Cloze Encounters:

The Impact of Pirated Data Access on LLM Performance *

Stella Jia
UC Berkeley
stellajia@berkeley.edu

Abhishek Nagaraj
UC Berkeley & NBER
nagaraj@berkeley.edu

December 17, 2024

Preliminary: Do Not Cite

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 the performance of LLMs. We test the performance of leading foundation models (GPT3.5, GPT4, Claude and LLaMA) 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 and employ an instrumental variables strategy that instruments for Books3 inclusion via publication year. We find significant improvements in LLM performance 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.

*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. Any opinions and conclusions expressed herein are those of the authors only, and any errors are our own.

1 Introduction

The rise of generative AI and foundation models have captured the attention of consumers, businesses and policymakers alike. The ability of these models to work with structured and unstructured data, combined with their accessibility to users without technical expertise leads to a lot of value. For example, LLMs have shown to significantly enhance and supplement human capabilities in writing, idea generation, programming, images etc (Noy and Zhang, 2023; Boussioux et al., 2024). Further, in a business context, these have shown to aid in customer support (Brynjolfsson et al., 2023) and enhance productivity of knowledge workers (Dell'Acqua et al., 2023).

Despite their positive economic impacts on users, their reliance on data, especially copyrighted content raises significant legal and economic questions. There are three related issues. First, many copyright owners are up in arms about the unauthorized use of their content. For instance, The New York Times filed a lawsuit against OpenAI for using its content without permission, providing numerous examples of New York Times content being reproduced verbatim by ChatGPT.[1] How the potential use of such copyrighted content affects LLM performance is hard to determine. Second, while some copyright holders are suing foundation model developers over the unlicensed use of their material, many other content providers are willing to license content to foundation model developers. For example, Reddit has partnered with OpenAI and Google to grant access to its content for a reported $60 million/year, enabling their models to stay current on recent topics.[2][3] Insofar as such licensing is set to continue, determining appropriate compensation for specific training data and understanding the relationship between data access and model performance remain challenging. Third, from a forensic point of view, while it is clear that previous generations of these models were trained on copyrighted and pirated material,[4] the lack of transparency around training data (even in more "open" models like Meta's LLAMA) have led to intense speculation about which content these models are trained on, and what content is excluded. This uncertainty and lack of transparency has fueled significant legal disputes in the field and created challenges of economic and copyright policy seeking to balance the rights of content creators for fair compensation and those of developers for fair use.

In order to tackle these questions, we need better empirical evidence on how access to copyrighted content might influence models in real-world settings. Despite significant research in computer science on the abilities of foundation models, we lack concrete evidence on the role of access to training data on model performance. To what extent are models using pirated and copyrighted material in their training and how

[1] Other high-profile legal actions in this domain include Getty Images v. Stability AI, the Authors Guild v. OpenAI, and Kadrey v. Meta, among others.

[2] https://openai.com/index/openai-and-reddit-partnership/

[3] https://www.cbsnews.com/news/google-reddit-60-million-deal-ai-training/

[4] https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/

does access to such information affect model performance?

From a theoretical point of view it is not clear how access to any specific piece of copyrighted content affects model performance. On one hand, model creators claim that any one source of content has very little bearing on performance and so their willingness to pay for such material is low. Further, direct access might not be required because content is often replicated or summarized in different sources of training data–this would mean that even without direct access, a model is able to learn about a particular source in a reliable way (Carlini et al., 2022; Chang et al., 2023). On the other, 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). What is needed are concrete 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. Unlike past work in copyright policy that uses open data and citations to link the use of data to performance (Nagaraj and Reimers, 2023), there exists no data or research design that helps answer this question easily. This is the central challenge that we tackle in this project.

Understanding the relationship between a model's access to training data and its ability to learn from it is a complex challenge. First, model developers rarely disclose the specific sources used during training, making it nearly impossible to determine what data was included. Additionally, systematically quantifying a model's knowledge of a particular source is inherently difficult. Even if we could measure 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 challenges. First to measure the performance of a model as it relates a particular piece of content, we first build on a recent technique from the machine learning literature (Chang et al., 2023; Petroni et al., 2019) that helps us quantify to what extent a model is able to learn about a given source. Specifically, this method called "name cloze" tests a model's ability to answer factual questions about a given book based on missing information. The "name cloze" test originates from a common test of comprehension in psychology known as the Cloze procedure (Taylor, 1953). This metric allows us to reliably compare a model's knowledge about any given book and is our measure of model performance.

