BOOKS, COPYRIGHT, AND THE FUTURE OF FOUNDATION MODELS: AN EMPIRICAL STUDY OF FICTION IN PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

The use of books to train large language models has prompted numerous lawsuits against commercial AI providers, yet the reasons behind such data choices 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 on creative writing tasks, with contemporary fiction yielding gains beyond those from corpora based on books in the public domain. We then examine what makes fictional text distinct from other pretraining sources, and examine whether synthetic data can recover its benefits. We find that models we trained on our synthetic data recipe achieve performance comparable to models trained on contemporary fiction, while making it harder to establish provenance through membership inference and data extraction attacks. Together, these experiments offer the largest empirical study to date of contemporary 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 (LLMs) 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; Team, 2025a; 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 bring light to 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 distinct data sources, exhibiting systematic differences in their structural properties, narrative conventions, and communicative goals (Cutter, 1904; Altmann et al., 2014; Mar & 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

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

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. As some data may contain contentious or sensitive material, we describe our data handling practices, safeguards, and governance processes in our ethics statement (§6). 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 compared to models trained without books, and after being instruction-tuned, increased competence on creative writing tasks.

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

III. 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.

IV. We consider a novel synthetic data generation pipeline in which copyrighted books are never trained on directly, but instead serve as LLM input for generating diverse synthetic data. In the models we studied, this largely recovers the benefits of training on the original contemporary fiction, while making provenance substantially harder to establish via membership inference or data extraction. We discuss the associated ethical implications in Section 6.

Taken together, our results empirically demonstrate that books data can substantially improve model performance for particular writing tasks, but that books data is not homogeneous, 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., 2024), 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.

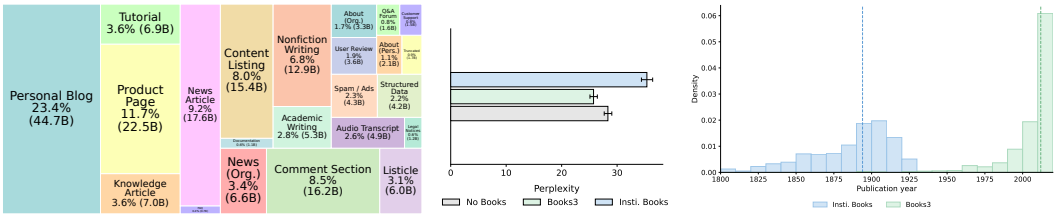
Harvard Institutional Books (Cargnelutti et al., 2025)[1]: A corpus of nearly one million digitized books that are in the public domain, intended to provide high-quality data for training large language models. The books were originally digitized from Harvard Library's participation in the Google Books 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., 2023). This is a widely used proprietary corpus, and in our analysis, we found $\sim$99% of books are potentially under copyright (more details in D.3).

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'* in Wettig et al. (2025)), we draw nested web subsets of 185B, 180B, and 175B tokens, with each smaller corpus contained within the next largest: the 180B-token corpus is a subset of the 185B-token corpus, and the 175B-token corpus is a subset of the 180B-token corpus. Format distribution over the corpora is shown in Fig. 1a. 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 (INSTITUTIONAL BOOKS) or 5B tokens from BOOKS3 (more details in Section D.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 (BOOKS3 + INSTITUTIONAL BOOKS). All resulting corpora contain 185B tokens, allowing us to compare models trained with no books, public-domain books, contemporary books, and a mixture of both while keeping the total token budget fixed.

Pre-training 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,

[1]Institutional Books provided by the Institutional Data Initiative with source material contributed by Harvard Library, available at https://institutional.org/institutional-books



(a) (b) (c)

Figure 1: (a) Distribution across web-organizer formats of the non-creative writing corpus spanning 191.4B tokens. All non-creative writing mixes are sampled from this mixture. (b) Story generation perplexity on WritingPrompts. BOOKS3 achieves the lowest perplexity. (C) Publication-year distributions of INSTITUTIONAL BOOKS (median 1894) and BOOKS3 (median 2012) fiction.

corresponding to approximately 185B tokens, with a fixed sequence length of 4096 tokens. Training used 4 nodes of either B200 or H200 GPUs. Additional pre-training details are outlined in App. E.1.

Post-training We post-train every base model with an identical two-stage pipeline consisting of supervised fine-tuning (SFT) followed by direct preference optimization (DPO) (Rafailov et al., 2024), replicating the OLMo-2-0425-1B & Tülu-3 recipe (Walsh et al., 2025; Lambert et al., 2025) using the `open-instruct` library (Olmo et al., 2025). Further details are outlined in App. E.2.

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).

We evaluate our instruction-tuned models along three axes: 1) **General instruction following**: we use the official pairwise templates and configurations of length-controlled AlpacaEval 2.0 (Taori et al., 2023), WildBench (Lin et al., 2024), and MT-Bench (Zheng et al., 2023), adapted to compare our model pairs. 2) **Creative writing**: a pairwise adaption of EQ-Bench Creative Writing v3 (Paech, 2024) together with absolute scores from the LitBench reward model (Fein et al., 2026), which reports high agreement with human judgements. 3) **Non-creative writing**: we construct prompt sets across six different categories such as correspondence, opinion/review, argumentative essays, advice/how-to pieces, blog/discussion articles, and rewriting/editing. We evaluate the generations using Prometheus 2 (Kim et al., 2024) (a pairwise LLM judge) and a length-controlled GPTScore (Fu et al., 2024) win-rate, together with automatic evaluation metrics GRUEN (Zhu & Bhat, 2020) and UniEval (fluency & coherence) (Zhong et al., 2022). Further details are in App. H.

3 WHAT ROLE DO BOOKS PLAY IN PRETRAINING DATA?

It is increasingly common to include books as 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, reported containing $\sim$ 6B tokens of book data. Among commercial models, GPT-3 was reported to be trained on 67B 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., 2023). 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.

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 THE NO BOOKS CONDITION

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

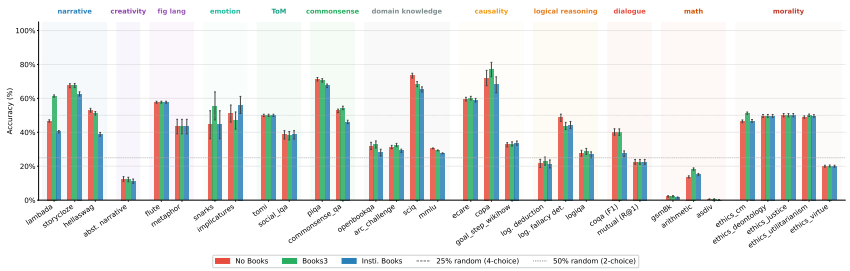


Figure 2: Base model evaluation across 12 skill categories comparing the three model variants. BOOKS3 outperforms both NO BOOKS and INSTITUTIONAL BOOKS in narrative tasks (e.g., LAMBADA, HellaSwag), while performance on most other categories remains comparable. Dashed/dotted lines indicate random baselines for 4-/2-choice tasks while error bars denote standard error.

models. To study this, we compare two models trained on the same base corpus, where one model replaces part of the corpus with Books3 (BOOKS3) and the other does not (NO BOOKS).

Fig. 2 shows the comparison between the two models. Across most general-purpose tasks, the two models perform comparably, 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 broad discourse context and long-range dependencies. Most notably, the BOOKS3 model shows a large improvement (14.7%) on LAMBADA (Paperno et al., 2016), a task that tests long-range language coherence. Consistent with this, as seen in Fig. 1b, 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 an identical recipe, as mentioned in Section 2, and evaluate them on general instruction-following and writing benchmarks (Tab. 1). We find that the BOOKS3 instruction-tuned model here significantly outperforms the NO BOOKS baseline on creative writing tasks such as LitBench and EQ-Bench. These results are consistent with our human evaluation study in App. I. Taken together with the base-model results in Fig. 1b, these results suggest that the creative-writing benefits associated with book pretraining persist, and may be amplified, after instruction tuning. On non-creative writing (judged by Prometheus 2), the BOOKS3 model reaches parity with NO BOOKS on correspondence (e.g., writing an email), opinion/review, and rewriting/editing, and trails only slightly on the remaining categories such as argumentative essays, advice/how-to pieces, and blog / discussion articles. This near-parity is corroborated when measuring the generation quality using automatic evaluation metrics such as GRUEN (intended to capture grammaticality, structure, focus, and coherence) and UniEval (coherence and fluency). This suggests that using BOOKS3 in pretraining may leave the fluency, grammaticality, and coherence of non-creative prose largely intact. BOOKS3 slightly harms performance compared to NO BOOKS on general instruction following tasks such as Alpaca, WildBench, and MT-Bench (win rates shown in Tab. 1).

Taken together, these results show that training on contemporary fiction improves literary and creative text capabilities while showing parity on some downstream non-creative writing skills. This raises a natural question: are contemporary (*often* copyrighted) fiction books uniquely valuable, or can public-domain books offer similar benefits?

3.2 COPYRIGHTED VS. PUBLIC DOMAIN BOOKS

Having established that training models on contemporary fiction improves downstream creative writing over the NO BOOKS baseline, we now ask whether public domain books can offer similar utility. We compare the model trained on BOOKS3 against the model trained on books from the INSTITUTIONAL BOOKS corpus, where all other training data is held constant.

First, we examine the base models. Amongst base models, the BOOKS3 model outperforms INSTITUTIONAL BOOKS across all narrative coherence tasks (Fig. 2). Notably, INSTITUTIONAL BOOKS performs worse than even the NO BOOKS baseline in this category, suggesting that these books actively hurt coherence modeling. The INSTITUTIONAL BOOKS model also underperforms BOOKS3 on emotion recognition (snarks), commonsense reasoning (commonsense_qa), and causal reasoning (copa). This pattern extends to WritingPrompts, where INSTITUTIONAL BOOKS achieves the highest perplexity (Fig. 1b).

Table 1: **Instruction-tuned evaluation.** Pairwise columns report win rates (%) with values above 50 favoring the first model; absolute scores are higher-is-better. BOOKS3 consistently outperforms INSTITUTIONAL BOOKS and outperforms NO BOOKS on creative writing, while trailing NO BOOKS on general instruction following. For Prometheus 2, bold subcategories indicate tasks where the BOOKS3–NO BOOKS win rate exceeds 50% for at least one seed. Prometheus 2 and GPTScore aggregate all six non-creative writing categories. Values are mean ± std. deviation over three seeds.

