Books, Copyright, and the Future of Foundation Models: An Empirical Study of Fiction in Pretraining

Anonymous ACL submission

Abstract

The use of books to train large language models has prompted numerous lawsuits against commercial AI providers, yet the reasons such data choices matter remain largely opaque outside industry labs. We ask a simple question: what role does fiction play during pretraining? To answer it, we pretrain 1B-parameter models from scratch under controlled token budgets on data mixes that include web text alone, web text with public-domain books, web text with contemporary fiction, and web text with both. Our results show that fiction substantially improves downstream model performance, with contemporary fiction yielding gains beyond those from public-domain corpora. We then examine what makes fictional text distinct from other pretraining sources, and test whether synthetic data can recover its benefits. Together, these experiments offer the largest empirical study to date of copyrighted fiction in model pretraining, contributing evidence to debates over data governance, creators' rights, and how particular data domains shape AI capabilities.

1 Introduction

The strong performance of modern large language models depends on pretraining at scale on vast text datasets (Gunasekar et al., 2023; Longpre et al., 2024). At the same time, leading commercial AI companies have become increasingly opaque about the composition, provenance, and licensing status of the datasets used to train their frontier models (Wan et al., 2025; OpenAI, 2023; Comanici et al., 2025; Singh et al., 2026), leaving the scientific community with limited insight into how these frontier models are being built and which data sources underpin their capabilities. This opacity also hinders independent assessment or auditing of these model development pipelines.

As the competition intensifies to develop frontier models that will outperform counterpart systems on standard benchmarks, multiple lawsuits have alleged that providers are resorting to practices such as training on copyrighted books obtained from pirated sources (O'Nan et al. v. Databricks, 2024; Kadrey v. Meta Platforms, 2025; Lyon v. Adobe, 2025), including the Books3 corpus (Presser, 2020). Anthropic, a leading model provider behind the Claude series of models (Anthropic, 2025), purportedly spent tens of millions of dollars to acquire and scan books as part of *Project Panama* (Schaffer et al., 2026). Intriguingly, reports from two leading frontier model providers, Anthropic and OpenAI, indicate that writing tasks rank among the most frequent applications of frontier models (Handa et al., 2025; Chatterji et al., 2025).

In this work, we aim to shed light on this landscape. We systematically investigate how the inclusion of data sources containing contemporary fiction, such as Books3 (Reisner, 2025), in pretraining data influences the downstream capabilities of large language models. We focus on fiction in isolation as fiction and non-fiction constitute fundamentally different data sources with distinct structures and characteristics (Altmann et al., 2014; Mar and Rain, 2015). These differences motivate studying fiction as a distinct data source, rather than as an undifferentiated component of a larger corpora of 'books'. To study the effect of pretraining on data containing fiction, we train 1B-parameter models under controlled token budgets on data mixes featuring: (a) broad web text with no fiction, (b) web text combined with fiction books in the public domain, (c) web text combined with contemporary fiction books, and (d) web text combined with fiction drawn from the public-domain and contemporary fiction. Because some data may contain contentious or sensitive material, we describe our data handling practices, safeguards, and governance processes in our ethics statement (§7). Next, to understand how this fiction data differs from other commonly available data sources, we train

predictors to distinguish fiction data from other types of web text. Finally, we ask if it is possible to construct a *synthetic data generation procedure for pretraining* that confers similar benefits in the training pipeline. We show:

I. Models trained on contemporary fiction from the Books3 corpus show more narrative coherence on base model evaluations, and after being instruction-tuned, increased competence on a range of instruction-following tasks.

II. Models trained on contemporary fiction far outperform models trained on fiction in the public domain from the Harvard Books Corpus across base model and downstream instruct-model evaluation. 'Books' are not a homogeneous data source, and different corpora produce different capabilities.

III. While base-model evaluations do not fully reflect the relative differences between models, models display qualitatively different behaviors after being instruction-tuned.

IV. We study which features distinguish fiction from other common data sources used in pretraining. We find that the most discriminative features include those associated with narrative coherence, figurative language, and dialog.

V. We consider a novel synthetic data generation pipeline, where contentious data is not included directly in training, but is provided as input to an LLM to generate diverse synthetic data. We show that this procedure can largely confer the benefits of training on contemporary fiction, while showing lower surface similarity to the source content.

Taken together, our results empirically demonstrate that books data can substantially improve model performance, but that books data is not homogenous, shedding light on the incentives that may be shaping training data acquisition in frontier model development.

2 Methodology

Datasets We utilize the following corpora for pretraining dataset curation, as detailed below:

WebOrganizer (Wettig et al., 2025): A 200B-token corpus derived from the DCLM pool (Li et al., 2024a), annotated with `topic` and `format` labels. The `topic` annotation is intended to capture the document's subject matter, while `format` denotes its stylistic or structural category.

Harvard Institutional Books (Cargnelutti et al., 2025): A corpus of nearly one million digitized books that are in the public domain, intended to

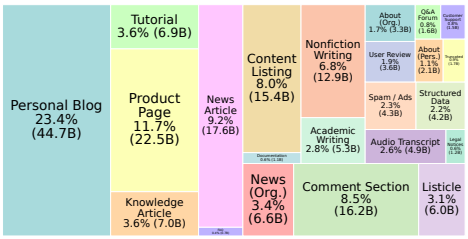


Figure 1: Distribution across web-organizer formats of the non-creative writing corpus spanning 191.4B tokens. All non-cw mixes are sampled from this mixture.

provide high-quality data for training large language models. The books were originally digitized from Harvard Library's participation in the Google Books project, and the corpus was compiled with funding from Microsoft and OpenAI.

Books3 (Gao et al., 2020; Presser, 2020) A dataset consisting of a mix of fiction and non-fiction books and included as a component of the Pile (Gao et al., 2020), that was used to train the Pythia Models (Biderman et al., 2023), GPT-NeoX-20B (Black et al., 2022), Bloomberg-GPT (Wu et al., 2023), and Llama-1 (Touvron et al., 2023a).

We construct a family of 185B-token pretraining corpora built around WebOrganizer (Wettig et al., 2025). From the approximately 191B available WebOrganizer tokens, after excluding documents tagged as Creative Writing (defined as a page containing a *'short story, chapters from a novel, poem or song lyrics'*), we draw nested web subsets of 185B, 180B, and 175B tokens. Format distribution over the corpora is shown in Fig. 1. This enables us to construct the 185B-token subset that defines the **No Books** condition. Using the 180B-token subset, we construct two book-augmented corpora by adding either 5B tokens from Harvard Institutional Books 1.0 (Cargnelutti et al., 2025) or 5B tokens from Books3 (more details in Section C.2). Finally, using the 175B-token subset, we construct a mixed-book condition by adding 5B tokens from Harvard Institutional Books and 5B tokens from Books3, a widely used proprietary corpus. All resulting corpora contain 185B tokens, allowing us to compare models trained with no books, public-domain books, proprietary books, and a mixture of both while keeping the total token budget fixed.

Pretraining We pretrain all models from scratch with GPT-NeoX (Andonian et al., 2023), using an architecture closely matched to Llama-3.2-1B (Meta, 2024). Each model is trained for 22,100 steps, corresponding to approximately 185B tokens, with a fixed sequence length of 4096 tokens. Training uses 4 nodes of either B200 or H200 GPUs,

depending on availability, with a per-GPU batch size of 32 and 2 gradient-accumulation steps per update. We use an initial learning rate of 4×10^{-4}, warm up for the first 5% of training, and then apply cosine decay to 10% of the initial rate by the final step. We borrow these hyper-parameters from Wei et al. (2026). Each run requires approximately 1,380 H200 GPU-hours or 850 B200 GPU-hours.

Instruction Tuning We instruction-tune each base model on the Tulu v2 SFT (Lambert et al., 2025) mixture for a single epoch using a cosine learning rate schedule with a peak learning rate of 2×10^{-5} and a warmup ratio of 0.03. We use an effective batch size of 128 for these runs.

Evaluation We evaluate our base models on a broad set of capabilities including narrative understanding, creativity, figurative language, emotion recognition, theory of mind, commonsense and logical reasoning, domain knowledge, long-context modeling, causality, dialogue understanding, morality, math, and coding. Most base model evaluations are performed using LM-Evaluation-Harness (Gao et al., 2024) & OLMES (Gu et al., 2025). For instruction-tuned models, we evaluate on AlpacaEval (Taori et al., 2023), MT-Bench (Zheng et al., 2023), LitBench (Fein et al., 2026), and WritingBench (Wu et al., 2025). Additional evaluation details in Appendix G.

3 What Role Do Books Play in Pretraining Data

It is increasingly common to include books as a part of pretraining corpora. While the majority of models trained by commercial companies rarely disclose the proportion of books used in their training data, several open-source initiatives have documented their use. Notably, Pile (Gao et al., 2020), one of the first dedicated open pretraining corpora, contained approximately 15% books (measured in bytes). Similarly, Dolma (Soldaini et al., 2024), an open-source dataset of three trillion tokens built to train large language models, contains around 6B tokens of book data. Among commercial models, GPT-3 was reported to be trained on 67 billion tokens of books (Brown et al., 2020), and Llama included books with a sampling proportion of 4.5% in its 1.4 trillion token training corpus (Touvron et al., 2023b). These examples provide indirect evidence that books are widely regarded as a high-quality data source for pretraining. However, it remains unclear what exactly are the effects of including book data in the training pipeline of LLMs.

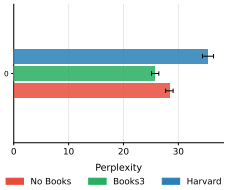


Figure 2: Story generation perplexity on WritingPrompts. Books3 achieves the lowest perplexity.

A prominent characteristic of these models is their wide range of capabilities. Since this work focuses on understanding what book data contributes to large language model pretraining, we naturally study its influence across this broad capability spectrum. To this end, we analyze the effect of training with and without book data on a diverse set of skills introduced in Section 2.

3.1 Comparing Books3 with No Books Baseline

We start by asking what benefits contemporary, often copyrighted, books in the training data might offer compared to training without any books at all. This question is directly relevant to understanding why many organizations have relied on copyrighted books when training production language models. To study this, we compare two models trained on the same base corpus, where one model replaces part of the corpus with Books3 and the other does not.

