BabyLM-Tiny: Comparing Text Subjects for Low-Resource Pretraining

Nikita Goryetiy and Moritz Ladenburger and Lukas Edman
Technical University of Munich

Abstract

Language model pretraining traditionally relies on massive datasets and computational resources, raising questions about data efficiency and developmental plausibility. This paper investigates pretraining under the BabyLM-Tiny setting, which restricts training to only 1M words, focusing on how textual genre, dataset size, and linguistic complexity shape learning. We introduce a systematic comparison across child-focused speech, encyclopedic text, literary genres, and dialogue, using a DeBERTa-v3-small model trained from scratch. In addition, we highlight the trade-offs between data continuity and diversity and outline broader implications for efficient and sustainable language modeling. Our code is publicly available on GitHub[1].

1 Introduction

1.1 Background and Motivation

Recent advances in language modeling have been primarily driven by increasing the number of parameters and training size at the cost of immense computing resources. However, this paradigm poses significant challenges for researchers with limited budgets for computation and raises fundamental questions regarding the data efficiency of natural language processing (NLP). To address these concerns, the BabyLM challenge was introduced, providing a benchmark for the development of language models trained on reduced amounts of data. The challenge restricts training datasets to 100 million words, the approximate equivalent of the linguistic input a human encounters by the time they reach the age of thirteen. The BabyLM-small track limits the dataset further to 10 million words. In this work, we extend this approach even further by proposing the BabyLM-Tiny setting, which reduces the training datasets to only 1 million words.

[1] https://github.com/moritzlad/BabyLM-Tiny

Data efficiency in natural language processing (NLP) is a critical concern that extends beyond computational limitations. As the AI research community increasingly emphasizes resource-efficient practices, it becomes more important to understand how competitive performance can be achieved while using less data. At the same time, research into data efficiency may provide deeper insights into the mechanisms of language acquisition and processing at a fundamental level. This could potentially help bridge the gap between human and artificial language learning.

Despite growing interest in efficient training methods, most prior work has concentrated on architectural improvements. By contrast, comparatively little attention has been given to the role of optimal text type selection for small-scale language model training. In particular, the question of whether certain genres or text types provide more effective learning signals remains largely unexplored in the context of severely limited training data.

The original BabyLM dataset consists primarily of child-directed text, including parent-child conversations, children's books, and movies. We examine the effects of training models on both these sub-datasets and newly created datasets sampled from books across multiple genres, Wikipedia articles, and movie subtitles. Beyond examining data diversity, we investigate whether language models mirror human learning patterns by benefiting from exposure to simple texts first, or whether training on complex datasets from the outset yields equally effective or superior performance.

1.2 Research Questions

This paper addresses three primary research questions. First, we examine how the quantity of training data affects model performance when working with severely constrained datasets. Second, we investigate which specific genres and domains are

most effective for limited data training scenarios. Finally, we explore whether combining different text types provides synergistic benefits that exceed the performance of individual text types alone.

2 Related Work

Data efficiency in language model pretraining has gained increasing attention in recent years. The BabyLM challenge has brought forth models that outperform other models trained on datasets with trillions of words. Approaches that were successful in past BabyLM competitions used different methods, including:

- **Curriculum Learning**: Starting with easy samples and gradually proceeding to harder ones, allowing the model to build on previous knowledge and develop a more robust understanding of patterns.
- **Knowledge distillation**: Creating a teacher model to help transfer knowledge to a smaller, simpler student model.
- **Data Preprocessing**: Using methods such as short sequences or individual sentences as training samples instead of the raw dataset as a whole.

(BabyLM Challenge organizers, 2025)

There also exist a number of artificially created datasets, such as TinyStories (Eldan and Li, 2023), a dataset generated using OpenAI's GPT-3.5 and GPT-4 models that consists of words a four-year-old would understand. This aligns with our question of whether simple genres (e.g., child-directed text) provide a better learning signal for low-resource language models.

Another paper related to the BabyLM-challenge (Haga et al., 2025) explores the Effect of Variation Sets on Language Model Training Efficiency. They use variation sets, described as "sets of consecutive utterances expressing a similar intent with slightly different words and structures," to improve model performance.