Instruction following *Pairwise win rate (%)*		**Books3 vs No Books**	**Inti. Books vs No Books**	**Books3 vs Insti. Books**
AlpacaEval		43.5 ± 2.2	25.7 ± 2.1	68.1 ± 0.8
WildBench		42.0 ± 1.7	21.0 ± 1.0	72.8 ± 2.0
MT-Bench		46.0 ± 2.0	23.8 ± 3.0	73.8 ± 3.3
Creative writing *Pairwise win rate (%)*		**Books3 vs No Books**	**Insti. Books vs No Books**	**Books3 vs Insti. Books**
EQ-Bench		78.6 ± 3.9	22.4 ± 5.0	91.1 ± 4.5
Absolute score	**No Books**	**Books3**	**Insti. Books**	**Books3 − Insti. Books**
LitBench	-3.16 ± 0.06	-2.92 ± 0.05	-3.38 ± 0.08	$+0.46 \pm 0.04$
Non-creative writing *Pairwise win rate (%)*		**Books3 vs No Books**	**Insti. Books vs No Books**	**Books3 vs Insti. Books**
Prometheus 2 (overall)		46.8 ± 0.1	35.7 ± 1.1	61.1 ± 1.0
Correspondence		47.8 ± 4.6	41.5 ± 4.8	58.5 ± 1.4
Opinion / review		46.9 ± 6.0	37.2 ± 6.9	57.4 ± 5.9
Argumentative		44.7 ± 4.5	31.9 ± 4.7	62.9 ± 1.9
Advice / how-to		45.5 ± 2.7	32.5 ± 1.6	64.5 ± 2.5
Blog / discussion / article		47.2 ± 0.5	32.8 ± 2.4	63.3 ± 3.3
Rewriting / editing		48.5 ± 3.1	37.8 ± 1.1	60.5 ± 3.9
GPTScore		43.3 ± 4.2	37.2 ± 2.4	57.0 ± 1.7
Absolute score	**No Books**	**Books3**	**Insti. Books**	**Books3 − Insti. Books**
GRUEN	0.368 ± 0.022	0.345 ± 0.022	0.280 ± 0.008	$+0.065 \pm 0.028$
UniEval	0.926 ± 0.003	0.913 ± 0.003	0.873 ± 0.004	$+0.040 \pm 0.003$

Next, we examine the instruction-tuned models. As seen in Tab. 1, models trained on contemporary books achieved considerably better performance than those trained on books from the INSTITUTIONAL BOOKS corpus on creative writing tasks (a win rate of 91.1% on EQ-Bench, and an absolute score increase of 0.46 on LitBench). These models also outperform models trained on public-domain books in other kinds of writing tasks, and demonstrate improved instruction-following capability on the instruction-following benchmarks. These results are consistent with our human evaluation study in App. I. These results offer evidence that books from BOOKS3 provide a significant advantage over public domain books from the INSTITUTIONAL BOOKS corpus on coherence, writing, and instruction-following tasks, and that training on INSTITUTIONAL BOOKS not only fails to match BOOKS3, but in all of the tasks demonstrates worse performance than *not training on any books at all*. We also observe that the model trained on the mixture of BOOKS3 and INSTITUTIONAL BOOKS outperforms the model trained only on INSTITUTIONAL BOOKS, while still lagging behind models trained on a single or double pass (5/10B tokens) over BOOKS3. See F.1 for more details.

We consider the following candidate explanations for this gap. The first is digitization errors: since books from the INSTITUTIONAL BOOKS corpus are digitized from physical holdings in the Harvard Library's collections, they may contain significant Optical Character Recognition (OCR) errors. We considered this explanation less likely, as we used a post-processed version of the corpus that has passed through an extensive cleaning pipeline Cargnelutti et al. (2025). To verify this, we manually inspect 100 snippets of 100 words each, drawn from 100 distinct INSTITUTIONAL BOOKS documents. None contain unintelligible text; however, 16% exhibit layout artifacts, such as page numbers inserted mid-sentence, which may introduce noise for language modeling. The second is temporal distribution shift. The two corpora differ substantially in publication date (Figure 1c; see Section D.3): the median publication year is 2012 for BOOKS3 and 1894 for INSTITUTIONAL

Table 2: **Perplexity comparison, BOOKS3 vs. INSTITUTIONAL BOOKS.** (a) Open-weight base models on 1,000 sequences sampled from the BOOKS3 and INSTITUTIONAL BOOKS corpora (mean perpelxity). (b) Our BOOKS3- and INSTITUTIONAL BOOKS-trained 1B models, per-prompt mean perplexity on the instruction text of each benchmark.

(a)

Model	Books3	Institutional Books
Qwen3-1.7B-Base	17.83	20.64
Llama-3.2-1B	16.94	22.59
OLMo-2-0425-1B	17.37	25.78

(b)

Benchmark	Books3	Institutional Books
WildBench	180.1	258.8
LitBench	49.1	63.2
Non-CW (aggregated)	97.7	126.5

BOOKS, a difference of 118 years. The feature-based classifiers trained in Section 3.3 support this account: archaic word density is a high-importance feature for INSTITUTIONAL BOOKS (Fig. 8) but not for BOOKS3. To further test this explanation, we measure how well models trained on each corpus predict contemporary text. We compute the perplexities of models trained on BOOKS3 and INSTITUTIONAL BOOKS data on WildBench instructions, which are drawn from real user-chatbot conversations in WildChat Zhao et al. (2024). We observe that the INSTITUTIONAL BOOKS model exhibits much higher perplexity than the BOOKS3 model on these realistic user requests(Tab. 2a). Similarly, we compute the perplexities on instructions from the creative and non-creative writing benchmarks, observing similar trends. Conversely, we measure how far each corpus lies from the text distribution of modern language models by computing the perplexity of sequences sampled from each corpus under three modern base models: Qwen3-1.7B-Base, Llama-3.2-1B, and OLMo-2-0425-1B (setup details in Appendix H.3). Across all three models, BOOKS3 sequences have lower perplexity, with a relative reduction of at least 15.77% (Tab. 2a). Genre composition may also contribute: in a manual inspection of 100 randomly sampled INSTITUTIONAL BOOKS books, we find 5% are verse novels, i.e., works of narrative fiction written in verse. In contrast, we find no such cases in an identical inspection of 100 randomly sampled BOOKS3 books.

In summary, under matched training conditions, models pretrained on contemporary books (BOOKS3) outperform models pretrained on public-domain books from INSTITUTIONAL BOOKS on the writing and instruction-following benchmarks we evaluate, and this advantage widens after instruction tuning. These findings offer one possible explanation for the demand for contemporary fiction such as BOOKS3 as pretraining data.

3.3 WHAT DISTINGUISHES BOOK DATA FROM OTHER PRETRAINING DATA SOURCES?

Given that adding book data improves coherence and literary generation (Section 3), we ask what properties distinguish books from other common pretraining sources. We compare books against six common pretraining sources: general web text, Reddit, math, code, academic papers, and news, using 62 features spanning narrative coherence and long-range dependency, theory of mind and dialogue, emotion, lexical diversity, figurative language, syntax, style, compressibility, and readability (full details in Appendix C.2). We fit an L1-regularized logistic regression to select discriminative features, then train XGBoost classifiers to distinguish book from non-book text. We train separate classifiers for BOOKS3 and INSTITUTIONAL BOOKS, while keeping the non-book distribution fixed. Books are strongly distinguishable from other pretraining sources, but BOOKS3 and INSTITUTIONAL BOOKS differ in what makes them distinctive. Among semantic features, the most important categories for both corpora are those associated with theory of mind and dialogue, narrative coherence and long-range dependency, and figurative language. BOOKS3 places greater weight on the first two: maximum coreference-chain length and named-entity density rank highly, consistent with long character-driven narratives. BOOKS3 additionally exhibits greater lexical diversity. In contrast, figurative language is more prominent for INSTITUTIONAL BOOKS, with archaic-word density among its strongest features, consistent with the older makeup of public-domain text. Among stylistic and readability features, readability is particularly informative for BOOKS3, while INSTITUTIONAL BOOKS is more strongly characterized by stylistic features such as semicolons, em-dashes, parentheses, and varied punctuation. With all features, classifiers distinguish BOOKS3 and INSTITUTIONAL BOOKS from non-book text with 97.3% and 99.1% accuracy, respectively. For an extended discussion please refer to App. C.

4 IS IT FEASIBLE TO REPLICATE THE VALUE OF BOOKS DATA?

Our preceding experiments indicate that contemporary fiction books can be a high-quality data source for improving model performance on writing tasks. In particular, contemporary fiction yields larger gains than public-domain fiction, and its inclusion is associated with improvements in long-range coherence and creative writing quality. To this end, we explore whether it is possible to construct synthetic data for model pretraining, that can replicate the value of natural fiction book data. Such a procedure might be lucrative for providers, especially as 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), or leak membership details (Shokri et al., 2017). Even if high-quality book data is lawfully acquired, models may regurgitate portions of its contents verbatim. Providers may therefore seek alternatives that provide similar utility with reduced risk of regurgitation or membership leakage. However, synthetic data derived from contentious material has the potential to complicate data provenance, since it may bear low resemblance to the original source content.

4.1 PRETRAINING ON SYNTHETIC FICTION

We investigate whether synthetic fiction can match the pretraining utility of natural contemporary fiction. Our goal is to generate data that is similar in distribution (e.g., topics, diversity, etc.) to high-quality contemporary fiction.