Figure 3 shows the comparison between the two models. Across most general-purpose tasks, the two models perform at roughly the same level, suggesting that swapping 5B tokens of web text for book data does not meaningfully hurt general capabilities. The gains from Books3 are concentrated in tasks that require understanding or generating extended, coherent text. Most notably, Books3 shows a large improvement (14.7%) on LAMBADA (Paperno et al., 2016), a task that tests long-range language coherence. This aligns with the intuition that the narrative text in the books trains the model to track context over longer spans. Consistent with this, as seen in Figure 2, the Books3 model also achieves lower perplexity than the no-books model on long-form literary text, such as human-written stories from WritingPrompts (Fan et al., 2018).

To further probe whether these differences carry over after instruction tuning, we fine-tune both base models using the setup mentioned in Section 2 and evaluate them on writing-focused bench-

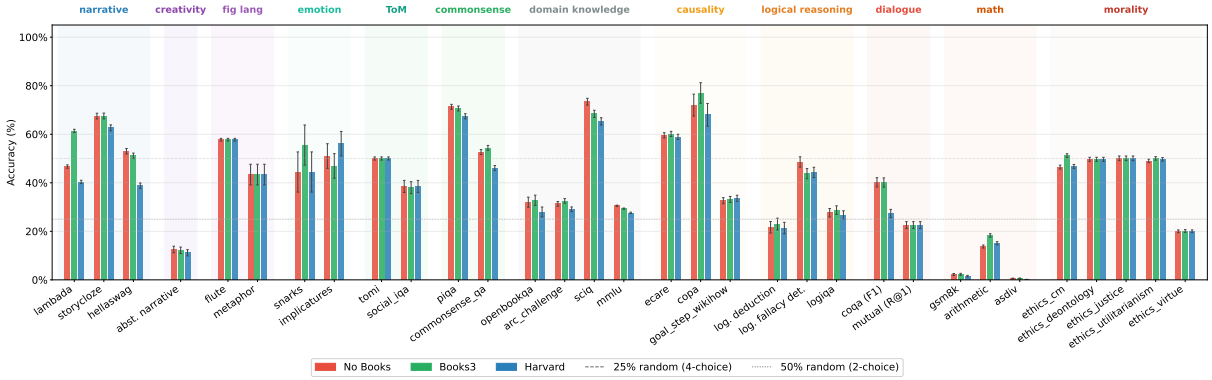


Figure 3: Base model evaluation across 12 skill categories comparing the three model variants. Books3 outperforms both No Books and Harvard in narrative tasks (e.g., LAMBADA, HellaSwag), while performance on most other categories remains comparable across models. Dashed and dotted lines indicate random baselines for 4-choice and 2-choice tasks, respectively. Error bars denote standard error.

marks. We find that Books3 outperforms the no-books baseline on LitBench (Fein et al., 2026), MT-Bench (Zheng et al., 2023) writing and roleplay, WritingBench (Wu et al., 2025) literature and arts, and Alpaca (Li et al., 2023; Dubois et al., 2024) win rate (shows in Table 1). The gap here is notably larger than what we observe in base model evals, suggesting that books have a compounding effect when combined with instruction tuning.

Taken together, these results show that training on contemporary fiction improves literary and creative text capabilities while leaving most general skills unchanged. This raises a natural question: are contemporary books uniquely valuable, or can public-domain books offer similar benefits?

3.2 Copyrighted vs. Public Domain Books

Having established that training on contemporary fiction improves over the no-books baseline, we now ask whether public domain books provide similar utility. We compare the model trained on Books3 against the model trained on the Harvard Institutional Books Corpus, where all other training data is held constant. At the base model level, Books3 outperforms Harvard across all narrative coherence tasks (Figure 3). Notably, Harvard performs worse than even the no-books baseline in this category, suggesting that these books actively hurt coherence modeling. Harvard also underperforms Books3 on emotion recognition (snarks), commonsense reasoning (commonsense_qa), and causal reasoning (copa). This pattern extends to WritingPrompts, where Harvard achieves the highest perplexity (Figure 2).

As seen in Table 1, these trends amplify substantially after instruction tuning. On LitBench and WritingBench, Books3 outperforms Harvard by margins of 0.21 and 0.25, respectively. The delta is most significant on Alpaca, where the model trained on Books3 achieves a 46.3% higher win rate, indicating that the advantage of copyrighted books extends well beyond literary tasks to general instruction-following ability. Section D.1 reports our findings on training using 10B book tokens.

Taken together, these results suggest that books from Books3 provide a significant advantage over public domain books from the Harvard corpus for both coherence and text generation, and that Harvard's public domain corpus not only fails to match Books3 but in several tasks falls below the no-books baseline. We hypothesize that this may be due to quality degradation introduced during OCR done during construction of the dataset, as well as the dated nature of the books, which were largely published in the 19th and 20th centuries.

We also observe that Books3 has a stronger effect on the model after instruction tuning. While Books3 shows moderate gains over Harvard at the base model level, this gap grows much larger after instruction tuning. Our results suggest that the choice of book data during pretraining has a larger impact on the final model than base model evaluations alone would suggest.

4 How does Books data look different from other common pretraining sources

In Section 3, we showed that Books3 data matches the no-books baseline on general tasks while outperforming it on coherence and literary text generation. Since these gains come directly from the training data, a natural question is: what makes book data structurally different from other sources?

Eval	No Books	Books3	Harvard	Delta (Books3 − Harvard)
AlpacaEval (win rate)	28% ± 17%	**46.97**% ± 21.65%	0.44% ± 0.28%	+46.53%
AlpacaEval (len. controlled)	6.38% ± 0.5%	**8.92**% ± 0.74%	0.085% ± 0.006%	+8.835%
MT-Bench (overall)	**1.075**	**1.075**	1.025	+0.05
Writing	1.05	**1.1**	1.0	+0.1
Roleplay	1.15	**1.25**	1.15	+0.1
Extraction	1.0	1.0	1.0	0.0
Reasoning	1.0	1.0	1.0	0.0
Math	1.0	1.0	1.0	0.0
Coding	1.0	1.0	1.0	0.0
STEM	**1.2**	1.0	1.05	−0.05
Humanities	1.2	**1.25**	1.0	+0.25
LitBench	−0.95 ± 0.04	**−0.76** ± 0.04	−0.97 ± 0.04	+0.21
WritingBench (overall)	1.844 ± 0.052	**1.845** ± 0.047	1.593 ± 0.041	+0.252
Academic & engineering	1.983 ± 0.135	**2.046** ± 0.108	1.800 ± 0.111	+0.246
Finance & business	1.818 ± 0.101	**1.826** ± 0.104	1.647 ± 0.085	+0.179
Politics & law	**1.742** ± 0.101	1.674 ± 0.118	1.518 ± 0.085	+0.156
Literature & arts	1.641 ± 0.125	**1.710** ± 0.120	1.368 ± 0.081	+0.342
Education	**1.988** ± 0.128	1.894 ± 0.130	1.622 ± 0.116	+0.272
Advertising & marketing	**1.956** ± 0.175	1.944 ± 0.129	1.575 ± 0.114	+0.369

Table 1: Instruction-tuned model evaluation across three models: No Books, Books3 (copyrighted), and Harvard (public domain). Books3 outperforms Harvard at all tasks and surpasses the No Books baseline on the majority of the tasks. Bold indicates the best result per row. The Delta column shows the advantage of Books3 over Harvard. LitBench and WritingBench report mean ± 95% CI; Alpaca reports mean ± standard error. Note that the absolute scores are lower as our models undergo lightweight instruction tuning for a single epoch; however, the relative ordering of Books3 > No Books > Harvard is consistent across tasks.

To study this, we compare book data against six common pretraining domains: general web text, Reddit[1], math, code, academic papers, and news. For general web text, we follow Shi et al. (2026) and use the baseline version of DCLM (Li et al., 2024b) that excludes book data. News data is obtained from DCLM using the domain classifier from Wettig et al. (2025). Reddit, math, code, and academic paper data are drawn from the corresponding domain-specific subsets of Dolma (Soldaini et al., 2024).

To capture the properties that distinguish book data from other sources, we design 62 features across 9 categories: narrative coherence and long-range dependency, theory of mind & dialogue, emotion, lexical diversity, figurative language, syntactic structure, stylistic structure, compressibility, and readability. We refer the reader to the complete descriptions of these categories in Appendix E Section E.2.

We train a logistic regression classifier with L1-regularization to identify the most discriminative features, discarding uninformative ones. We then train an XGBoost classifier using only the selected features. Both classifiers are trained on 100K documents: 50K from book data and 50K from non-book sources (10K each from Reddit, academic papers, general web text, and news; 5K each from math and code). We train two versions of each classifier: one using Books3 as the positive class and one using Harvard, keeping the negative class fixed in both cases. This lets us compare how copyrighted and public domain books each differ from non-book sources, and ultimately, how they differ from one another.

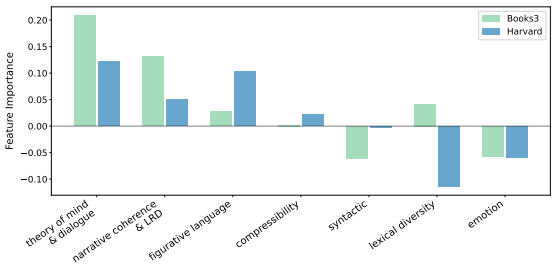


Figure 4: Feature coefficients from training classifiers for Books3 and the Harvard corpus. Books3 exhibits stronger theory-of-mind and dialogue features, greater narrative coherence and long-range dependencies, and higher lexical diversity, all of which likely contribute to improved downstream utility.

4.1 Semantic Features

We first train the classifier using only semantic feature categories, excluding stylistic and readability features that capture surface-level writing patterns.

[1]https://www.reddit.com/

This isolates the semantically meaningful properties that distinguish book data from other sources.

