NBER WORKING PAPER SERIES

# CLOZE ENCOUNTERS:
# THE IMPACT OF PIRATED DATA ACCESS ON LLM PERFORMANCE

Stella Jia
Abhishek Nagaraj

Working Paper 33598
http://www.nber.org/papers/w33598

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2025

Srushti Pawar and Nancy Ma provided invaluable research assistance. This work was not funded by any external organization, although we do thank OpenAI for research credits that were used to execute a part of this research. We are very grateful to Alex Reisner, David Bamman and participants at the Berkeley-Haas Macro Research Lunch and the Data Innovation Lab for useful discussions about this project. Any opinions and conclusions expressed herein are those of the authors only, and any errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w33598

Cloze Encounters: The Impact of Pirated Data Access on LLM Performance
Stella Jia and Abhishek Nagaraj
NBER Working Paper No. 33598
March 2025
JEL No. K24, O36

## ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable capabilities in text generation, but their performance may be influenced by the datasets on which they are trained, including potentially unauthorized or pirated content. We investigate the extent to which data access through pirated books influences LLM responses. We test the performance of leading foundation models (GPT, Claude, Llama, and Gemini) on a set of books that were and were not included in the Books3 dataset, which contains full-text pirated books and could be used for LLM training. We assess book-level performance using the "name cloze" word-prediction task. To examine the causal effect of Books3 inclusion we employ an instrumental variables strategy that exploits the pattern of book publication years in the Books3 dataset. In our sample of 12,916 books, we find significant improvements in LLM name cloze accuracy on books available within the Books3 dataset compared to those not present in these data. These effects are more pronounced for less popular books as compared to more popular books and vary across leading models. These findings have crucial implications for the economics of digitization, copyright policy, and the design and training of AI systems.

Stella Jia
University of California, Berkeley
stellajia@berkeley.edu

Abhishek Nagaraj
Haas School of Business
University of California, Berkeley
2220 Piedmont Ave
Berkeley, CA 94720
and NBER
nagaraj@berkeley.edu

# 1 Introduction

Large Language Models (LLMs) have emerged as one of the most transformative technologies of our time, demonstrating remarkable capabilities in understanding and generating human-like text across a wide range of applications. For example, LLMs have shown to significantly enhance and supplement human capabilities in writing, idea generation, programming and image generation (Noy and Zhang, 2023; Boussioux et al., 2024; Ludwig et al., 2025; Tranchero et al., 2024; Cui et al., 2024; Zhou and Lee, 2024). Further, in a business context, these have shown to aid in customer support, market research and enhance productivity of knowledge workers (Brynjolfsson et al., 2023; Dell'Acqua et al., 2023; Brand et al., 2023; Jaffe et al., 2024).

While these models have demonstrated substantial economic value, their reliance on copyrighted and potentially pirated content during the training process has raised significant legal, economic, and technical issues. On the legal side, many copyright owners are up in arms about the unauthorized use of their content (New York Times Company v. Microsoft Corp., 2023; Getty Images (US) Inc. v. Stability AI, 2023; Authors Guild v. OpenAI Inc., 2023). For instance, the New York Times Company v. Microsoft Corp. (2023) lawsuit cited numerous examples of New York Times content being reproduced verbatim by ChatGPT. On the economic side, many content providers are willing to license content to foundation model developers but determining the value of such data access is challenging. For example, Reddit has partnered with OpenAI and Google to grant access to its content for a reported $60 million/year, although there are no well-established frameworks to determine the appropriate valuation.[1][2] On the technical side, there are numerous questions relevant to the design of LLMs themselves, we do not fully know the extent to which these models "memorize" training inputs from certain data sources or truly generalize (Nasr et al., 2023; Carlini et al., 2022) and the role of specific training data sources on model performance (Park et al., 2023; Elazar et al., 2023). Despite scattered disclosure in this domain,[3] the lack of transparency (Bommasani et al., 2023) around training data (even in more "open" models like Meta's Llama series) have led to intense speculation about which content these models are trained on, and what content is excluded.

The answers to many of these questions depend on a key elasticity: the effect of access to copyrighted content on model responses. In particular, models are reported to have trained on "shadow libraries," which are large collections of copyrighted material that are otherwise not freely accessible. Understanding their role of such training data on model performance is particularly relevant. From a theoretical point of view it is not clear how access to specific sources of copyrighted content affects model performance. On one hand, it is reasonable to infer that any one source of content has very little bearing on performance because of the

[1] https://openai.com/index/openai-and-reddit-partnership/
[2] https://www.cbsnews.com/news/google-reddit-60-million-deal-ai-training/
[3] https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/

sheer volume and diversity of data (Gao et al., 2020; Dodge et al., 2021) used in training. Models may not even need direct access, since content is often replicated elsewhere in training data through summaries and references (Carlini et al., 2022; Chang et al., 2023). On the other hand, as evidenced by an active licensing market for such data, it is clear that training data and content are the lifeblood of generative AI foundation models and are therefore very valuable in determining model performance (Azoulay et al., 2024). Beyond generalities, these theoretical predictions are unlikely to apply uniformly, varying significantly depending on the substitutability of content and across different types of LLMs with different architectures and training approaches.

What are needed are empirical estimates that examine the effect of access to certain bits of (copyrighted) content to foundation models and the effects of such access to model performance in production settings. However, despite legal and theoretical writing around digitization and copyright (Lemley and Casey, 2020; Lemley, 2024; Yang and Zhang, 2024; de Rassenfosse et al., 2025), and estimates from computer scientists from carefully constructed "toy" experiments (Carlini et al., 2022; Park et al., 2023), there exists no data or research design that helps answer this question in production settings. Theoretical work is helpful to identify useful elasticities while the generalization of toy models to large-scale models remains an open question given the large difference in model performance based on model size (Kaplan et al., 2020). This is the central challenge that we tackle in this project.

There are two barriers to overcome: measurement and identification. On the measurement side, model developers rarely disclose the specific sources used during training, making it nearly impossible to determine what data was included. Further, there are no uniform metrics of model performance, especially as they relate to their knowledge of specific sources of content. On the identification side, even if we could measure performance and variations in access to training data, direct comparisons between accessible and inaccessible materials would be problematic. Access is often correlated with other factors that influence learning, such as the availability of similar content and the ease of learning across different content types. For instance, a model's better performance on public domain books compared to more obscure works might reflect these confounding factors rather than the mere presence or absence of the data in training. Addressing this requires a research design that introduces quasi-exogenous variation in access to similar types of content and enables the linkage of this variation to specific measures of the model's knowledge for a given source.

In this project, we provide a research design and empirical framework that tackles these dual challenges. We examine the effect of access to a large number of books through the "Books3" dataset of pirated books on multiple LLMs in production settings. First, to measure the performance of a model as it relates a particular book, we first build on the "name cloze" technique that helps us quantify to what extent a model is able to answer questions about a given source from the machine learning literature (Chang et al., 2023; Petroni

et al., 2019). Specifically, this technique tests a model's ability to answer factual questions about a given book based on missing information. For example, given a passage like *"Yeah? Well maybe satyr emotions work differently than human emotions. Because you're wrong. I don't care what he thinks. [MASK] pulled his feet up onto the branch,"* (from *The Lightning Thief* by Rick Riordan) the model would need to correctly identify "Grover" as the missing name. This provides a concrete, measurable way to assess how well a model has learned and retained specific factual information from books in its training data.

For identification, we leverage our institutional knowledge of the availability of training data for LLMs in the recent past. In particular, we examine the "Books3" dataset posted online by a single individual who bundled about 195,000 copyrighted books (in an act of benevolent piracy) for the express purpose of training LLMs.[4] As we will show, the year of publication of a book had a strong predictive effect on whether or not a particular book was included in the Books3 dataset, which provides a potential natural experiment to examine the effect of access to training data on model performance. In particular, we build a sample of 12,916 books (of which 6,349 were included in Books3 and the rest were not). Using the year of publication as an instrumental variable (IV) for Books3 selection, we examine the causal effect of data access on model performance. We examine how inclusion in the Books3 dataset affects model performance across four sets of leading models (GPT class models by OpenAI, Claude by Anthropic, Gemini by Google and Llama by Meta). We also examine how these effects vary for books that are more vs. less popular as determined by their reviews on the Goodreads book review platform.

