and 0.0337 show that this association depends partly on the matched author–title relation. Together, these results link title-score discrimination to the correctness of author metadata within the audited component. They do not establish complete training membership. In a separate query-held-out citation experiment, adding the behavioral vector changes SPECTER2 AUC by 0.00013. Corpus-label discrimination and downstream utility are distinct properties, and evaluating both requires explicit document identities and split units.

Reproducibility Statement

The anonymous supplementary material contains sampling, matching, identifier-audit, probing, and evaluation code with compact audited outputs; complete per-paper outputs will be released upon publication. The identifier sensitivity analysis retains complete matched pairs using metadata alone and preserves the originally scored distractors. The citation experiment records its sampled pairs and query-grouped partitions separately from the component-auditing cohort. The primary seed is 2027; citation split sensitivity uses seeds 2027–2046. Identifier filtering and the downstream equivalence tolerance are post-hoc analyses. The shuffled-author control uses a frozen within-year derangement and reports all four model–distractor cells.

Ethics Statement

The study uses publicly released corpora, open-weight models, and public bibliographic metadata. A behavioral corpus-association score requires interpretation against the stated component scope and error rates, rather than as a verdict on an individual document's training membership. The citation result is a predictive comparison, not a privacy guarantee.

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

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.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.5281/zenodo.5297715, 2021.

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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://arxiv.org/abs/2402.03216v1. arXiv preprint, version 1.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, 2020. doi: 10.18653/v1/2020.acl-main.207.

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.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

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.

Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, John Kirchenbauer, Tom Goldstein, Brian Bartoldson, Bhavya Kailkhura, and Tyler Murray. The Common Pile v0.1:

An 8TB dataset of public domain and openly licensed text. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1918. Main Conference, Datasets and Benchmarks Track.

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.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. doi: 10.1002/asi.20591.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11670–11688, 2022. doi: 10.18653/v1/2022.emnlp-main.802.

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. URL https://arxiv.org/abs/2205.01833.

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.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, 2023. doi: 10.18653/v1/2023.emnlp-main.338.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. doi: 10.48550/arXiv.2501.00656.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

A Sampling and Label Audit Details

The component-auditing cohort originates from a 10,000-work arXiv-source sample and is reduced to 1,482 metadata-matched pairs, then 1,222 identifier-consistent pairs. The citation analysis uses a separate earlier 5,199-paper cohort with available vectors; its original title-occurrence labels are not used as training-membership truth.

B Matched Evaluation Details

Metadata distractor ranking weights venue mismatch by 1.5, year gap by 2, relative title-length gap by 1, relative author-count gap by 0.5, and log-citation gap by 0.25. Hard-distractor TF–IDF uses at most 50,000 features, minimum document frequency two, and sublinear term frequencies. It searches the nearest 160 titles before eligibility filtering.

C Resampling Scope

Probe intervals condition on the evaluated pairs and fixed candidate sets; they do not measure sensitivity to new corpus snapshots. Citation intervals resample query clusters because links from the same paper are dependent. These resampling units answer different questions.