Our generation pipeline conditions on books from the Books3 dataset. From each book in the Books3 dataset, we first remove any boilerplate text that isn't fiction by discarding the first 1000 words and the final ∼5% of words. Let W denote the number of remaining words. We partition the book into $N = \lfloor W/1500 \rfloor$ contiguous segments of ∼1,500 words each, which serve as pseudo-chapters. Then, we generate a pseudo-genre and plot summary for the book: we extract three 500-word excerpts from different positions in the book: top (first 500 words), middle (500 words starting at word $W/2$), and end (bottom 500 words). We then feed the title and these three excerpts to Llama 3.1 8B Instruct (Llama Team, 2024) to generate a free-form genre and plot summary. Next, from each chapter $n \in \{1, \ldots, N\}$, we extract the first 500 words to get a local excerpt. Using the generated genre, generated plot summary, and extracted excerpt, all drawn from the same book, we craft three different prompt templates as shown in Appendix G. We generate a 1000–2000-word synthetic chapter using Gemma 3 12B Instruct (Team, 2025b).

Since a single book follows a particular narrative across thousands of words, this methodology allows us to generate multiple chapters from the same book. 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. To mitigate the risk of model collapse associated with homogeneous synthetic data (Shumailov et al., 2024), we aim to induce diversity from the combination of genres, summaries, excerpts, prompt templates, and sampling parameters. Finally, we remove any synthetically generated chapters that share at least a 50-word verbatim overlap with any chapters in the original BOOKS3 corpus. Additional details are outlined in Appendix G. We pretrain on a 185B-token mixture comprising 5B tokens of synthetic fiction and the same 180B tokens of non-book data used in our other experiments (following identical training).

Results show that the *models we trained on synthetic data achieve comparable performance to the models trained on data from* BOOKS3 *across all instruction following, creative, and non-creative writing benchmarks* (Tab. 3). The largest gains are on creative writing: the SYNTHETIC model achieves a win rate of $79.7\% \pm 6.8$ against BOOKS3, and achieves a score of -1.72 ± 0.23 against BOOKS3's -2.92 ± 0.05 on LitBench. On non-creative writing tasks, the SYNTHETIC model achieves a win rate of $53.6\% \pm 3.1$ on Prometheus 2 and $52.3\% \pm 1.8$ on GPTScore vs. BOOKS3. Furthermore, it achieves a score of 0.394 ± 0.02 against BOOKS3's 0.345 ± 0.02 on GRUEN, and 0.925 ± 0.006 against BOOKS3's 0.913 ± 0.003 on UniEval. On instruction-following tasks, the SYNTHETIC model achieves win rates of $53.0\% \pm 2.6$ on AlpacaEval, $56.5\% \pm 1.4$ on WildBench, and $54.1\% \pm 0.6$ on MT-Bench against BOOKS3. On 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 (Fig. 13). 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 speculate that the diversity in our data generation process helps mitigate this phenomenon.

Taken together, these results suggest that synthetic data may present a viable alternative to natural copyrighted fiction for training language models with strong writing capabilities, particularly for

Table 3: Synthetic instruct model evaluation expressed as a pair-wise win rate % against BOOKS3, No Books, and Harvard. A win rate $> 50\%$ means the first model in the pair wins. Synthetic model outperforms BOOKS3 at all tasks and surpasses the No Books baseline on the creative writing task, MT-Bench, and Prometheus 2 (highlighted in bold); while trailing No Books only slightly on WildBench and AlpacaEval. Both non-creative writing evals (Prometheus 2, GPTScore) aggregate all six writing categories. All numbers are mean $\pm$ standard deviation across three seeds.

Category	Eval	Synthetic vs Books3	Synthetic vs No Books	Synthetic vs Insti. Books
Instruction following	AlpacaEval	53.0 ± 2.6	46.1 ± 2.1	70.5 ± 1.6
	WildBench	56.5 ± 1.4	49.2 ± 3.1	76.3 ± 0.6
	MT-Bench	54.1 ± 0.6	$\mathbf{50.8 \pm 3.1}$	74.4 ± 2.2
Creative writing	EQ-Bench	79.7 ± 6.8	$\mathbf{89.1 \pm 9.8}$	92.2 ± 6.8
Non-creative writing	Prometheus 2	53.6 ± 3.1	$\mathbf{51.0 \pm 2.5}$	64.6 ± 2.6
	Correspondence	50.0 ± 3.3	$\mathbf{52.0 \pm 5.9}$	58.1 ± 2.7
	Opinion / review	54.5 ± 11.2	49.7 ± 5.6	60.3 ± 8.5
	Argumentative	54.7 ± 2.0	$\mathbf{52.2 \pm 4.2}$	68.5 ± 1.0
	Advice / how-to	55.9 ± 2.4	49.0 ± 2.2	69.7 ± 2.4
	Blog / discussion / article	54.8 ± 1.7	$\mathbf{52.7 \pm 2.2}$	69.4 ± 1.7
	Rewriting / editing	51.6 ± 4.0	$\mathbf{50.4 \pm 3.2}$	62.3 ± 2.3
	GPTScore	52.3 ± 1.8	45.3 ± 1.5	58.9 ± 1.2

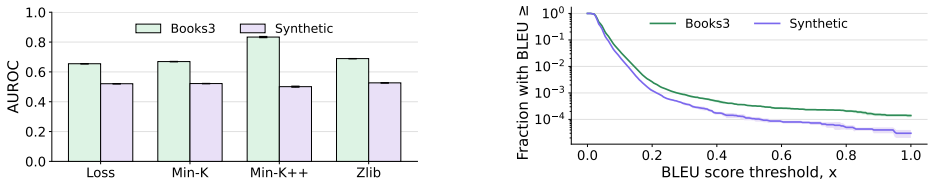
instruction-following and generative writing tasks. At the same time, synthetic data poses a serious risk of obscuring data provenance, as we demonstrate below.

4.2 DATA EXTRACTION & MEMBERSHIP INFERENCE: SYNTHETIC VS. COPYRIGHTED PRETRAINING

We study the SYNTHETIC model's susceptibility to membership inference (Shokri et al., 2017) and training data extraction attacks (Carlini et al., 2021) targeting the original BOOKS3 training data, relative to model trained directly on BOOKS3. For membership inference, we specifically run Loss-based (Yeom et al., 2018), MinK% (Shi et al., 2024), MinK%++ (Zhang et al., 2025) & ZLib (Carlini et al., 2021) using *open-unlearning* (Dorna et al., 2025). Each model is evaluated on the same 1,000 BOOKS3 training documents (members) and 1,000 BOOKS3 validation documents (non-members). As shown in Fig. 3a, models trained on synthetic versions of the text are substantially less susceptible to membership inference attacks. Across all context lengths, their AUROC values remain close to random-chance performance, whereas models trained directly on BOOKS3 become increasingly distinguishable as the context length grows, which is inline with observations in Wei et al. (2026).
For training-data extraction, we randomly draw 100K unique 100-token sequences from the BOOKS3 training set. For each sequence, we feed the first 50 tokens as a prefix to each base model and sample five 50-token continuations (top-k=40). Next, we compute the BLEU score (Papineni et al., 2002; Ippolito et al., 2023) between each continuation and the 50-token ground-truth suffix and keep the maximum over the five (a best-of-5 attack). For each BLEU threshold, we plot the fraction of sequences whose (best) BLEU matches or exceeds it, averaged over three training seeds (Fig. 3b). The BOOKS3 model consistently reproduces more copyrighted BOOKS3 text than the SYNTHETIC model across all thresholds and seeds. Since memorization increases with model size and data repetition (Carlini et al., 2023), we expect this separation to grow at larger scales.

5 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., 2024). 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



(a) Membership leakage

(b) Memorization.

Figure 3: (a) **Membership inference attack on BOOKS3 training text**: models trained on synthetic text are less susceptible than those trained directly on BOOKS3 (b) **Memorization of BOOKS3 training text.** Fraction of 100k sequences with reconstruction BLEU $\geq x$. The y-axis is log-scaled and values show mean over three seeds ± 1 SD. BOOKS3 $>$ SYNTHETIC at every threshold.

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 & 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 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 & 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. Further discussion is provided in App. B

6 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 and that reduces membership inference performance to near-chance, 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.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding, the generation of synthetic data (as outlined in Sec. 4), and editing of portions of the manuscript. Generative AI was used to assist with rephrasing, primarily to improve the clarity of text. We have not used generative AI tools to generate any of this work's main content, ideas, claims, provide feedback on research methodology or experiments, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. All AI-assisted writing was reviewed and verified, and AI-assisted code was verified and tested for correctness by two authors. The rest of the required disclosure tasks such as using AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and assist with translation are not applicable to this work.

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.

We would also like to discuss the ethical considerations surrounding procedures such as the synthetic data procedure we outline in Section 4. Our results show that model providers could potentially transform copyrighted books into synthetic training data that substantially improves downstream performance, while leaving little trace of the source works. They also appear to leave auditing through membership inference largely ineffective. Such procedures do appear to reduce the risks of verbatim regurgitation, and this might be presented as a safeguard. However, this doesn't address the underlying concerns at play: the works are still used without authors' consent, the synthetic data may still use non-literal aspects of the original data, and the resulting model trained on it could generate content that substitutes for, or dilutes the market for, the original works. Further, since our findings suggest that auditing approaches based on detecting memorization can be circumvented, meaningful accountability may require providers to disclose their training sources rather than relying on external post-hoc auditing. In this work, we chose to describe this synthetic data generation procedure because it is simple to implement, and we believe that researchers, auditors, authors, and policymakers are better served by knowing it is possible.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of our results, all data, code, and execution environments shall be made publicly available. A detailed description of our LLM training setup is provided in Section 2, dataset statistics are reported in Section 2 & Appendix D and evaluation details are reported in Appendix H.

REFERENCES

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

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, Phil Wang, and Samuel Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch. https://www.github.com/eleutherai/gpt-neox, 9 2023. Version 2.0.0.

Anthropic. Claude, 2025. URL https://claude.ai. Large language model.

Anthropic. System card: Claude sonnet 4.6. https://www.anthropic.com/claude-sonnet-4-6-system-card, February 2026. Accessed: [insert date you accessed it].

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Mo Bavarian, Heewoo Jun, Nikolas A. Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle. *ArXiv*, abs/2207.14255, 2022. URL https://api.semanticscholar.org/CorpusID:251135268.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024. URL https://arxiv.org/abs/2404.05961.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, 2024. URL https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.

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

Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988.

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

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

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

Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The 2024 foundation model transparency index, 2025. URL https://arxiv.org/abs/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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911, 2014. URL https://api.semanticscholar.org/CorpusID:34985620.