We find that direct access to the full text of a book Books3 significantly enhances performance on the name cloze task. The most pronounced improvements are observed in GPT-3.5 Turbo and GPT-4.0, compared to the Llama 3.1 series and the Claude Haiku and Gemini series of models. The "Books3 effect" is about a 21-23% increase in accuracy for GPT class models as compared in the 7-9% range for the Llama-70b and Claude and Gemini models. This result is at least in part due to our finding that the GPT class models perform about 5 percentage points worse on the name-cloze task across all books as compared to rival models. Interestingly, we find no detectable Books3 effect for the smaller Llama model (Llama-8B) as compared to its larger counterpart (Llama-70B), suggesting that training data effects depend on model size. Next, we examined how book popularity affects accuracy, hypothesizing that works with more substitutes would show a weaker Books3 effect. Our findings indicate that the Books3 effect is much larger for less popular books as compared to less popular books. This points to a significant role for training data substitution, which reduces the effect of direct access to source material on model performance. In the end, our results provide the first set of empirical results that concretely tie access to digital content through a pirated database (Books3) to a direct measure of model performance. [5]

[4] `https://x.com/theshawwn/status/1320282151645073408?lang=en`

[5] Interestingly, while Llama is confirmed to have been trained on Books3, our work suggests that other leading models are relying on this data source as well.

Our work contributes to multiple literature's in law & economics and computer science. First, while while a significant body of work has empirically studied the economics of copyright and piracy in digital settings (Oberholzer-Gee and Strumpf, 2007; Nagaraj and Reimers, 2023; Reimers, 2016; Rob and Waldfogel, 2006; Watson, 2017; Waldfogel, 2012; Moser, 2005; Smith and Telang, 2009; Chiou and Tucker, 2017), we extend that work to examine how these issues apply for the development of generative AI models. This work has contrasted how a lack of digital access and legal restrictions can constrain reuse. We show how the performance of generative AI systems might also depend on the piracy status of the source material. Further, we uncover important institutional details (such as the Books3 dataset) and apply novel tools like the cloze task, that provides useful frameworks to help further research in this area.

In parallel, while the cloze method has been used in prior computer science research to assess comprehension (Petroni et al., 2019; Dhingra et al., 2017; Chang et al., 2023) in synthetic experiments and creating cloze evaluation datasets (Mostafazadeh et al., 2017; Onishi et al., 2016; Petroni et al., 2019), we apply it to a real-world policy and economic questions. By using a dataset under legal scrutiny (Books3) and applying the cloze method to production models, we investigate the relationship between input and output in LLMs on a specific source to assess model performance. Further, there is some recent work on TDA (training data attribution) where researchers typically build their own smaller models with access to different types of training data or use smaller models to assess their performance (Park et al., 2023; Akyürek et al., 2022; Carlini et al., 2022). While these studies are insightful in that they can offer broad measures of performance and experimentally vary training data sources, unfortunately they cannot speak to the impact of training data access on performance in real-world production systems. Given the variation we found with the model size, our work provides some complementary evidence from production models that are much larger in size, albeit using a more narrow definition of performance. Further, we add to the toolkit by showing how applied microeconometric approaches could complement existing methods in CS to examine this question.

Our work also has direct implications for ongoing legal and policy debates. First, we provide some of the first evidence that suggesting a link between the Books3 dataset and model performance, a topic that has received much speculation.[6] We provide the first empirical evidence linking these data to model performance. Second, our results provide some real-world empirical estimates on the effect of training data on model performance which might be important for key economic calculations in this area. For example, we can refute the claim that any single source has a negligible effect on model performance, even in large production models. However, these effects do vary significantly, across both models and notably content popularity. These findings therefore have implications for pricing of data licenses, and support more bargaining power for content providers who provide more unique content. They also have implications for copyright policy that might decide on to what extent the use of copyrighted data by model developers constitutes fair use

[6] https://www.wired.com/story/battle-over-books3/

(Lemley and Casey, 2020; Samuelson, 2023; Henderson et al., 2023). For model developers, our work provides the seeds of new potential benchmarks that would help them assess their dependence on pirated data sources. Additionally, our work provides a useful framework for developers to evaluate and select optimal training datasets for future model development.

It is useful to note a few limitations of our work. First, our work provides only an indirect link between pirated data sources and model performance. We can only indirectly infer the reliance on certain data sources. Transparency from model developers about training data would provide more direct evidence. Second, the "name cloze" test is a relatively narrow test of model performance. Even though there is work to suggest that performance on cloze tests is related to general intelligence or model utility (Petroni et al., 2019), more work is needed to establish the link in our setting. Further, we simply show that Books3 inclusion affects model performance – not that it affects sales or any downstream economic outcome that might cannibalize the market for the original content producers. We also do not speak to questions around "fair use" that might render reliance on such data to be legal.

With these caveats in mind, the paper proceeds as follows. In Section 2, we provide some basic details on the role of training data in LLMs, the Books3 data specifically, and the name cloze test. Then in Section 3, we describe our data collection and sampling strategy and our IV research design. Section 4 describes our key results and 5 concludes.

# 2 Empirical Setting

## 2.1 Background

**A. Training Data and LLMs:** Generative AI models, particularly large language models (LLMs), typically undergo a two-stage training process: pre-training and post-training (Brown et al., 2020). The pre-training phase is crucial and resource-intensive, focused on exposing the model to massive datasets to learn general language patterns and world knowledge. For text-based LLMs, this involves training on vast quantities of text data to enable the model to predict the next token in a sequence and the composition of these pre-training datasets is a critical factor influencing model capabilities (Brown et al., 2020). Estimates suggest that pre-training datasets can be extraordinarily large, ranging from terabytes to petabytes of text data, encompassing trillions of tokens. For instance, the Llama-3 set of models was trained on over 15 trillion tokens (Dubey et al., 2024). Despite its importance, developers are often guarded about the specific data sources used for pre-training, even for models released with open weights. Significant variation is suspected across different models on this dimension (Bommasani et al., 2023).

Even though we don't know what data is used to train specific models, reports suggest a few common sources

that include both public domain and copyrighted content (Granite Team, 2024; Touvron et al., 2023). These sources notably include datasets like Common Crawl (Dodge et al., 2021), a vast repository of web data, USPTO records on patents, and notably The Pile, a massive 825GB dataset. This last dataset was created as a part of a movement to create open rivals to popular closed-sourced models and the initiative was spearheaded by EleutherAI, aiming to democratize access to high-quality training data (Gao et al., 2020). The Pile, includes sources like Wikipedia, academic papers, and various internet archives. Overall training data used in modern production models encompass a wide range of text, from publicly available web pages to social media and news articles.

**B. Books3:** Of the many key datasets rumored to be used in LLM training, a key one is Books3. Books3 is inspired by the Books1 and Books2 datasets, which are part of the BookCorpus dataset introduced in 2015 (Zhu et al., 2015), and was used in training many foundational research models including the GPT-1 and GPT-2 models from OpenAI (Radford, 2018; Radford et al., 2019; Brown et al., 2020; Kenton and Toutanova, 2019). This BookCorpus dataset (that predates Books3) contains 7,185 unique books, including many copyrighted works. Books3's emergence is closely tied to the broader movement in the AI research community to create large-scale, diverse datasets for training increasingly powerful language models that includes the creation of the Pile dataset.

Within this context, Shawn Presser, a figure associated with the data-sharing community and potentially involved with EleutherAI, compiled Books3 in October 2022.[7]. The goal of Books3 was to recreate data presumably being used by leading model developers to train their AI models, including the GPT-3 model released by OpenAI. Books3 expanded previous efforts significantly in scale, reportedly encompassing around 196,000 books. Very little is known about precisely how Presser selected these books, but these works were not digitized for the express purpose of being included in this dataset.[8] Instead Presser relied on a "shadow" library, *Bibliotik*, which itself hosted a vast amount of pirated ebooks and packaged these books into a format that could be used to train LLMs.

What does the Books3 dataset contain exactly? Thanks to a few journalistic efforts, we have good metadata on the books included (even though the dataset with the full text of the books itself is no longer publicly available owing to copyright restrictions). According to one source,[9] the dataset includes number of books by popular authors like Michael Pollan, Stephen King and Zadie Smith. To give a sense of the scale, Books3 includes over 30,000 titles from Penguin Random House and its imprints, 14,000 from HarperCollins, 7,000 from Macmillan and 1,800 from Oxford University Press.[10] We will use this metadata and describe them further shortly. However, in Figure 2, we provide a distribution of the publishing year of the books included

[7] `https://www.wired.com/story/battle-over-books3/`
[8] We did reach out to Presser to get more information on this process, but did not receive a response.
[9] `https://aicopyright.substack.com/p/the-books-used-to-train-llms`
[10] `https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/`

in the Books3 database. Books between the years 2000-2019 represent almost 80% of the entire dataset. Even within these years, there is significant variation. Almost 14% of all books come from just two years - 2013 and 2014, with a large majority coming between 2012 and 2020. These years likely represent years when Bibliotik was particularly active and effective in archiving pirated copies of books. We will exploit this variation as a core part of our analysis.