Figure 4 shows the category importance for both classifiers. Book data is distinguishable from non-book data across multiple categories in both cases. The three most important categories are theory of mind and dialogue, narrative coherence and long-range dependency, and figurative language. Both classifiers show positive coefficients for these categories. However, Books3 shows higher coefficients than Harvard for theory of mind & dialogue and narrative coherence & long-range dependency, indicating that these properties are more prominent in Books3 text. As shown in Figure 10, the most dominant individual feature for Books3 is max coreference chain length, which measures how consistently the same entity is referred to throughout a document. This feature is much less prominent for Harvard. Named entity density is also more important for Books3, capturing how many person names appear in a passage and reflecting the richer character-driven narratives in modern fiction. Furthermore, Books3 contains more unique words than Harvard, as reflected in the lexical diversity category, suggesting a richer and more varied vocabulary than public domain books.

Harvard, on the other hand, shows a higher positive coefficient for figurative language than Books3. As shown in Figure 12, archaic word density is one of the most dominant features for Harvard and is nearly absent in Books3. This feature measures how often old-fashioned English words appear in the text (e.g., *"thee"*, *"thou"*, *"alas"*). Our observations are inline with the books being largely published in the 18th and 19th centuries. Since books often only become public domain after several decades of publishing, public domain corpora will continue to have this issue of vocabulary drift.

This reflects a fundamental difference between copyrighted and public-domain books: Books3 contains contemporary fiction with stronger narrative coherence and lexical diversity, while Harvard books contain more figurative language typical of older prose. Both classifiers separate book from non-book text with high accuracy: 96.4% for Books3, with 0.97 precision and 0.96 recall on the book class, and 97.3% for Harvard, with 0.97 precision and 0.98 recall.

4.2 Stylistic and Readability Features

We next add stylistic and readability features to the classifier to compare the two book corpora in writing style and text complexity.

We find readability to be a representative feature of Books3 text (Figure 9). Specifically, Flesch-Kincaid grade level is its top individual feature (Figure 13), which estimates text complexity based on average sentence length and syllable count per word. This shows that Books3 text tends to have longer sentences than the Harvard text. We also find that Dale-Chall readability has a negative coefficient for Books3, indicating that Books3 uses more common and easy-to-understand words. We find that Harvard scores higher on the stylistic category with a positive coefficient, while stylistic features carry a negative coefficient for Books3. Harvard books show higher usage of semicolons, em-dashes, parentheses, and varied punctuation, which are characteristic of older prose writing. Adding stylistic and readability features improves classifier performance for both. The Books3 classifier achieves 97.3% accuracy with 0.98 precision and 0.97 recall. The Harvard classifier reaches 99.1% accuracy with 0.99 precision and 0.99 recall.

The feature patterns we observe also align with the downstream performance differences reported in Section 3.2. Books3 text is richer in narrative coherence & long-range dependency, dialogue, and lexical diversity, which may help the model perform better on narrative coherence tasks and creative writing benchmarks such as LitBench and WritingBench. These results also suggest that modern contemporary fiction, which is predominantly found in copyrighted book corpora like Books3, could provide a stronger signal for the coherence and creative writing skills tested by modern benchmarks than public domain text does.

5 Is it feasible to replicate the value of books data

From our results so far, we know that contemporary fiction books are a high-quality data source for improving model performance on coherence and writing tasks. We have also observed empirically that contemporary books provide greater utility than public domain books, and that modern contemporary fiction is important for teaching coherence and writing skills to the model. To this end, we explore whether it is possible to construct a procedure for generating synthetic data for model pretraining, that can replicate the value of natural fiction book data. Such a procedure might be lucrative for providers: Large language models are known

to memorize and reproduce portions of their training data, including copyrighted works (Chen et al., 2024; Cooper et al., 2025). Even if high-quality book data is legally sourced, it could be reproduced by the model. The provider may therefore want an alternative that provides similar utility with reduced risk of regurgitation. However, such a procedure can also potentially break down the pipelines of data provenance, since such synthetic data may bear low resemblance to the original source content.

5.1 Synthetic Pretraining

We study whether it is possible to generate high-quality synthetic fiction that gives similar utility as natural fiction books. Our goal is to generate data that is similar in distribution to high-quality contemporary fiction. To achieve this, we use a variety of metadata to guide the generation of book chapters. Specifically, we use Gemma 3 12B Instruct (Team et al., 2025) as our generator and combine genre, plot summary, and excerpts from Books3 to guide the generation of book chapters. Additional details are outlined in Appendix F.

To train our models, we combine the 5B tokens of synthetic book data with the same 180B tokens of non-book data used in our other experiments, following the same setup described in Section 2.

5.2 Synthetic Data as an Alternative

Our results show that the model trained on synthetic data outperforms the model trained on Books3 on several writing benchmarks (Table 2). On WritingBench, the synthetic model outperforms Books3 on five out of six subtasks, winning on finance, politics, literature, education, and advertising. The synthetic model also outperforms Books3 on MT-Bench overall (1.119 vs 1.075) and MT-Bench writing (1.5 vd 1.1), indicating that synthetically generated fiction does better on modern writing benchmarks. On the two tasks where Books3 still leads, the synthetic model finishes second in both cases.

On the base model evaluations, the synthetic model outperforms the model trained on the Harvard book corpus and comes close to the performance of the Books3 model on LAMBADA (Figure 8). More importantly, the synthetic data does not degrade the model utility on any of the tasks. This finding is particularly notable given prior results showing that training on synthetic data can lead to model collapse (Shumailov et al., 2024). We hypothesize that the diversity in our data generation process helps mitigate this phenomenon.

Taken together, these results suggest that synthetic data is a viable alternative to natural copyrighted fiction for training language models with strong writing capabilities, particularly for instruction-following and generative writing tasks.

6 Related Work

Training Data and Model Behavior Pretraining data strongly shapes large language model capabilities and risks. Prior work shows that source selection, filtering, deduplication, and documentation affect downstream behavior and social harms (Bender et al., 2021; Dodge et al., 2021). More recent work studies data mixtures directly, including methods for optimizing domain weights and evaluating curation recipes under controlled budgets (Xie et al., 2023; Li et al., 2024a). We build on this data-centric literature by isolating fictional books and comparing public-domain, copyrighted, and synthetic fiction under fixed pretraining budgets.

Transparency, Control, and Copyright Concerns over unconsented training have motivated methods for detecting training-data use, measuring memorization, and tracing model outputs back to source corpora (Carlini et al., 2021, 2023; Liu et al., 2025). Complementary efforts, such as opt-out registries, provenance metadata, and dataset licensing frameworks, aim to express permissions before training occurs (Longpre et al., 2023). These technical questions intersect with ongoing legal and policy disputes over whether using copyrighted works for model training is transformative, fair use, or otherwise permissible (Samuelson, 2023; Lemley, 2024). Recent empirical work sharpens these concerns: fine-tuning can reactivate latent verbatim recall of copyrighted books, bypassing alignment and output filters and producing long passages from held-out works (Liu et al., 2026). Complementary studies find that author-specific fine-tuning can produce non-verbatim literary prose that readers sometimes prefer to expert human writing, raising market-substitution questions for creative labor and fair-use analysis (Chakrabarty et al., 2025; Chakrabarty and Dhillon, 2026). Because training corpora are often opaque, it remains difficult to assess how much particular classes of copyrighted works contribute to model capabilities. Our work addresses this gap by measuring the functional role of copyrighted fiction in pretraining.

Synthetic Data as a Substitute Synthetic data

Eval	No Books	Books3	Harvard	Synthetic
AlpacaEval (win rate)	$28\% \pm 17\%$	$\mathbf{46.97}\% \pm 21.65\%$	$0.44\% \pm 0.28\%$	$39.77\% \pm 21.42\%$
AlpacaEval (len. controlled)	$6.38\% \pm 0.5\%$	$\mathbf{8.92}\% \pm 0.74\%$	$0.085\% \pm 0.006\%$	$8.82\% \pm 0.64\%$
MT-Bench (overall)	1.075	1.075	1.025	**1.119**
Writing	1.05	1.1	1.0	**1.5**
Roleplay	1.15	**1.25**	1.15	1.2
Extraction	1.0	1.0	1.0	1.0
Reasoning	1.0	1.0	1.0	1.0
Math	1.0	1.0	1.0	1.0
Coding	1.0	1.0	1.0	1.0
STEM	**1.2**	1.0	1.05	1.05
Humanities	1.2	**1.25**	1.0	1.2
LitBench	-0.95 ± 0.04	$\mathbf{-0.76} \pm 0.04$	-0.97 ± 0.04	-0.82 ± 0.05
WritingBench (overall)	1.844 ± 0.052	1.845 ± 0.047	1.593 ± 0.041	$\mathbf{1.901} \pm 0.050$
Academic & engineering	1.983 ± 0.135	$\mathbf{2.046} \pm 0.108$	1.800 ± 0.111	1.985 ± 0.127
Finance & business	1.818 ± 0.101	1.826 ± 0.104	1.647 ± 0.085	$\mathbf{1.914} \pm 0.102$
Politics & law	1.742 ± 0.101	1.674 ± 0.118	1.518 ± 0.085	$\mathbf{1.784} \pm 0.107$
Literature & arts	1.641 ± 0.125	1.710 ± 0.120	1.368 ± 0.081	$\mathbf{1.768} \pm 0.123$
Education	1.988 ± 0.128	1.894 ± 0.130	1.622 ± 0.116	$\mathbf{1.994} \pm 0.144$
Advertising & marketing	1.956 ± 0.175	1.944 ± 0.129	1.575 ± 0.114	$\mathbf{2.008} \pm 0.153$

Table 2: Instruction-tuned model evaluation across four models: No Books, Books3 (copyrighted), Harvard (public domain), and Synthetic (LLM-generated fiction). Bold indicates the best result per row. LitBench and WritingBench report mean $\pm$ 95% CI; AlpacaEval reports mean $\pm$ standard error. Note that absolute scores are lower as our models undergo lightweight instruction tuning for a single epoch.

is increasingly used to reduce reliance on scarce, costly, or legally encumbered human-written text. TinyStories shows that generated stories can teach small models grammar and narrative structure (Eldan and Li, 2023), while later work such as the Phi series and Cosmopedia demonstrates the value of high-quality synthetic data in broader pretraining mixtures (Gunasekar et al., 2023; Ben Allal et al., 2024). At the same time, synthetic data can inherit biases, stylistic narrowing, and contamination from the generator model, and recursive training on generated text can reduce diversity or cause model collapse (Yu et al., 2023; Shumailov et al., 2024). We therefore evaluate synthetic fiction as a possible substitute for copyrighted fiction, measuring both downstream utility and verbatim reproduction risk.

