Pretraining a Japanese-Only Large Language Model for Studying Second Language Acquisition

Shiho Matta[1] Yin Jou Huang[1] Fei Cheng[1] Takashi Kodama[2]
Hirokazu Kiyomaru[2] Yugo Murawaki[1]
[1]Kyoto University [2]NII-LLMC
{matta,huang,feicheng,murawaki}@nlp.ist.i.kyoto-u.ac.jp
{tkodama,kiyomaru}@nii.ac.jp

Abstract

We introduce Dango, a decoder-only LLM pretrained to approximate an L1 Japanese speaker for studying L1–L2 transfer. Existing multilingual LLMs are trained on mixed-language data, which obscures sequential L1–L2 effects. We show that Japanese web corpora contain substantial incidental English that can compromise L1-only pretraining. We term this issue L2 contamination and propose a strict filtering pipeline to mitigate it. Using the filtered corpus, we train Dango. Dango matches Japanese proficiency of a multilingual baseline while remaining comparatively weak in other languages. We release Dango and our data processing code to enable controlled SLA research and more faithful L2-learner simulators.[1)]

1 Introduction

Second language acquisition (SLA) refers to acquiring a second language (L2) after one has already learned a first language (L1) natively. As humans develop proficiency in an L2, their L1 can strongly shape the L2 acquisition, sometimes facilitating progress and sometimes causing interference [1]. A natural computational approach is to use language models to study and reproduce these transfer effects, but this immediately raises a mismatch: modern large language models (LLMs) achieve remarkably strong multilingual performance, often surpassing typical human L2 ability [2]. Typically, these LLMs are trained on multilingual corpora simultaneously [3, 4] rather than in an L1–L2 sequential setting. This mismatch makes them difficult to simulate L1–L2 transfer.

Existing transfer-oriented LMs that simulate an L1–L2 acquisition process are not yet practical simulators of human L2 speakers. Most work uses encoder-only architectures [5, 1], which cannot support open-ended generation, or very small decoder models (e.g., 30M GPT-2 scale) [6], which tend to be limited in instruction following and interactive generation. This constrains progress toward controllable L2-speaker simulators for applications such as learner support [7] and teacher training [8].

To enable human-like SLA research with more modern LLMs, we pretrain a decoder-only LLM on a Japanese web corpus to simulate an L1 Japanese speaker, and treat English as the target L2. Concretely, we train a 1.8B-parameter Llama 2–style decoder Transformer, **Dango**, which is large enough to support instruction tuning and open-ended generation, enabling future analysis of transfer-driven behaviors in interactive settings.

A key challenge in scaling L1 Japanese pretraining is that Japanese web corpora inevitably contain English exposure. Japanese speakers routinely encounter and use English through loanwords, proper nouns, and code-switching [9], which appear in news articles, blogs, and Wikipedia pages crawled from the web. While some of this English is minimal and everyday, other instances are inappropriate for our L1-only pretraining goal: for example, full English passages or parallel translations (see Appendix B). We refer to such unintended exposure as **L2 contamination**. Without explicitly auditing and controlling for it, it is difficult to attribute subsequent L2 behavior to controlled L2 learning rather than to pretraining exposure.

To quantify and mitigate L2 contamination, we build a strict filtering pipeline (Figure 1) and apply it to a large Japanese web corpus, finding that roughly 30% of documents contain contamination. We remove contami-

1) https://github.com/mattashiho233/dango

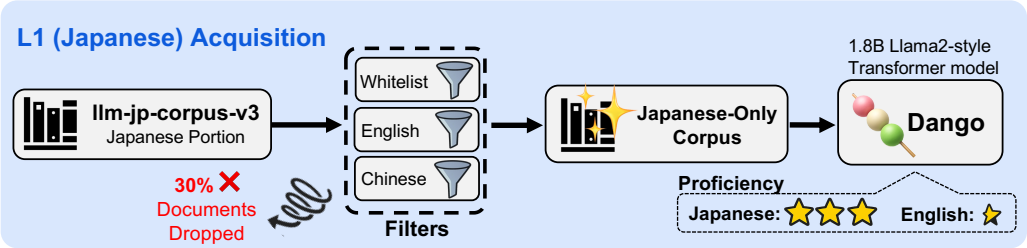


Figure 1: An overview of our proposed methodology.

nated English material while preserving minimal, everyday English fragments that Japanese speakers commonly encounter (see Appendix B for allowed examples). Using this filtered corpus, we pretrain **Dango** as an L1 Japanese model, providing a practical foundation for studying L1–L2 transfer with modern, instruction-tunable LLMs.