Next, we leverage a notable piracy event to examine the impact of easier access to a large number of

copyright books on a model's ability to perform name cloze tasks about that particular publication. To find such exogenous variation, we rely on our institutional knowledge of the availability of training data in the recent past. In particular, we examine a specific dataset "Books3" that was made available by a single individual who bundled about 200k copyrighted books (in an act of benevolent piracy) for the express purpose of training LLMs.[5] 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 us with a potential natural experiment to examine the effect of ease of access on a model's performance as measured by the name cloze procedure. In particular, we build a sample of about 12,000 books (of which about 6000 were included in Books3) and use the year of publication in an instrumental variables strategy to examine the causal effect of access to the ability of LLMs to learn about training data. We examine how inclusion in the Books3 dataset affects model performance across three sets of leading models (GPT class models by OpenAI, Claude by Anthropic 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 through inclusion in the Books3 dataset significantly enhances name cloze task performance. The most pronounced improvements are observed in GPT-3.5 Turbo and GPT-4.0, compared to the Claude 3.5 Haiku model and LLaMA. The "Books3" effect is about a 20% increase in accuracy for GPT class models as compared to 6% for Claude and 8% for LLaMA. Interestingly, we find no detectable Books3 effect for the smaller LLaMA model (LLaMA-8B) as compared to its larger counterpart (LLaMA-70B). When looking across popularity we were interested in evaluating how the prevalence of substitutes for a work influences accuracy, based on the hypothesis that a higher number of substitutes would diminish the significance of the Books3 effect. Our findings indicate that the Books3 effect is much smaller for 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. [6]

While a significant body of work has looked at the economics of copryright and piracy in digital settings Oberholzer-Gee and Strumpf (2007), 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 will help to further research in this area.

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

[6] 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.

In parallel, while the cloze method has been used in prior computer science research to assess comprehension (Petroni et al., 2019; Mostafazadeh et al., 2017; Dhingra et al., 2017) in synthetic experiments, we apply it to a real-world policy and economic settings. By using a dataset under legal scrutiny (Books3) and applying the name cloze method, we investigate the relationship between input and output in large language models (LLMs) on a specific source to assess model performance. We also provide a new research design that uses the IV approach on the Books3 dataset, to provide causal estimates of data access on model performance.

Our work also has direct implications for ongoing legal and policy debates. Specifically, our findings provide empirical evidence that contradicts arguments that the use of copyrighted material during LLM training has a negligible impact on model performance. The observed positive effect of direct access to copyrighted content, even when controlling for substitution, underscores the importance of considering the value of original works in AI training. This has significant ramifications for the pricing of access to training data, particularly for less popular content with fewer substitutes. Our results also provide a framework for establishing and understanding the economic importance of training data to model performance that would be useful in adjudicating these disputes in other settings and help inform the development of copyright and related policies in the age of generative AI.

2 Empirical Setting

2.1 Background

LLMs rely on vast amounts of data, yet there is significant variation across models in their access to public domain and copyrighted materials (Granite Team, 2024; Radford et al., 2019). However, the extent of this variation remains unclear, as many closed-source models do not disclose the specifics of their training datasets. Despite the finite amount of data available on the internet, the growing interest in generative AI has intensified the search for more content. This pursuit is largely limited to digitized materials, as these are the primary sources for training such models. Common examples include news articles, social media posts, and large aggregated datasets like Common Crawl and The Pile. Recognizing this trend, knowledge institutions housing physical works are engaging in digitization efforts to contribute to AI development. [7]

The rising legal contentions between creators and developers highlights the importance of knowing what is in these models. Due to the uncertainty around what is in the training data, the common evidence of accusation points to 1) once open datasets like Books3 and 2) prompting the model to output content that is niche to a specific piece of copyrighted work. For models like LLaMa, which have confirmed the use of Books3 in training, this provides concrete legal evidence of using copyrighted material. In other legal

[7] https://institutionaldatainitiative.org/hello-world.html

cases, creators have pointed to an LLM's ability to mimic or summarize their work, not found elsewhere, as evidence of training on their books (e.g., Author Guild v. OpenAI).

For many creators, a core concern is the future economics of the creator industry. AI has the potential to enter the market (e.g. Amazon seeing a spike in AI-generated books) and is able to generate content with the characters and settings of prior work and turn it into new things. This poses a threat to authors. However it remains unclear to what extent AI would have on impacting the author market.