3 Datasets

We create a variety of datasets from different sources. The original BabyLM Dataset consists mostly of child-focused text. The Wiki datasets allow us to research the effects of content on model performance. Through the Gutenberg datasets, we are able to investigate the effects of genre, and through the Open Subtitle Datasets, the effects of different kinds of dialogue.

3.1 BabyLM Dataset

The BabyLM Dataset consists of six linguistically diverse sub-datasets: BNC (British English dialogue), CHILDES (child-directed speech transcripts), Project Gutenberg (public domain books), OpenSubtitles (movie dialogue), Simple English Wikipedia (encyclopedia articles), and Switchboard (telephone conversations). We develop a sampling pipeline that extracts 10,000-word chunks uniformly at random from each sub-dataset until reaching the target dataset size. For select sub-datasets (Simple Wiki, Open Subtitles, and Gutenberg), we create datasets with 100k, 300k, 600k, 1 million, and 1.2 million words to study the effects of dataset size on model performance.

3.2 Wiki Datasets

To research the effect of domain-specific content on model performance, we sample topic-focused datasets from Wikipedia articles across five domains: Society, Quantum Mechanics, Linguistics, History, and Culture. We recursively traverse Wikipedia's category tree and sample whole articles until reaching the desired dataset size of 1 million words. These five topics were selected because they span both technical and general knowledge with varying linguistic complexity. We recursively traversed Wikipedia's category hierarchies, sampling complete articles until reaching 1 million words per topic. This approach ensures coherent, topic-specific content while maintaining the natural article structure.

3.3 Gutenberg Datasets

To further investigate the effects of genre selection, we also create genre-specific datasets from Project Gutenberg's public domain collection of books. (Gutenberg, n.d.) We curate lists of books in the six popular genres (Sci-Fi/Fantasy, Romance, (Shakespearean) Drama, Mystery, Non-Fiction, Youth/Young Adult), capped at 1 million words each.

Additionally, we investigate the impact of text continuity by sub-sampling the Gutenberg component of BabyLM with varying book completeness levels. We create four 1M-word datasets containing 25%, 50%, 75%, and 100% complete books, respectively, to assess whether longer continuous passages improve model performance compared to fragmented text chunks.

Percentage	# Books	Avg. Words per Book
25%	91 books	10,989 words
50%	42 books	23,810 words
75%	30 books	33,333 words
100%	23 books	43,478 words

3.4 Open Subtitles Datasets

To examine the effect of conversational and cinematic dialogue on model performance, we construct a domain-specific dataset using subtitle transcripts from the Sublens-20M dataset, which is derived from the OpenSubtitles resource (Eden et al., 2022; Eden, 2022). First, we query The Movie Database (TMDb) API[2] to retrieve curated lists of films across six popular genres: Action, Comedy, Documentary, History, Romance, and Science Fiction. For each genre, we sampled approximately 100 films, ensuring coverage across different decades and popularity levels.

Using IMDb identifiers, we aligned these titles with the Sublens-20M subtitle collection, extracting the corresponding subtitle transcripts. From these, we generate genre-specific datasets of 1.0M words each. Non-English files and duplicates were filtered, and multiple encodings were tested to ensure robust text extraction. The resulting datasets capture naturalistic dialogue and narrative structure, complementing the more formal styles of Wikipedia and Gutenberg.

4 Methodology

4.1 Experimental Setup

As our model, we choose DeBERTa v3 (He et al., 2021), a state-of-the-art Masked Language Model (MLM) developed by Microsoft. BERT-style models demonstrated strong performance and architectural advantages in last year's BabyLM challenge. (Hu et al., 2024) We train the model (He et al., 2021) using the Hugging Face Transformers framework. The model architecture and tokenizer were based on the `microsoft/deberta-v3-small` configuration, with model weights initialized from scratch. Training was conducted for 10 epochs on sequences of up to 64 tokens, optimized with AdamW using a cosine learning rate schedule, mixed-precision training, and an effective batch size of 64. Evaluations and checkpoints were recorded at regular intervals to monitor performance throughout training.