**C. The Reaction to Books3 and Copyright Challenges to LLMs:** As mentioned before, most commercial foundation model providers (including many "open" ones) do not disclose the specific data they trained on, including their potential use of the Books3 dataset. Having said that, suspecting such use, copyright holders have been notably opposed to any use of their works in training such models, citing copyright infringement. In the case of Books3, the Authors Guild, which represents authors in the US, has been notably vocal in its opposition. For example, it has launched a class-action lawsuit against OpenAI[11] and mentions Books3 specifically claiming *"The similarities in the sizes of Books2 and Books3, and the fact that there are only a few pirate repositories on the Internet that allow bulk ebook downloads, strongly indicates that the books contained in Books2 were also obtained from one of the notorious repositories discussed above."*

These cases are emblematic of a whole rash of cases being currently litigated by numerous content creators (including entities like the New York Times, legal database providers WestLaw and image data providers like Getty Images). The outcome of these cases will come down to a number of factors, including notably whether the use of such data constitutes "fair use." The fair use defense does not debate that foundation model developers used copyrighted content in training their models, but instead on whether it was lawful for them to do so. Our paper does not directly tackle this question. However, the fair use defense hinges on a key factor to which we can speak to, notably "amount and substantiality" of the work taken. In other words, fair use depends on the question of whether a significant amount of source text was used in the training of foundation models. For example, whether the model was trained on the entire raw text of a book, or merely its summaries available on third-party sources. Insofar as this information is not publicly known, it is important to have methods to determine to what extent certain data sources form a part of the training corpus. There are other key factors that help assess fair use, including whether or not the derivative use, including the (potentially negative) effect of the use upon the potential market for or value of the copyrighted work. Our work has less to say on this point directly, although the name cloze test could be seen as a rough proxy of substitutability or general performance under some strong assumptions.

Beyond copyright claims, a parallel development has been the birth of a nascent market for licensing content. For example, the academic publisher Taylor and Francis is on record licensing its content covering over

[11] https://authorsguild.org/app/uploads/2023/09/Authors-Guild-OpenAI-Class-Action-Complaint-Sep-2023.pdf

3000 academic journals for about $10 million for the training of LLMs.[12] Similarly, Reddit is reportedly licensing all its content to Google for about $60M per year, and Wiley is being paid about $23M for one-time access to previously published academic papers.[13] Beyond these valuations, exactly how one determines the economic value of access to certain bits of training data on the performance of LLM systems remains hard to assess. Our work provides one framework to begin answering this question.

## 2.2 The Name Cloze Test: Linking Data Access with LLM Performance

Past studies of the effect of data access on follow-on use in contexts such as academic texts or Wikipedia (Nagaraj, 2018; Biasi and Moser, 2021), often relies on natural linkages between the follow-on text and the source material via citations. Citations are hard to obtain with LLMs - partially because LLM designers do not include citations from model responses to source text, and partially because such linkages are quite hard to obtain (even for the determined model developer), given very unclear linkages between model outputs and inputs. For example, if an LLM is trained on a sentence like "Jane Austen wrote Pride and Prejudice," it learns not just about Jane Austen and Pride and Prejudice, but also about sentence structure, verb conjugation, and the relationship between authors and books from that source. Therefore, if the model later correctly answers the question "Who wrote Pride and Prejudice?", it's difficult to definitively say that this response is solely because of that single sentence in its training data. To further complicate matters, multiple input sources might have the same information - so how can one attribute any specific response to any specific piece of text?

To get around this challenge, we focus on a very specific, albeit narrow measure of LLM performance that is designed to elicit its knowledge about very specific (and relatively obscure) pieces of knowledge that are likely to depend on direct access to the source text of books. Specifically, we reference the "Name Cloze" test (Chang et al., 2023) which proceeds as follows. For each book we randomly select 100 passages from the full text that includes a named entity and mask this "name" token. We then ask the model to predict the missing token. We do not explicitly reference the source book while doing so. We score every response as "true" or "false" based on the ground truth, allowing us to score every model using an average of its performance across the 100 passages. See Figure 1, Panel B for a detailed example and the prompt we use.

This method has a number of advantages. First, it explicitly links model outputs to specific books, allowing us to assess performance on a per-book basis. Second, it circumvents restrictions that many models have in place to prevent the output of copyrighted material, as we do not explicitly reference the original books in our prompts. Third, because we have ground truth (the actual names), we can objectively score model responses at the book level, and this procedure can be consistently applied across any language model. Fourth, this test

[12] https://www.insidehighered.com/news/faculty-issues/research/2024/07/29/taylor-francis-ai-deal-sets-worrying-pre
[13] https://www.monda.ai/blog/ultimate-list-of-data-licensing-deals-for-ai

is inspired by "cloze" tests, which have a deep history in psychology and education. Cloze tests, originating in the 1950s (Taylor, 1953), are a well-established method for measuring language comprehension in humans and have long been considered a valuable tool in psychological assessments of intelligence and cognitive closure. Cloze comes from the word "closure" which is used in Gestalt psychology to refer to the human tendency to see an unfinished template as a whole by filling it in with something it's familiar with. This mimics the observed tendency of LLMs to memorize content and output things based on highest probability. Despite its merits, the name cloze test has limitations including the requirement for full text access, its narrow focus on name recognition as a performance measure, and its indirect relevance to broader economic questions like content substitution or a broad measure of intelligence or capability. Despite these challenges, our research methodology will adopt this measure as a proxy of LLM performance, given its attractiveness in answering the research question at hand.

# 3 Data

## 3.1 Research Design

An ideal experiment would take a large sample of comparable books, and make a subset of them available for training LLMs (treatment set). We would then assess the performance of these models at the book level for treatment and control books. We could train multiple models in this way, but doing so at the scale of production models would be prohibitively expensive. Our research design tackles two related challenges in implementing such an experiment: (a) linking source texts to LLM performance and (b) finding causal variation in access to training data.

On the first point, the name cloze methodology provides an elegant way to link inputs to outputs in a practical manner. Chang et al. (2023) shows how this test could be leveraged to understand memorization in a sample of about 500 books. We expand this exercise considerably and link it to piracy status. On the second point, we can exploit the uneven coverage of books in the Books3 dataset for identification. Specifically, a random book is much more likely to be in the Books3 dataset if it was published in 2018 as compared to 2022, to name one example. Specifically, we can use the shares of books in the Books3 dataset by publication year as an instrument for Books3 status. Insofar as this instrument is directly unrelated to model performance, it would meet the exclusion restriction and quality as a valid instrument. There are reasons to think that publication year would affect model performance through other channels for example popularity or quality. Having said that after controlling for these factors, we do not see any reason why a specific year (such as 2018 vs 2022) should have any meaningful effect on model performance, except through inclusion in the Books3 dataset which has a very particular pattern through which books are included.

### 3.2 Data

To assess LLM performance via name cloze, we constructed a dataset of 12,916 books. Our data collection strategy was comprehensive, leveraging multiple sources and proceeded in several steps and is described in detail in Figure 1 Panel A. We provide a summary below.

First, we searched ISBNdb, an online database of book metadata to build a large set of target books. Since the entire database is not available to download, we searched for publication years from 1800 to 2023, and for a whole host of common stop words like "and", "the" etc. and related words such as "journey", "time", "adventure" etc.[14] Through this process we obtained metadata on a total of 794,081 books. In parallel, we obtained metadata on the books included in the Books3 dataset from a online archive that contains this list,[15] and tagged each book in this sample as whether or not it was included in the Books3 dataset.

Next, for each of these books we searched Goodreads, which contains additional information of each of the titles in our base database. Specifically we matched on title and author and found matches for 518,684 books from the ISBNdb dataset and 195,977 from the Books3 dataset. This step gave us additional information such as number of reviews, number of ratings, average rating, etc. Then, since we need the full text of books for the name cloze algorithm, we queried Libgen, a shadow library providing access to books, articles, and other copyrighted content. Since Libgen does not contain the full text of every book, from the list of 518,684 books we were able to successfully scrape 16,161 books.

Then, we ran the 16,161 books through the name cloze procedure as detailed in Chang et al. (2023). Specifically, we use the code created by Chang et al. (2023)[16] to preprocess the full text of a book by running named entity recognition (NER) to extract people, facilities, geo-political entities, locations, vehicles, and organizations (e.g. Percy Jackson, the house, London, the river, the car, the army) from passages of 40–60 tokens that contain exactly one named entity. Each extracted entity is then masked, and the passage is formatted as a prompt requesting a language model to predict the missing entity. For each book, we sample a set of 100 such passages, and we create a following prompt at the passage level.