Additionally, determining the extent of infringement to preserve innovation in the face of emerging technologies has repeatedly been a difficult issue. In the face of GAI, there are some traits that are different to prior technologies like its ability to create new work using existing IP with limited control from the user of the tool. Due to the uncertainty around the technical and legal space of GAI, we need a framework to reason about its interaction with copyright.

2.2 Books3 and Testing Performance

Our study focuses on the contentious Books3 dataset, which contains approximately 200,000 books, many of them copyrighted. Books3 is one of the few large-scale text corpora of books, following earlier datasets such as BookCorpus (Zhu et al., 2015), Books1, and Books2. Notably, Books3 has been cited in multiple lawsuits and confirmed to be used in some models, though its inclusion in others remains ambiguous. Earlier corpora, including Books1 and Books2, have been referenced in the development of OpenAI's GPT models (Brown et al., 2020). Due to escalating legal concerns, these datasets are no longer publicly accessible. For our study, we obtained the full list of books in Books3 from a source who accessed the dataset prior to its restriction. This dataset serves as the foundation for our experiment.

We leverage a cloze procedure to assess how well an LLM knows a particular book. Cloze tests originated as a way to measure language comprehension (Taylor, 1953). 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. We build on this concept of "closure" through a "name cloze test" (Chang et al., 2023) where for each book we randomly select 100 passages and mask a name token. The motivation is that the LLM should only be able to know the right answer if the direct text of the book is in the training data. We use the number of correct passages of the 100 as a measure of how well the LLM knows a book.

3 Data and Research Design

3.1 Data

The data creation procedure involved integrating two primary datasets—Books3 and Non-Books3—to construct a comprehensive book corpus for analysis. The process began by obtaining Books3 metadata from a journalist with access to the dataset on metadata for Books3. To enrich metadata, HTML book profiles for the 195,000 titles were scraped from Goodreads, providing additional descriptive fields such as titles, authors, and ISBNs.

To extend the corpus, Non-Books3 data comprising 400,000 additional books was generated using a multi-step approach. Keywords, filtered by publication year, were searched on ISBNdb to retrieve candidate titles, authors, and ISBNs. These entries were cross-validated to exclude overlap with the Books3 dataset. Metadata for 500,000 books was further scraped from Goodreads and systematically organized into a structured CSV file. A subset of these titles (12,600 books) was then sampled, with text files downloaded from Libgen, yielding two key datasets that we combined: Non-Books3 sample data (6,700 books) and Books3 sample data (5,900 books).

We then implemented the name-cloze method, for which we need access to the full text of a book. This pipeline is designed to evaluate language models' ability to predict proper names in masked passages extracted from literary texts. The process begins with selecting a book, such as "Lord of the Flies", and randomly extracting 100 masked passages where a single proper name is replaced with a placeholder. These masked passages are then converted into prompts that explicitly instruct the model to predict the proper name based on the surrounding context. The prompt emphasizes that the correct prediction must be a proper noun, consisting of a single word, and not a pronoun or generic term. To ensure compliance, the instructions penalize predictions like pronouns (e.g., he, she, us) or other ambiguous terms. Examples provided include sentences where masked names like "Ponyboy" or "Gerty" must be inferred. The prepared prompts are then sent to large language model APIs, which return predictions for the masked names. This pipeline enables a systematic evaluation of a model's capability to identify and generate contextually appropriate proper names, offering insights into its understanding of narrative and character-specific information. For each book we note the accuracy of the response and focus on the average as our key dependent variable.

The summary statistics of our final sample is in Table 1. We have a roughly even distribution of Books3 and out of Books3. The raw average scores of each of the models we tested on are in the second panel in Table 1 where smaller models LLaMA 8B have the lowest overall average while LLaMA 70B and Claude have the highest. Most of the books in this sample are Nonfiction.

3.2 Research Design

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.