Experiments show that Dango matches the Japanese proficiency of a multilingual baseline, llm-jp-3 [10], while demonstrating significantly lower proficiency in English, Chinese, and other languages. We release Dango and our code to enable controlled L1–L2 transfer studies, supporting future work on L2 contamination and faithful learner simulation.

2 Methodology

We model L1 (Japanese) acquisition by pretraining a decoder-only LLM on a strictly filtered corpus. Section 2.1 describes the corpus origin and filtering procedures, and Section 2.2 details the pretraining setup. Linguistic proficiency assessment methods are provided in Section 2.3.

2.1 Corpus and Filtering Method

To bootstrap our study, we build upon an established baseline by adopting the training corpus from the llm-jp project, specifically the llm-jp-corpus-v3[2] version. This corpus is intentionally multilingual, consisting of Japanese, English, Chinese and Korean, as well as code data. We only used the Japanese portion, which mostly consists of Japanese Common Crawl[3] and Japanese WARP web-crawled PDF data.[4] The data was already cleaned and deduplicated[5], however, we noticed that it contained a significant amount of English content.

We built a text filtering pipeline that removes non-Japanese content from the corpus, with English as the primary target. The intent is to reduce the model's exposure to English grammar and long, well-formed English sentences, while still allowing benign signals such as short everyday words, common phrases, and proper nouns that often appear in Japanese text.

The pipeline combines a whitelist filter with two blocklist filters (English and Chinese) applied at both the document and line level. The **whitelist filter** defines a character inventory we consider acceptable for Japanese-centric text, which includes Japanese kana, common CJK[6] characters, punctuation, symbols, and other frequently occurring marks. Documents containing characters outside this range beyond 0.1% will be dropped entirely. This filter effectively handles most languages that do not use CJK or Latin characters, such as Russian. Then, **blocklist filters** (English and Chinese filters) are applied to remove the lines (separated by "\n") that contain too many English (Latin) or Chinese characters. For example, the English filter will remove a *line* in a document if any of the following conditions hold: (i) it contains more than 20 Latin characters; (ii) the ratio of Latin to non-Latin characters exceeds 40%; or (iii) it contains more than four consecutive Latin words separated by spaces. For Chinese, we constructed a Chinese (primarily Simplified Chinese) character list derived from the Unihan BMP list[7] and removed any line containing these characters. Furthermore, if more lines than a certain threshold (5% for English, 0.1% for Chinese) are removed in a document, we discard the document because the context may be damaged due to filtering.

Applying the full pipeline removes approximately 30% of documents from the original corpus, highlighting the need for strict filtering when building a Japanese L1 pretraining dataset. After filtering, the corpus's total token count shrank from 592B to 376B. Detailed filtering statis-

2) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3/-/tree/main
3) https://commoncrawl.org.
4) National Diet Library (Japan). NDL Web Archiving Project (WARP), https://warp.ndl.go.jp/.
5) Refer to the original repository for the pre-processing details.
6) Chinese Hanzi, Japanese Kanji, and Korean Hanja.
7) https://unicode.org/charts/unihan.html

tics are provided in Table 1 in Appendix.

2.2 Pretraining an L1-Only LLM

We pretrained a 1.8B-parameter Llama 2–style decoder-only Transformer from scratch using the filtered corpus. We adopted the training framework of llm-jp-3 models open-sourced by the llm-jp project.[8] We trained Dango on 100B tokens from the filtered corpus. Prior work on compute-optimal training suggests that a 1.8B-parameter model is optimally trained on roughly 35B tokens [11]. We trained for approximately 3 times the tokens because Dango's Japanese generalization continued to improve beyond 35B tokens (Figure 2). Training was stopped at 100B tokens due to budget constraints. More pretraining details are provided in Appendix A.