> You have seen the following passage in your training data. What is the proper name that fills in the [MASK] token in it? This name is exactly one word long, and is a proper name (not a pronoun or any other word). You will be penalized if you predict pronouns like him, he, she, us, etc. Please predict only proper nouns! You must make a guess, even if you are uncertain.
>
> Example:

[14] We obtained this list through iterative prompting from the GPT-4 model.
[15] `https://web.archive.org/web/20230928061137/https://battle.shawwn.com/books3-metadata.jsonl`
[16] `https://github.com/bamman-group/gpt4-books`

Input: “Stay gold, [MASK], stay gold.”

Output: <name>Ponyboy</name>

Input: “The door opened, and [MASK], dressed and hatted, entered with a cup of tea.”

Output: <name>Gerty</name>

Input: [insert selected passaged with masked token]

Output:

This prompt is then submitted to multiple models - OpenAI’s GPT 3.5 and 4o models (via the API interface), to the Llama 3.1 8B and 405B models (via Openrouter), Google’s Gemini 2.0 Flash model and Anthropic’s Claude 3 Haiku model via their API interfaces. For each book-model combination we record accuracy as the proportion of correct predictions out of total queries. We aim to use 100 prompts per book, although for some books, we obtain fewer responses. Our accuracy score can thus be interpreted as the percent of times models correctly predicted the masked token on the name cloze task. After processing each book through this procedure and mapping the results back to our metadata in a CSV file, we encountered minor data losses and mismatches, resulting in a final sample of 12,916 books.

### 3.3 Summary Statistics

The summary statistics of our final sample is in Table 1. We have a roughly even distribution of Books3 and control books. The top half of this table presents key covariates at the book level, including total ratings and total reviews, where the median values are 68 and 7 respectively, although the means are much higher, suggesting a considerable variation in reader engagement across books. The average rating for books in the sample is 3.55 and ratings range from 0 (indicating no ratings) to a perfect score of 5. Publication years span a vast range but these books are pretty recent, with a median publication year of 2004. In terms of genre, a significant majority (71%) of the books are classified as nonfiction, highlighting a focus on informational or educational content in the books sampled.

The lower section of the table shows performance metrics of various LLMs evaluated against this dataset. Two facts are worth noting. Overall, these scores can be interpreted as a percent of prompts for which LLMs got the right response. Interpreted this way, none of the models get more than 17% of responses right on our quite stringent test. By comparison, humans score 0% on similar tasks (Chang et al., 2023). However, these scores also do vary considerably. Claude shows a higher mean performance score (14%) compared to other models, with Llama 70B also showing strong performance (15%). The smaller Llama 8B model has a much lower score compared to the rest, although its likely trained on a similar training set. This points to the importance of model size in shaping the overall accuracy for this task, beyond just the

training data set. We code these books into 4 popularity bins based on number of Goodreads reviews, with the books roughly split between 0-10,10-100, 100-1000, and 1000+ reviews. These are the popularity bins we will use in our analysis. Examples of very popular books (1000+ reviews) in the our sample include J.D Salinger's "Catcher in the Rye" (in the Books3 dataset) and "Life, the Universe and Everything" by Douglas Adams (not in Books3). In our analysis, we will now explore to what extent these accuracy scores differ systematically between the Books3 and the control set. We will also explore how the Books3 effect varies based on the popularity bins we assigned.

### 3.4 Validating the IV

The ability of an LLM to learn about a source is subject to bias due to the presence of confounding variables like the frequency of a source in the dataset or the popularity of the book. If LLMs do better or worse on Books3 books vs controls, this might just be because Books3 books are highly selected and are more likely to occur in the training data.

As previewed, we will use the skew in Books3 books by publication year to get around this issue. Figure 2, Panel A provides a histogram that illustrates the distribution of publication years for books in the Books3 dataset, consisting of 195,997 books. A distinct pattern emerges, showing a sharp increase in the number of books published starting around 2007, peaking between 2011 and 2014, and then declining sharply after 2018. This distribution is quite specific and concentrated, forming a clear surge in book publications over a relatively narrow time span. In contrast, the overall incidence of books published in the literary world typically follows a more smooth and continuous distribution, with publications spread out more evenly across time. This concentrated pattern in Books3 suggests specific process focused on books from a particular era, possibly influenced by time period when the original books were sourced.

Using this data, we calculate our instrument $Share_t$ for each publication year as follows, which will be the key instrumental variable we use in the IV analysis.

$$Share_t = \frac{\text{\# of books in Books3 with first publication year } t}{\text{\# of books in Books3}}$$

Panel B of Figure 2, shows the strong correlation between a given books Books3 status in our sample of 12,916 books and this $Share_t$ variable, showing the strength of the instrument in our context. This figure is calculated by binning the IV in 8 groups and calculating the percent of Books3 books in our sample in each category. The correlation between the two variables is 0.46 and a regression shows a positive and strong relationship between the two.

# 4 Results

## 4.1 Model-Free Evidence

Before we turn to the IV analysis, Figure 3 provides some model-free evidence to suggest a link between Books3 inclusion and LLM performance. Here, for each of the six models we study, we plot the relationship between the publication year of books in our sample and the accuracy of various large language models (LLMs). The analysis focuses on books split into two groups: those more likely to be included in the Books3 dataset (2011–2020) and those outside this window. Note we are not using our data on Books3 inclusion directly here, but rather simply the publication year. By comparing LLM performance across these publication windows, we aim to identify patterns and infer the influence of the Books3 dataset on model accuracy.

Panel (a) shows the chart for GPT-3.5 with the accuracy scores of the model on the y-axis and the publication years of books on the x-axis. Each dot here is the average accuracy for all books in a given publication year. Within the 2011–2020 window, where books are more likely to be in the Books3 dataset, there is a clear jump in model accuracy. Accuracy increases notably within this period, reaching its peak, before dropping sharply for books published outside this window (post-2020 and pre-2011). This pattern highlights a strong alignment between the dataset's content and the model's performance.

We see a similar pattern across all other charts with few variations. The pattern for Llama 8b is less pronounced compared to GPT-3.5 and GPT-4. While there is a slight improvement in accuracy during the 2011–2020 window, the increase is less dramatic, and the post-2020 decline is more moderate. This suggests a weaker reliance on the Books3 dataset. For Claude, Llama 70b and Gemini the results are similar. Across all figures, there is a clear pattern of improved accuracy for books published during the 2011–2020 window, likely reflecting the inclusion of these books in the Books3 dataset. We will now investigate this pattern quantitatively using a regression framework.

## 4.2 Baseline IV Estimates

Using the $Share_i$ as an IV, we estimate the first-stage of the IV as:

$$1(Books3)_{it} = \alpha + \beta \times Share_t + \Sigma_i t1(popularity)_i + \epsilon_i$$

where the outcome $1(Books3)_{it}$ is an indicator measuring Books3 membership for book $i$ published in year $t$ and $Share_t$ is the IV defined before. The estimated coefficient on $\beta$ is positive (7.7) and significant (t-stat 59) and the magnitude of the F-test value is strong (3,383.6), validating the relevance of the instrument as a key predictor of Books3 inclusion. This can also be see in Figure 2 Panel B.

Now, in the second stage, we estimate regressions where we predict the accuracy scores with the instrumented Books3 variable for each model separately. Results are presented in Table 2. The coefficients indicate the marginal effect of Books3 on model accuracy, and the statistical significance levels are provided. All regressions are based on 12,916 observations and coefficients are estimated with heteroskedasticity robust standard errors.

Across all but one model (Llama 8b), Books3 inclusion has a positive and significant effect on accuracy scores. For GPT3.5 Turbo and GPT 4o, the coefficients are are 0.0284 and 0.0231, respectively, suggesting a strong positive association with comprehension of the source material. Against their mean performance, this is about a 23.4% increase and 21.0% increase respectively. Llama 70b shows a smaller effect of about an 8.8% increase in performance relative to the baseline. In contrast, Llama 8b yields a near-zero estimate, indicating no measurable impact. The estimates for Claude and Gemini translate to a 7.3% and 6.9% effect respectively.

