AI DISCLOSURES PROJECT

WORKING PAPER SERIES

Beyond Public Access in LLM Pre-Training Data

Non-public book content in OpenAI's Models

Sruly Rosenblat
Program Associate
Social Science Research Council

Tim O'Reilly
Program Director
Social Science Research Council

Ilan Strauss
Program Director
Social Science Research Council

100 YEARS | SOCIAL SCIENCE RESEARCH COUNCIL

WORKING PAPER NO. 1
2025/04

About The AI Disclosures Project

Led by technologist Tim O'Reilly and economist Ilan Strauss, the AI Disclosures Project addresses the potentially harmful societal impacts of AI's unrestrained commercialization. By improving corporate and technological transparency and disclosure mechanisms, it aims to ensure that economic incentives don't compromise safety or equity, and avoid fostering excessive risks. Disclosures are vital for well-functioning markets yet remain lacking in AI. Just as financial disclosure standards fostered robust securities markets, standardized AI disclosures can build trust, expedite adoption, and spur innovation. Through research, collaboration, and policy engagement, the AI Disclosures Project aims to develop a systematic framework for meaningful "Generally Accepted AI Management Principles." The project is generously funded by the Omidyar Network, Alfred P. Sloan Foundation, and Patrick J. McGovern Foundation.

ISSN: 3067-1361
DOI: 10.35650/AIDP.4111.d.2025

This working paper can be referenced as follows:
Rosenblat, Sruly, Tim O'Reilly, and Ilan Strauss. "Beyond Public Access in LLM Pre-Training Data: Non-public book content in OpenAI's Models." SSRC AI Disclosures Project Working Paper Series (SSRC AI WP 2025-04), Social Science Research Council, April 2025.
https://www.ssrc.org/publications/beyond-public-access-in-llm-pre-training-data-non-public-book-content-in-openais-models/

100 YEARS | SOCIAL SCIENCE RESEARCH COUNCIL

Website: AI Disclosures Project
Substack: Asimov's Addendum
X: @AIDisclosures

Beyond Public Access in LLM Pre-Training Data

Non-public book content in OpenAI's Models

Sruly Rosenblat *[1], Tim O'Reilly[1,3], and Ilan Strauss[1,2]

[1]AI Disclosures Project, Social Science Research Council

[2]Institute for Innovation and Public Purpose, University College London

[3]O'Reilly Media

Abstract

Using a legally obtained dataset of 34 copyrighted O'Reilly Media books, we apply the DE-COP membership inference attack method to investigate whether OpenAI's large language models were trained on copyrighted content without consent. Our AUROC scores show that GPT-4o, OpenAI's more recent and capable model, demonstrates strong recognition of paywalled O'Reilly book content (AUROC = 82%), compared to OpenAI's earlier model GPT-3.5 Turbo. In contrast, GPT-3.5 Turbo shows greater relative recognition of publicly accessible O'Reilly book samples. GPT-4o Mini, as a much smaller model, shows no knowledge of public or non-public O'Reilly Media content when tested (AUROC $\approx$ 50%). Testing multiple models, with the same cutoff date, helps us account for potential language shifts over time that might bias our findings. These results highlight the urgent need for increased corporate transparency regarding pre-training data sources as a means to develop formal licensing frameworks for AI content training.

Keywords: Membership Inference Attacks, Large Language Models, Copyright Issues, Data Access Violations, Pre-Training Data, Architecture of Participation.

*Varying Contributions. Sruly Rosenblat: Compute, statistical analysis and AUROC method, appendix, graphs, and tables. Ilan Strauss: Paper write-up, structure, core findings, policy discussion. Tim O'Reilly: Topic conceptualization and research design (public vs. non-public data). Isobel Moure: Policy discussion section.

We gratefully acknowledge funding support from the Omidyar Network, Alfred P. Sloan Foundation, McGovern Foundation, and the O'Reilly Foundation, without which this work would not have been possible. We also extend our full appreciation to Andrew Odewahn for compiling the O'Reilly dataset and to Anshuman Suri and André Duarte for their helpful feedback on our initial draft. Thank you to Isobel Moure for edits. All mistakes are solely our own. Corresponding author is: yrosenblat@ssrc.org. The code for this paper can be found at: https://github.com/AI-Disclosures-Project/Detecting-Access-Violations-in-a-LLMs-Pre-Training-Data.

Contents

1 Introduction: Identifying access violations

Large Language Models (LLMs) require incredible amounts of public and non-public data to learn human language (called the 'pre-training' stage). Yet the origins and legal status of this pre-training data remains largely undisclosed by the corporations who gather and use it (OpenAI 2023a; Anthropic 2023). Several high-profile legal proceedings indicate that major AI companies may train on non-public, often illegally obtained, content (New York Times 2023; Roth 2024; Belanger 2025). In response, AI companies are calling for model pre-training to be exempt from copyright obligations (OpenAI 2025; Whitwam 2025). If adopted, copyright holders and content creators may be unable to sustain themselves and their creations, with profound implications for the survival of the internet's traffic-driven business model (Blaszczyk et al. 2024; Knibbs 2025; Durantaye 2025).

This paper examines whether non-publicly accessible (non-public) copyrighted O'Reilly Media books were included in the training datasets of OpenAI's GPT series of models. Each O'Reilly book contains both publicly accessible, free-to-use preview content, and non-public, effectively paywalled content. This allows us to see whether OpenAI primarily trained its models on publicly available data or if it circumvented paywall restrictions and used non-public data (Figure 1).

We employ the DE-COP membership inference attack method (Shokri et al. 2017; Duarte et al. 2024) to test whether a model can reliably differentiate between human-authored (O'Reilly Media) texts from paraphrased LLM versions of the text that we generate. If it can, then the model might have prior knowledge of the text from its training (Duarte et al. 2024). By systematically probing a model's knowledge of texts published before and after its training cutoff dates, we can estimate the probability of particular O'Reilly Media book extracts having been included in a model's training data.

We test OpenAI's GPT-3.5 Turbo, GPT-4o Mini, and GPT-4o models across 13,962 paragraphs from 34 O'Reilly books for access violations, distinguishing between public and non-public extracted from the same books. On the basis of AUROC scores calculated for each of the 34 O'Reilly books, where 50% reflects a random chance of being trained on, **we find that**:

1. ***The role of non-public data in OpenAI's model pre-training data has increased significantly over time***. OpenAI's more recent and capable GPT-4o model shows strong recognition of paywalled O'Reilly book content (82% AUROC score), while OpenAI's GPT-3.5 Turbo, trained around two years prior, does not (AUROC score just above 50%).

2. ***GPT-4o exhibits far stronger recognition of non-public O'Reilly book content compared to publicly accessible samples***, with AUROC scores of 82% (non-public) vs 64% (public). We would expect the opposite, since public data is more easily accessible and repeated across the internet. This highlights the value-add of paywalled high-quality data to a model's training.

3. ***Earlier OpenAI models may have been more selective in their training data, using predominantly publicly available content***. In contrast to GPT-4o, GPT-3.5 Turbo shows greater relative recognition of publicly accessible O'Reilly book samples than non-public ones, with 64% (public) vs. 54% (non-public) AUROC scores.

4. ***Smaller models are harder to test accurately***. We find that GPT-4o Mini, with the same training cutoff as GPT-4o, was not trained on non-public O'Reilly data, and shows similarly low recognition of public book data. This may reflect its inability, as a smaller model, to remember text compared to GPT-4o, a much larger (by parameter count) model (Meeus et al. 2024).