[2] https://www.themoviedb.org/settings/api

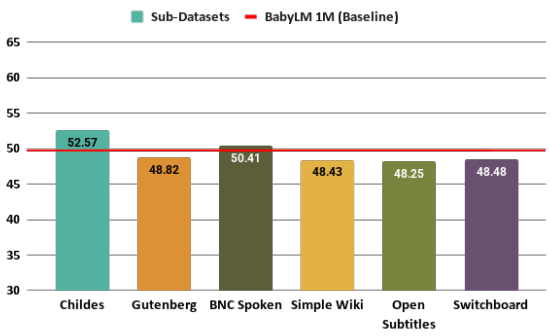

Figure 1: Model performance across BabyLM sub-datasets at 1.0M words compared to the BabyLM baseline. Only Childes and BNC Spoken maintain a small improvement.

4.2 Evaluation Framework

We use the provided BabyLM evaluation pipeline, especially BLiMP (Benchmark of Linguistic Minimal Pairs), to evaluate model performance.
BLiMP consists of 67 individual datasets, each containing 1,000 minimal pairs of sentences. A minimal pair comprises two sentences that are minimally different but contrast in grammatical acceptability: One sentence is grammatical while the other is ungrammatical, isolating specific phenomena in syntax, morphology, or semantics. The benchmark is organized into 12 broad linguistic phenomena categories, including:

- Quantifiers (Distribution restrictions for quantifiers)
- Irregular Forms (Irregular morphology on English past participles)
- Filler Gap Dependency (Dependencies arising from phrasal movement in wh-questions)

We use the *zero-shot-fast* pipeline, which allows for fast evaluation and uses a smaller set of evaluation samples.

5 Results and Discussion

5.1 Comparative Performance of BabyLM Subsets

As an initial step, we compared the performance of individual BabyLM sub-datasets trained at 1.0M words against the BabyLM 1.0M mixed baseline (Figure 1). Interestingly, some individual text types showed consistently higher performance than the mixed BabyLM corpus of the same size. This suggests that genre-specific specialization can in certain cases yield more effective linguistic generalization than heterogeneous mixtures.

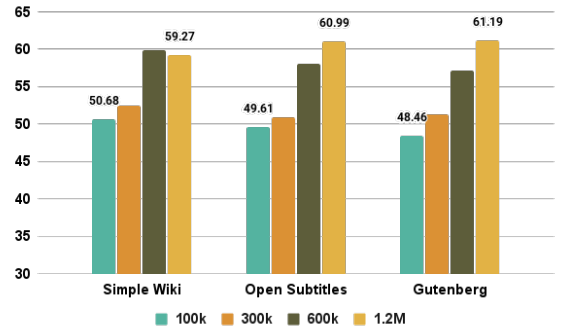

Figure 2: Model performance across Simple Wiki, Open Subtitles, and Gutenberg datasets at different corpus sizes (100k–1.2M). Performance improves with size but saturates around 600k–1.0M words.

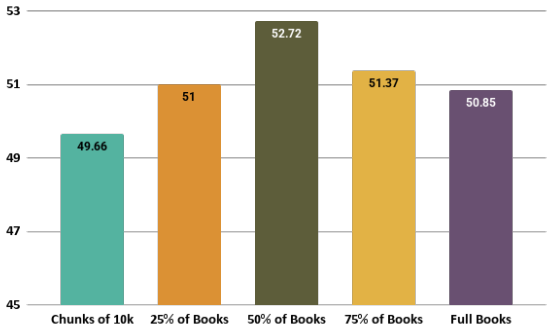

Figure 3: Model performance across different context lengths on the Gutenberg dataset. Performance peaks at 50% of books before declining due to reduced textual diversity.

From this comparison, we selected Gutenberg, Simple Wiki, and OpenSubtitles for further analysis, as they represent text types that people are regularly exposed to in everyday life. We did not extend experiments with BNC Spoken, since identifying genre-specific variation within conversational dialogue is less straightforward compared to books, encyclopedic texts, or film scripts.

As expected, increasing dataset size generally improves model quality (Figure 2). However, the effect is not strictly linear across all text types. Different genres improve at different rates: some display a smoother upward trend, while others exhibit sharp jumps in performance. Notably, for Simple Wiki, the best result is achieved already at 600k words, slightly outperforming the 1.2M version. This indicates diminishing returns and suggests that both genre and data efficiency, rather than sheer volume, are key factors in small-scale pretraining.