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

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

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

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. Stealing part of a production language model, 2024. URL https://arxiv.org/abs/2403.06634.

Tuhin Chakrabarty and Paramveer S. Dhillon. Can good writing be generative? expert-level ai writing emerges through fine-tuning on high-quality books, 2026. URL https://arxiv.org/abs/2601.18353.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7139–7159, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.481. URL https://aclanthology.org/2022.emnlp-main.481/.

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

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7312–7327, 2023.

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

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, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15134–15158, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.844. URL https://aclanthology.org/2024.emnlp-main.844/.

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

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dorrit Cohn. *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton University Press, 1978. ISBN 9780691063690. URL http://www.jstor.org/stable/j.ctvzxx9vd.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

A. Feder Cooper, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Mark Lemley, Daniel E. Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=SUoe3q5gXU.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1406–1416, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.130. URL https://aclanthology.org/2020.acl-main.130/.

Charles Ammi Cutter. *Rules for a dictionary catalog*. US Government Printing Office, 1904.

Edgar Dale and Jeanne S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28, 1948. ISSN 15554023. URL http://www.jstor.org/stable/1473169.

Peter Deutsch and Jean-Loup Gailly. Zlib compressed data format specification version 3.3. *RFC*, 1950:1–11, 1996. URL https://api.semanticscholar.org/CorpusID:57003200.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 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*, pp. 1286–1305, 2021.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics, 2025. URL https://arxiv.org/abs/2506.12618.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.33. URL https://aclanthology.org/2022.acl-long.33/.

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

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *International Conference on Learning Representations*, volume 2024, pp. 7735–7790, 2024.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

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

David Elson, Nicholas Dames, and Kathleen McKeown. Extracting social networks from literary fiction. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre (eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 138–147, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1015/.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. LitBench: A benchmark and dataset for reliable evaluation of creative writing. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7740–7755, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.362. URL https://aclanthology.org/2026.eacl-long.362/.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRcYFm8vyB.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

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

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, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

T. Givón. Topic continuity in discourse: An introduction. 1983. URL https://api.semanticscholar.org/CorpusID:140765004.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations, 2025. URL https://arxiv.org/abs/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. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Skyler Hallinan, Jaehun Jung, Melanie Sclar, Ximing Lu, Abhilasha Ravichander, Sahana Ramnath, Yejin Choi, Sai Praneeth Karimireddy, Niloofar Mireshghallah, and Xiang Ren. The surprising effectiveness of membership inference with simple n-gram coverage. *arXiv preprint arXiv:2508.09603*, 2025.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*, 2025.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah Smith. Data mixture inference attack: Bpe tokenizers reveal training data compositions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8956–8983. Curran Associates, Inc., 2024. doi: 10.52202/079017-0285. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10e6dfea9a673bef4a7b1cb9234891bc-Paper-Conference.pdf.

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

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß (eds.), *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.3. URL https://aclanthology.org/2023.inlg-main.3/.

Inc. Kadrey v. Meta Platforms, 2025. URL https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv03417/415175/598/. Order denying plaintiffs' motion for partial summary judgment and granting Meta's cross-motion on fair use grounds.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset, 2018. URL https://arxiv.org/abs/1810.09305.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. Longlamp: A benchmark for personalized long-form text generation, 2024. URL https://arxiv.org/abs/2407.11016.

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, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 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)*, pp. 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://www.aclweb.org/anthology/D19-1598.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. 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, 2024. Full version forthcoming in Journal of the Copyright Society.

Mark Lemley. How generative ai turns copyright law upside down. *Science and Technology Law Review*, 25(2), Jun. 2024. doi: 10.52214/stlr.v25i2.12761. URL https://journals.library.columbia.edu/index.php/stlr/article/view/12761.

Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dav, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: in search of the next generation of training sets for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL https://arxiv.org/abs/2406.04770.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld, Rock Yuren Pang, Pang Wei Koh, Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, and Jesse Dodge. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 178–188,

Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.18. URL https://aclanthology.org/2025.acl-demo.18/.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020. URL https://arxiv.org/abs/2007.08124.

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

AI@Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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. The data provenance initiative: A large scale audit of dataset licensing attribution in ai, 2023. URL https://arxiv.org/abs/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. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.179. URL https://aclanthology.org/2024.naacl-long.179/.

Inc. Lyon v. Adobe, 2025. URL https://dockets.justia.com/docket/california/candce/5:2025cv10732/461340. Class action complaint alleging direct copyright infringement;.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. Coherence boosting: When your pretrained language model is not paying enough attention, 2022. URL https://arxiv.org/abs/2110.08294.

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

Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data, 2023. URL https://arxiv.org/abs/2303.03919.

Philip M. McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392, 2010. doi: 10.3758/BRM.42.2.381.

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

Meta. 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/, 2024.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers, 2021. URL https://arxiv.org/abs/2106.15772.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon, 2013. URL https://arxiv.org/abs/1308.6297.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories, 2016. URL https://arxiv.org/abs/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. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vjel3nWP2a.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Brighter than gold: Figurative language in user generated comparisons. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2008–2018, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1215. URL https://aclanthology.org/D14-1215/.

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

Inc. et al. O'Nan et al. v. Databricks, 2024. URL https://www.courtlistener.com/docket/68325564/onan-v-databricks-inc/. 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.

OpenAI. GPT-4 technical report. *arXiv preprint arxiv:2303.08774*, 2023. URL https://arxiv.org/pdf/2303.08774.pdf.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. F-coref: Fast, accurate and easy to use coreference resolution. In Wray Buntine and Maria Liakata (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 48–56, Taipei, Taiwan, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-demo.6. URL https://aclanthology.org/2022.aacl-demo.6/.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. CoEdIT: Text editing by task-specific instruction tuning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5274–5291, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.350. URL https://aclanthology.org/2023.findings-emnlp.350/.

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.

Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula, and Yejin Choi. Information-guided identification of training data imprint in (proprietary) large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1962–1978, 2025.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), November 2016. ISSN 2193-1127. doi: 10.1140/epjds/s13688-016-0093-1. URL http://dx.doi.org/10.1140/epjds/s13688-016-0093-1.

Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge, 2019. URL https://arxiv.org/abs/1808.07042.

Alex Reisner. The unbelievable scale of AI's pirated-books problem. *The Atlantic*, March 2025. URL https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/.

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

Ted Ruffman, Lance Slade, and Elena Crowe. The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development*, 73(3):734–751, 2002. doi: 10.1111/1467-8624.00435.

Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023. doi: 10.1126/science.adi0656. URL https://www.science.org/doi/abs/10.1126/science.adi0656.

Aaron Schaffer, Will Oremus, and Nitasha Tiku. Anthropic 'destructively' scanned millions of books to build Claude. *The Washington Post*, January 2026. URL https://www.washingtonpost.com/technology/2026/01/27/anthropic-ai-scan-destroy-books/.

Elena Semino and {M. H.} Short. *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing.* Routledge, 2004. ISBN 0415286697.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

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, Pang Wei Koh, Hannaneh Hajishirzi, Ali Farhadi, and Sewon Min. FlexOLMo: Open language models for flexible data use. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=1rUj9ZN6Bz.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models . In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, May 2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

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

Damien Sileo. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*, 2023. URL https://arxiv.org/abs/2301.05948.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/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, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

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

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. doi: https://doi.org/10.1002/asi.21001. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 613–624, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

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

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025a. URL https://arxiv.org/abs/2507.06261.

Gemma Team. Gemma 3 technical report, 2025b. URL https://arxiv.org/abs/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. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In Regina Barzilay and Mark Johnson (eds.), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 680–690, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1063/.

Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352, 1998. URL https://api.semanticscholar.org/CorpusID:30607737.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. The 2025 foundation model transparency index, 2025. URL https://arxiv.org/abs/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. Hubble: a model suite to advance the study of LLM memorization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZfdnZhOP0k.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation, 2025. URL https://arxiv.org/abs/2502.10341.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 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*, 2023. URL https://openreview.net/forum?id=6hZIfAY9GD.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. Animacy encoding in English: Why and how. In *Proceedings of the Workshop on Discourse Annotation*, pp. 118–125, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-0216/.

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

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-kURL https://arxiv.org/abs/2404.02936.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

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. 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, 2023. Curran Associates Inc.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.131. URL https://aclanthology.org/2022.emnlp-main.131/.

Wanzheng Zhu and Suma Bhat. GRUEN for evaluating linguistic quality of generated text. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 94–108, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.9. URL https://aclanthology.org/2020.findings-emnlp.9/.

APPENDIX INDEX

- **Appendix A: Limitations**
- **Appendix B: Extended Related Work**
- **Appendix C: Books Data v/s Common Pretraining Sources**
- **Appendix C: Books Data v/s Common Pretraining Sources**
  - C.1 Extended: What Distinguishes Book Data From Other Pretraining Data Sources?
  - C.2 More Details on Book Data Classifier
    - C.2.1 Feature Descriptions for Book Data Classifier
    - C.2.2 Feature Coefficients
- **Appendix D: Pretraining Dataset Statistics**
  - D.1 Dataset Statistics
  - D.2 Curating Fiction Books Dataset for Pretraining
- **Appendix E: Training Details**
  - E.1 Pre-training Details
  - E.2 Post-training Details
- **Appendix F: Extended Evaluations Results**
  - F.1 Training on Additional Book Tokens
  - F.2 Impact of Books on Coding Capabilities
  - F.3 Training using Synthetic Fiction Data
- **Appendix G: More Details on Synthetic Data Construction**
  - G.1 Prompt Templates
  - G.2 Controlling for Verbatim Regurgitation
- **Appendix H: Additional Evaluation Details**
  - H.1 Base Model Evaluations
  - H.2 Instruct Model Evaluations
  - H.3 Perplexity Evaluations
- **Appendix I: Human Eval**

A 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.

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., 2024; 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 & Durme, 2023; Elazar et al., 2024; 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; Chang et al., 2023; Duarte et al., 2024; Nasr et al., 2025; Ravichander et al., 2025; Hallinan 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 & 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 & 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 Books Data v/s Common Pretraining Sources

C.1 Extended: What Distinguishes Book Data From Other Pretraining Data 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?