Such access violations might have occurred via the LibGen database, as all of the O'Reilly books tested were found in it. By way of robustness, we show that although newer LLMs have an improved ability to distinguish human-authored from machine-generated language *regardless of whether a particular text was trained on*, this does not reduce our method's ability to classify data as being trained on or not.

Our study design accounts for the potential of time-specific differences in language to bias our results (Duan et al. 2024; Debeshee et al. 2024), which can arise because we split our sample (of potentially trained on and so in-sample, vs. not trained on and so out-of-sample) by date. Such bias can occur if the DE-COP test mistakes language that the model is simply "familiar" with (due to temporal shifts) for content the model was trained on. To ensure that this bias does not drive our findings, we test two models (GPT-4o and GPT-4o Mini) that were both trained on data from the same period. Because these two models show notably different results, time-specific effects are unlikely to be the determining factor.

Our study contributes to research on detecting unauthorized data usage in AI training (Mattern et al. 2023; Shi et al. 2023b; Jingyang Zhang et al. 2024) by using legally sourced non-public copyrighted books, that allows us to detect access violations, such as paywall circumvention. In contrast, earlier studies use mostly public datasets when identifying what models were trained on (Shi et al. 2023a; Duarte et al. 2024; Duan et al. 2024).

Our findings highlight the need for stronger accountability in AI companies model pre-training process. Liability provisions that incentivize improved corporate transparency in disclosing data provenance (O'Reilly 2024; O'Reilly et al. 2025) may be an important step to facilitating commercial markets for training data licensing and remuneration (Thornhill 2025). Membership inference attacks can help pressurize model developers to negotiate such agreements. But by itself is insufficient, especially given its limited efficacy against smaller models, more advanced models, and models with certain post-training features (Satvaty et al. 2024; Jie Zhang et al. 2024; Balaji 2024).

Section 2 outlines our books dataset and DE-COP and AUROC methods. Section 3 presents our findings. Section 4 discusses their policy implications for establishing formal commercial markets for content creator training data. Appendix A contains a more details on our sample and analysis.

2 Data and Methods

This section first details our O'Reilly dataset of 34 books and explains how its division into publicly accessible vs non-public (effectively paywalled) book samples enables us to accurately identify access violations in a model's pre-training. Next, we describe our method, which involves first testing the model's recognition of paragraphs from O'Reilly Media books, using the DE-COP membership inference attack method; before, secondly, testing the validity and meaning of these findings for each of the 34 books using more robust AUROC scores.

2.1 Data: Public vs. non-public book data

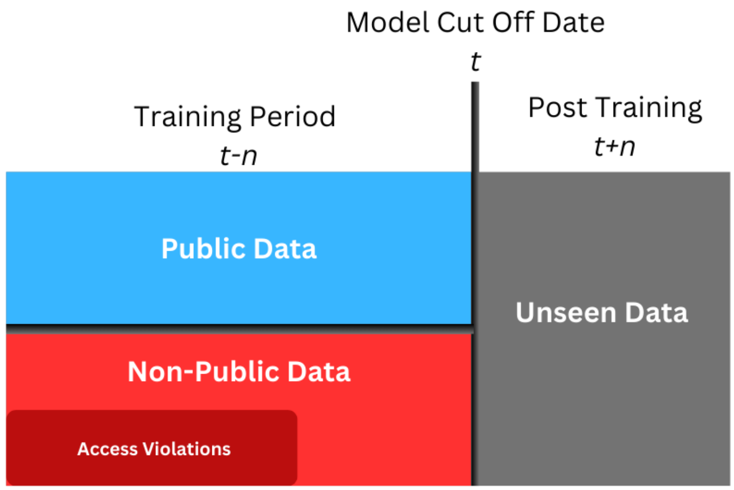
Our dataset contains 34 copyrighted O'Reilly Media books lent to us, that we then split into a total of 13,962 paragraphs. Paragraphs are used to calculate the initial mean DE-COP score, one for each book, from which a single AUROC Score is then calculated across all books for each of OpenAI's models.

The O'Reilly Media books dataset has the unique quality of containing both non-public (behind a paywall) and public (freely available) text within the same book. This allows us to differentiate between instances where a model was trained exclusively on public data and cases where access violations may have occurred. We define public text as any content made available by O'Reilly Media for content previews – specifically the first 1,500 characters of each chapter as well as the entirety of chapters one and four. All other O'Reilly text we define as non-public.

To accurately measure the performance of the DE-COP membership inference attack method (discussed below), paragraph samples must be divided into two distinct categories, that in practice we can only approximate: data known to be included in the model's pre-training dataset and those known to be excluded. In our case we designate books published before the model's training cutoff ($t-n$) as *possibly* in-dataset (previously seen and trained on) samples, and books published after a model's training cutoff ($t+n$) as *known* out-of-dataset samples that the model could not have been trained on (see Figure 1). "Access violations" are defined as the subset of non-public book paragraphs, published during the model's training period, that we identify as being used for training.

We categorize books published before October 2023 (for GPT-4o and GPT-4o Mini), and before September 2021 (for GPT-3.5 Turbo) as *potentially in-dataset* ($t-n$), and those books published after these model training cutoff dates as *out-of-dataset* ($t+n$), where t is the model training cutoff date (October 2023 and September 2021, respectively). This date is defined by the model developer as the last date that the model's pre-training dataset contains data for.

Figure 1. We split our sample of O'Reilly books by time period & accessibility.



Note: Data published prior to a model's training completion ($t-n$) may have been trained on. Data published after a model's training cutoff ($t+n$) is known to not be in the model's training data. Any portion of non-public data found to be included in a model's training would constitute an access violation (bottom left square).

Our method of splitting our sample between potentially-in-dataset ($t-n$) and known out-of-dataset ($t+n$) *by date* may introduce "temporal bias" into our findings (Duan et al. 2024; Debeshee et al. 2024), and in turn provide us with misleadingly high AUROC scores. This occurs when features in the data changes over time, creating distinguishable patterns between training and testing datasets split by time periods. Data then can be separately identified by an LLM based solely on the language varying with time – in our case into potentially-in-dataset ($t-n$) and known out-of-dataset ($t+n$) data – with no actual prior knowledge of the text itself.[1]

To account for this we test two different GPT models (GPT-4o and GPT-4o Mini) that were trained during the same period, and ideally on the same data, such that if our tests show very different AUROC results then time-specific causes are unlikely to be decisive. This study design helps isolate prior model knowledge of the data as the primary driver of our results.[2]

[1]Temporal bias is when the ability to infer membership through DE-COP, or any related method, is confounded by time-dependent changes in the data, rather than by genuine evidence that a particular example was (or was not) in the training set. Similarly, stylistic bias captures biases that arise from shifts in how data "looks" or is distributed (e.g., changes in vocabulary, writing style, or domain). Both these biases can appear if one naively splits data by time period, for instance, using older data for training and newer data for testing, without making any associated adjustments. In other words, during the DE-COP test the model might mistake familiar vs. unfamiliar language, for familiar vs. unfamiliar content they were trained on.

[2]This differences we observe cannot be explained by GPT-4o simply being better at distinguishing human-authored from

We carefully filter the dataset to avoid any ambiguous cases, such as second edition books with potentially minor changes on the previous editions published during the training period, that risk contaminating our "unseen" classification. Additionally, to minimize edge cases where publication dates might overlap with training cutoffs, we excluded books published during a model's cutoff year from tests involving that specific model. Consequently, not every model was tested on precisely the same set of books (see Appendix A.1).