These findings collectively suggest that inclusion in Books3 significantly enhances name cloze performance for certain LLMs, particularly those developed by OpenAI (GPT models), while effects are weaker or negligible for smaller models like Llama 8B. Note that the non-GPT models (except Llama 8b) have significantly higher baseline name cloze performance (about 15% compared to 11%), but their sensitivity to Books3 is much smaller. For the smaller model we test (Llama 8b) the baseline performance is low, but the Books3 effect is negligible. This variation highlights model-specific dependencies on training data composition and potential differences in underlying architectures.

### 4.3 Robustness Checks

While the baseline results are instructive, this section presents robustness checks to validate them further. In particular, Table 3 presents results from three specific alternate models.

In the "OLS" columns, the model uses ordinary least squares (OLS) regression where the variable "Books3" is a binary indicator for whether a book's publication year falls between 2011 and 2020. This is meant to provide a regression analog for the data presented in Figure 3. These results provide a straightforward comparison to assess the impact of Books3 on accuracy without additional assumptions of IV. The "IV" columns replicate the baseline instrumental variable (IV) specification, testing whether the results hold under alternative setups. The first includes popularity fixed effects (coded into four categories based on the number of reviews on Goodreads), while the other incorporates these fixed effects as well as control variables such as the number of reviews, ratings, and average rating, also sourced from Goodreads. These adjustments aim to account for variations in book popularity and quality, ensuring that the observed effects are not driven by these factors.

Overall, the results across these models demonstrate the robustness of the relationship between Books3 and model accuracy. The coefficients in the OLS and IV models remain consistent in magnitude and statistical significance with the baseline estimates, confirming that the inclusion of Books3 during the training period is a significant predictor of LLM performance. The additional controls and fixed effects slightly reduce the size of the coefficients in some cases, suggesting that book popularity and related factors partially mediate the effect. However, the persistence of significant coefficients across specifications underscores the robustness of the original findings.

### 4.4 Heterogenous estimates

Finally, when looking at the baseline results, we are interested not just in the overall effects of Books3 inclusion, but also to examine whether these effects vary based on the inherent popularity of the book. Observers have suggested that the value of a particular source depends on the extent to which a particular source is a unique source of that content. For example, in its response to the NY Times lawsuit,[17] OpenAI claimed that many examples of "regurgitations" could be linked to articles where there were "multiple third-party" websites that had the same content. The overall conjecture is that model performance depends on the popularity of the underlying content. We test this idea directly in this analysis.

Specifically, in Figure 4, we present visual estimates from IV models similar to Table 2, except we estimate these on separate samples of books by the four popularity bins (0-10,10-100,100-1000 and 1000+) as split by the number of book reviews. We present these estimates (four per model) for each of the models in the six panels of Figure 4.

These findings indicate that the Book3 effect grows and is larger for less popular books as compared to more popular books. The natural explanation is that for popular books, the model likely encounters a wide range of substitutes in its training data like summaries, reviews, or other derivative texts. These substitutes may dilute the effect of having the original book in the dataset, as the model might learn redundant or less precise patterns. Less popular books with fewer substitutes, however, might allow the model to leverage the original text more effectively, as it provides unique, high-value input. Therefore for these less popular books the value of Books3 inclusion is higher. However for the most popular books, we find a *negative* effect of Books3 inclusion on performance. This effect is puzzling, and suggests that for these most popular books access to full-text might hurt rather than help performance, at least on name cloze tasks. Comparing results across models does not suggest substantial heterogeneity.

Combined, our results provide three takeaways. (a) Whether or not a book was included in the Books3 dataset has a causal effect on LLM performance, (b) the Books3 effect is significant across all model families,

[17] https://openai.com/index/openai-and-journalism/

although the magnitude varies and is negligible for the smaller models like LLAMA 8B and (c) the effects get stronger for less popular (and therefore more unique) content.

# 5 Discussion

Our study provides some of the first empirical evidence linking access to pirated digital content, specifically through the Books3 dataset, to enhanced performance of large language models (LLMs). Using a novel research design that leverages variation in book inclusion based on publication year, we demonstrate that direct access to the full text of copyrighted works significantly improves LLM performance on name cloze tasks. The effects are most pronounced for models like GPT-3.5 Turbo and GPT-4o, while smaller models, such as LLaMA 8B, exhibit negligible impacts. Additionally, we find that the Books3 effect diminishes for more popular works, likely due to the availability of substitutes in training data, underscoring the role of unique content in driving performance improvements.

These findings hold significant implications for ongoing debates surrounding copyright policy, digital piracy, and AI model development. We propose one test for content providers to examine the likely inclusion of their work in training data corpora. We also show how leading production models vary in their likely reliance on the Books3 dataset. Our work also holds implications for pricing models for training data, by helping link specific inputs to performance, although an exact dollar value is hard to determine. Finally, for the work in computer science on training data attribution, we show how natural experimental and applied microeconometric frameworks could be useful to evaluate production models. This could complement current methods which are largely on experimental models of smaller scale.

The preliminary results highlight key limitations and opportunities for future work. First, while informative, the name cloze test offers a relatively narrow measure of LLM performance and does not directly assess broader aspects of utility or comprehension common in benchmarks like MMLU (Hendrycks et al., 2020). Further, it remains unclear how this metric relates to the critical economic questions surrounding substitution and the potential for LLM outputs to cannibalize the market for original source material, as our test provides only an indirect measure of this elasticity. Including more tasks, such as summarization or critical analysis at the book level could provide further insights. Further, a key economic elasticity is the extent to which use of specific training data sources affect the market for the original work. Our paper does not speak to this issue, but future work could bring in additional data to shed some light. We hope to address these limitations in ongoing and future work.

Overall, by highlighting the measurable value of access to pirated works, our results call for a deeper evaluation of this topic with production models. Future work should explore broader economic and legal ramifications, as well as develop transparency standards for model training data, to ensure a balanced and

equitable approach to innovation and intellectual property in the digital era.

# 6 References


Akyürek, E., T. Bolukbasi, F. Liu, B. Xiong, I. Tenney, J. Andreas, and K. Guu (2022): "Towards tracing factual knowledge in language models back to the training data," *arXiv preprint arXiv:2205.11482*.

Authors Guild v. OpenAI Inc., e. a. (2023): vol. Case No. 1:23-cv-08292-SHS.

Azoulay, P., J. L. Krieger, and A. Nagaraj (2024): "Old Moats for New Models: Openness, Control, and Competition in Generative AI," Working Paper 32474, National Bureau of Economic Research.

Biasi, B. and P. Moser (2021): "Effects of copyrights on science: Evidence from the wwii book republication program," *American Economic Journal: Microeconomics*, 13, 218–260.

Bommasani, R., K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang (2023): "The Foundation Model Transparency Index," *arXiv preprint arXiv:2310.12941*.

Boussioux, L., J. N. Lane, M. Zhang, V. Jacimovic, and K. R. Lakhani (2024): "The Crowdless Future? Generative AI and Creative Problem-Solving," *Organization Science*, 35.

Brand, J., A. Israeli, and D. Ngwe (2023): "Using LLMs for Market Research," Working Paper 23-062, Harvard Business School Marketing Unit, available at SSRN.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020): "Language models are few-shot learners," *Advances in neural information processing systems*, 33, 1877–1901.

Brynjolfsson, E., D. Li, and L. R. Raymond (2023): "Generative AI at Work," Tech. Rep. 31161, National Bureau of Economic Research, Cambridge, MA.

Carlini, N., D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang (2022): "Quantifying Memorization Across Neural Language Models," *arXiv preprint*, arXiv:2202.07646.

Chang, K. K., M. Cramer, S. Soni, and D. Bamman (2023): "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4," in *Proceedings of EMNLP*.

Chiou, L. and C. Tucker (2017): "Content aggregation by platforms: The case of the news media," *Journal of Economics & Management Strategy*, 26, 782–805.

Cui, K. Z., M. Demirer, S. Jaffe, L. Musolff, S. Peng, and T. Salz (2024): "The Productivity Effects of Generative AI: Evidence from a Field Experiment with GitHub Copilot," Working Paper, we are providing

a preview of a project that analyzes two field experiments with 1,974 software developers at Microsoft and Accenture to evaluate the productivity impact of Generative AI.

de Rassenfosse, G., A. B. Jaffe, and J. Waldfogel (2025): "Intellectual property and creative machines," *Entrepreneurship and Innovation Policy and the Economy*, 4, 47–79.

Dell'Acqua, F., E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (2023): "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Tech. Rep. 24-013, Harvard Business School, Boston, MA.

Dhingra, B., K. Mazaitis, and W. W. Cohen (2017): "Quasar: Datasets for Question Answering by Search and Reading," *arXiv preprint*, arXiv:1707.03904v2 [cs.CL].

Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner (2021): "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," *arXiv preprint arXiv:2104.08758*.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. (2024): "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*.

Elazar, Y., N. Kassner, S. Ravfogel, A. Feder, A. Ravichander, M. Mosbach, Y. Belinkov, H. Schütze, and Y. Goldberg (2023): "Measuring Causal Effects of Data Statistics on Language Model's 'Factual' Predictions," .

Gao, L., S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. (2020): "The pile: An 800gb dataset of diverse text for language modeling," *arXiv preprint arXiv:2101.00027*.

Getty Images (US) Inc. v. Stability AI, I. (2023): .

Granite Team, I. (2024): "Granite 3.0 Language Models," .

Henderson, P., X. Li, D. Jurafsky, T. Hashimoto, M. A. Lemley, and P. Liang (2023): "Foundation models and fair use," *Journal of Machine Learning Research*, 24, 1–79.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2020): "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*.

Jaffe, S., N. P. Shah, J. Butler, A. Farach, A. Cambon, B. Hecht, M. Schwarz, and J. Teevan (2024): "Generative AI in Real-World Workplaces," Technical Report MSR-TR-2024-29, Microsoft Research.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020): "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*.

Kenton, J. D. M.-W. C. and L. K. Toutanova (2019): "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, Minneapolis, Minnesota, vol. 1.

Lemley, M. (2024): "How generative AI turns copyright law upside down," *Science and Technology Law Review*, 25.

Lemley, M. A. and B. Casey (2020): "Fair learning," *Tex. L. Rev.*, 99, 743.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025): "Large language models: An applied econometric framework," Tech. rep., National Bureau of Economic Research.

Moser, P. (2005): "How do patent laws influence innovation? Evidence from nineteenth-century world's fairs," *American economic review*, 95, 1214–1236.

Mostafazadeh, N., N. Chambers, A. Louis, and M. Roth (2017): "LSDSem 2017 Shared Task: The Story Cloze Test," in *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, Association for Computational Linguistics, 46–51.

Nagaraj, A. (2018): "Does copyright affect reuse? Evidence from Google Books and Wikipedia," *Management Science*, 64, 3091–3107.

Nagaraj, A. and I. Reimers (2023): "Digitization and the Market for Physical Works: Evidence from the Google Books Project," *American Economic Journal: Economic Policy*, 15, 428–458.

Nasr, M., N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee (2023): "Scalable extraction of training data from (production) language models," *arXiv preprint arXiv:2311.17035*.

New York Times Company v. Microsoft Corp., e. a. (2023): vol. Case No. 1:23-cv-11195.

Noy, S. and W. Zhang (2023): "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Preprint*, submitted March 6.

Oberholzer-Gee, F. and K. Strumpf (2007): "The Effect of File Sharing on Record Sales: An Empirical Analysis," *Journal of Political Economy*, 115.

Onishi, T., H. Wang, M. Bansal, K. Gimpel, and D. McAllester (2016): "Who did what: A large-scale person-centered cloze dataset," *arXiv preprint arXiv:1608.05457*.

Park, S. M., K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry (2023): "Trak: Attributing model behavior at scale," *arXiv preprint arXiv:2303.14186*.

Petroni, F., T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel (2019): "Language Models as Knowledge Bases?" in *Proceedings of EMNLP 2019*.

Radford, A. (2018): "Improving language understanding by generative pre-training," .

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019): "Language models are unsupervised multitask learners," *OpenAI blog*, 1, 9.

Reimers, I. (2016): "Can private copyright protection be effective? Evidence from book publishing," *The journal of law and economics*, 59, 411–440.

Rob, R. and J. Waldfogel (2006): "Piracy on the high C's: Music downloading, sales displacement, and social welfare in a sample of college students," *The Journal of Law and Economics*, 49, 29–62.

Samuelson, P. (2023): "Fair use defenses in disruptive technology cases," *UCLA Law Review, Forthcoming*.

Smith, M. D. and R. Telang (2009): "Competing with free: The impact of movie broadcasts on DVD sales and Internet piracy," *Mis Quarterly*, 321–338.

Taylor, W. L. (1953): "Cloze Procedure: A New Tool for Measuring Readability," *Journalism Quarterly*, 30, 415–433.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023): "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*.

Tranchero, M., C.-F. Brenninkmeijer, A. Murugan, and A. Nagaraj (2024): "Theorizing with large language models," Tech. rep., National Bureau of Economic Research.

Waldfogel, J. (2012): "Copyright protection, technological change, and the quality of new products: Evidence from recorded music since Napster," *The journal of law and economics*, 55, 715–740.

Watson, J. (2017): "What is the value of re-use? complementarities in popular music," *Complementarities in Popular Music (September 1, 2017) .NET Institute Working Paper*.

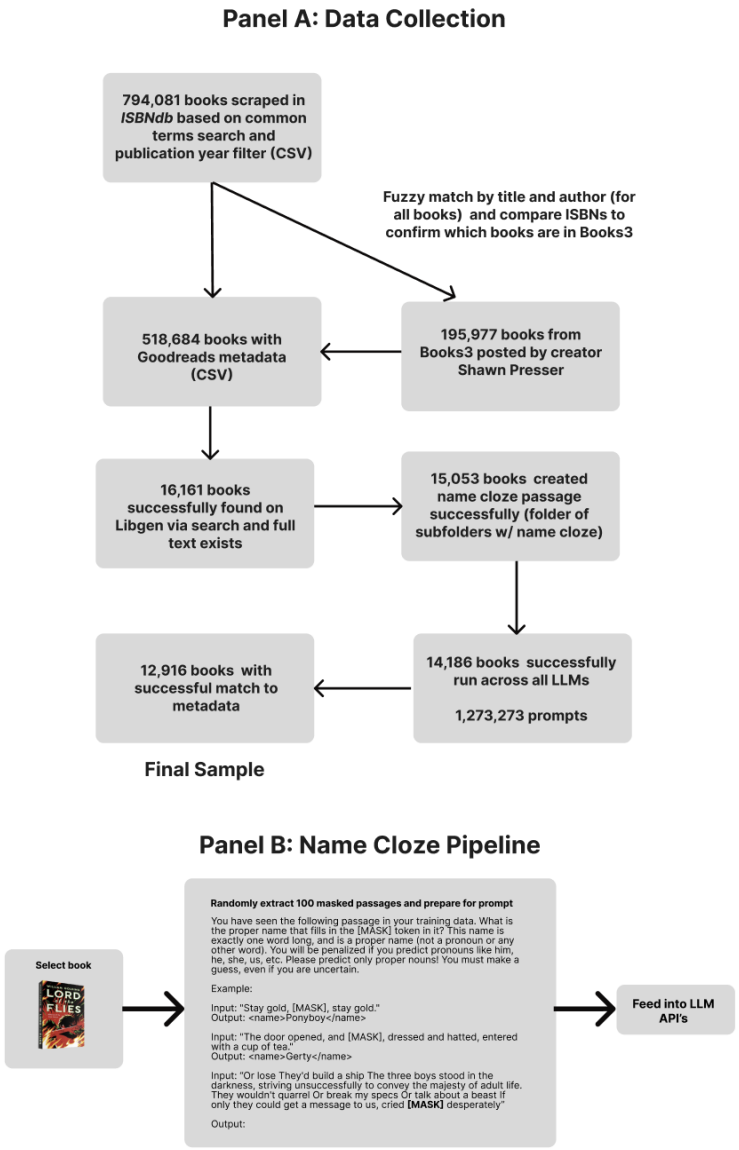
Yang, S. A. and A. H. Zhang (2024): "Generative ai and copyright: A dynamic perspective," *arXiv preprint arXiv:2402.17801*.

Zhou, E. and D. Lee (2024): "Generative artificial intelligence, human creativity, and art," *PNAS Nexus*, 3, pgae052.

Zhu, Y., R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015): "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 19–27.