To study this, we compare book data against six common pretraining domains: general web text, Reddit, 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., 2024) 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 C.2 Section C.2.1.

We train a logistic regression classifier with L1-regularization to identify the most discriminative features and 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 INSTITUTIONAL BOOKS, 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.

Semantic Features We first train the classifier using only semantic feature categories, excluding stylistic and readability features that capture surface-level writing patterns. 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 INSTITUTIONAL BOOKS 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 6, 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 INSTITUTIONAL BOOKS. 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 INSTITUTIONAL BOOKS, as reflected in the lexical diversity category, suggesting a richer and more varied vocabulary than public domain books.

INSTITUTIONAL BOOKS, on the other hand, shows a higher positive coefficient for figurative language than BOOKS3. As shown in Figure 8, archaic word density is one of the most dominant features for INSTITUTIONAL BOOKS 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 INSTITUTIONAL BOOKS 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 INSTITUTIONAL BOOKS, with 0.97 precision and 0.98 recall.

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 (Figure 5).

We find readability to be a representative feature of BOOKS3 text (Figure 5). Specifically, Flesch-Kincaid grade level is its top individual feature (Figure 9), 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 INSTITUTIONAL BOOKS 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 INSTITUTIONAL BOOKS scores higher on the stylistic category with a positive coefficient, while stylistic features carry a negative coefficient for BOOKS3. INSTITUTIONAL 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 INSTITUTIONAL BOOKS 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 EQ-Bench. 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.

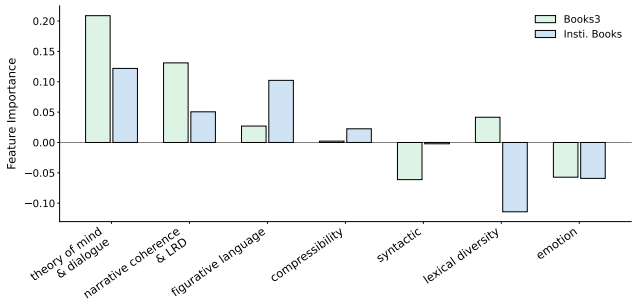


Figure 4: Feature coefficients from training classifiers for BOOKS3 and the INSTITUTIONAL BOOKS corpus. BOOKS3 exhibits stronger theory-of-mind and dialogue features, greater narrative coherence and long-range dependencies, and higher lexical diversity, all of which likely contribute to improved downstream utility.

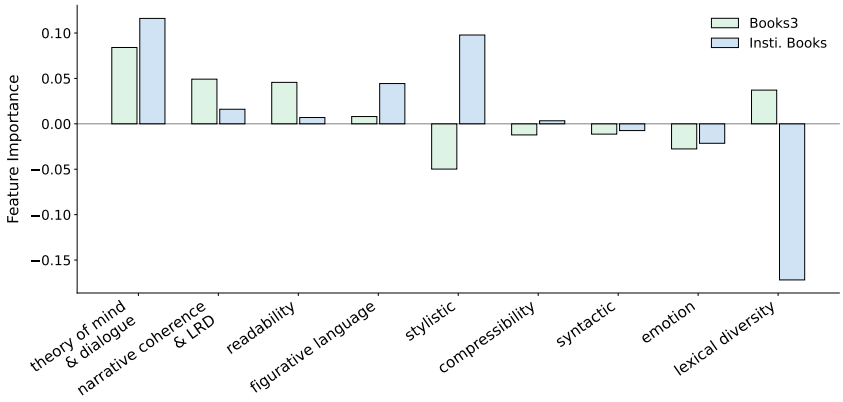


Figure 5: Feature coefficients from training classifiers for Books3 and the Harvard corpus using semantic, stylistic, and readability feature categories.

C.2 MORE DETAILS ON BOOK DATA CLASSIFIER

This section reports details about our classifier training setup. Section C.2.1 presents the description of all the features used in training, and Section C.2.2 showcases feature importance coefficients while training the classifier with a mix of semantic, stylistic, and readability features.

C.2.1 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 (Malkin et al., 2022).
Avg. Coreference Chain Length	Mean number of mentions per entity (Givón, 1983; Otmazgin et al., 2022).
Max Coreference Chain Length	Length of the longest coreference chain, usually corresponding to the protagonist (Givón, 1983; Otmazgin et al., 2022).
Max Entity Re-mention Distance	Furthest token gap between a noun and its pronoun (Givón, 1983; Otmazgin et al., 2022).
Named Entity (PER) Density	Count of person entities normalized by total words (Honnibal et al., 2020).
Unique Character Count	Total number of distinct named actors in the text (Honnibal et al., 2020).
Temporal Marker Density	Frequency of words indicating passage of time (e.g., *suddenly, later, meanwhile*) (Biber, 1988; Honnibal et al., 2020).
Spatial Preposition Density	Frequency of words anchoring spatial relationships (e.g., *beneath, across, through*) (Biber, 1988; Honnibal et al., 2020).
Past Tense Verb Ratio	Proportion of verbs referring to past actions out of all verbs (Biber, 1988; Honnibal et al., 2020).
Theory of Mind & Dialogue	
Dialogue Percentage	Ratio of text spoken by characters versus narration (Elson et al., 2010).
Theory of Mind Markers	Frequency of constructions where one character models another's mental state (e.g., *he thought the stranger was afraid*) (Premack & Woodruff, 1978; Ruffman et al., 2002).
Direct/Indirect Speech Ratio	Ratio of explicit quoted speech to reported speech (Cohn, 1978; Semino & Short, 2004).
1st-Person Pronoun Ratio	Density of first-person pronouns (*I, me, my, we, us*) (Biber, 1988).
2nd-Person Pronoun Ratio	Density of second-person pronouns (*you, your, yourself*) (Biber, 1988).
3rd-Person Pronoun Ratio	Density of third-person pronouns (*he, she, they, them*) (Biber, 1988).
Impersonal Pronoun Ratio	Density of abstract pronouns (*it, one, this, that*) (Biber, 1988).
Pronoun-to-Noun Ratio	Ratio of pronouns to nouns, measuring reliance on context versus explicit naming (Biber, 1988).
Cognitive Verb Density	Frequency of thinking and feeling verbs (e.g., *thought, realized, wondered, felt*) (Biber, 1988).
Communication Verb Density	Frequency of speaking verbs (e.g., *said, shouted, whispered, muttered*) (Biber, 1988).
Animacy Ratio	Proportion of actions performed by living versus non-living subjects (Zaenen et al., 2004).
Emotion	
Emotional Arc Shape	Measures the emotional trajectory by computing the amplitude and direction of sentiment across 10 equal text chunks using VADER (Hutto & Gilbert, 2014; Reagan et al., 2016).
Sentiment Volatility	Standard deviation of VADER sentiment scores across 5-sentence rolling windows (Hutto & Gilbert, 2014).

Table 4

Feature	Description
Avg. Sentiment Subjectivity	Mean subjectivity score across the document using TextBlob[2].
Positive Emotion Lexicon Density	Frequency of joy and trust words from the NRC Emotion Lexicon (Mohammad & Turney, 2013).
Negative Emotion Lexicon Density	Frequency of fear, anger, and sadness words from the NRC Emotion Lexicon (Mohammad & Turney, 2013).
Surprise Lexicon Density	Frequency of surprise words from the NRC Emotion Lexicon (Mohammad & Turney, 2013).
Motion Verb Density	Frequency of physical movement verbs (e.g., *ran, walked, jumped*) (Levin, 1993).
Stative/Active Verb Ratio	Ratio of stative verbs (*be, seem, appear*) to action verbs (Biber, 1988; Levin, 1993).
Present Tense Verb Ratio	Proportion of present tense verbs out of all verbs (Biber, 1988; Honnibal et al., 2020).
Lexical Diversity	
MTLD	Length-agnostic measure of vocabulary diversity (McCarthy & Jarvis, 2010).
Hapax Legomena Ratio	Proportion of words appearing exactly once in the document (Tweedie & Baayen, 1998).
Lexical Density	Ratio of content words (nouns, verbs, adjectives, adverbs) to total words (Biber, 1988).
Adjective Density	Frequency of adjectives per total words (Biber, 1988).
Adverb Density	Frequency of adverbs per total words (Biber, 1988).
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 (Biber, 1988).
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* (Niculae & Danescu-Niculescu-Mizil, 2014).
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 (Turney et al., 2011; Brysbaert et al., 2014).
Syntactic Structure	
Mean Dependency Tree Depth	Average syntactic complexity measured by dependency tree depth per sentence (Honnibal et al., 2020).
Max Dependency Tree Depth	Depth of the most heavily nested sentence in the document (Honnibal et al., 2020).
Subordinate Clause Density	Frequency of subordinating conjunctions (e.g., *because, although*) (Biber, 1988).
Coordinate Conjunction Density	Frequency of coordinating conjunctions (e.g., *and, but, or*) (Biber, 1988).
Inverted Syntax Frequency	Frequency of sentences where the object precedes the subject, common in poetic writing.

[2]https://textblob.readthedocs.io/en/dev/

Table 4

Feature	Description
POS Bigram Entropy	Shannon entropy of adjacent part-of-speech tag pairs, measuring grammatical unpredictability (Shannon, 1948; Honnibal et al., 2020).
Stylistic Structure: features motivated from Stamatatos (2009)	
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.
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 (Shannon, 1948).
zlib Compression Ratio	Compression-based measure of text redundancy (Deutsch & Gailly, 1996).
Readability	
Flesch-Kincaid Grade Level	Estimates the years of education required to understand the text based on syllables per word and words per sentence (Kincaid et al., 1975).
Dale-Chall Readability Score	Evaluates vocabulary complexity by measuring the proportion of words outside a baseline dictionary of 3,000 common English words (Dale & Chall, 1948).

Table 5: Complete list of features used to distinguish book data from other pretraining sources, organized by category.

C.2.2 Feature Coefficients

Figure 6 and Figure 8 show the feature coefficients for the Books3 and Institutional Books classifiers trained on semantic features, and Figure 7 and Figure 9 show coefficients for the Books3 and Institutional Books classifiers trained on semantic, stylistic, and readability features.

D Pretraining Dataset Statistics

D.1 Dataset Statistics

Fig. 10 showcases the distribution of tokens across documents used during training. Table 6 additionally outlines some descriptive statistics about the data.