2.2 Methods: DE-COP and AUROC

We first use the **DE-COP** membership inference attack (Duarte et al. 2024) to determine whether a particular data sample was part of a target model's training set. This works by quizzing an LLM with a multiple choice test to see whether it can identify original human-authored O'Reilly book paragraphs from machine-generated paraphrased alternatives that we present it with. If the model frequently correctly identifies the actual (human-generated) book text (for books published during the model's training period) then this likely indicates prior model recognition (training) of that text. The statistic used is the 'guess rate' and the baseline against which this is evaluated is the guess rate on books published after the model finished training, where 'guess rate' is the percentage of correctly identified human written text for all 24 paragraph permutations.[3] We test GPT-3.5 Turbo, GPT-4o Mini, and GPT-4o in this manner and use Claude 3.5 Sonnet[4] to generate the paraphrased, machine versions, of the text.

The second step in our study is to use the DE-COP quiz scores or 'guess rates' generated above, to calculate **AUROC Scores** (Area Under the Receiver Operating Characteristic). This evaluates whether there's a meaningful difference ('separability') between how a model handles content that it was potentially trained on versus content published after its training was

AI-generated text than GPT-4o Mini, as AUROC here measures the difference in knowledge within the same model between books published prior to and after training completion. Even assuming a model has perfect identification capability, if it had not been trained on any of the test samples, we would expect an AUROC score of approximately 50%.

[3]The models are first quizzed on nearly every paragraph in each of the 34 O'Reilly books tested – with each paragraph being used to generate 24 quizzes to eliminate position bias, representing every possible permutation of the four texts being positioned differently (one human-authored text and three machine-generated paraphrases). A 'paragraph' here refers to adjacent sentences of around 120 words, they were not necessarily formatted as paragraphs in the source material.

[4]This differs from the paraphrase model used in Duarte et al. (2024), we chose Claude 3.5 Sonnet as it was the most powerful model from Anthropic at the time of testing.

Table 1. Sample Size & Average Word Count by Model and Data Split

Model	Data-Split	Sample Size (n)	Average Word Count
GPT-4o	Public	1,965	112
	Non-Public	8,997	113
	Potentially In-Dataset	8,985	113
	Out-of-Dataset	1,977	110
GPT-4o Mini	Public	1,968	112
	Non-Public	9,005	113
	Potentially In-Dataset	8,991	113
	Out-of-Dataset	1,982	110
GPT-3.5 Turbo	Public	1,929	113
	Non-Public	6,171	113
	Potentially In-Dataset	2084	114
	Out-of-Dataset	6016	113

Note: Sample sizes (in paragraphs) and average word counts across different data splits for each model. Potentially in-dataset represents data published prior to a model's cutoff date; out-of-dataset represents data published afterward.

completed. AUROC measures a classifier's ability to distinguish between two classes, with scores ranging from 0 to 1, with 0.5 representing random chance and values closer to 1 indicating a strong ability to accurately 'discriminate' (i.e., classify) between the two classes (or categories). In our case, AUROC measures the ability to separate books that may have been trained on (class/ category 1), from books the model could not have seen (class / category 2). A high AUROC score, therefore, implies that the model was trained on many of the books in our dataset published prior to the model's cutoff date. We calculate AUROC scores at both the paragraph and book levels, though our primary finding is at the book level.[5]

3 Findings: Did OpenAI train on copyrighted books?

We present our core findings below, based on robust AUROC scores that take 34 O'Reilly Media books and first calculate DE-COP guess rates for public and non-public book paragraphs. Next,

[5]Our results are similar to Puerto et al. (2024), who finds that aggregating results over larger data units significantly enhances the performance of membership inference attacks, our book level AUROC scores calculated on the mean DE-COP scores for each book were often significantly higher than AUROC done on the paragraph level.

we use the mean DE-COP guess rate for each book, from which AUROC scores are calculated for each LLM (one per model) across all books. We run and test the various LLMs via Python (Google Colab) using OpenAI and Anthropic's batch API processing (Appendix A.3 and A.4).

In what follows an AUROC score of 50% indicates model performance during testing, for book content recognition, of no better than random chance (equivalent to flipping a coin); while test scores approaching 100% suggest near-perfect classification ability (between potentially in-dataset and out-of-dataset samples) – based on the previously estimated DE-COP guess rate.

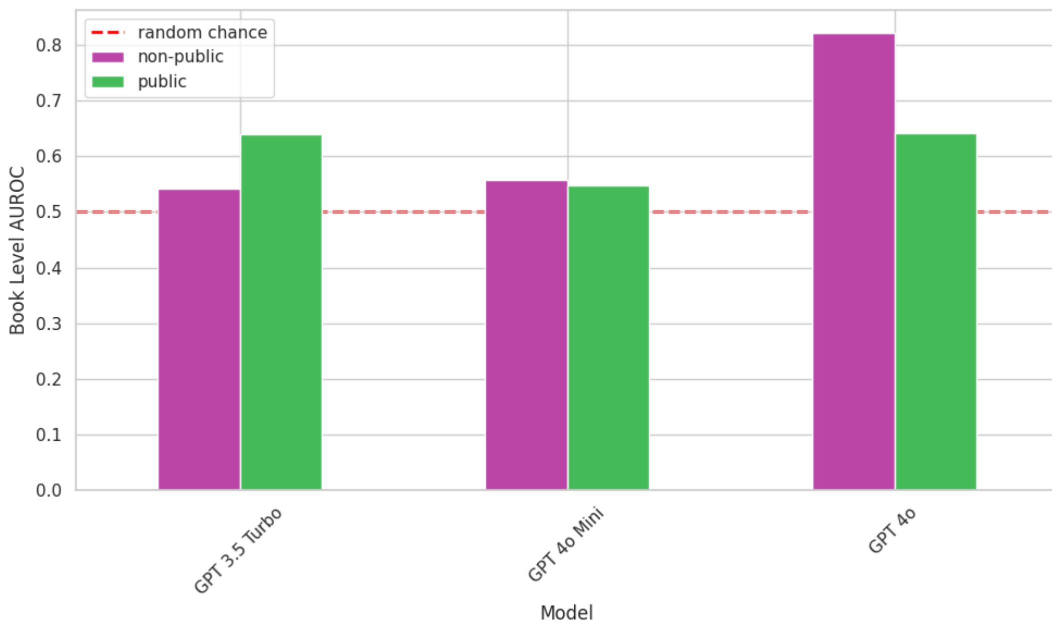
We find that the role of non-public data in OpenAI's model pre-training data has increased significantly over time. Figure 2 shows that OpenAI's more recent and capable GPT-4o model shows strong recognition of paywalled O'Reilly book content (82% book level AUROC score), while OpenAI's GPT-3.5 Turbo, with a training cutoff two years prior in September 2021, does not (AUROC score just above 50%). This indicates a strongly improved model ability to distinguish between non-public books that were potentially included in the training dataset and those published after the model's pre-training cutoff. GPT-4o's 82% AUROC score suggests that the model recognizes, and so has prior knowledge of, many non-public O'Reilly books published prior to its training cutoff date (of September 2023).

Secondly, Figure 2 also shows that ***GPT-4o exhibits far stronger recognition of non-public O'Reilly book content compared to publicly accessible samples***, with AUROC scores of 82% (non-public) vs 64% (public). We would expect the opposite, since public data is more easily accessible and repeated across the internet.[6] This points to the fact that high-quality, frequently paywalled, data is vital for effective model training.