An approach we took to isolate the effect of Books3 inclusion is using the natural pattern present in the publication year of books in Books3. Books3 contains a spike in books with publication year between 2014 to 2018. We leverage this spike to create an shares instrumental variable which involves taking the proportion of Books3 books that exist in a publication year. The spike is unique to Books3 as seen in Figure 2 where we see that the distribution of books not in Books3 is shifted to the left. Additionally, the F-statistics for our instrument in Table 3 is over 3,000. We implement this in our design through an approach similar to the shift shares instrumental variable (Borusyak et al., 2024). Using this IV approach, we can then estimate the causal effect of Books3 inclusion on name cloze accuracy.

$$Share_i = \frac{\text{\# of books in Books3 for publication year of } i}{\text{\# of books in Books3}}$$

$$Y_i = \alpha + \beta \times (IsBooks3_i \sim Share_i) + \epsilon_i$$

4 Results

We first examine the effect of Books3 inclusion on name cloze accuracy in Table 4 which present estimates from instrumental variable (IV) regressions examining the effect of inclusion in the *Books3* dataset on the performance of five different large language models (LLMs). Across all models, Books3 inclusion shows varying levels of significance. For GPT3.5 Turbo and GPT 4o, the coefficients are highly significant at the 1% level, with values of 0.0214 and 0.0234, respectively, suggesting a strong positive association with comprehension of the source material. Against a baseline of about 0.11, this is over a 20% increase in performance. The Claude model also exhibits a positive effect, with a coefficient of 0.0083, significant at the 10% level. In contrast, LLaMA 8B yields a near-zero estimate (0.0003), indicating no measurable impact, while LLaMA 70B shows a smaller but statistically significant effect (0.0124) at the 1% level.

The consistency of the F-test value (3,846.8) across all regressions underscores the strength of the instru-

mental variable approach, validating the relevance of the instrument. These findings collectively suggest that inclusion in Books3 significantly enhances comprehension performance for certain LLMs, particularly those developed by OpenAI (GPT models), while effects are weaker or negligible for others like LLaMA 8B. This variation highlights model-specific dependencies on training data composition and potential differences in underlying architectures.

When looking across popularity in Table 4 we are interested in evaluating how the prevalence of substitutes for a work influences accuracy, based on the hypothesis that a higher number of substitutes would diminish the significance of the Books3 effect. Our findings indicate that the Book3 effect grows and is larger for less popular books as compared to more popular books. A potential 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.

Overall 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, 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 robust 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. By highlighting the measurable value of access to original works, our results call for a reevaluation of compensation frameworks for content creators in the age of generative AI. 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.

The preliminary results highlight key limitations and opportunities for future work. The current comprehension metric, based on 100 masked passages, may not fully capture LLM performance. Expanding tasks, such as summarization for Nonfiction or distinguishing knowledge of main versus minor characters in Fiction, could provide more nuanced insights. We hope to address these limitations in ongoing and future work.

6 References

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.

Borusyak, K., P. Hull, and X. Jaravel (2024): "A Practical Guide to Shift-Share Instruments," Working Paper 33236, National Bureau of Economic Research.

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.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, and D. Amodei (2020): "Language Models are Few-Shot Learners," *arXiv preprint*.

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*.

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].

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. 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.

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.

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*.

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

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

Table 1: Summary Statistics

Statistic	Mean	St. Dev.	Median	Min	Max	N
Book Characteristics						
Books3	0.465	0.499	0	0	1	11,896
Total ratings	2,222.986	8,452.635	68	0	99,856	11,896
Total reviews	174.772	658.501	7	0	12,810	11,896
Average rating	3.476	1.244	3.850	0.000	5.000	11,896
Publication year	1,994.618	22.873	2,005	1,788	2,020	11,896
Fiction	0.127	0.334	0	0	1	8,653
Nonfiction	0.873	0.334	1	0	1	8,653
LLM Results						
GPT 3.5 Turbo	0.110	0.097	0.083	0.000	0.800	11,896
GPT 4o	0.109	0.101	0.080	0.000	0.750	11,896
Claude	0.143	0.117	0.120	0.000	0.806	11,896
LLaMA 8B	0.035	0.042	0.020	0.000	0.410	11,896
LLaMA 70B	0.148	0.120	0.120	0.000	0.800	11,896

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 8,653 of the 11,896 in the sample have genre listed and of the 8,653 close to 90% are nonfiction.

Table 2: Popularity Distribution

	Popularity	Total Ratings	Count
1	Very High	1000+	2,319
2	High	100-1000	2,921
3	Medium	10-100	3,471
4	Low	0-10	3,185

Note: This table presents the distribution of popularity, measured using the total number of ratings on Goodreads as a proxy. The distribution is roughly even across each popularity category.