# 7 Tables and Figures

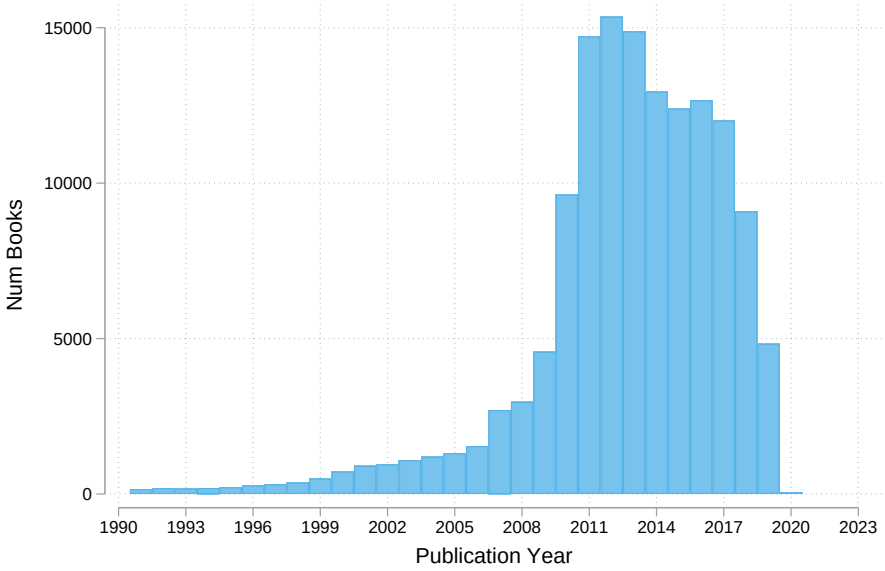
Figure 1: Data Collection and Experiment Pipeline



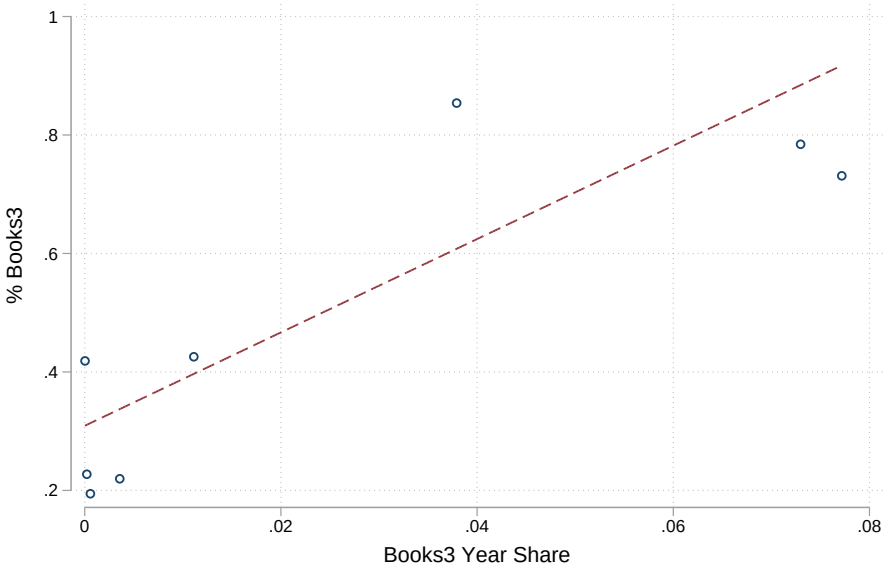

**Note:** Panel A provides a detailed overview of the data collection process for the treatment group (Books3) and control group (Non-Books3), covering the collection of metadata and full text files. Panel B presents an example passage from a book, illustrating how it is sent to the API along with the exact prompt.

Figure 2: Books3 Data Distribution by Publication Year

(a) Books3 Publication Years (195,997 books)