GPT-4o's high familiarity with O'Reilly Media books likely reflects a deliberate effort by OpenAI to train on the O'Reilly book dataset. However, some of this familiarity could have been acquired through more benign means – for example, excerpts from these books may have entered the dataset via user queries.

[6]Another possible explanation for why non-public text is more represented in GPT-4o's training, despite being harder to access, is that our public subset tends to be more formulaic (see Table 3), potentially making it less distinctive and memorable to the model due to its lower perplexity (Meeus et al. 2024).

Figure 2. AUROC Scores Showing Model Recognition of Pre-Training Data



Note: Showing book level AUROC scores (n=34) across models and data splits (see Table 1 for sample sizes). Book level AUROC is calculated by averaging the identification rates of all paragraphs within each book and running AUROC on that.

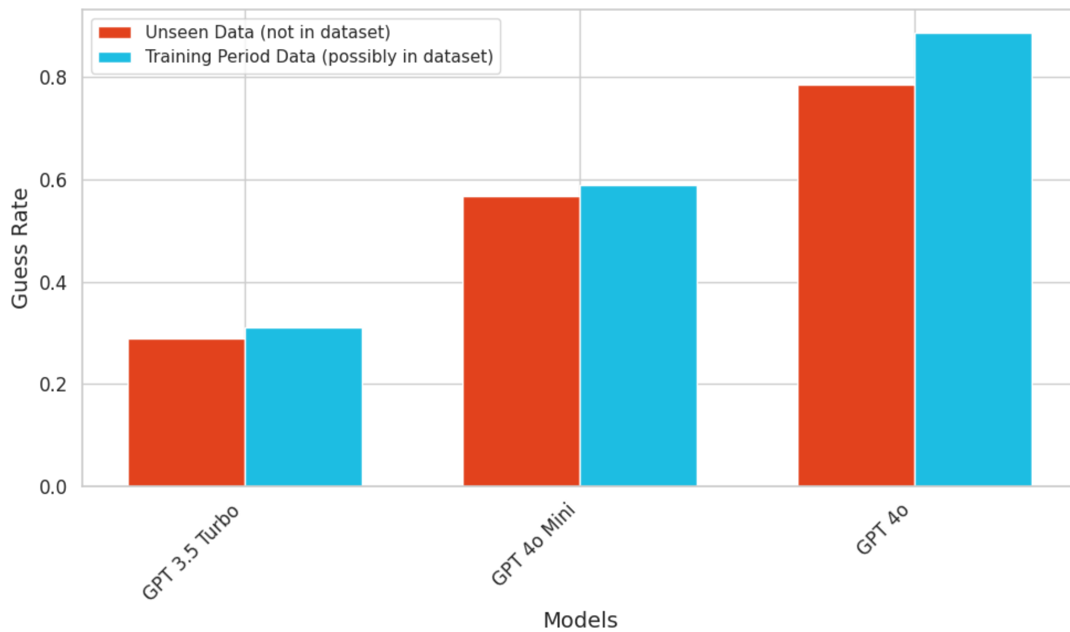
Thirdly, Figure 2 show that ***earlier OpenAI models may have been more selective in their training data, using predominantly publicly available content***. Since in contrast to GPT-4o, GPT-3.5 Turbo, with a training cutoff two years prior, shows greater relative recognition of publicly accessible O'Reilly book samples compared to non-public ones, with a 64% non-public AUROC score vs. 54% for public. This trend seems to align with broader AI industry patterns, whereby the pursuit of larger, more diverse, and higher quality, training datasets may override considerations of access restrictions and copyright implications (New York Times 2023; Belanger 2025).

3.1 Robustness and limitations

One reason for the above findings, and a limitation of our study, may be that *smaller models are harder to test accurately in membership inference attacks*. We find that GPT-4o Mini, with the same training cutoff as 4o, was not trained on non-public O'Reilly data, and shows similarly low recognition of public book data too (Figure 2). GPT-4o Mini recorded AUROC scores of

55% on public data and 56% on non-public data, both near random chance. This may not reflect its inherit knowledge of text, as per its training, but instead GPT-4o Mini's inability, as a smaller model, to remember text compared to 4o, a much larger model by parameter count (Carlini et al. 2022).[7]

Figure 3. DE-COP Guess Rate Improves: More capable models identify human text even when not trained on it.



Note: DE-COP guess rates (i.e., identification rates) pooled across all books for OpenAI models. Red bars represent unseen data published after a model finished training ($t+n$), and blue bars represent data published before the cutoff date ($t-n$) that is suspected to be in the training set. See Table 2 for more.

Second, we note that improving LLM capabilities can make the identification of pre-training data through membership inference attacks more difficult. As per Figure 3, we find that OpenAI's models ability to correctly identify human-authored text, among paraphrased LLM alternatives, improves with model capabilities, *even for texts the model could not been trained on* – published after the model's training cutoff. Figure 3 shows the baseline DE-COP identification rate on books published after the model's training cutoff (unseen books). This increased from 31% for GPT-3.5 Turbo (training finished September 2021), to 57% for GPT-4o

[7]OpenAI does not disclose model sizes but GPT-4o Mini is smaller than GPT-4o and presumably smaller than GPT-3.5 Turbo.

Mini (training finished October 2023), and to 78% for GPT-4o (training finished October 2023).

Once the baseline guess rate ('identification rate') exceeds 96%, the difference between potentially in-dataset and out-of-dataset paragraphs could become undetectable at the paragraph level. For now, however, the gap remains sufficiently large to reliably separate the two categories when calculating AUROC score, particularly when aggregating results at the book level.

A final limitation is that our book level AUROC estimates are uncertain, with high bootstrapped confidence intervals. This is likely due to the small sample size though rather than the estimator's efficacy. Since as the sample size increases certainty increases. For GPT-4o, meaningful paragraph level AUROC scores with tighter confidence intervals arise with the larger estimated (paragraph level) sample size. (Table 5, Appendix A.2).

4 Discussion: Towards functional content AI marketplaces?

Although the evidence presented here on model access violations is specific to OpenAI and O'Reilly Media books, this is likely a systemic issue, and our findings aim to support changes in data collection and usage practices across AI model developers.

Our findings suggest that current AI model development practices may be creating what O'Reilly (2024) describes as an "extractive dead end", creating not just a legal challenge but an existential one for the internet's content ecosystem. The economic implications of uncompensated training data usage extend beyond individual copyright holders to the broader sustainability of professional content creation. If AI companies extract value from a content creator's produced materials without fairly compensating the creator, they risk depleting the very resources upon which their AI systems depend (ibid.). This dynamic creates a tragedy of the commons.[8] If left unaddressed, uncompensated training data could lead to a downward spiral in the internet's content quality and diversity. As revenue streams for professional content cre-

[8] As Longpre et al. (2024) notes: "in less than a year, $\sim$ 5% of the tokens in C4 and other major corpora have recently become restricted by robots.txt. And nearly 45% of these tokens now carry some form of restrictions from the domain's Terms of Service."

ation diminish, fewer resources will be dedicated to producing the high-quality, accurate, and diverse human content that AI systems rely on for training – and inference.

Our key finding, that OpenAI trained their more advanced GPT-4o model on non-public data, is only preliminary based on a small sample of books and subject to the above methodological caveats. Membership inference attacks of a model's outputs are not a substitute for detailed model cards that disclose and disaggregate the sources of model training data (Mitchell et al. 2019; Gebru et al. 2021). However, requiring smaller companies to sift through their pre-training dataset and individually identify the sources for each of their training inputs is unrealistic without tools and standards designed for this purpose.