Additional discussion is provided in Appendix B

7 Discussion and Conclusion

In this work, we aim to bring clarity to the ongoing debate surrounding the use of books in training large language models. Our findings reinforce previous observations by Longpre et al. (2024) that books can be beneficial in large language model training. However, our findings are nuanced: 'books data' should not be treated as a monolithic category. We find substantial variation in utility across sources: for example, corpora of books in the public domain that were studied in our work can actually negatively impact language modeling performance. These findings raise important questions for the future of data curation in LLM training. Although there is growing investment in constructing new legally-permissible corpora of content in the public-domain, it remains unclear to what extent such corpora are actually being incorporated into frontier training pipelines, and where they will bring greatest benefits. Moreover, due to legal and competitive reasons, there is reduced transparency around the training data practices of state-of-the-art models which makes it more difficult to understand or audit these processes externally. Importantly, this reduced visibility should not be interpreted as evidence that the use of contentious data has declined; rather, such data may still be incorporated to a similar extent as before, but with less public transparency into these processes. Finally, we show that data curation processes may become increasingly complex in the future with provenance becoming harder to track: in our work, we find that it is possible to construct high-quality synthetic pretraining data that bears negligible surface similarity to the source content, while conferring similar benefits to large language model pretraining. We hope our work contributes to our shared scientific understanding of the role data composition and provenance play in downstream model capabilities.

Ethical Considerations

To better understand the role and impact of contentious training data such as copyrighted fiction in the development of large language models, we conduct controlled experiments in which models are trained on such data solely for the purpose of measuring and analyzing their relative effects on model behavior and performance. We impose strict controls on the use and retention of these artifacts, involving the Books3 dataset. In this study, Books3 and any derived artifacts are used solely for non-commercial research purposes and are not incorporated into any deployed or publicly released systems. Specifically, we do not release, distribute, or deploy any model trained on Books3, including fine-tuned models, intermediate checkpoints, or derivative variants. In addition, Books3 data is not shared with or processed through third-party commercial APIs such as the APIs associated with frontier proprietary models, including within the synthetic-data generation workflows described in this study. Access to Books3 is restricted to the study authors. All copies of Books3 held by the authors for this study, including backups and derived processing artifacts, will be deleted by May 1, 2029. All models trained on Books3 as part of this study will be deleted by May 1, 2030.

Limitations

Our findings should be interpreted in the context of the study's limitations. First, our experiments study a particular subset of contentious training data: copyrighted fiction represented by Books3. Therefore, they may not generalize to other forms of copyrighted content. Moreover, frontier language models are pretrained at substantially larger scales than is currently feasible in academia, both in terms of parameter count and training tokens. As a result, the quantitative effects we observe may differ under large-scale industrial training regimes.

Second, while we measure behavioral and performance differences associated with the inclusion of specific fiction datasets, our experiments do not establish uniquely attributable causal mechanisms for the observed behaviors. Additionally, some measured effects may arise from interactions between Books3 and the broader pretraining corpus rather than from Books3 in isolation. We also study a limited set of data mixes, and results may vary if the relative proportion of data from books, or other aspects of data composition are varied.

Third, our study considers only a limited set of sources for fiction. In the public domain, there may exist corpora of fiction that are more valuable than the ones we have evaluated in this study, and conversely, there may exist copyrighted or contemporary fiction datasets whose effects differ considerably from those observed here. Further, either of these corpora may be accessible at a much larger scale than what was used in this study. Accordingly, our results should not be interpreted as establishing any kind of universal conclusion about fiction data as a category.

Finally, ethical considerations constrain the extent to which artifacts from this study can be retained or publicly distributed. In accordance with these constraints, models trained on data configurations that are deemed sensitive will not be publicly released or deployed by the authors.

References

Ulrike Altmann, Isabel C Bohrn, Oliver Lubrich, Winfried Menninghaus, and Arthur M Jacobs. 2014. Fact vs fiction—how paratextual information shapes our reading processes. *Social cognitive and affective neuroscience*, 9(1):22–29.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, and 2 others. 2023. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch. `https://www.github.com/eleutherai/gpt-neox`. Version 2.0.0.

Anthropic. 2025. Claude. Large language model.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Mo Bavarian, Heewoo Jun, Nikolas A. Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *ArXiv*, abs/2207.14255.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, and 1 others. 2022. GPT-NeoX-20B: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2025. The 2024 foundation model transparency index. *Preprint*, arXiv:2407.12929.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Matteo Cargnelutti, Catherine Brobston, John Hess, Jack Cushman, Kristi Mukk, Aristana Scourtas, Kyle Courtney, Greg Leppert, Amanda Watson, Martha Whitehead, and Jonathan Zittrain. 2025. Institutional books 1.0: A 242b token dataset from harvard library's collections, refined for accuracy and usability. *Preprint*, arXiv:2506.08300.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *ICLR*.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. 2024. Stealing part of a production language model. *Preprint*, arXiv:2403.06634.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Tuhin Chakrabarty and Paramveer S. Dhillon. 2026. Can good writing be generative? expert-level ai writing emerges through fine-tuning on high-quality books. *Preprint*, arXiv:2601.18353.

Tuhin Chakrabarty, Jane C Ginsburg, and Paramveer Dhillon. 2025. Readers prefer outputs of ai trained on copyrighted books over expert human writers. *arXiv preprint arXiv:2510.13939*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. 2024. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15134–15158, Miami, Florida, USA. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

A. Feder Cooper, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Mark Lemley, Daniel E. Ho, and Percy Liang. 2025. Extracting memorized pieces of (copyrighted) books from open-weight language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. 2024. De-cop: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Desmond Elliott. 2026. Generative ai archaeology. In *ICLR Blogposts 2026*. Https://iclr-blogposts.github.io/2026/blog/2026/genai-archaeology/.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2026. LitBench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, Rabat, Morocco. Association for Computational Linguistics.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. 2024. Logits of API-protected LLMs leak proprietary information. In *First Conference on Language Modeling*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. Olmes: A standard for language model evaluations. *Preprint*, arXiv:2406.08446.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, and 1 others. 2025. Which economic tasks are performed with

ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah Smith. 2024. Data mixture inference attack: Bpe tokenizers reveal training data compositions. In *Advances in Neural Information Processing Systems*, volume 37, pages 8956–8983. Curran Associates, Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Inc. Kadrey v. Meta Platforms. 2025. Order denying plaintiffs' motion for partial summary judgment and granting Meta's cross-motion on fair use grounds. [link].

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. 2024. Talkin' 'bout AI generation: Copyright and the generative-AI supply chain. In *Proceedings of the 2024 Symposium on Computer Science and Law (CSLAW '24)*. ACM. Full version forthcoming in Journal of the Copyright Society.

Mark Lemley. 2024. How generative ai turns copyright law upside down. *Science and Technology Law Review*, 25(2).

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, and 1 others. 2024a. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024b. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, and 11 others. 2025. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 178–188, Vienna, Austria. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Xinyue Liu, Niloofar Mireshghallah, Jane C Ginsburg, and Tuhin Chakrabarty. 2026. Alignment whack-a-mole: Finetuning activates verbatim recall of copyrighted books in large language models. *arXiv preprint arXiv:2603.20957*.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing attribution in ai. *Preprint*, arXiv:2310.16787.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Inc. Lyon v. Adobe. 2025. Class action complaint alleging direct copyright infringement;. [link].

Raymond A Mar and Marina Rain. 2015. Narrative fiction and expository nonfiction differentially predict verbal ability. *Scientific Studies of Reading*, 19(6):419–433.

Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. *Preprint*, arXiv:2303.03919.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre De Montjoye. 2024. Copyright traps for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 35296–35309.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers. *Preprint*, arXiv:2106.15772.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *Preprint*, arXiv:1604.01696.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Inc. et al. O'Nan et al. v. Databricks. 2024. Class action complaint alleging direct copyright infringement against Mosaic ML and vicarious infringement against Databricks for training MPT and DBRX large language models on copyrighted books via RedPajama/Books3 datasets. [link].

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arxiv:2303.08774*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Shawn Presser. 2020. Books3. https://twitter.com/theshawwn/status/1320282149329784833.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Preprint*, arXiv:1808.07042.

Alex Reisner. 2025. The unbelievable scale of AI's pirated-books problem. *The Atlantic*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Pamela Samuelson. 2023. Generative ai meets copyright. *Science*, 381(6654):158–161.

Aaron Schaffer, Will Oremus, and Nitasha Tiku. 2026. Anthropic 'destructively' scanned millions of books to build Claude. *The Washington Post*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Weijia Shi, Akshita Bhagia, Kevin Farhat, Niklas Muennighoff, Jacob Morrison, Evan Pete Walsh, Dustin Schwenk, Shayne Longpre, Jake Poznanski, Allyson Ettinger, Daogao Liu, Margaret Li, Mike Lewis, Wen tau Yih, Dirk Groeneveld, Luca Soldaini, Kyle Lo, Noah A. Smith, Luke Zettlemoyer, and 4 others. 2026. FlexOLMo: Open language models for flexible data use. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Damien Sileo. 2023. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry,

Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. 2025. The 2025 foundation model transparency index. *Preprint*, arXiv:2512.10169.

Johnny Wei, Ameya Godbole, Mohammad Aflah Khan, Ryan Yixiang Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, and Robin Jia. 2026. Hubble: a model suite to advance the study of LLM memorization. In *The Fourteenth International Conference on Learning Representations*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pretraining data curation. *Preprint*, arXiv:2502.10341.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *Preprint*, arXiv:2503.05244.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. DoReMi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A LLM Usage

We use LLMs for help with writing of the paper and also as coding assistants. In both cases we independently verify the outputs.