D.2 Curating Fiction Books Dataset for Pretraining

Harvard: To extract fiction books from the Harvard Institutional Book Corpus 1.0 (Cargnelutti et al., 2025), we first extract English-only books using the `language_gen=eng` filter. Then, we feed title, author, publication year, and subject catalogs to Gemma-3-4B-it (Team, 2025b) and ask it to predict whether the book is a work of fiction. We also combine books marked as fiction from the Institutional Books' official metadata [3]. Around 7,900 books (≈1B tokens) are obtained using the metadata and 31,173 books (≈4B tokens) are obtained using our classifier.

[3] https://huggingface.co/datasets/institutional/institutional-books-hl-metadata

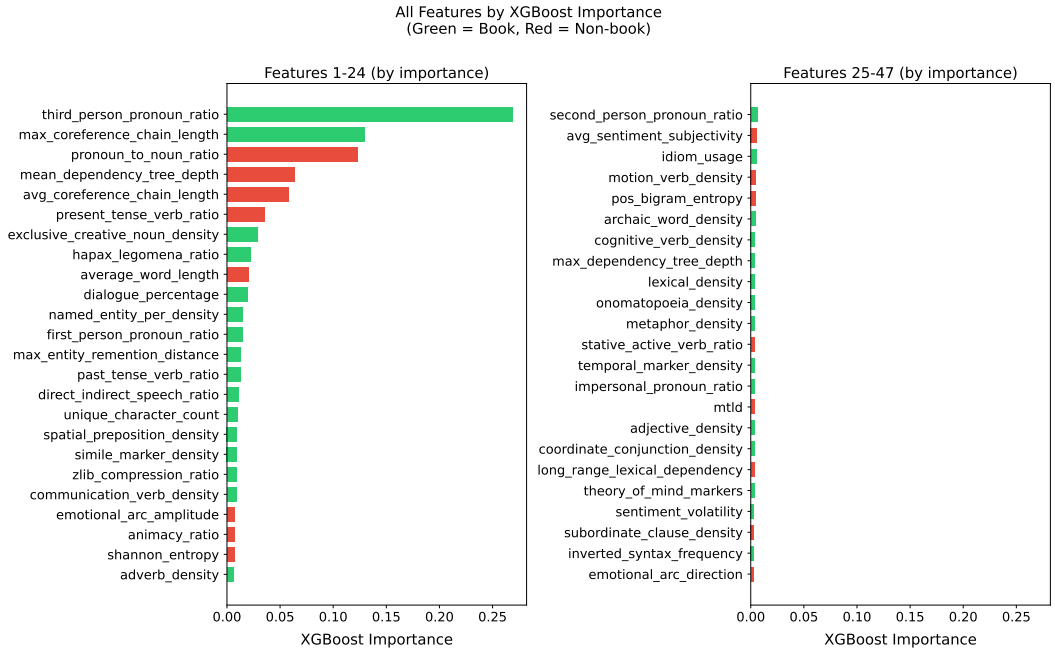


Figure 6: Feature coefficients for the Books3 classifier trained on semantic features.

Table 6: 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

Books3: 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.

Precision Analysis: We also conduct a precision analysis of our fiction classifiers. We draw a random sample of 100 classifier-labeled fiction books from each corpus and check each book's genre against authoritative web sources. We find the precision on BOOKS3 to be around 90% and on INSTITUTIONAL BOOKS to be 88% (including verse novel and drama (imaginative literature) as fiction).

D.3 ANALYZING THE AGE & COPYRIGHT STATUS OF BOOKS

Age: For BOOKS3, we draw a random sample of 2,000 books and look up each title and author in Open Library. We keep the 848 books for which a record is found *and* its author matches the book's author, and record the year of first publication. For books in INSTITUTIONAL BOOKS corpus, we use the year metadata provided by the authors.

Copyright Status for Books3: We use the definition provided by the U.S. Copyright Office[4]. Specifically, we treat books published after January 1, 1978 as copyrighted works. We find that at most 0.9% are clearly in the public domain, so roughly 99% of the sampled books are potentially under copyright.

[4]https://www.copyright.gov/help/faq/faq-duration.html

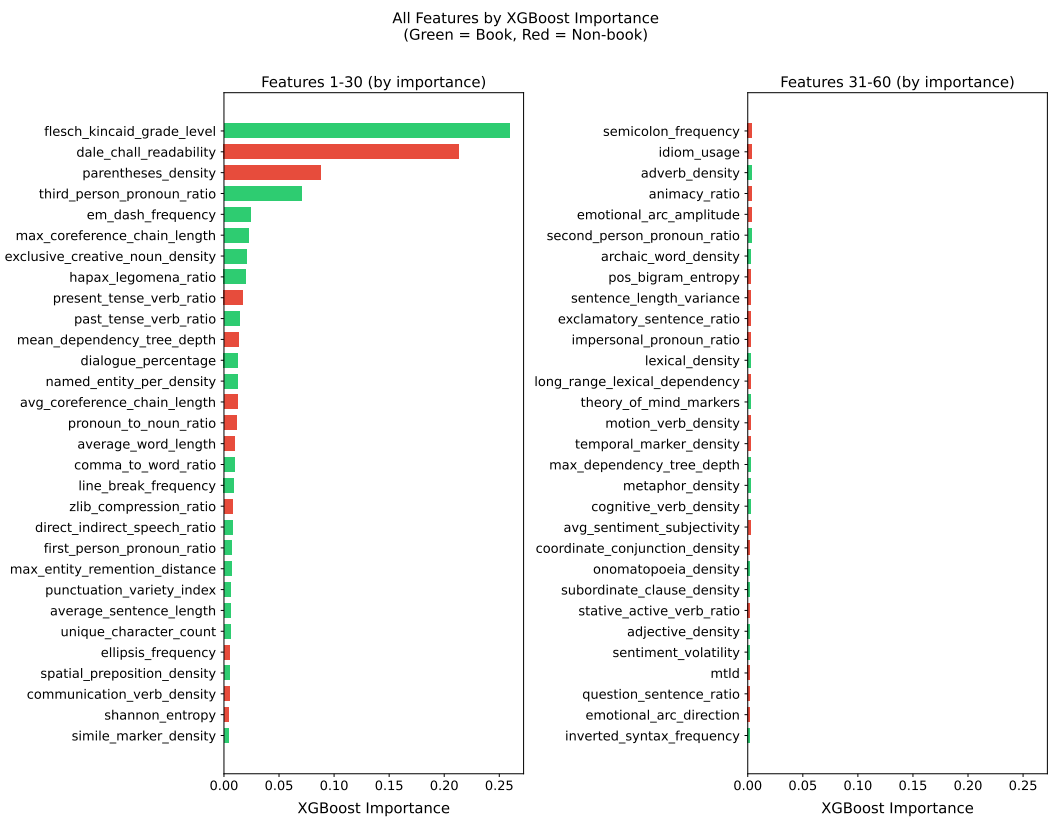


Figure 7: Feature coefficients for the Books3 classifier trained on semantic, stylistic, and readability features.

E Training Details

E.1 Pre-training Details

During pre-training, we use 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, and we train 14 1B models (spanning six different data configurations, with three seeds each for our four core models).

E.2 Post-training Details

For SFT we train on the Tülu-3 SFT mixture (Walsh et al., 2025) for 2 epochs with a learning rate of 2×10^{-5}, a linear schedule with 0.03 warmup ratio, maximum sequence length 4096, and an effective batch size of 128. We then apply DPO on the OLMo-2 1B preference mixture (Walsh et al., 2025) for 1 epoch with a learning rate of 1×10^{-6}, a linear schedule with 0.1 warmup ratio, the length-normalized DPO loss ($\beta = 5$), maximum sequence length 2048, and effective batch size 128. Both stages use the Tülu chat template with a prepended BOS token, bfloat16 mixed precision, and DeepSpeed ZeRO-3 sharding. All hyperparameters are held identical across the No Books, Books3, Institutional Books, and Synthetic models.

F Extended Evaluations Results

F.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 (2xBooks3), and (2) a mixed corpus consisting of 5B tokens each from Books3 and Harvard (Books3 +Institutional Books). For these runs we use the 175B token subset of the

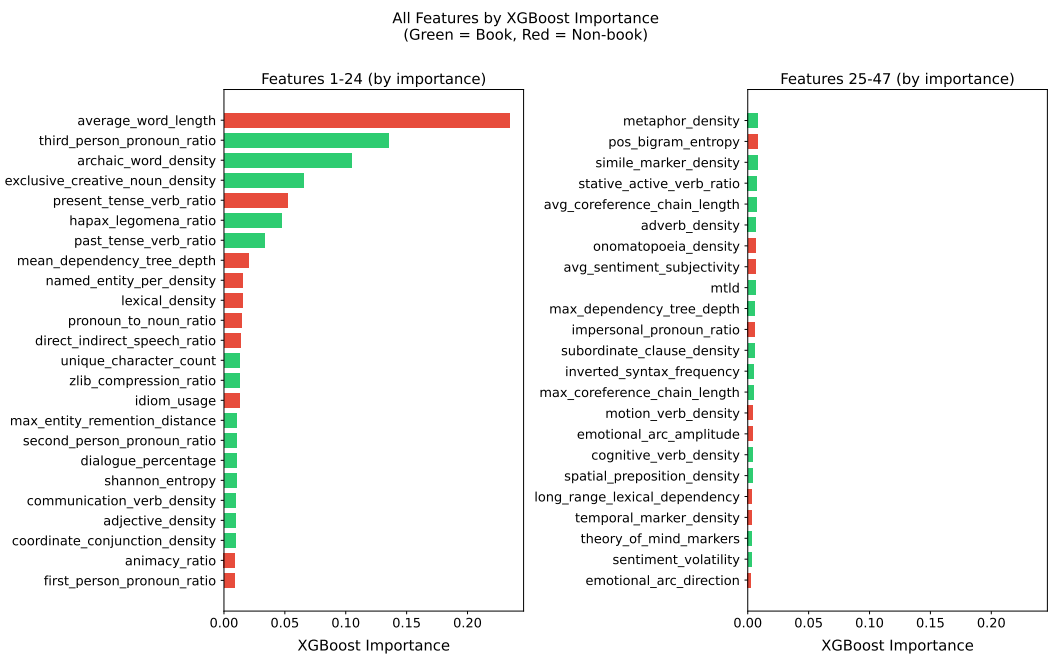


Figure 8: Feature coefficients for the Harvard classifier trained on semantic features.