Common Corpus (Langlais et al. 2024), a large pre-vetted training dataset, is one way around this issue. By centralizing the data cleaning process and providing verifiable pre-training data as a common public good, datasets like Common Corpus could enable smaller firms to train models on non-proprietary data, and easily facilitate disclosure (ibid.). Specialized data auditing companies are already arising but limited in what they can achieve without specific standards (O'Reilly et al. 2025).

Disclosure requirements may come into force in Europe, which could trigger broader requirements and standards in global AI markets. The European Union (EU) AI Act requires developers of general purpose models to "draw up and make publicly available a sufficiently detailed summary of the content used for training" (European Union 2024). This section of the act will not come into full force until 2026, and it is unclear what a "sufficiently detailed summary" will entail. Meanwhile, model developers in the United States are lobbying for model pre-training to be exempt from copyright obligations and for the US government to shield US firms from EU regulation (OpenAI 2025). However, this disclosure requirement from the EU AI Act could still help trigger a positive disclosure-standards cycle if properly specified and enforced.

Ensuring that IP holders know when their work has been used in model training represents a crucial first step toward establishing AI markets for content creator data. Technical methods for this are still in their infancy (Grosse et al. 2023; Zhao et al. 2024). But when applied to specific types of content, such as music, these methods seem to achieve better results, with at least one new music platform already apparently being able to attribute AI generated music outputs to specific music training inputs (Paine 2025).

Despite evidence that AI companies most likely obtained data illegally for their model training (Belanger 2025), this has not stopped a relatively sizable market from emerging in which AI model developers do pay for content, including through licensing deals (Paul et al. 2024).[9] Intermediaries have arisen to facilitate the purchasing of training data by AI model developers, obtain consent from the data providers, strip out personally identifiable information (PII), and ultimately split earnings with the content providers. *Defined.ai*, for example, licenses data to a range of companies including Google, Meta, Apple, Amazon and Microsoft, selling photos for $1 all the way to $300 per hour for longer films (ibid.). Yet AI model developers are still spending most of their money on paying for labeled data for the model's fine-tuning stage – a type of data that is not readily scraped from the internet. Providers of such data generate substantial profits. Scale.AI, a leading provider, was valued at $13.8 billion last year (Sawers 2024).

Licensing and payments to content creators would likely be far more common if liability assignment on data usage existed across the AI value chain and internet. The impact of this could be substantial, since, as OpenAI noted to the House of Lords Communications and Digital Select Committee inquiry: "it would be impossible to train today's leading AI models without using copyrighted materials" (OpenAI 2023b). Music company licensing deals with AI companies – still under discussion – may be illustrative of what is possible when greater liability assignment in the AI value chain exists for data usage – given that data owners are fairly concentrated and have bargaining power (Castelvecchi 2024). In late 2024, *Musical AI* and *Beatoven.ai* began building what is described as "the music industry's first fully licensed, rightsholder-compensating, generative AI platform trained on copyrighted music and other audio", with compensation for content creators' training data based on software that attributes

[9] For example: "A Shutterstock competitor, Freepik, told Reuters it had struck agreements with two large tech companies to license the majority of its archive of 200 million images at 2 to 4 cents per image... OpenAI, an early Shutterstock customer, has also signed licensing agreements with at least four news organizations, including The Associated Press, opens new tab and Axel Springer".

generations to training data (Paine 2025).

Even absent clear liability assignment, model training continues on data that companies have not provided their express permission to train on. Strong evidence exists that AI companies violate robots.txt guidelines and train on content from websites regardless (Longpre et al. 2024). In response, *Cloudflare* has created a new product called *AI Labyrinth* to protect websites from unauthorized access by AI bots. *Miso.ai* has similar research and products underway.

5 Conclusion

Using 34 proprietary O'Reilly Media books lent to us, this study provides unique empirical evidence that OpenAI's GPT-4o was likely trained on non-public, copyrighted content. By employing the DE-COP membership inference attack, we found that GPT-4o achieved a high book level AUROC score on non-public content (82%) – even higher than on publicly available O'Reilly media book content (64%), indicating likely prior recognition of this content from pre-training.

Moreover, the role of non-public data in OpenAI's model pre-training data has likely increased significantly over time. OpenAI's more recent and capable GPT-4o model shows strong recognition of paywalled O'Reilly book content (82% AUROC score), while OpenAI's GPT-3.5 Turbo, trained two years prior, does not (AUROC score just above 50%). Earlier OpenAI models may have been more selective in their training data, using predominantly publicly available content. In contrast to GPT-4o, GPT-3.5 Turbo, with a training cutoff around two years prior, shows greater relative recognition of publicly accessible O'Reilly book samples than non-public ones, with 64% (non-public) vs. 54% (public) AUROC scores.

Although the evidence presented here is specific to OpenAI and O'Reilly Media books, this is likely a systemic issue, and our findings aim to provoke changes in data collection and usage practices cross AI model developers. Meta allegedly trained their models on LibGen – a massive corpus of pirated books (Meta 2024; Belanger 2025). Anthropic allegedly used 'the pile' dataset for its training, which also contains many pirated books (Roth 2024). Given the need for high-quality paywalled data to ensure AI models are smart and kept up to date, training on such data will be a necessity for the foreseeable future (OpenAI 2023b). This means that structured markets for such data still have time, and a need, to arise.

If left unaddressed, the current disregard for IP rights could ultimately harm AI developers themselves, even if its use is ruled legally permissible. Sustainable ecosystems need to be designed so that both creators and developers can benefit from generative AI. Otherwise, model developers are likely to rapidly plateau in their progress, especially as newer content becomes produced less and less by humans. Liability regimes may be the big push required to help form viable marketplaces for various types of model training – and inference – content.

References

Anthropic (Nov. 2023). *Claude 3 Model Card*. Tech. rep. Accessed: 2024-12-04. Anthropic. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Balaji, Suchir (Oct. 2024). *When Does Generative AI Qualify for Fair Use?* Accessed: 2024-12-04. URL: https://suchir.net/fair_use.html.

Belanger, Ashley (Feb. 12, 2025). ""Torrenting from a corporate laptop doesn't feel right": Meta emails unsealed". *Ars Technica*. URL: https://arstechnica.com/tech-policy/2025/02/meta-torrented-over-81-7tb-of-pirated-books-to-train-ai-authors-say/.

Blaszczyk, Matt, Geoffrey McGovern, and Karlyn D. Stanley (Nov. 2024). *Artificial Intelligence Impacts on Copyright Law*. Accessed: 2024-12-04. URL: https://www.rand.org/pubs/perspectives/PEA3243-1.html.

Carlini, Nicholas et al. (2022). "Quantifying memorization across neural language models". *The Eleventh International Conference on Learning Representations*.

Castelvecchi, Davide (2024). "AI Companies Are Finally Being Forced to Cough Up for Training Data". *MIT Technology Review*. Accessed: 2025-03-26. URL: https://www.technologyreview.com/2024/07/02/1094508/ai-companies-are-finally-being-forced-to-cough-up-for-training-data/.

Debeshee, Das, Zhang Jie, and Tramèr Florian (June 2024). "Blind Baselines Beat Membership Inference Attacks for Foundation Models". *arXiv*. Accessed: 2024-12-04. URL: https://arxiv.org/abs/2406.16201.