B Extended Related Work

Studying Training Data and Its Effects Pretraining data is now understood to be a central determinant of large language model behavior. Prior work has shown that corpus construction choices such as source selection, filtering, deduplication, and documentation affect both downstream capabilities and social harms (Bender et al., 2021; Dodge et al., 2021; Gao et al., 2020). Domain-adaptive pretraining further demonstrates that continued exposure to particular domains can substantially change model performance on related tasks (Gururangan et al., 2020). More recent work studies pretraining mixtures directly, including methods for optimizing domain weights and benchmarking data curation recipes under fixed model and compute budgets (Xie et al., 2023; Li et al., 2024a; Wettig et al., 2025). These studies motivate treating books and fiction not as interchangeable text, but as data domains whose contribution can be measured experimentally. Our work builds on this data-centric view but differs in focusing specifically on fictional books and on separating public-domain books from copyrighted fiction under controlled pretraining budgets.

Tools to Help Creators Growing concern over unconsented training has led to technical work on transparency, attribution, and creator control. Membership-inference and pretraining-data-detection methods attempt to determine whether a given document was included in a model's training data, including likelihood-based approaches and copyright-oriented benchmarks for books (Carlini et al., 2021, 2023; Shi et al., 2024; Duarte et al., 2024). Related systems such as Data Portraits and OLMoTrace aim to make training-data overlap and output-to-training-corpus matches more inspectable at scale (Marone and Durme, 2023; Liu et al., 2025). These tools are important for auditing, but they are generally retrospective: they help identify possible inclusion or memorization after data collection and model training have occurred.

Platform-level approaches such as opt-out registries, crawler restrictions, provenance metadata, and dataset licensing frameworks seek to express permissions before training occurs (Longpre et al., 2023). For authors and book publishers, however, analogous technical protections remain less mature. This gap makes empirical evidence about the marginal value of copyrighted fiction especially important: without understanding what models gain from such data, it is difficult to design meaningful licensing, opt-out, attribution, or compensation mechanisms.

Copyright and Large Language Models The use of copyrighted works in LLM training has become a major legal and policy issue. Lawsuits by authors, publishers, and media organizations have challenged the use of copyrighted books and articles in model training, while model developers have argued that training is transformative or otherwise protected under fair-use doctrines (Samuelson, 2023; Lemley, 2024). Legal scholarship has emphasized that copyright questions arise throughout the generative-AI supply chain, including data acquisition, dataset distribution, model training, model release, and generated outputs (Lee et al., 2024). These debates are complicated by the opacity of training corpora and by uncertainty about how specific works affect model capabilities.

Technical work on memorization and extraction is closely connected to these legal questions. Language models can memorize and sometimes reproduce rare or duplicated training sequences, with memorization increasing as model size, duplication, and prompt context increase (Carlini et al., 2021, 2023). Copyright-trap and book-extraction studies further examine whether copyrighted text can be detected in, or elicited from, trained models (Meeus et al., 2024; Nasr et al., 2025). Recent empirical work sharpens these concerns: fine-tuning can reactivate latent verbatim recall of copyrighted books, bypassing alignment and output filters and producing long passages from held-out works (Liu et al., 2026). Complementary studies find that author-specific fine-tuning can produce non-verbatim literary prose that readers sometimes prefer to expert human writing, raising market-substitution questions for creative labor and fair-use analysis (Chakrabarty et al., 2025; Chakrabarty and Dhillon, 2026). Our work is complementary: rather than asking only whether copyrighted books were included or reproduced, we ask what functional role fictional books play in pretraining and whether copyrighted fiction provides benefits beyond public-domain alternatives.

Synthetic Data Generation in Pretraining Synthetic data has become a prominent strategy for improving model training while reducing reliance on scarce, expensive, or legally encumbered human-written data. Work such as TinyStories shows that carefully generated synthetic stories can teach small models coherent grammar and narrative structure (Eldan and Li, 2023). The Phi model series and related efforts demonstrate that textbook-quality synthetic data can improve small language models in code and reasoning-heavy settings (Gunasekar et al., 2023). Large-scale synthetic corpora such as Cosmopedia extend this idea to broader pretraining-style mixtures by generating textbooks, stories, blog posts, and instructional text from seed data (Ben Allal et al., 2024).

At the same time, synthetic data is not a straightforward replacement for human-authored corpora. Generated text may inherit biases, omissions, stylistic narrowing, and contamination from the generator model and its prompts (Yu et al., 2023). Recursive training on model-generated data can also degrade diversity and cause model collapse when synthetic data displaces the original human distribution (Shumailov et al., 2024). Our experiments therefore study synthetic fiction as a potential substitute for copyrighted fiction while directly measuring both downstream utility and verbatim reproduction risk.

Shedding Light on Frontier Lab Practices Frontier model development is increasingly conducted behind closed doors: model providers often disclose high-level capabilities and safety results while withholding details about training data, filtering rules, data mixture weights, compute, architecture, and post-training procedures (Bommasani et al., 2025). This opacity has motivated a line of work sometimes described as "generative AI archaeology," where researchers infer hidden properties of deployed or partially open models from observable artifacts (Elliott, 2026). Examples include inferring training-data mixtures from BPE tokenizer merge lists (Hayase et al., 2024), detecting or extracting memorized training examples (Carlini et al., 2021; Nasr et al., 2025), identifying possible training-procedure signatures from model representations (BehnamGhader et al., 2024), and recovering architectural information or output-layer structure from API-accessible logits (Finlayson et al., 2024; Carlini et al., 2024). In contrast, open-science efforts such as Pythia & OLMo make training data, code, checkpoints, or documentation available so that claims about training decisions can be examined directly (Biderman et al., 2023; Olmo et al., 2025). Our work fits within this broader effort to make consequential model-development practices legible, but takes a complementary approach: rather than reverse-engineering a closed model after deployment, we run controlled pretraining experiments that isolate one disputed practice—the inclusion of copyrighted fiction—and measure its effect on downstream behavior.

C Pretraining Dataset Statistics

C.1 Dataset Statistics

Fig. 5 showcases the distribution of tokens across documents used during training. Table 3 additionally outlines some descriptive statistics about the data.

C.2 Curating Fiction Subset

To extract fiction books from the Harvard Institutional Book Corpus 1.0 (Cargnelutti et al., 2025), we feed title, author, publication year, and subject catalogs to Gemma-3-4B-it (Team et al., 2025) and ask it to predict whether the book is a work of fiction. To tag fiction books in Books3, we feed a 5000-character excerpt along with the title to Gemma-3-4B-it and ask it to tag with a yes or no option.

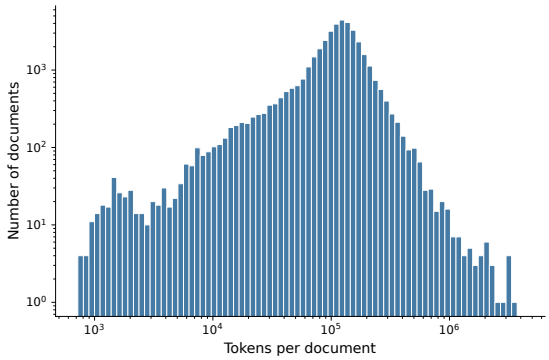
D Extended Evaluations Results

D.1 Training on Additional Book Tokens

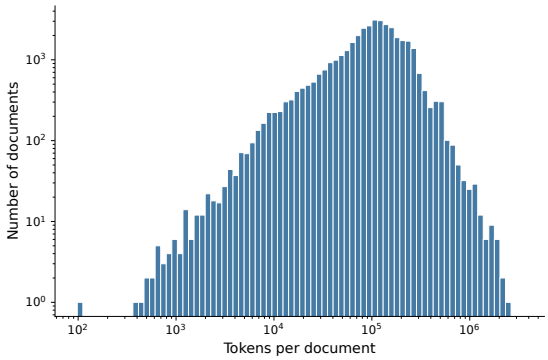
We train 1B dense models on 10B book tokens under two settings: (1) two passes over 5B tokens from Books3, and (2) a mixed corpus consisting of 5B tokens each from Books3 and Harvard. For these runs we use the 175B token subset of the WebOrganizer without creative writing corpus. Figure 6 and Table 4 present the results for base and instruct versions.

D.2 Impact of Books on Coding Capabilities

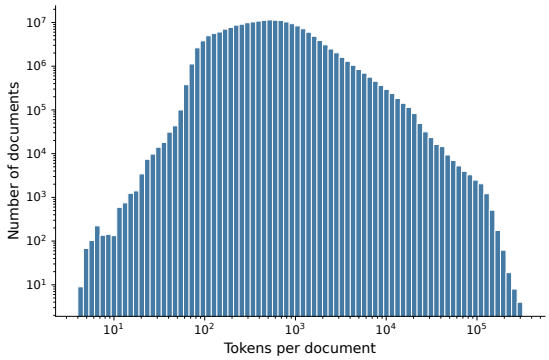
As shown in Figure 7, we observe that the no-books model outperforms Books3 on coding (Figure 7). On HumanEval (infill) (Bavarian et al., 2022), HumanEval (Chen et al., 2021), and MBPP (Austin et al., 2021), the no-books model scores notably higher, suggesting that displacing web data with fiction books comes at a cost to code-related capabilities. This tradeoff is worth noting when con-



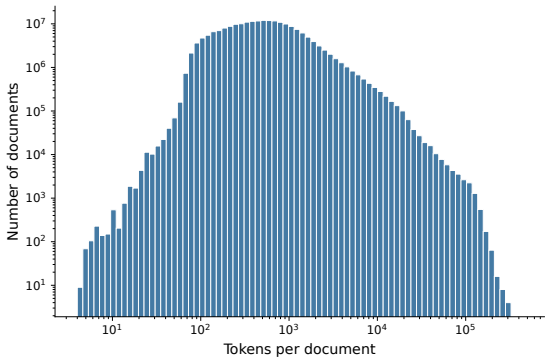
(a) Books3 (5B Tokens)



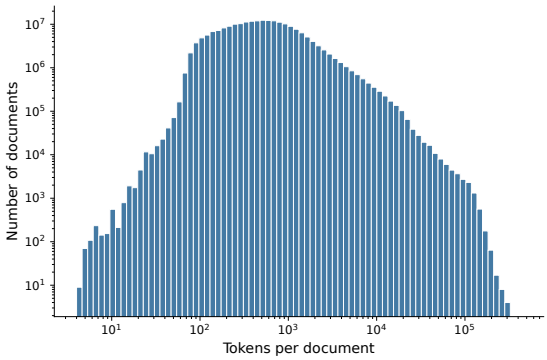
(b) Harvard (5B Tokens)



(c) WebOrganizer without Creative Writing (175B Tokens)



(d) WebOrganizer without Creative Writing (180B Tokens)



(e) WebOrganizer without Creative Writing (185B Tokens)

Figure 5: Histograms showcasing distribution of tokens across documents from different corpora used during training.