We adopted the pretrained multilingual tokenizer from the released version of llm-jp-3 models.[9] This tokenizer uses a vocabulary of 99,584 tokens spanning Japanese and English, with additional coverage of Chinese and Korean. We acknowledge that initializing the model with L2 subword knowledge from the outset is less human-like. We leave experiments with alternative, more human-like tokenization assumptions to future work.

2.3 Linguistic Proficiency Assessments.

We evaluate the model's linguistic proficiency on various languages, focusing on Japanese and English.

2.3.1 Assessing Japanese Proficiency.

We evaluate the model on llm-jp-eval-v1.4.1[10] to measure Japanese instruction-following capability and intrinsic Japanese language knowledge. This few-shot benchmark [12] uses Japanese prompts that specify task requirements and expected outputs across diverse Japanese-language tasks, such as natural language inference and multiple-choice QA. We report an average score across tasks, excluding programming, math, and translation tasks, which are less aligned with our goal of evaluating Japanese proficiency.

We also adopt JBLiMP [13], a BLiMP [14] style benchmark for grammatical knowledge. We adopted the data from JBLiMP and report the performance under the MultiBLiMP [15] framework and report the model's Japanese proficiency alongside llm-jp-eval. See details about BLiMP tests and MultiBLiMP in Section 2.3.2.

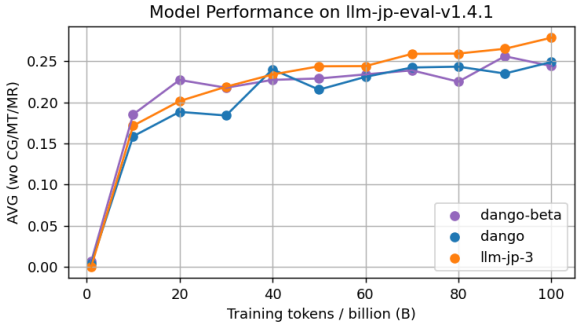


Figure 2: Model performance on llm-jp-eval-v1.4.1.

2.3.2 Assessing English Proficiency.

We adopt MultiBLiMP [15], a BLiMP [14] style multilingual benchmark for grammatical knowledge. In a BLiMP test, models are evaluated using minimal pairs: two nearly identical sentences that differ only in a targeted grammatical property, where one sentence is well-formed and the other contains a specific violation (e.g., *The cats sleep* vs. **The cats sleeps*). The model "passes" a test item if it assigns higher probability to the grammatical sentence than to its ungrammatical counterpart, and overall performance is reported as accuracy across items (typically with chance at 50%). We used the English part in MultiBLiMP for English proficiency assessment.

2.3.3 Assessing Other Languages.

We evaluated Chinese proficiency using ZhoBLiMP [16] to assess the effectiveness of the Chinese filter. We also tested Russian in MultiBLiMP, as it uses non-Latin and non-Japanese characters, to verify the effectiveness of the whitelist filter.

3 Experiments and Results

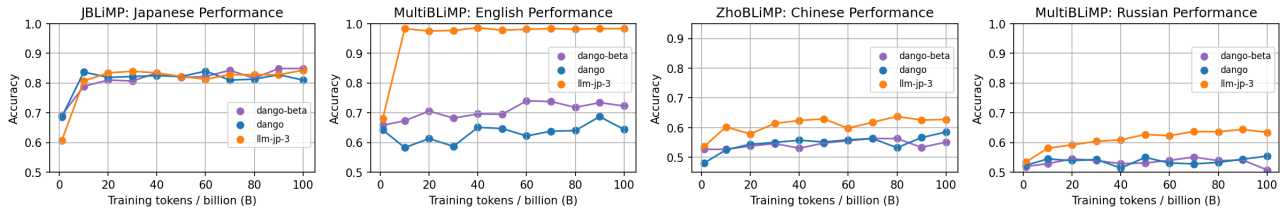
In this section, we share detailed experimental settings and evaluation results for Dango.