Duan, Michael et al. (Feb. 2024). "Do Membership Inference Attacks Work on Large Language Models?" *arXiv preprint arXiv:2402.07841*. Accessed: 2024-12-04. URL: https://arxiv.org/pdf/2402.07841.

Duarte, André V et al. (2024). "De-cop: Detecting copyrighted content in language models training data". *arXiv preprint arXiv:2402.09910*.

Durantaye, Katharina de la (2025). "Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI". *IIC-International Review of Intellectual Property and Competition Law*, pp. 1–34.

European Union (June 2024). *Artificial Intelligence Act, Recital 107*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

Gebru, Timnit et al. (2021). "Datasheets for Datasets. Documentation to facilitate communication between dataset creators and consumers". *Communications of the ACM* 64 (12).

Grosse, Roger et al. (2023). "Studying large language model generalization with influence functions". *arXiv preprint arXiv:2308.03296*.

Knibbs, Kate (2025). "Thomson Reuters Wins First Major AI Copyright Case in the US". *WIRED*. URL: https://www.wired.com/story/thomson-reuters-ai-copyright-lawsuit/.

Langlais, Pierre-Carl, Anastasia Stasenko, and Catherine Arnett (Nov. 2024). "Releasing the largest multilingual open pretraining dataset". *Hugging Face Blog*. URL: https://huggingface.co/blog/Pclanglais/two-trillion-tokens-open.

Longpre, Shayne et al. (2024). "Consent in crisis: The rapid decline of the ai data commons". *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mattern, Justus et al. (2023). "Membership inference attacks against language models via neighbourhood comparison". *arXiv preprint arXiv:2305.18462*.

Meeus, Matthieu et al. (2024). "Copyright traps for large language models". *arXiv preprint arXiv:2402.09363*.

Meta (Jan. 2024). *Defendant Meta Platforms Inc.'s Answer to First Consolidated Amended Complaint*. URL: https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.72.0_1.pdf.

Mitchell, Margaret et al. (2019). "Model Cards for Model Reporting". *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229.

New York Times (Dec. 2023). *The New York Times Company v. OpenAI, Inc.* Accessed: 2024-12-04. URL: https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

O'Reilly, Tim (June 2024). *How to Fix 'AI's Original Sin'*. Accessed: 2024-12-04. URL: https://www.oreilly.com/radar/how-to-fix-ais-original-sin/.

O'Reilly, Tim et al. (Apr. 2025). *Gaps in Corporate AI Controls: Policy, Practice, and Standards in AI Markets*. Working Paper. AI Disclosures Project. URL: https://www.ssrc.org/programs/ai-disclosures-project/research/.

OpenAI (Mar. 2023a). *GPT-4 System Card*. Tech. rep. Accessed: 2024-12-04. URL: https://cdn.openai.com/papers/gpt-4-system-card.pdf.

— (2023b). *OpenAI–written evidence (LLM0113)*. https://committees.parliament.uk/writtenevidence/126981/pdf/. House of Lords Communications and Digital Select Committee inquiry: Large language models.

— (Mar. 2025). *Response to OSTP/NSF RFI: Notice Request for Information on the Development of an Artificial Intelligence (AI) Action Plan*. Tech. rep. OpenAI. URL: https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf.

Paine, Andre (2025). "Musical AI and Beatoven.ai to Build Fully Licensed Artificial Intelligence Platform for Music Creation". *Music Week*. URL: https://www.musicweek.com/digital/read/musical-ai-and-beatoven-ai-to-build-fully-licensed-artificial-intelligence-platform-for-music-creation/091000.

Paul, Katie and Anna Tong (Apr. 2024). "Inside Big Tech's underground race to buy AI training data". *Reuters*. URL: https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

Puerto, Haritz et al. (2024). "Scaling Up Membership Inference: When and How Attacks Succeed on Large Language Models". *arXiv preprint arXiv:2411.00154*.

Roth, Emma (Aug. 2024). "Authors sue Anthropic for training AI using pirated books". *The Verge*. URL: https://www.theverge.com/2024/8/20/24224450/anthropic-copyright-lawsuit-pirated-books-ai.

Satvaty, Ali, Suzan Verberne, and Fatih Turkmen (2024). "Undesirable Memorization in Large Language Models: A Survey". *arXiv preprint arXiv:2410.02650*.

Sawers, Paul (May 2024). *Data-labeling startup Scale AI raises $1B as valuation doubles to $13.8B*. URL: https://techcrunch.com/2024/05/21/data-labeling-startup-scale-ai-raises-1b-as-valuation-doubles-to-13-8b.
Shi, Weijia et al. (2023a). *Detecting Pretraining Data from Large Language Models*. arXiv: 2310.16789 [cs.CL].
— (2023b). "Detecting pretraining data from large language models". *arXiv preprint arXiv:2310.16789*.
Shokri, Reza et al. (2017). "Membership inference attacks against machine learning models". *2017 IEEE symposium on security and privacy (SP)*. IEEE, pp. 3–18.
Thornhill, John (2025). "Help is coming in the AI copyright wars". *Financial Times*. URL: https://www.ft.com/content/b98979ba-6ae7-4490-97a9-127381440b1f.
Whitwam, Ryan (2025). "Google joins OpenAI in pushing feds to codify AI training as fair use". *Ars Technica*. URL: https://arstechnica.com/google/2025/03/google-agrees-with-openai-that-copyright-has-no-place-in-ai-development/.
Zhang, Jie et al. (2024). "Membership inference attacks cannot prove that a model was trained on your data". *arXiv preprint arXiv:2409.19798*.
Zhang, Jingyang et al. (2024). "Min-k%++: Improved baseline for detecting pre-training data from large language models". *arXiv preprint arXiv:2404.02936*.
Zhao, Haiyan et al. (2024). "Explainability for large language models: A survey". *ACM Transactions on Intelligent Systems and Technology* 15.2, pp. 1–38.

A Appendix

A.1 Additional Details About Our Dataset

We tested OpenAI's models on a total of 34 books, but not all books were used for every model. The table below lists the books used and their publication dates. For each model, we excluded any data published in the year the model completed its training from our testing.

Table 2. Detailed information about the books included in our dataset.

Title	Date	GPT-3.5 Turbo Paragraph Count	GPT-4o Mini Paragraph Count	GPT-4o Paragraph Count
97 Things Every Information Security Professional Should Know	2021-09-14	—	315	314
AI-Powered Business Intelligence	2022-06-10	239	397	396
Advancing into Analytics	2021-04-18	—	157	157
Applied Machine Learning and AI for Engineers	2022-11-10	329	353	353
Azure Cookbook	2023-06-29	42	—	—
Building Green Software	2024-03-11	226	416	414
Building Knowledge Graphs	2023-06-26	160	—	—
Building Recommendation Systems in Python and JAX	2023-12-11	311	—	—
Building Solutions with the Microsoft Power Platform	2023-01-06	283	—	—
C# 8.0 in a Nutshell	2020-05-12	335	335	334
Cloud Native Go	2021-04-20	—	358	358
Communicating with Data	2021-10-03	—	446	446
Continuous Deployment	2024-07-25	584	584	582
Data Quality Fundamentals	2022-09-02	447	447	447
Deciphering Data Architectures	2024-02-07	363	477	477
Delta Lake: Up and Running	2023-10-17	187	—	—
DevOps Tools for Java Developers	2022-04-15	304	467	464
Distributed Tracing in Practice	2020-04-14	323	578	578
FastAPI	2023-11-13	79	—	—
Genomics in the Cloud	2020-04-08	479	767	767