Table 3: Descriptive statistics of the evaluated corpora, including total token count, document count, and per-document token distribution metrics.

Corpus	Token Count	Number of Documents	Average Tokens per Document	Max Tokens per Document	Min Tokens per Document	Median Tokens per Document
Books3	5B	39851	125473.28	4157394	722	115792
Harvard	5B	38795	128883.07	3430212	98	105028
WebOrganizer	175B	198750225	880.50	319227	4	484
	180B	204425335	880.51	436154	4	484
	185B	210106644	880.50	436154	4	484

Eval	No Books	Books3	2xBooks3	Books3 + Harvard
AlpacaEval (win rate)	$28\% \pm 17\%$	$\mathbf{46.97}\% \pm 21.65\%$	$37.66\% \pm 21.37\%$	$4.21\% \pm 3.99\%$
AlpacaEval (len. controlled)	$6.38\% \pm 0.5\%$	$\mathbf{8.92}\% \pm 0.74\%$	$8.24\% \pm 0.66\%$	$1.76\% \pm 0.12\%$
MT-Bench (overall)	**1.075**	**1.075**	1.062	1.031
Writing	1.05	1.1	**1.15**	1.0
Roleplay	1.15	**1.25**	1.1	1.1
Extraction	**1.0**	**1.0**	**1.0**	**1.0**
Reasoning	1.0	1.0	**1.05**	1.0
Math	**1.0**	**1.0**	**1.0**	**1.0**
Coding	**1.0**	**1.0**	**1.0**	**1.0**
STEM	**1.2**	1.0	1.1	1.1
Humanities	1.2	**1.25**	1.1	1.05
LitBench	-0.95 ± 0.04	-0.76 ± 0.04	$\mathbf{-0.66} \pm 0.04$	-0.89 ± 0.04
WritingBench (overall)	1.844 ± 0.052	$\mathbf{1.845} \pm 0.047$	1.797 ± 0.047	1.651 ± 0.041
Academic & engineering	1.983 ± 0.135	$\mathbf{2.046} \pm 0.108$	1.844 ± 0.122	1.770 ± 0.105
Finance & business	1.818 ± 0.101	$\mathbf{1.826} \pm 0.104$	$\mathbf{1.826} \pm 0.090$	1.670 ± 0.083
Politics & law	1.742 ± 0.101	1.674 ± 0.118	$\mathbf{1.767} \pm 0.099$	1.656 ± 0.102
Literature & arts	1.641 ± 0.125	$\mathbf{1.710} \pm 0.120$	1.660 ± 0.112	1.499 ± 0.097
Education	$\mathbf{1.988} \pm 0.128$	1.894 ± 0.130	1.925 ± 0.128	1.650 ± 0.106
Advertising & marketing	$\mathbf{1.956} \pm 0.175$	1.944 ± 0.129	1.789 ± 0.135	1.642 ± 0.117

Table 4: Instruction-tuned model evaluation across four models: No Books, Books3 (copyrighted), 2xBooks3 (double the amount of Books3), and Books3 + Harvard (copyrighted and public domain combined). Bold indicates the best result per row. LitBench and WritingBench report mean $\pm$ 95% CI; AlpacaEval reports mean $\pm$ standard error. Note that absolute scores are lower as our models undergo lightweight instruction tuning for a single epoch.

sidering including books in the general-purpose pretraining mixture.

D.3 Training using Synthetic Fiction Data

Figure 8 shows the performance of the model trained on synthetic fiction books with the rest of the models. We find that the synthetically trained model outperforms the model trained on the Harvard Book Corpus on the narrative coherence category and comes close to the model trained on Books3 on LAMBADA.

E More Details for Book Data Classifier

This section reports details about our classifier training setup. Section E.2 presents the description of all the features used in training, and Section E.1 showcases feature importance coefficients while training the classifier with a mix of semantic, stylistic, and readability features.

E.1 Combining Semantic, Stylistic and Readability Features

Figure 9 shows the feature importance coefficients for the Books3 and Harvard Corpus classifier when training with a combination of semantic, stylistic, and readability features.

Figure 6: Base model evaluation across 12 skill categories comparing the three model variants: Books3 (5B tokens), 2xBooks3 (10B tokens), and Books3+Harvard (10B tokens). Books3 and 2xBooks3 outperform Books3+Harvard in narrative tasks (e.g., LAMBADA, HellaSwag), while performance on most other categories remains comparable across models. Dashed and dotted lines indicate random baselines for 4-choice and 2-choice tasks, respectively. Error bars denote standard error. Except for the narrative coherence category, we report the average across all tasks within a single category.

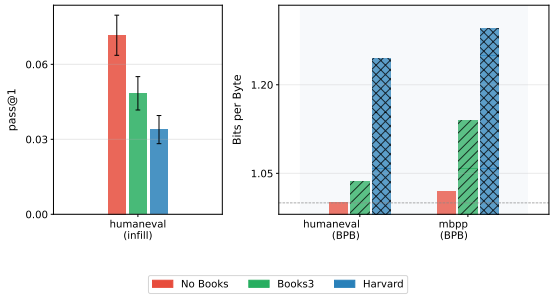


Figure 7: Coding evaluation on HumanEval (pass@1) and bits-per-byte (BPB) on HumanEval and MBPP. No Books achieves the best performance, followed by Books3 and Harvard. Error bars denote 95% CI for PPL metrics and standard error for coding.

Figure 8: Base model evaluation across 12 skill categories comparing four model variants. The model trained on synthetic data outperforms the model trained on the Harvard corpus on narrative coherence tasks, and comes close to the model trained on Books3. Dashed and dotted lines indicate random baselines for 4-choice and 2-choice tasks, respectively. Error bars denote standard error.

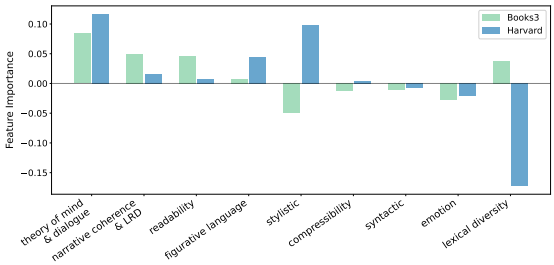


Figure 9: Feature coefficients from training classifiers for Books3 and the Harvard corpus using semantic, stylistic, and readability feature categories.

E.2 Feature Descriptions for Book Data Classifier

Feature	Description
Narrative Coherence & Long-Range Dependency	
Long-Range Lexical Dependency	How far back in the text the final content word first appears; high values indicate long-range context is needed to predict the ending.
Avg. Coreference Chain Length	Mean number of mentions per entity.
Max Coreference Chain Length	Length of the longest coreference chain, usually corresponding to the protagonist.
Max Entity Re-mention Distance	Furthest token gap between a noun and its pronoun.
Named Entity (PER) Density	Count of person entities normalized by total words.
Unique Character Count	Total number of distinct named actors in the text.
Temporal Marker Density	Frequency of words indicating passage of time (e.g., *suddenly, later, meanwhile*).
Spatial Preposition Density	Frequency of words anchoring spatial relationships (e.g., *beneath, across, through*).
Past Tense Verb Ratio	Proportion of verbs referring to past actions out of all verbs.
Theory of Mind & Dialogue	
Dialogue Percentage	Ratio of text spoken by characters versus narration.
Theory of Mind Markers	Frequency of constructions where one character models another's mental state (e.g., *he thought the stranger was afraid*).
Direct/Indirect Speech Ratio	Ratio of explicit quoted speech to reported speech.
1st-Person Pronoun Ratio	Density of first-person pronouns (*I, me, my, we, us*).
2nd-Person Pronoun Ratio	Density of second-person pronouns (*you, your, yourself*).
3rd-Person Pronoun Ratio	Density of third-person pronouns (*he, she, they, them*).
Impersonal Pronoun Ratio	Density of abstract pronouns (*it, one, this, that*).
Pronoun-to-Noun Ratio	Ratio of pronouns to nouns, measuring reliance on context versus explicit naming.
Cognitive Verb Density	Frequency of thinking and feeling verbs (e.g., *thought, realized, wondered, felt*).
Communication Verb Density	Frequency of speaking verbs (e.g., *said, shouted, whispered, muttered*).
Animacy Ratio	Proportion of actions performed by living versus non-living subjects.
Emotion	
Emotional Arc Shape	Measures the emotional trajectory by computing the amplitude and direction of sentiment across 10 equal text chunks using VADER.
Sentiment Volatility	Standard deviation of VADER sentiment scores across 5-sentence rolling windows.
Avg. Sentiment Subjectivity	Mean subjectivity score across the document using TextBlob.
Positive Emotion Lexicon Density	Frequency of joy and trust words from the NRC Emotion Lexicon.