5.2 Context Length and Genre Effects in Gutenberg Texts

In the previous section, we downsampled the BabyLM Gutenberg sub-dataset by repeatedly extracting randomly sampled chunks of 10,000 words from individual books. This approach prompted us to investigate how different context lengths affect model performance when the total dataset size remains fixed.

Model performance initially improves as we increase chunk size, reaching a peak when using 50% of each book, before declining with larger portions (Figure 3). This pattern reflects two competing factors that influence learning outcomes. On the one hand, using larger portions of individual books enables the model to better capture contextual dependencies. On the other hand, since we maintain a constant dataset size of 1 million words, larger chunks necessarily reduce the number of different books included, thereby limiting the linguistic and thematic diversity available for training.

This trade-off between contextual depth and textual diversity suggests that optimal performance requires balancing these competing demands rather than simply maximizing context length.

We also compare the model performance when training on whole books from different genres (Figure 4). We can see that all models fall in a similar range in the overall scores, but show very different strengths and weaknesses in the individual sub-categories. Overall, Sci-Fi/Fantasy scores best with an overall score of 54.2% and Mystery scores worst at 49.0%. Some notable outliers in categories include Sci-Fi/Fantasy, with outstanding performance in the category Filler-Gap Dependency. We attribute this to the fact that Sci-fi and Fantasy literature often features intricate world-building and technical explanations, which lead to more complex sentence structures compared to genres like romance, which prioritize emotional narratives and dialogue. Early modern English Literature outperforms the other genres in quantifiers. We performed a frequency analysis with a list of 104 common quantifiers. Interestingly, early modern English literature has the lowest total count of quantifiers in the dataset, but features a richer variety and more evenly distributed usage of quantifiers compared to the other datasets. This brings us to the conclusion that again, diversity of data is an important factor in dataset creation.

Early modern English literature also outperforms

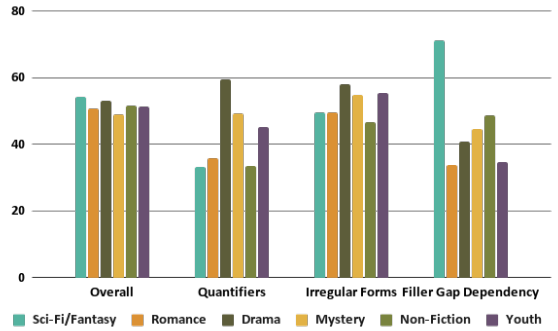

Figure 4: The figure shows model performance across literary genres, with Sci-Fi/Fantasy achieving the highest overall score (54.2%) while demonstrating distinct strengths in different linguistic categories by genre.

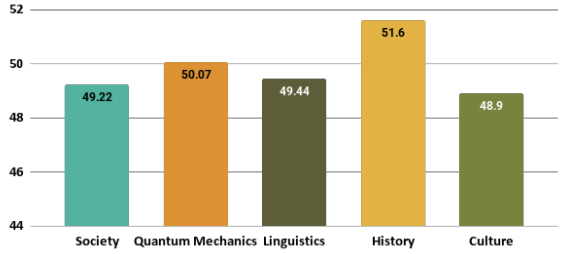

Figure 5: Wikipedia articles about History lead to the highest overall scores, followed by, surprisingly, quantum mechanics.

the other genres in irregular forms. The richer variety of verb forms is responsible for this effect.

5.3 Topic and Complexity Effects in Wikipedia Articles

We chose Wikipedia articles to look at the impact of text complexity and content, as these articles are generally written in a similar style. As seen in Figure 5, our five chosen topics (Society, Quantum Mechanics, Linguistics, History, and Culture) all performed similarly in the overall score. The model trained on quantum mechanics articles shows surprisingly strong performance, especially since the text includes a lot of formulas and technical descriptions. In Figure 6, we can see that the quantum mechanics dataset outperforms in the categories *Filler Gap Dependency* and *NPI licensing*, while falling short in *bindings* compared to the other categories.

We think the strong performance in filler gap dependency is due to the many examples of long sentences with relative clauses like "An orbital is completely different from a total stationary state, **which** is a many-particle state requiring . . . ".

A similar argument can be made for NPI licensing. Technical descriptions frequently employ environments that license Negative Polarity Items (NPIs). For example, in physics texts we encounter sentences such as "...there is **no impact** on the measurement" or "...no physical theory of hidden local variables can **ever** reproduce all the predictions."