Title	Date	GPT-3.5 Turbo Paragraph Count	GPT-4o Mini Paragraph Count	GPT-4o Paragraph Count
Leading Lean	2020-01-23	301	486	486
Learning Digital Identity	2023-01-10	478	—	—
Natural Language Processing with Spark NLP	2020-06-25	135	271	271
Policy as Code	2024-07-09	235	335	334
Practical Natural Language Processing	2020-06-17	292	410	410
Programming C# 10	2022-08-05	1059	1538	1538
Prompt Engineering for Generative AI[10]	2024-05-16	262	304	304
RESTful Web API Patterns and Practices Cookbook	2022-10-17	276	444	444
Scaling Machine Learning with Spark	2023-03-09	291	—	—
Security and Microservice Architecture on AWS	2021-09-08	—	452	452
Software Architecture: The Hard Parts	2021-10-25	—	404	404
The Customer-Driven Culture: A Microsoft Story	2020-03-10	219	366	366
Web API Cookbook[11]	2024-03-28	87	109	109
Web Accessibility Cookbook	2024-06-17	123	170	170

Note: For GPT-4o, we use a sample of 11,375 paragraphs across 26 books, of which 9,300 are non-public and 2,075 are public. Similarly, for GPT-4o Mini we use 11,386 paragraphs (9,308 non-public and 2,078 public) across 26 books. Finally, GPT-3.5 Turbo used 8,449 paragraphs, with 6,410 non-public and 2,039 public paragraphs across 28 books.

In our study any books published in 2023 were excluded from tests involving GPT-4o and GPT-4o Mini, while books published in 2021 were omitted from any tests involving GPT-3.5 Turbo.

Table 3 displays the most common three-word phrases in the public and non-public datasets. These phrasing differences reflect that the public text is typically extracted from the first 1500 words of each chapter (with the exception, being chapters one and four of each book

[10] All paragraphs in *Prompt Engineering for Generative AI* were used to calibrate results but not for testing.
[11] All paragraphs in *Web API Cookbook* were used to calibrate results but not for testing

Table 3. There is a noticeable difference in phrasing between public and non-public text.

Public Split		Non-Public Split	
Phrase	Occurrences	Phrase	Occurrences
in this chapter	138	as well as	455
as well as	115	one of the	449
one of the	99	be able to	444
be able to	89	the number of	436
a lot of	85	you want to	399

Note: Shows most frequent phrases of each data split. The most common phrase in the public dataset introduces a chapter, likely because the public split primarily consists of the first 1,500 characters of each chapter.

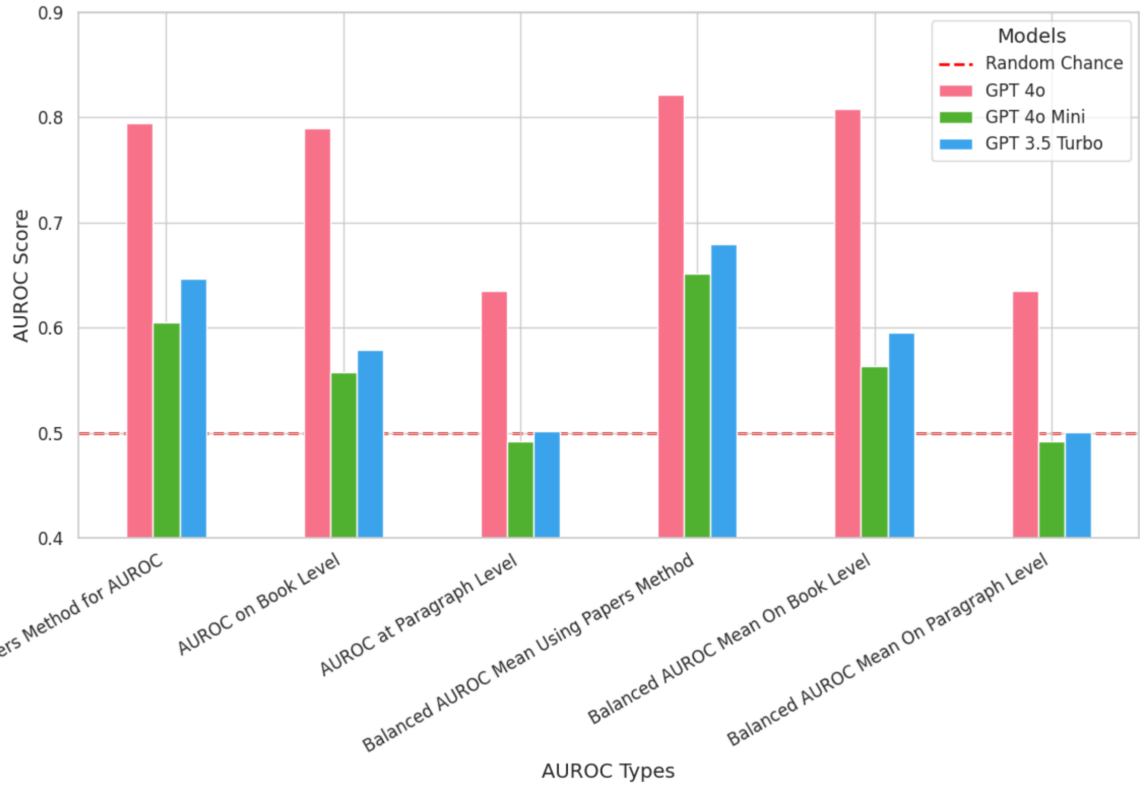
where the entire chapter is public). The public portion of the dataset contains more language introducing a chapter as it mostly consists of the first 1,500 words of each chapter.

A.2 AUROC Results

We found that there are various ways to calculate AUROC scores, some of which can lead to significantly different results (see Figure 4 and Table 4). For example, Duarte et al. (2024) calculated their scores at the book level by first computing the mean guessing score across all paragraphs within each book. They then used these book level guessing rates to determine an optimal threshold, ultimately converting each book into a binary prediction based on whether it exceeded this threshold. This approach appeared to give a boost over doing the AUROC calculation directly without threshholding first (shown in Figure 4). We calculated AUROC scores using the following methods:

- **Papers Method for AUROC** refers to the AUROC score approach from the DE-COP study (ibid.). First, the optimal threshold is determined. Next, each book is given a binary value based on the optimal threshold. Finally, the AUROC score is computed on these binary predictions.
- **Book Level AUROC** is calculated by averaging the identification rates across all paragraphs in a book and then computing the AUROC score using these averages.

Figure 4. AUROC score is highly dependent on the data scale and method it is measured with.



Note: Despite measuring many different AUROC variations we always had a similar pattern of GPT-4o demonstrating the most knowledge, followed by GPT-3.5 Turbo and finally GPT-4o Mini showing the least recognition. See Tables 1 4, 5 and 6.

- **Paragraph Level AUROC** uses the identification rate for individual paragraphs to compute the AUROC score.
- **Balanced AUROC scores** are derived similarly to the other AUROC methods but are calculated from 100 subsets, each containing equal proportions of data from before and after the cutoff date. The mean scores from these subsets are then reported.

Unless otherwise specified, the AUROC scores reported in this paper are book level AUROC. Throughout the paper we refer repeatedly to book level AUROC. However, when testing for robustness we found that the 95% bootstrapped confidence interval was very large at the book level (see Table 5).

Table 4. All AUROC Metrics by Data Split and Model.