Table 5

Feature	Description
Negative Emotion Lexicon Density	Frequency of fear, anger, and sadness words from the NRC Emotion Lexicon.
Surprise Lexicon Density	Frequency of surprise words from the NRC Emotion Lexicon.
Motion Verb Density	Frequency of physical movement verbs (e.g., *ran, walked, jumped*).
Stative/Active Verb Ratio	Ratio of stative verbs (*be, seem, appear*) to action verbs.
Present Tense Verb Ratio	Proportion of present tense verbs out of all verbs.
Lexical Diversity	
MTLD	Length-agnostic measure of vocabulary diversity.
Hapax Legomena Ratio	Proportion of words appearing exactly once in the document.
Lexical Density	Ratio of content words (nouns, verbs, adjectives, adverbs) to total words.
Adjective Density	Frequency of adjectives per total words.
Adverb Density	Frequency of adverbs per total words.
Exclusive Creative Noun Density	Frequency of nouns that appear only in creative writing across the training corpora.
Average Word Length	Mean number of characters per word.
Figurative Language	
Idiom Usage	Frequency of non-literal phrases matched against an idiom dictionary.
Simile Marker Density	Frequency of explicit comparisons using *like* or *as*.
Onomatopoeia Density	Frequency of sound-imitating words (e.g., *crash, buzz, hiss, bang*).
Archaic Word Density	Frequency of older English forms (e.g., *thee, thou, whilst, alas*).
Metaphor Density	Frequency of verb-object or adjective-noun pairs where a concrete modifier is applied to an abstract concept (e.g., *shattered hopes*), measured using the Brysbaert Concreteness Lexicon.
Syntactic Structure	
Mean Dependency Tree Depth	Average syntactic complexity measured by dependency tree depth per sentence.
Max Dependency Tree Depth	Depth of the most heavily nested sentence in the document.
Subordinate Clause Density	Frequency of subordinating conjunctions (e.g., *because, although*).
Coordinate Conjunction Density	Frequency of coordinating conjunctions (e.g., *and, but, or*).
Inverted Syntax Frequency	Frequency of sentences where the object precedes the subject, common in poetic writing.
POS Bigram Entropy	Shannon entropy of adjacent part-of-speech tag pairs, measuring grammatical unpredictability.
Stylistic Structure	
Parentheses Density	Frequency of parenthetical asides per total words.
Comma-to-Word Ratio	Frequency of commas, indicating clause-heavy descriptive writing.
Punctuation Variety Index	Shannon entropy of the distribution of punctuation marks.
Semicolon Frequency	Frequency of semicolons per sentence.

Table 5

Feature	Description
Ellipsis Frequency	Frequency of ellipses per total words.
Em-Dash Frequency	Frequency of em-dashes per total words.
Line Break Frequency	Frequency of line breaks per total words.
Sentence Length Variance	Variance in word count across sentences.
Average Sentence Length	Mean number of words per sentence.
Question Sentence Ratio	Proportion of sentences ending with a question mark.
Exclamatory Sentence Ratio	Proportion of sentences ending with an exclamation mark.
Compressibility	
Shannon Entropy	Information-theoretic measure of text complexity.
zlib Compression Ratio	Compression-based measure of text redundancy.
Readability	
Flesch-Kincaid Grade Level	Estimates the years of education required to understand the text based on syllables per word and words per sentence.
Dale-Chall Readability Score	Evaluates vocabulary complexity by measuring the proportion of words outside a baseline dictionary of 3,000 common English words.

Table 6: Complete list of features used to distinguish book data from other pretraining sources, organized by category.

E.3 Feature Coefficients

F Synthetic Data Construction

To obtain genres and plot summaries, we provide the book title and three 500-word excerpts from each book, taken from the top, middle, and bottom of the book, to Llama 3.1 8B Instruct (Grattafiori et al., 2024) and generate free-form genres and plot summaries. We also compute the ROUGE-L (Lin, 2004) score between each summary and the corresponding excerpts, and regenerate the summary if the score exceeds a threshold of 0.5. This check ensures that summaries do not contain verbatim sequences from the original books. We also extract a 500-word excerpt from each chapter, assuming a chapter length of roughly 1500 words.

Using the genre, plot summary, and excerpt, all drawn from the same book, we generate a 1000–2000 word synthetic chapter. Since a single book follows a particular narrative across thousands of words, our methodology allows us to generate multiple chapters from the same book while preserving overall narrative flow. To generate each chapter, we randomly sample from temperature values of 0.7, 0.9, and 1.1, where all chapters from the same book use the same temperature but values differ across books, and use top-p = 0.95. The diversity in our generation process comes from the combination of genres, summaries, excerpts, and sampling parameters, which is important since models trained on homogeneous synthetic data are prone to model collapse (Shumailov et al., 2024).

Finally, we remove any chapters that overlap with the Books3 training data. To do this, we build a hash array over all Books3 documents using a sliding window of 50 words. We then apply the same sliding window to each synthetic chapter and remove any chapter with one or more 50-word verbatim overlaps. We generate 6B tokens with a 1B buffer to obtain an effective 5B tokens of synthetic data. Of approximately 450k generated chapters, around 18k are removed due to verbatim overlap.

G Additional Evaluation Details

We run the following evaluations on our base models:

Evaluations from LM-Eval-Harness We use LM-Eval-Harness for LAMBADA (Paperno et al., 2016), StoryCloze (Mostafazadeh et al., 2016), FLUTE (Chakrabarty et al., 2022), Tomi (Sileo, 2023; Le et al., 2019), OpenBookQA (Mihaylov et al., 2018), ECare (Du et al., 2022), SciQ (Welbl et al., 2017), COPA (Roemmele et al., 2011), LogiQA (Liu et al., 2020), COQA (Reddy et al., 2019), MuTual (Cui et al., 2020), HumanEval (infilling) (Bavarian et al., 2022), ASDiv (Miao et al., 2021) & Hendrycks ETHICS (Hendrycks et al., 2021a)

Evaluations from OLMES We use OLMES for ARC Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), CSQA (Talmor et al., 2019), MMLU (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), Codex HumanEval (3-Shot) (Chen et al., 2021) & MBPP (3-Shot) (Austin et al., 2021)

Evaluations from BigBench We conduct BigBench evaluations (Srivastava et al., 2022) without using the frameworks above.

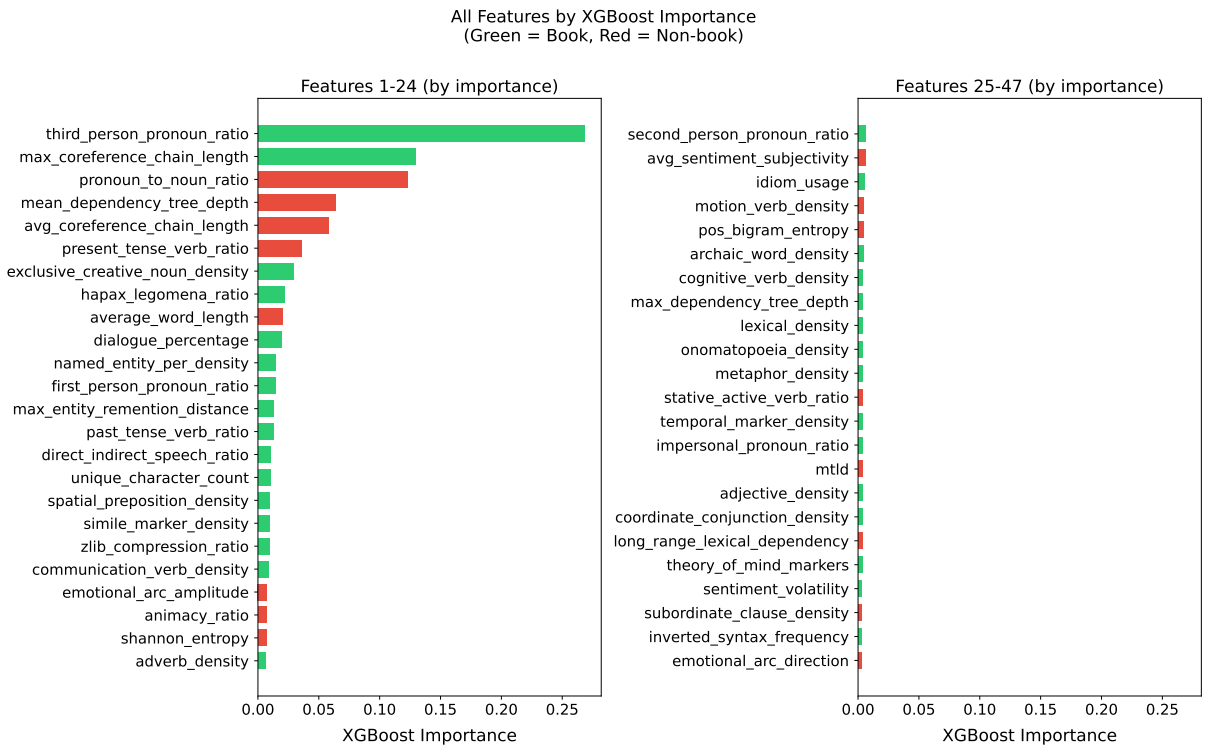


Figure 10: Feature coefficients for the Books3 classifier trained on semantic features.