WebOrganizer without creative writing corpus. Figure 11 and Table 7 present the results for base and instruct versions. We observe that 2xBOOKS3 model outperforms BOOKS3 +INSTITUTIONAL BOOKS model on all evals. We also find that the BOOKS3 model with a single pass on 5B tokens has a better performance against NO BOOKS baseline as compared to the model trained with two passes (10B tokens). This indicates that doing additional passes on the books token may hurt the model performance. However, we find that BOOKS3 +INSTITUTIONAL BOOKS performs better than INSTITUTIONAL BOOKS on win-rates against NO BOOKS baseline, indicating that mixing contemporary fiction into archaic public-domain fiction can help improve the performance.

F.2 IMPACT OF BOOKS ON CODING CAPABILITIES

As shown in Figure 12, we observe that the NO BOOKS model outperforms BOOKS3 on coding (Figure 12). 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 considering including books in the general-purpose pretraining mixture.

F.3 TRAINING USING SYNTHETIC FICTION DATA

Figure 13 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 INSTITUTIONAL BOOKS Book Corpus on the narrative coherence category and comes close to the model trained on BOOKS3 on LAMBADA.

G MORE DETAILS ON SYNTHETIC DATA CONSTRUCTION

G.1 PROMPT TEMPLATES

System Prompt

```
You are a novelist. When given a writing task, respond only
with the story text itself. Never begin with 'Here is',
'Sure', 'Certainly', or any other preamble or commentary.
Start immediately with the first sentence of the narrative
and end when the scene is complete.
```

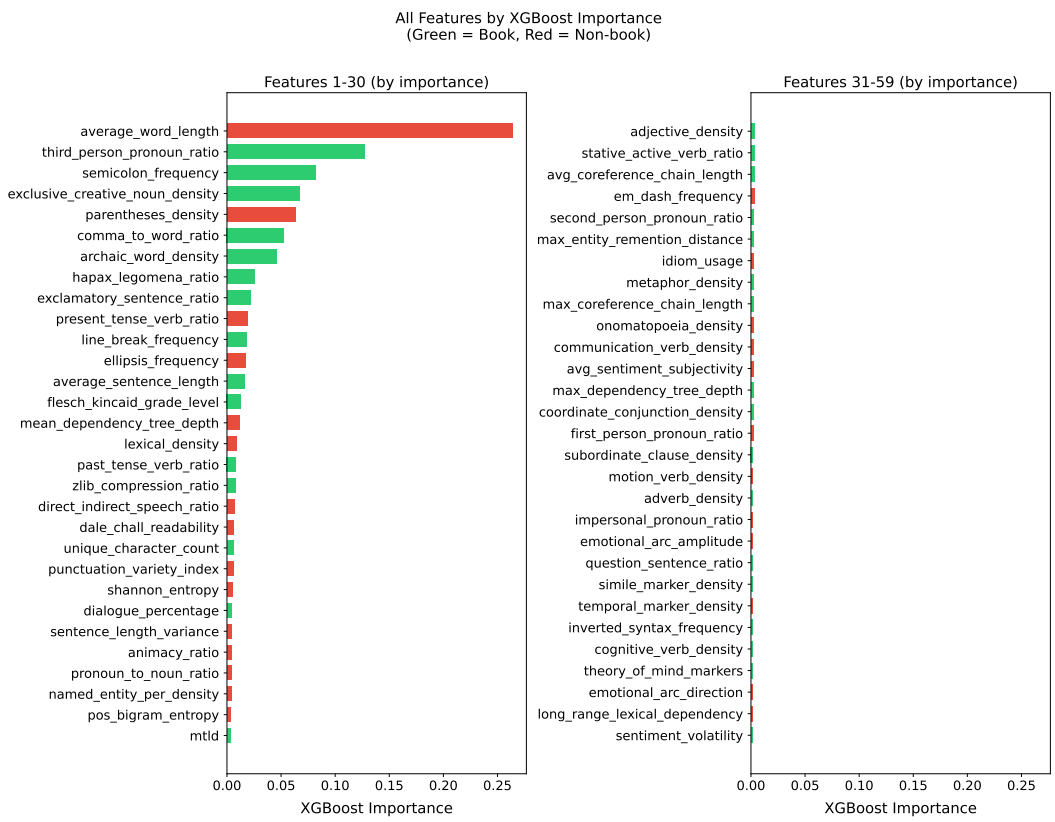
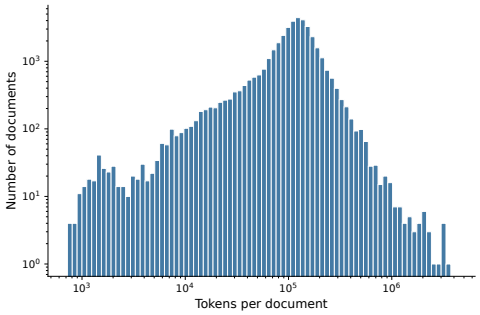


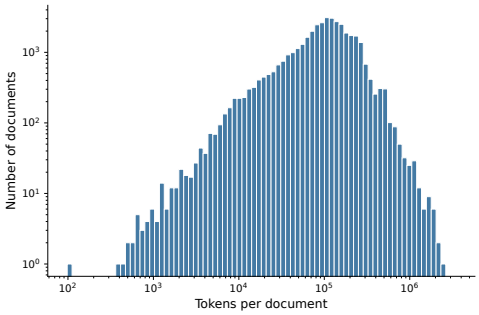
Figure 9: Feature coefficients for the Harvard classifier trained on semantic, stylistic, and readability features.

User Prompt Templates

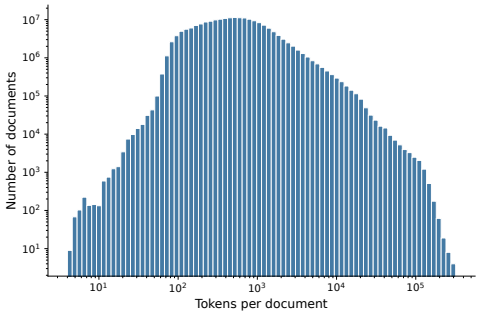
```
[Template 1]
Genre: {genre}

Overall plot: {summary}

Here is an excerpt from this part of the novel for style and
context:
{local_excerpt}

Write an original {min_words}-{max_words} word chapter that
fits this narrative
world and prose style. Do not copy from the excerpt.

[Template 2]
You are writing a chapter for a {genre} novel.

Story summary: {summary}

Excerpt from this section of the book:
{local_excerpt}

Write a new {min_words} to {max_words} word chapter that
matches the voice and
```



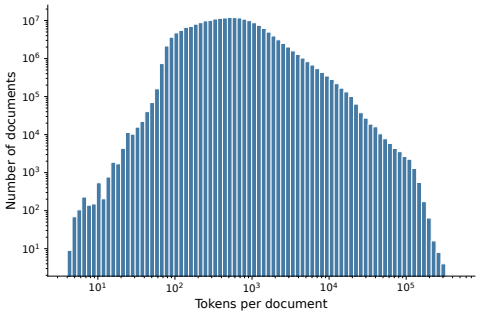
(a) Books3 (5B Tokens)



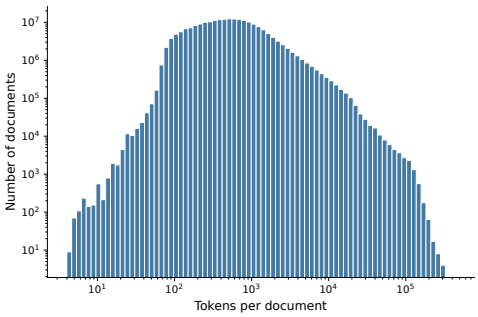
(b) Harvard (5B Tokens)



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



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



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

Figure 10: Histograms showcasing distribution of tokens across documents from different corpora used during training.

Table 7: Instruction-tuned evaluation of four fiction pretraining variants against the No Books baseline: BOOKS3 (contemporary, 5B tokens), INSTITUTIONAL BOOKS (public-domain, 5B tokens), 2×BOOKS3 (two passes on Books3 fiction, 10B tokens), and BOOKS3 +INSTITUTIONAL BOOKS (contemporary plus public-domain, 5B tokens each). Pairwise columns report win rates (%) versus NO BOOKS. LitBench uses the creative-writing reward model and UniEval is the mean of fluency and coherence. Prometheus 2 and GPTScore aggregate all six non-creative writing categories. All results are for a single seed.

Instruction following *Pairwise win rate (%)*	**Books3**	**Insti. Books**	**2×Books3**	**Books3+Insti. Books**
AlpacaEval	42.1	27.5	36.2	29.7
WildBench	42.3	20.5	31.7	24.0
MT-Bench	47.5	20.3	31.2	25.0

Creative writing *Pairwise win rate (%)*	**Books3**	**Insti. Books**	**2×Books3**	**Books3+Insti. Books**
EQ-Bench	**78.1**	18.8	**54.7**	48.4

Absolute score	**No Books**	**Books3**	**Insti. Books**	**2×Books3**	**Books3+Insti. Books**
LitBench	−3.23	−2.93	−3.34	−2.91	−3.36

Non-creative writing *Pairwise win rate (%)*	**Books3**	**Insti. Books**	**2×Books3**	**Books3+Insti. Books**
Prometheus 2 (overall)	47.0	36.9	37.7	33.3
Correspondence	44.0	40.5	38.5	33.0
Opinion / review	46.5	43.6	42.2	46.5
Argumentative	49.1	37.2	31.2	27.9
Advice / how-to	42.4	30.6	36.2	24.4
Blog / discussion / article	47.3	31.5	35.9	31.6
Rewriting / editing	**51.7**	36.6	42.0	36.6
GPTScore	38.6	34.5	36.7	35.2

Absolute score	**No Books**	**Books3**	**Insti. Books**	**2×Books3**	**Books3+Insti. Books**
GRUEN	0.389	0.352	0.285	0.332	0.333
UniEval	0.928	0.916	0.873	0.889	0.872

```
style of the excerpt. Write original content - do not
reproduce the excerpt.

[Template 3]
Genre: {genre}
Summary: {summary}

Context excerpt:
{local_excerpt}

Write an immersive {min_words}-{max_words} word chapter for
this novel. Use the
excerpt as a style guide only. Write entirely new content.
```

Here {genre} is a free-text genre string, {summary} a book-level summary built from three fixed excerpts (beginning/middle/end), {local_excerpt} a sliding-window excerpt at the chapter's position, and {min_words}–{max_words} the 1000–2000 word bucket.