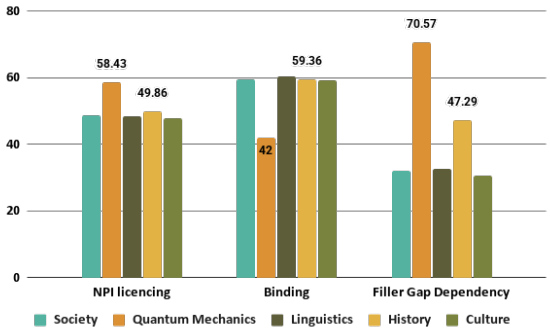

Figure 6: Performance across different Wikipedia training domains shows that quantum mechanics excels at NPI licensing and Filler Gap Dependencies but lags in bindings.

Relative clauses in quantum mechanics texts are often long and syntactically complex, which makes it harder for models to learn binding phenomena reliably. Even when bindings are present, the dependencies are buried inside dense technical definitions rather than simple sentence structures. For example, compare "Stationary **states** are quantum **states that are solutions** to the time-independent Schrödinger equation" with the simpler history sentence "Muhammad ibn al-Alqami . . . was a vizier to the last Abbasid **caliph**, **who** ruled until 1258."

5.4 Dialogue Genre Effects in OpenSubtitles

We compare model performance across different movie genres (Figure 7). The results show that all genres achieve overall scores within a relatively narrow range, but exhibit clear differences in their linguistic strengths and weaknesses. Comedy achieves the highest overall accuracy at 52.5%, followed by Romance (51.0%) and Sci-Fi (50.9%), while Documentary yields the lowest score at 48.9%.

Some notable outliers include Comedy and Romance, which perform particularly well across multiple linguistic categories, likely due to their conversational and emotionally rich dialogue. Such texts contain frequent pronoun use, turn-taking structures, and varied informal constructions, which may provide stronger learning signals for syntax,

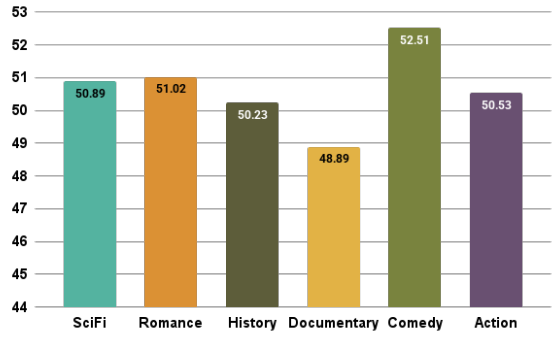

Figure 7: Model performance across subtitle genres. Comedy achieves the highest score, while Documentary performs worst.

semantics, and morphology. By contrast, Sci-Fi and Action achieve competitive overall scores but show distinct strengths in morphology and semantics, while underperforming in syntax-related phenomena. This pattern can be attributed to the prevalence of technical terminology, action-oriented commands, and descriptive passages, which enhance morphological variety but provide fewer examples of naturalistic conversational syntax. Interestingly, History and Romance show relatively balanced accuracy across most linguistic categories, suggesting that more human-centered and narrative-driven genres support broad generalization. Overall, these findings reinforce the hypothesis that exposure to informal, dialogue-rich text enables models to generalize better than training on structured or informational language alone.

6 Conclusion

We evaluated model performance across multiple text types under the BabyLM-Tiny setting. The results show that overall scores remain in a similar range, yet each dataset type exhibits distinct linguistic strengths and weaknesses. Among the BabyLM sub-datasets, books and subtitles perform best, confirming that the type of training text is as important as dataset size.

Some notable outliers include book datasets sampled at 50% completeness, which achieve optimal performance by balancing contextual depth with diversity. Although overall BLiMP scores are close across genres, Sci-Fi and Romance each show specific advantages in syntax and semantics, highlighting the role of genre-specific structures.

Wikipedia articles demonstrate comparable overall performance across topics, with history achieving the highest score and quantum mechanics excelling in specialized categories such as filler–gap dependency and NPI licensing. These results suggest that technical domains can provide unexpected benefits for certain linguistic phenomena despite their complexity.