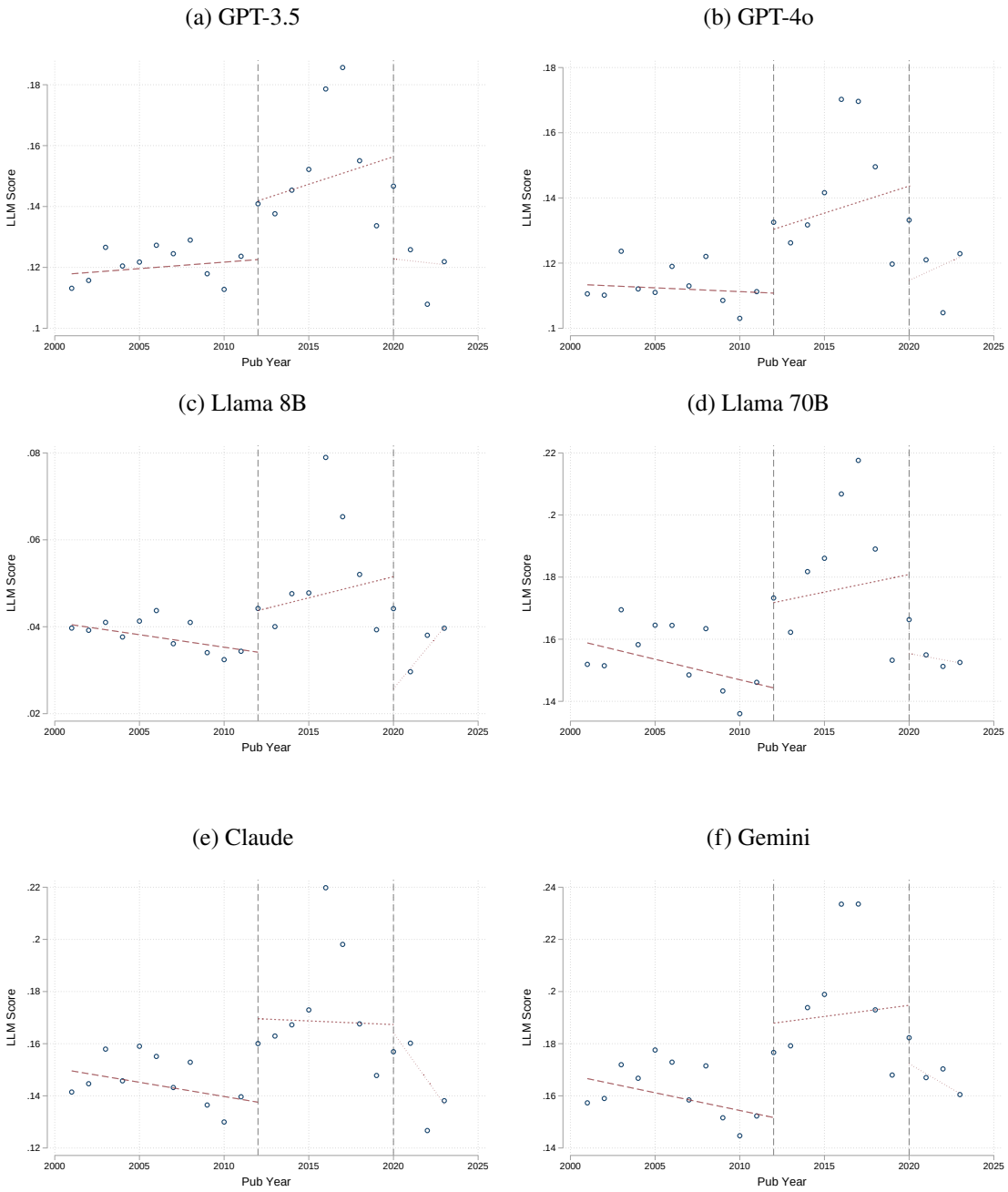



(b) Books3 Shares vs Incidence in Our Sample



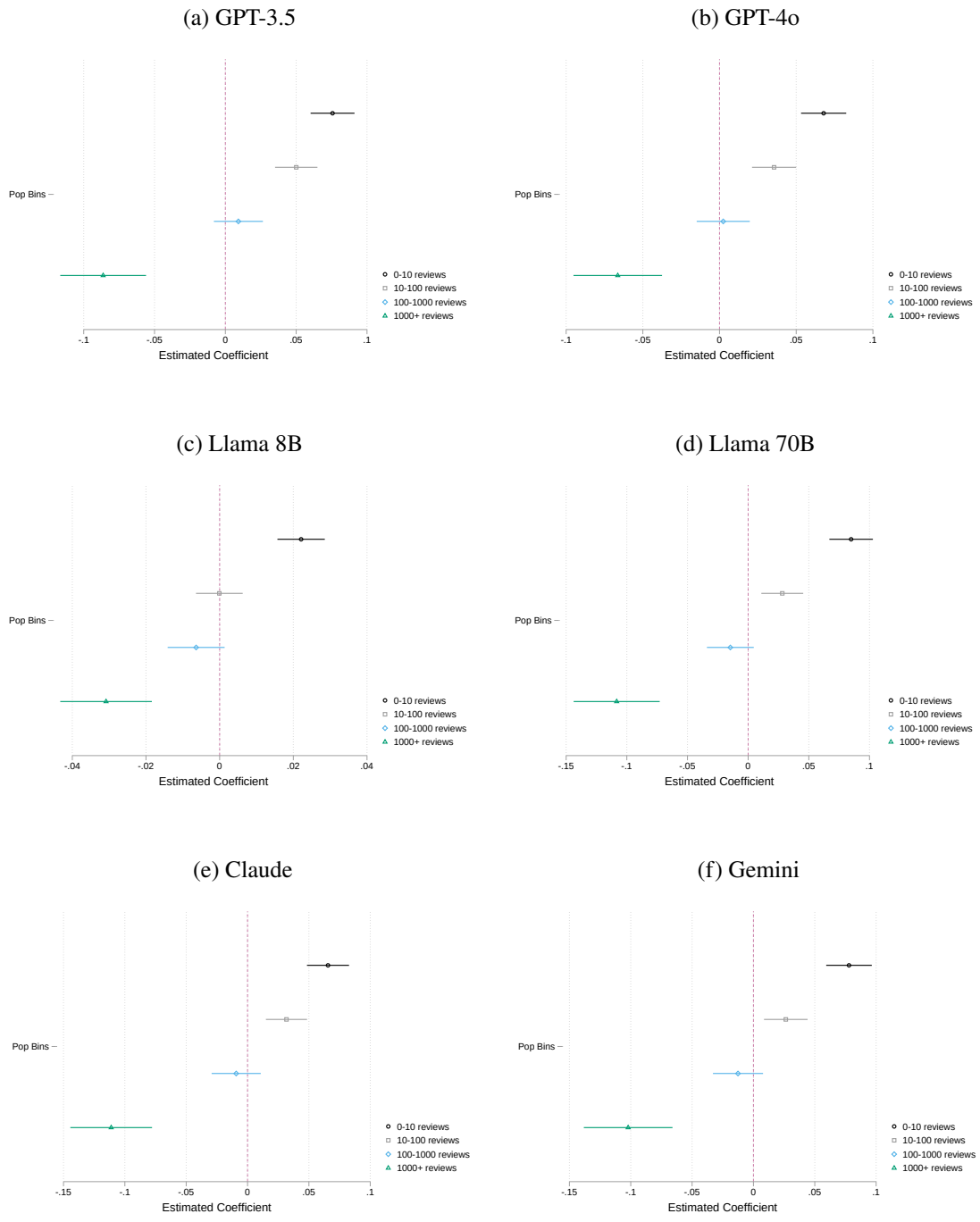

**Note:** Panel A provides a histogram of publication years in the full Books3 sample. Panel B provides a correlation between the instrument and the key endogenous variable, the dummy for Books3 in our sample.

Figure 3: Model-free Evidence: Publication Year vs Accuracy across LLMs




**Note:** This figure provides model-free evidence for the relationship between Books3 inclusion and model performance. We present binned scatter plots, where average performance from books from a given publication year are plotted for each model. We present fitted lines for books in the 2011-2020 period (with high Books3 representation) as compared to books outside of this window.

Figure 4: Books3 Effect By Book Popularity




**Note:** This figure summarizes estimates from 24 IV regressions estimating the effect of Books3 inclusion for four different popularity bins for each model in separate samples.

Table 1: Summary Statistics

| Statistic | Mean | St. Dev. | Median | Min | Max | N |
|---|---|---|---|---|---|---|
| **Book Characteristics** | | | | | | |
| Books3 | 0.49 | 0.50 | 0 | 0 | 1 | 12,916 |
| Total ratings | 15,515.35 | 154,672.60 | 87 | 0 | 6,147,566 | 12,916 |
| Total reviews | 656.67 | 4,956.55 | 9 | 0 | 178,631 | 12,916 |
| Average rating | 3.55 | 1.16 | 3.87 | 0.00 | 5.00 | 12,916 |
| Publication year | 1,989.16 | 48.97 | 2,004 | 1,000 | 2,023 | 12,886 |
| Fiction | 0.29 | 0.45 | 0 | 0 | 1 | 9,562 |
| 0-10 ratings | 0.24 | 0.43 | 0 | 0 | 1 | 12,916 |
| 10-100 ratings | 0.28 | 0.45 | 0 | 0 | 1 | 12,916 |
| 100-1000 ratings | 0.25 | 0.43 | 0 | 0 | 1 | 12,916 |
| 1000+ ratings | 0.23 | 0.42 | 0 | 0 | 1 | 12,916 |
| **LLM Results** | | | | | | |
| GPT 3.5 Turbo | 0.12 | 0.11 | 0.09 | 0.00 | 1.00 | 12,916 |
| GPT 4o | 0.11 | 0.11 | 0.08 | 0.00 | 1.00 | 12,916 |
| Claude | 0.14 | 0.12 | 0.12 | 0.00 | 1.00 | 12,916 |
| Llama 8B | 0.04 | 0.05 | 0.02 | 0.00 | 0.67 | 12,916 |
| Llama 70B | 0.15 | 0.13 | 0.12 | 0.00 | 1.00 | 12,916 |
| Gemini | 0.16 | 0.13 | 0.13 | 0.00 | 1.00 | 12,916 |

**Note:** This table presents summary statistics for the main data samples used in this study. For each book, we collected data from Goodreads on various characteristics like total number of ratings, total number of reviews, etc. Roughly half of the books in the sample are in Books3. Only 9,562 of the 12,916 in the sample have genre listed and of the 9,562 close to 70% are nonfiction.

Table 2: Baseline IV Regressions

| | (1) GPT3.5 Turbo | (2) GPT 4o | (3) Llama 8b | (4) Llama 70b | (5) Claude | (6) Gemini |
|---|---|---|---|---|---|---|
| Books3 | 0.0284*** | 0.0231*** | 0.000796 | 0.0133*** | 0.0104** | 0.0110** |
| | (0.00437) | (0.00422) | (0.00183) | (0.00502) | (0.00492) | (0.00521) |
| N | 12916 | 12916 | 12916 | 12916 | 12916 | 12916 |
| F-Stat | 3890.6 | 3890.6 | 3890.6 | 3890.6 | 3890.6 | 3890.6 |

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

**Note:** This table presents estimates from the main IV regression as described in the text at the book level. The main outcome variable is the average accuracy score on the cloze task at the book level for each of the six models studied in the paper. The F-Stat presented is from the first stage regression where Books3 membership for book from publication year $t$ is instrumented with the share of Books3 books from year $t$ of the total number of books in that dataset. Robust standard errors are presented.

Table 3: Robustness Checks

Panel A

| | GPT3.5 | | | GPT4o | | |
|---|---|---|---|---|---|---|
| | OLS | IV | IV | OLS | IV | IV |
| Books3 | 0.0307*** | 0.0292*** | 0.0296*** | 0.0290*** | 0.0227*** | 0.0226*** |
| | (0.00357) | (0.00451) | (0.00448) | (0.00343) | (0.00438) | (0.00434) |
| N | 12916 | 12916 | 12916 | 12916 | 12916 | 12916 |
| Pop. FE | No | Yes | Yes | No | Yes | Yes |
| Ctrls | No | No | Yes | No | No | Yes |

Panel B

| | Llama 8b | | | Llama 70b | | |
|---|---|---|---|---|---|---|
| | OLS | IV | IV | OLS | IV | IV |
| Books3 | 0.0108*** | 0.000338 | 0.000524 | 0.0309*** | 0.0149*** | 0.0153*** |
| | (0.00157) | (0.00192) | (0.00191) | (0.00398) | (0.00518) | (0.00516) |
| N | 12916 | 12916 | 12916 | 12916 | 12916 | 12916 |
| Pop. FE | No | Yes | Yes | No | Yes | Yes |
| Ctrls | No | No | Yes | No | No | Yes |

Panel C

| | Claude | | | Gemini | | |
|---|---|---|---|---|---|---|
| | OLS | IV | IV | OLS | IV | IV |
| Books3 | 0.0271*** | 0.0122** | 0.0122** | 0.0330*** | 0.0138** | 0.0141*** |
| | (0.00388) | (0.00507) | (0.00506) | (0.00422) | (0.00536) | (0.00532) |
| N | 12916 | 12916 | 12916 | 12916 | 12916 | 12916 |
| Pop. FE | No | Yes | Yes | No | Yes | Yes |
| Ctrls | No | No | Yes | No | No | Yes |

Standard errors in parentheses

$^{+}$ $p < 0.15$, $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

**Note:** This table presents robustness checks for the main results. The OLS column presents results from an OLS model where "Books3" is a dummy variable that equals one if the published year is between 2011 and 2020. In the IV columns, the specification is same as the baseline. "Pop FE" indicates popularity fixed effects where popularity is coded as a categorical variable that takes 4 values based on the number of reviews on Goodreads. Controls include: number of reviews, number of ratings and average rating as sourced from Goodreads. Robust standard errors.