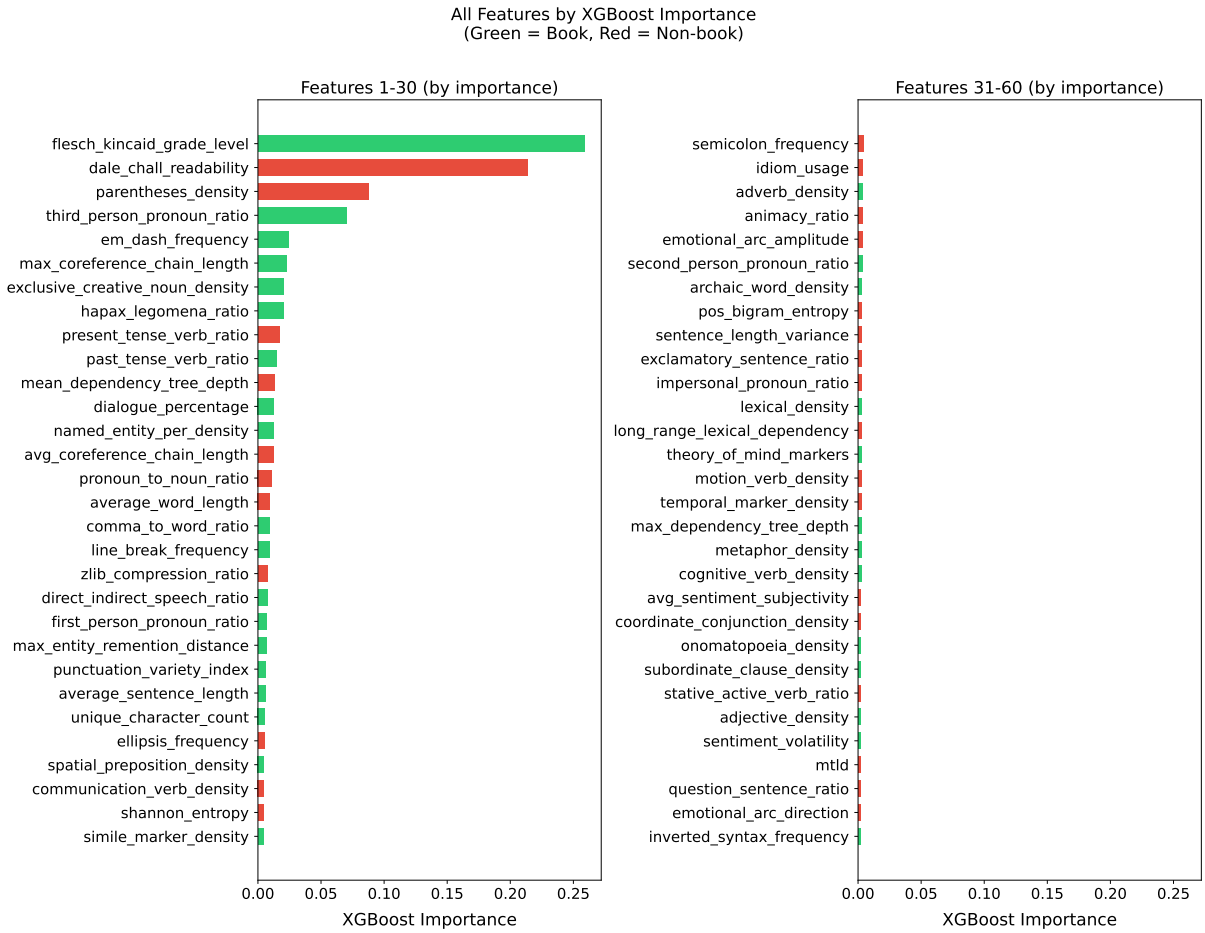


Figure 11: Feature coefficients for the Books3 classifier trained on semantic, stylistic, and readability features.

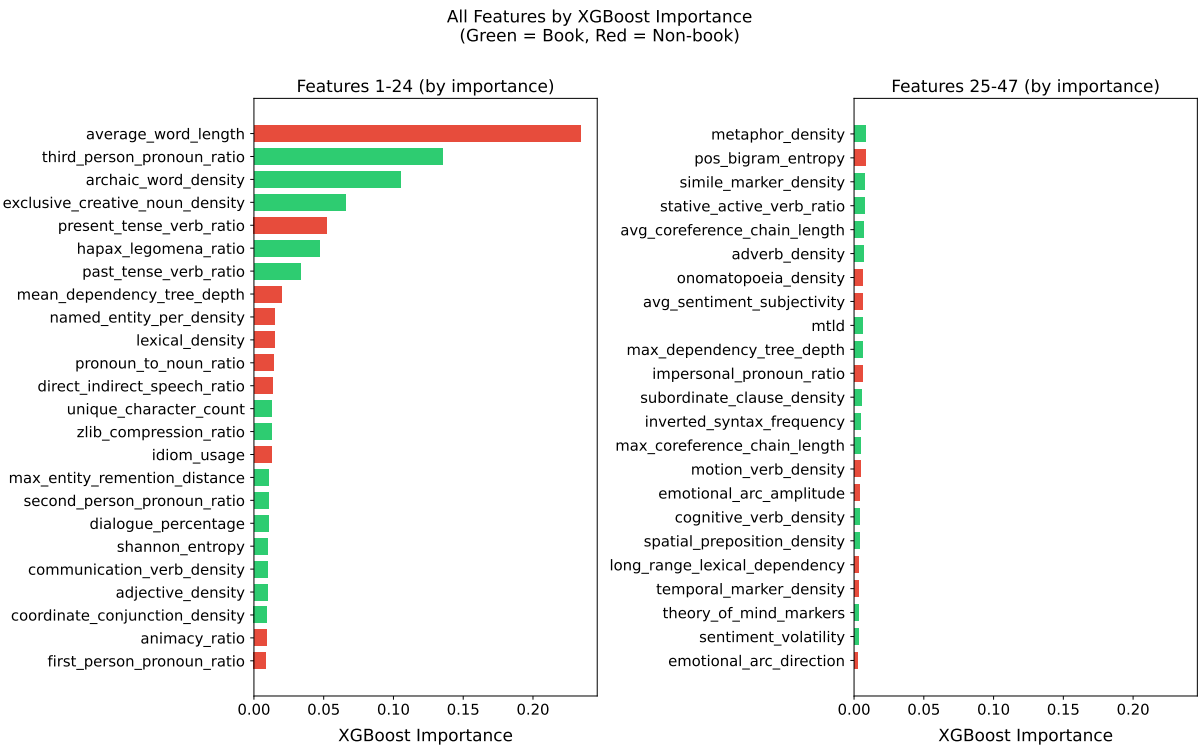


Figure 12: Feature coefficients for the Harvard classifier trained on semantic features.

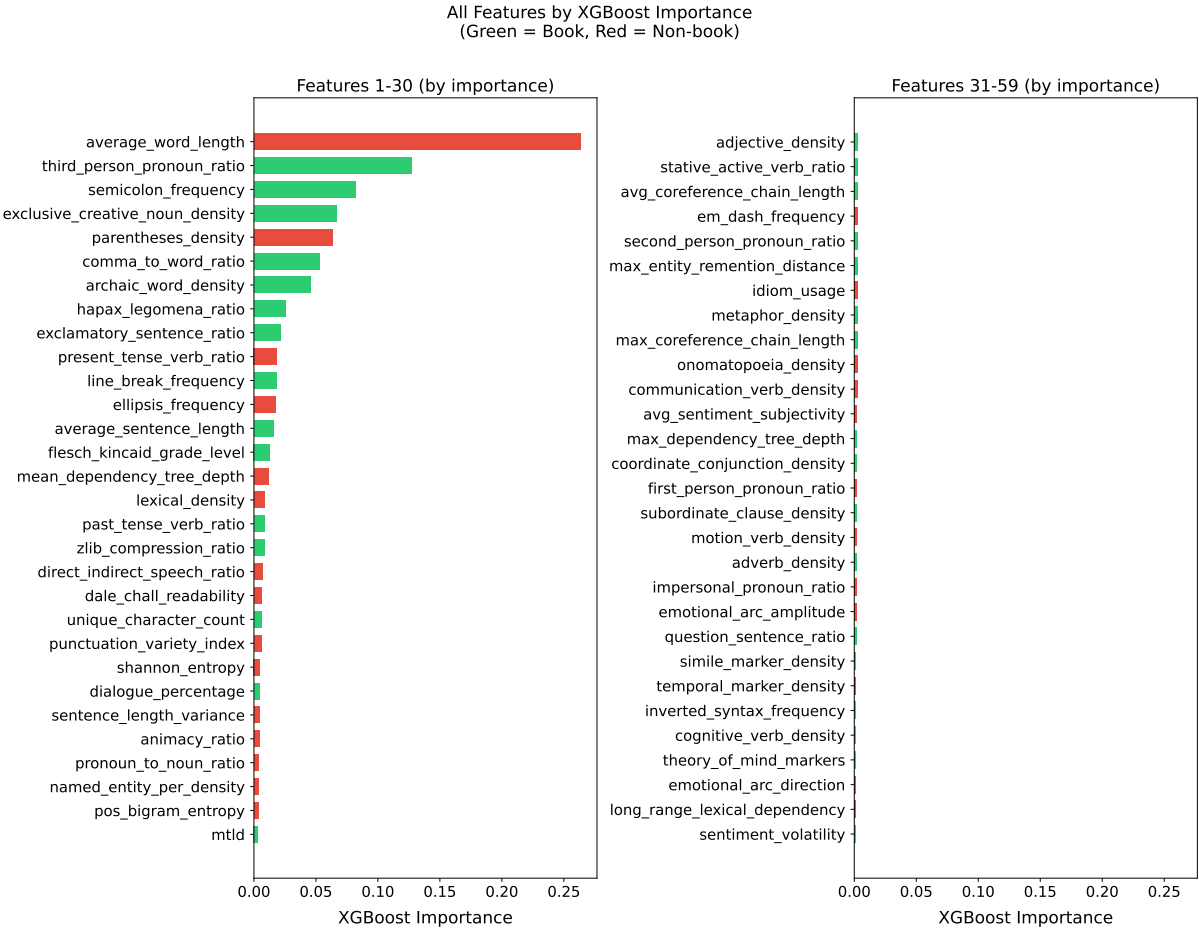


Figure 13: Feature coefficients for the Harvard classifier trained on semantic, stylistic, and readability features.

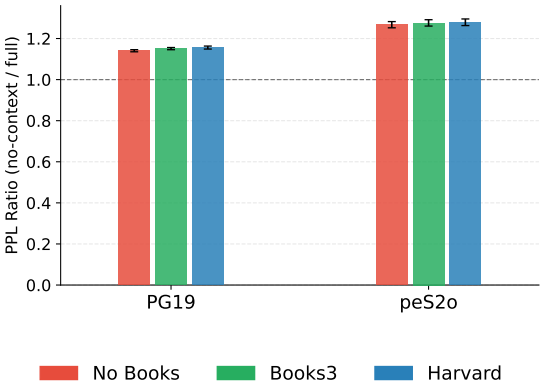


Figure 14: Context utilization ratio (PPL without context / PPL with full context) for all three models on PG19 and peS2o. A ratio above 1 indicates that the model benefits from longer context when predicting text. All three models show similar context utilization across both datasets.