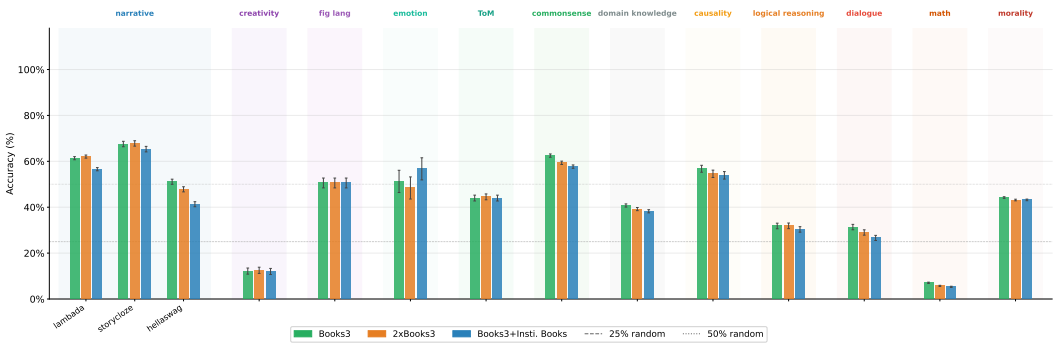


Figure 11: Base model evaluation across 12 skill categories comparing the three model variants: BOOKS3 (5B tokens), 2xBOOKS3 (10B tokens), and BOOKS3 +INSTITUTIONAL BOOKS (10B tokens). BOOKS3 and 2×BOOKS3 outperform BOOKS3 +INSTITUTIONAL BOOKS in narrative tasks (e.g., LAMBADA, HellaSwag), while performance 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.

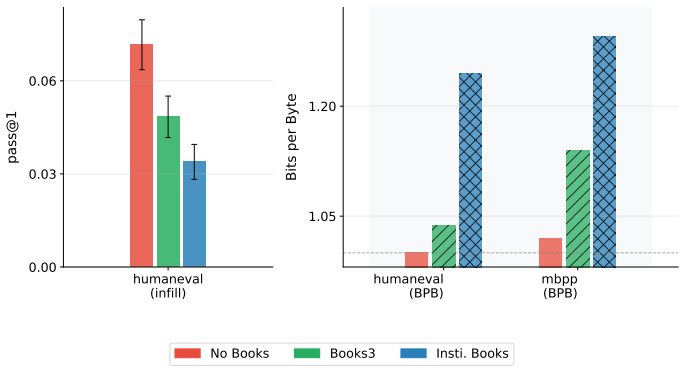


Figure 12: Coding evaluation on HumanEval (pass@1) and bits-per-byte (BPB) on HumanEval and MBPP. NO BOOKS achieves the best performance, followed by BOOKS3 and INSTITUTIONAL BOOKS. Error bars denote 95% CI for PPL metrics and standard error for coding.

G.2 CONTROLLING FOR VERBATIM REGURGITATION

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.
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.

H ADDITIONAL EVALUATION DETAILS

H.1 BASE MODEL EVALUATIONS

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),

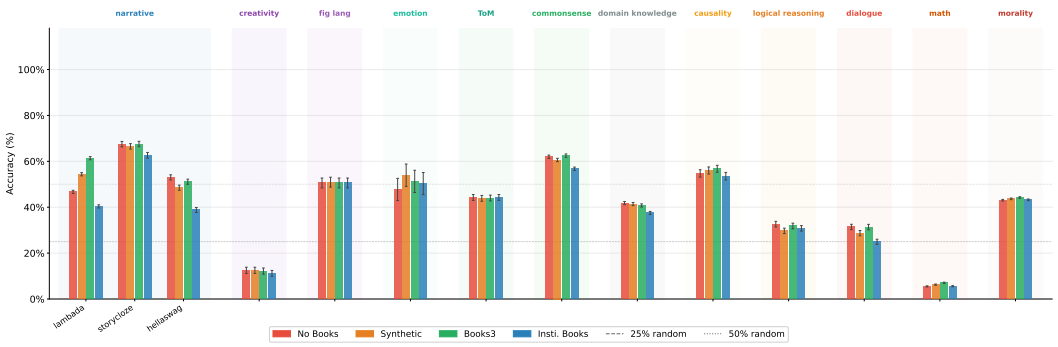


Figure 13: Base model evaluation across 12 skill categories comparing four model variants. The model trained on SYNTHETIC data outperforms the model trained on the INSTITUTIONAL BOOKS corpus on narrative coherence tasks, and comes close to the model trained on BOOKS3. Dashed and dotted lines indicate random baselines for 4-choice and 2-choice tasks, respectively. Error bars denote standard error.

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.

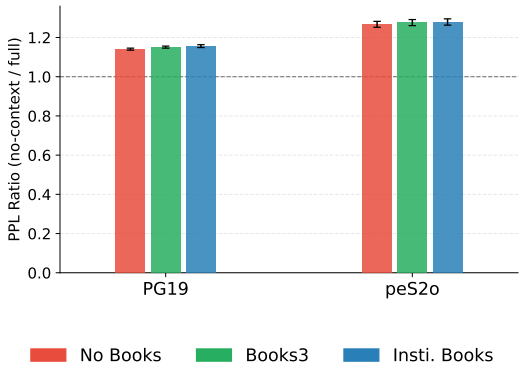


Figure 14: Context utilization ratio (PPL without context / PPL with full context) for all three models on PG19 and peS2o. A ratio above 1 indicates that the model benefits from longer context when predicting text. All three models show similar context utilization across both datasets.

H.2 INSTRUCT MODEL EVALUATIONS

Instruction Following : We use `temp=0.7` and `top_p=0.95` for decoding and allow the models to stop naturally until they hit `max_new_tokens=2048`. For all evals, the generations are blind and order randomized to avoid any position bias. For AlpacaEval 2.0, we compute length-controlled win rate using `top_logprobs=20`. We use single-turn examples and GPT 4.1 (OpenAI, 2025) as a judge across all instruction following evals.

Non-creative Writing : First, we create prompt sets spanning six categories from various sources, as shown in Table 8. Each category has 200 prompts. We report Prometheus 2 and length-controlled GPTScore (fluency & coherence) as pairwise win-rates, and GRUEN and UniEval (fluency & coherence) as absolute scores.

Category	Source
Correspondence	No Robots (Rajani et al., 2023), Dolly (Conover et al., 2023), WildChat (Zhao et al., 2024)
Opinion / review	LongLaMP (Kumar et al., 2024)
Argumentative	ChangeMyView (Tan et al., 2016)
Advice / how-to	WikiHow (Koupaee & Wang, 2018)
Blog / discussion / article	No Robots (Rajani et al., 2023), Dolly (Conover et al., 2023), WildChat (Zhao et al., 2024)
Rewriting / editing	CoEdIT (Raheja et al., 2023)

Table 8: Prompt sources for the six non-creative writing categories (200 prompts per category).

Table 9: Human evaluation and human-LLM agreement. Human WR and LLM WR are the human and LLM-judge win rates for the first model on the same items (ties count as 0.5). Agreement is the fraction of examples where the human and the LLM judge select the same winner. Each setting is judged by a single annotator.

Comparison	Eval (LLM judge)	n	Human WR (%)	LLM WR (%)	Agree (%)
BOOKS3 vs NO BOOKS	LitBench (reward model)	50	57	64	46
BOOKS3 vs INSTITUTIONAL BOOKS	LitBench (reward model)	97	73	69	46
	WildBench (GPT-4.1)	50	67	79	54
	Non-creative (Prometheus 2)	50	56	59	44

Creative Writing : We use Claude Sonnet 4.6 (Anthropic, 2026) as a judge for EQ-Bench v3.

H.3 PERPLEXITY EVALUATIONS

Perplexity Evaluation of Modern LLMs. We evaluate Qwen3-1.7B-Base (Yang et al., 2025), Llama-3.2-1B (Meta, 2024), and OLMo-2-0425-1B (Walsh et al., 2025), representing a diverse set of strong, similarly sized language models in the 1B-parameter range. We sample 1,000 sequences each from the BOOKS3 and INSTITUTIONAL BOOKS corpora and truncate each sequence to the first 1,024 tokens produced by the corresponding model's tokenizer to measure perplexity over. All models are evaluated using BF16 precision.

I HUMAN EVAL

To validate our automatic evaluations against a gold standard, we conduct a human study and measure how often human judgments agree with the LLM judges. For each example, we show a human annotator the prompt, two blind generations (one from each model), and a rubric, and ask them to select the better response, or a tie when the two are very close. For the same prompt, we record which model the LLM judge preferred. We then report, for each model pair, the human win rate for a model and the human-LLM agreement rate, defined as the fraction of examples on which the human and the LLM judge choose the same winner.

We run two studies. The first compares BOOKS3 against NO BOOKS on 50 LitBench prompts. The second compares BOOKS3 against INSTITUTIONAL BOOKS on LitBench, WildBench, and the non-creative writing categories (excluding math, coding, and data analysis). Each example is judged by a single annotator, and ties are counted as half a win for each model.

Table 9 shows the results. Humans prefer BOOKS3 over NO BOOKS on LitBench (57% win rate), and prefer BOOKS3 over INSTITUTIONAL BOOKS by a larger margin (73%). This matches the direction of our automatic metrics. The human and the LLM judge agree on 46% to 54% of individual examples, highest for the WildBench judge and lowest for the LitBench reward model. Put simply, the judges capture the overall preference between models well, but often disagree with humans on which single response is better.