Table 3: IV Regression Results

Group	GPT3.5 Turbo	GPT 4o	Claude	LLaMA 8B	LLaMA 70B
	(1)	(2)	(3)	(4)	(5)
Variables					
Books3	0.0214***	0.0234***	0.0083*	0.0003	0.0124***
	(0.0038)	(0.0039)	(0.0045)	(0.0016)	(0.0045)
Fit statistics					
F-test (1st stage), Books3	3,846.8	3,846.8	3,846.8	3,846.8	3,846.8

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: This table presents estimates from IV models run on five different LLMs evaluating the overall impact of inclusion in training data on comprehension of the source. The F-test is also included to evaluate the relevance of our IV.

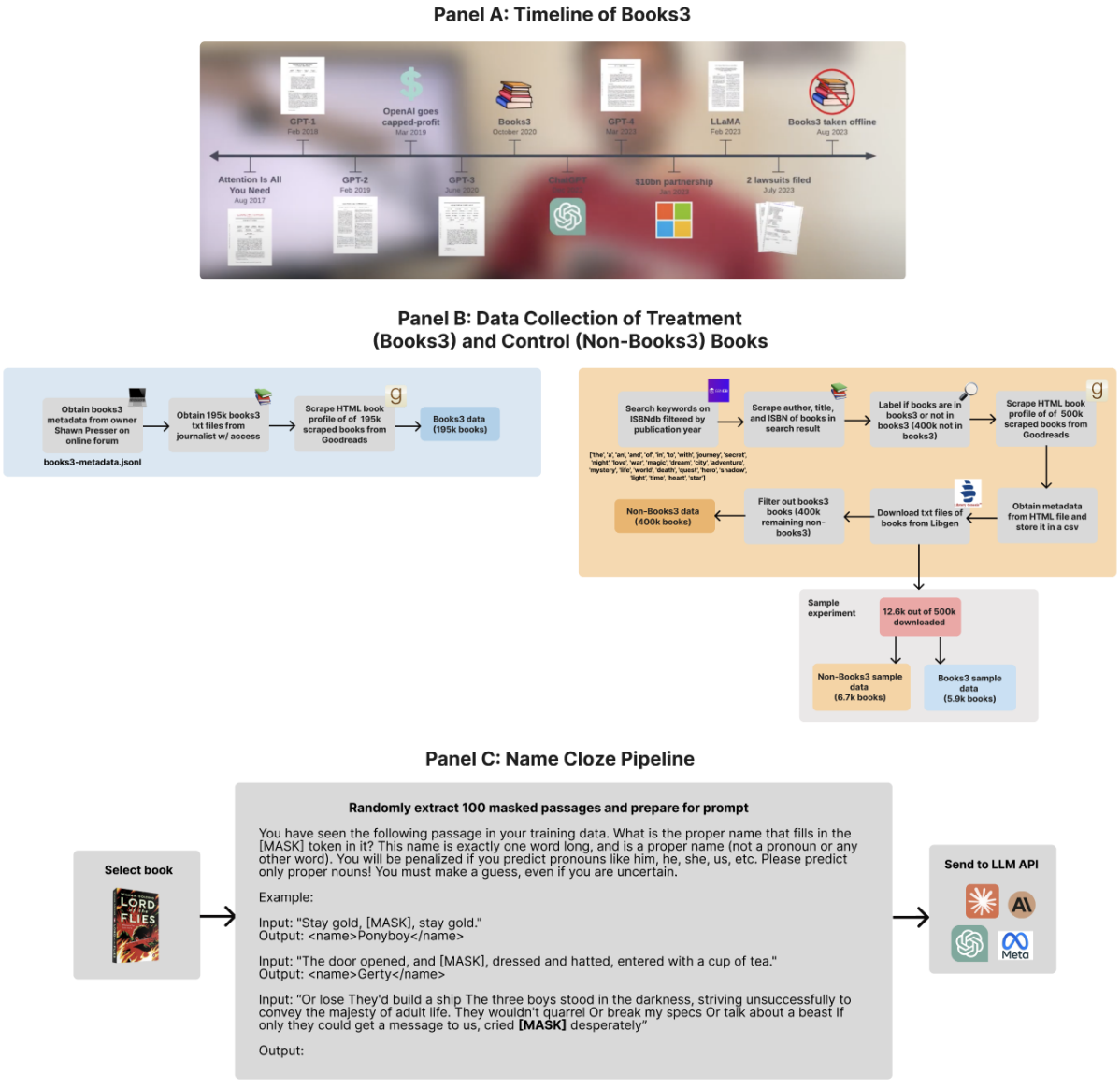
Table 4: IV of Books3 and Popularity Regression Results (base is Very High)