	GPT-4o	GPT-4o Mini	GPT-3.5 Turbo
All Paragraphs			
Papers Method for AUROC (Binary)	0.79	0.61	0.65
AUROC on Book Level	0.79	0.56	0.58
AUROC on Paragraph Level	0.63	0.49	0.50
Balanced AUROC Mean Using Papers Method	0.82	0.65	0.68
Balanced AUROC Mean on Book Level	0.81	0.56	0.60
Balanced AUROC Mean on Paragraph Level	0.64	0.49	0.50
Public Paragraphs			
Papers Method for AUROC (Binary)	0.69	0.64	0.67
AUROC on Book Level	0.64	0.55	0.64
AUROC on Paragraph Level	0.60	0.48	0.51
Balanced AUROC Mean Using Papers Method	0.71	0.65	0.69
Balanced AUROC Mean on Book Level	0.64	0.53	0.63
Balanced AUROC Mean on Paragraph Level	0.60	0.48	0.51
Non-Public Paragraphs			
Papers Method for AUROC (Binary)	0.84	0.66	0.62
AUROC on Book Level	0.82	0.56	0.54
AUROC on Paragraph Level	0.64	0.50	0.50
Balanced AUROC Mean Using Papers Method	0.84	0.67	0.63
Balanced AUROC Mean on Book Level	0.82	0.56	0.54
Balanced AUROC Mean on Paragraph Level	0.64	0.50	0.50

Note: Shows all the AUROC scores that we calculated (see Table 1 for sample sizes). Figure 4 visualizes this table.

This is likely attributable to our limited book count. We analyzed a sample of 32 books, each containing thousands of paragraphs. This small number of titles leads to a very wide bootstrapped confidence interval for the book level AUROC scores (see Table 5). Although not ideal, this outcome is expected if some books in our sample were part of the training data while others were not. Since we approximate the in-dataset and out-of-dataset groups with a small sample, any 'mislabeled' data – data that was assumed to be in-dataset but was not actually included – can disproportionately affect and skew the results.

In contrast, preforming bootstrap at the paragraph level – where data is pooled across books – yields significantly smaller bootstrapped confidence intervals across all models (see Table

Table 5. Book Level AUROC Scores with Bootstrapped Confidence Intervals by Data Split.

Model	Data-Split	Book Level AUROC
GPT-4o	All	0.79 (0.53, 0.96)
	Public	0.64 (0.36, 0.93)
	Non-Public	0.82 (0.60, 0.96)
GPT-4o Mini	All	0.56 (0.25, 0.84)
	Public	0.55 (0.20, 0.84)
	Non-Public	0.56 (0.28, 0.83)
GPT-3.5 Turbo	All	0.58 (0.33, 0.83)
	Public	0.64 (0.39, 0.86)
	Non-Public	0.54 (0.28, 0.77)

Note: We performed hierarchical bootstrapping using 1000 bootstraps over all books not published in the year of a model's cutoff date (see Table 2). To preform hierarchical bootstrapping we repeatedly sampled random books and then random paragraphs within each book. See Table 1 for sample sizes.

Table 6. Paragraph Level AUROC Scores with Bootstrapped Confidence Intervals by Data Split

Model	Data-Split	Paragraph Level AUROC
GPT-4o	All	0.63 (0.62, 0.65)
	Public	0.60 (0.57, 0.63)
	Non-Public	0.64 (0.63, 0.66)
GPT-4o Mini	All	0.49 (0.48, 0.51)
	Public	0.48 (0.45, 0.51)
	Non-Public	0.50 (0.48, 0.51)
GPT-3.5 Turbo	All	0.50 (0.49, 0.52)
	Public	0.51 (0.48, 0.54)
	Non-Public	0.50 (0.48, 0.51)

Note: We performed bootstrapping using 1000 bootstraps over all paragraphs not published in the year of a model's cutoff data. See Table 1 for sample sizes.

6), indicating that increasing the number of books would likely help tighten the book level bootstrapped confidence intervals.

A.3 Prompts and Settings Used for Tested Models

To test the model's ability to differentiate between member and non-member data, we used a modified version of the prompts presented in Duarte et al. 2024. An example prompt is shown below.

System Prompt:

```
You are a helpful assistant. You must answer using only the
provided options A, B, C, or D, you may not decline to answer.
```

User Prompt:

```
Question: Which of the following passages is verbatim from
the "97 Things Every Information Security Professional
Should Know" book by Christina Morillo?

Options:
A. Learn with a community. I've personally found that I've
developed new and existing skills much faster (by a
magnitude of months) when I've joined a community of
learners than when I was trying to learn the same skills
alone. By attending local and online user groups,
conferences, and other events, you can discover new
concepts, hone new skills, and network with possible
future colleagues. Also, in a community, you will gain a
more holistic perspective of information security and a
more complete picture of how others are managing
successful information security programs.

B. Connect with fellow learners. From my experience,
```

acquiring and improving abilities has been significantly quicker (saving several months) when participating in group learning compared to solo studying. Going to regional and virtual meetups, seminars, and similar gatherings helps you explore fresh ideas, develop capabilities, and build relationships with potential workmates. Furthermore, learning within a group provides broader insights into cybersecurity and better understanding of how various organizations implement effective security initiatives.

C. Join a learning group. Based on my observations, mastering both new and current abilities happens much more rapidly (reducing learning time by months) when I'm part of a learning circle versus studying independently. Through participation in area-based and internet-hosted gatherings, symposiums, and other meetings, you'll encounter different concepts, sharpen your abilities, and connect with prospective professional contacts. Additionally, group involvement offers deeper understanding of security practices and clearer insights into successful security program management across organizations.

D. Engage in collaborative learning. My personal journey shows that skill acquisition and enhancement occurs substantially faster (cutting months off learning time) within group settings rather than individual efforts. By taking part in both physical and digital group meetings, industry events, and related activities, you can learn new approaches, improve your capabilities, and establish

```
connections with future professional peers. Moreover,
group settings provide comprehensive knowledge about
information security and valuable examples of how different
teams run successful security operations.

Answer:
```

Model settings:

```
{
    "max_tokens": 1,
    "temperature": 0,
    "seed": 2319,
    "logprobs": True,
    "logit_bias": {32: +100, 33: +100, 34: +100, 35: +100},
    "top_logprobs": 20
}
```

The exact models tested were as follows: gpt-4o-2024-08-06, gpt-4o-mini-2024-07-18 and gpt-3.5-turbo-1106.

A.4 Prompts and Settings Used for Paraphrase model

We used Claude 3.5 Sonnet to generate paraphrases from the O'Reilly Media books. An example prompt is shown below.

User Prompt:

```
Rewrite this entire text (all sentences with no exception)
expressing the same meaning using different words. Aim to keep the
rewriting similar in length to the original text. Do it three
```

times. The text to be rewritten is identified as <Example A>.

Format your output as:

Example B: <insert paraphrase B>

Example C: <insert paraphrase C>

Example D: <insert paraphrase D>

-

Example A: In general, a soft trade-off exists between active learning that's useful for maximally improving your model globally and active learning that's useful for maximizing the likelihood that a user can and will rate a particular item. Let's look at one particular example that uses both.

Model settings:

```
{
    "temperature": 0.1,
    "model":"claude-3.5-sonnet"
}
```

AI DISCLOSURES PROJECT

100 YEARS SOCIAL SCIENCE RESEARCH COUNCIL

Social Science Research Council
300 Cadman Plaza West, 15th Floor
Brooklyn, NY 11201, USA