Finally, subtitle datasets illustrate the value of naturalistic, conversational language. Comedy and Romance perform particularly well across multiple categories, while Sci-Fi and Action emphasize morphology and semantics but lag in syntax. Overall, our findings reinforce the conclusion that exposure to diverse, dialogue-rich and contextually coherent text supports stronger generalization in low-resource pretraining.

7 Limitations

Our work has several limitations that should be considered when interpreting the results. First, we evaluate only a single model architecture, DeBERTa-v3-small, trained from scratch. While this ensures consistency across experiments, it limits the generality of our conclusions. Other architectures, such as causal language models or variants with modified attention mechanisms, may display different sensitivities to genre, dataset size, or context length. Exploring such alternatives would help to assess whether the trends we observe are architecture-dependent.

Second, the BabyLM-Tiny setting amplifies the effect of random variance. With just 1M tokens per run, small differences in initialization, sampling, or optimization can strongly influence performance. Although we standardized preprocessing and training pipelines, some results may still reflect noise rather than stable effects. Running additional seeds and averaging across replicates would provide more robust estimates.

Finally, our experiments are restricted to English data. Structural properties of English, including its relatively simple morphology and fixed word order, may limit the cross-linguistic scope of our findings. Languages with richer inflectional paradigms or freer word order might benefit differently from genre specialization or contextual continuity. Replication across multiple languages would therefore be an important step toward broader validation.

References

BabyLM Challenge organizers. 2025. Babylm challenge — sample-efficient pretraining on a developmentally plausible corpus. https://babylm.github.io/. Third BabyLM Challenge as a workshop at EMNLP 2025; call for papers and evaluation pipeline released.

Sagi Eden. 2022. Subtitlecf github repository. https://github.com/sagiede/SubtitleCF. Accessed: August 24, 2025.

Sagi Eden, Amit Livne, Oren Sar Shalom, Bracha Shapira, and Dietmar Jannach. 2022. Investigating the value of subtitles for improved movie recommendations. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 99–109.

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

Project Gutenberg. n.d. Project gutenberg. https://www.gutenberg.org/. Accessed: August 24, 2025.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2025. Babylm challenge: Exploring the effect of variation sets on language model training efficiency. *Preprint*, arXiv:2411.09587.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *Preprint*, arXiv:2412.05149.

Acknowledgments

We would like to express our sincere gratitude to **Lukas Edman** for his valuable guidance and continuous support throughout this project. His constructive feedback during our weekly meetings and his advice on the preparation of both our poster and this paper have been highly beneficial. We also greatly appreciate his assistance with questions related to the implementation.

Contributions

This work is the result of a collaborative effort among the authors. The individual contributions are as follows:

Paper

Nikita Goryetiy: Wrote the *Abstract*, *3.4 Open Subtitles Datasets*, *5.1 Comparative Performance of BabyLM Subsets*, *5.4 Dialogue Genre Effects in OpenSubtitles*, *6 Conclusion*, and *7 Limitations*, and revised and edited text across *1 Introduction* and other sections to improve clarity, consistency, and readability.

Moritz Ladenburger: Wrote *1 Introduction* (including *1.1 Background and Motivation* and *1.2 Research Questions*), *2 Related Work*, *3.1 BabyLM Dataset*, *3.2 Wiki Datasets*, *3.3 Gutenberg Datasets*, *4.1 Experimental Setup*, *4.2 Evaluation Framework*, *5.2 Context Length and Genre Effects in Gutenberg Texts*, and *5.3 Topic and Complexity Effects in Wikipedia Articles*; created all charts and figures.

Poster

Nikita Goryetiy: Created the content collaboratively with Moritz Ladenburger.

Moritz Ladenburger: Created the design and figures. Created the content collaboratively with Nikita Goryety.

Training & Dataset Curation

Nikita Goryetiy: Substantially improved and revised the model training pipeline from Lukas Edman. Set up the evaluation infrastructure. Created the sampling pipeline for the OpenSubtitle datasets. Was in charge of large parts of running the training and evaluation.

Moritz Ladenburger: Created the pipeline for downsampling of the BabyLM dataset and the pipelines for sampling the Wikipedia and Gutenberg datasets. Ran some parts of the training.