Group	GPT3.5 Turbo	GPT 4o	Claude	LLaMA 8B	LLaMA 70B
	(1)	(2)	(3)	(4)	(5)
Variables					
Books3	-0.0592***	-0.0536***	-0.0785***	-0.0248***	-0.0799***
	(0.0131)	(0.0135)	(0.0157)	(0.0054)	(0.0163)
High x Books3	0.0575***	0.0489***	0.0583***	0.0163**	0.0584***
	(0.0152)	(0.0158)	(0.0184)	(0.0064)	(0.0188)
Low x Books3	0.1333***	0.1339***	0.1476***	0.0489***	0.1701***
	(0.0150)	(0.0155)	(0.0177)	(0.0062)	(0.0184)
Medium x Books3	0.0962***	0.0879***	0.1117***	0.0224***	0.1064***
	(0.0151)	(0.0155)	(0.0180)	(0.0062)	(0.0185)
High	-0.0021	0.0032	0.0070	-0.0003	0.0073
	(0.0116)	(0.0119)	(0.0139)	(0.0049)	(0.0143)
Low	-0.0346***	-0.0375***	-0.0245*	-0.0161***	-0.0355***
	(0.0108)	(0.0110)	(0.0128)	(0.0045)	(0.0133)
Medium	-0.0074	-0.0019	2.54×10^{-5}	-0.0013	0.0039
	(0.0110)	(0.0113)	(0.0132)	(0.0047)	(0.0137)
Fit statistics					
Observations	11,896	11,896	11,896	11,896	11,896
Adjusted R^2	0.03002	0.02981	0.04607	0.01638	0.04363

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

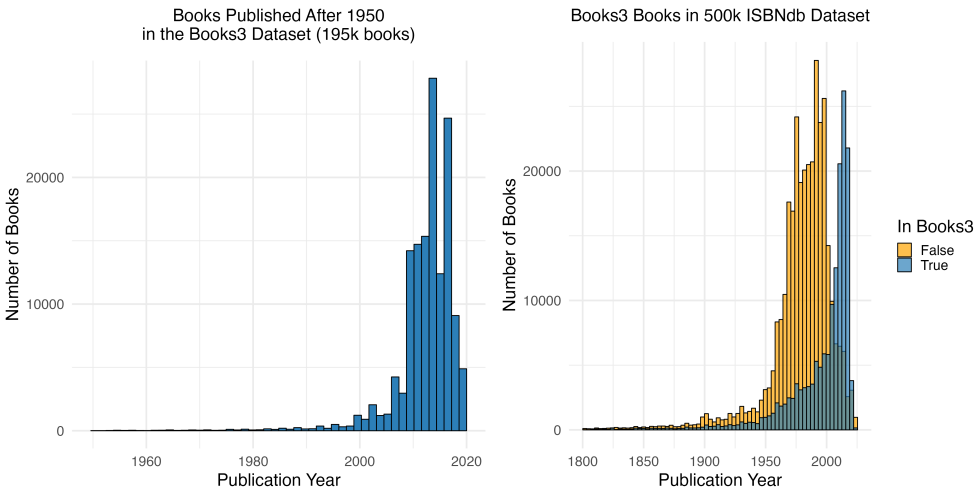
Note: This table displays estimates from IV models applied to five different LLMs, assessing the overall impact of inclusion in training data on comprehension of the source across books of varying popularity levels.

Figure 1: Data Collection and Experiment Pipeline



Note: Panel A outlines the timeline of Books3, created by Seth Lamancusa, including the dates of its creation and removal from public websites. Panel B 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 C presents an example passage from a book, illustrating how it is sent to the API along with the exact prompt.

Figure 2: Comparing Distribution of Books in Books3 with the 500k ISBNdb Dataset



Note: This figure compares the distribution of publication year of the full Books3 dataset (left) with the ISBNdb dataset (right). We then separately plot the Books3 and Non-Books3 within the ISBNdb dataset and identify a unique distribution of the Books3 books which matches the full Books3 distribution. Assuming our ISBNdb dataset is a sufficient representation of the "universe" of books, We use this to motivate a shift shares IV based on publication year.