3.1 Baselines

We include two baselines: the official llm-jp-3 model[11] and Dango-beta, a preliminary version of Dango. Dango-

8) https://github.com/llm-jp/scripts/tree/main/pretrain/scripts
9) https://huggingface.co/llm-jp/llm-jp-3-1.8b
10) https://github.com/llm-jp/scripts/tree/main/evaluation/installers/llm-jp-eval-v1.4.1

11) We obtained checkpoints trained up to 100B tokens by directly contacting the llm-jp team.



(a) Japanese (JBLiMP) (b) English (MultiBLiMP) (c) Chinese (ZhoBLiMP) (d) Russian (MultiBLiMP)

Figure 3: BLiMP-style benchmark performance during L1 pretraining.

beta was trained on filtered data with modified rules from Section 2.1: rule (i) (maximum Latin characters per line) was loosened from 20 to 100, and rule (iii) (removing four consecutive English words) was excluded.[12] Dango-beta is included to demonstrate the necessity of filtering out long and structured English sentences.

User:
(...4 shots...)
「それから 2 秒後には音が聞こえ、いろいろな物が振動し始めた」
(GT translation: And in two seconds, we heard the noise and things started to shake.)

Dango:
It is right you want to your whole life in Seattle.

Dango-beta:
I have heard of the sound of two seconds.

llm-jp-3:
Then I heard a sound and a number of things started to move.

Figure 4: Case study of English production on a llm-jp-eval translation task.

3.2 Linguistic Proficiency Assessments

We trace the linguistic proficiency of the models over a 10B-token training interval.[13]

Japanese Proficiency. On llm-jp-eval (Figure 2), Dango performs slightly below llm-jp-3 (-0.029) as training approaches 100B tokens, on par with Dango-beta.

On JBLiMP (Figure 3a), the Japanese-exclusive training of Dango and Dango-beta yields early gains over llm-jp-3 at 1B tokens. However, this gap closes over time, and all models converge to comparable performance. Overall, Dango attains Japanese proficiency on par with llm-jp-3.

English Proficiency. On MultiBLiMP English (Figure 3b), llm-jp-3 saturates by 10B tokens. In contrast, Dango stagnates near 0.65 even after 100B tokens, suggesting it fails to scale English grammatical competence even with additional pretraining.

Dango-beta demonstrates noticeably higher proficiency than Dango. This gap shows that filtering long, structured English sentences effectively hinders grammar acquisition.

Chinese Proficiency. On ZhoBLiMP (Figure 3c), both Dango and Dango-beta underperform llm-jp-3, scoring only slightly above chance. This confirms the effectiveness of the Chinese filter.

Other L2 Leakage. We evaluated Russian proficiency (Figure 3d) to test the exclusion of non-Latin/non-Japanese text. Both models performed near chance, supporting the whitelist filter's efficacy.

Case Study. We present a case study of model English generation in Figure 4.[14] The English generated by llm-jp-3 captured the meaning of the Japanese source text most accurately among the three. Dango-beta captured only a very limited portion; however, the grammar of the sentence itself is correct. Dango had great difficulty both in capturing the meaning and in producing grammatically correct English. The output remains largely n-gram-like, as only local fragments like "It is right," "you want to," and "your whole life" make sense. This behavior is likely induced by the four-consecutive-English-word filter.

4 Conclusions

We introduce Dango, a decoder-only LLM approximating an L1 Japanese speaker for controlled second language acquisition. Identifying "L2 contamination" (incidental English) in web corpora as a key obstacle, we propose a strict filtering pipeline to mitigate it. Dango matches the Japanese proficiency of a multilingual baseline while remaining comparatively weak in other languages. We release Dango and our code to support reproducible SLA studies and future learner simulators.

12) Dango-beta is trained on the Common Crawl section only.
13) In Figures 2 and 3, the leftmost points correspond to models trained on 1B tokens.
14) The case is from the alt-j-to-e task in llm-jp-eval.

Acknowledgments

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11763–11777, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 13171–13189, Singapore, December 2023. Association for Computational Linguistics.

[3] Teven Le Scao, Angela Fan, Christopher Akiki, et al. Bloom: A 176b-parameter open-access multilingual language model. **arXiv preprint arXiv:2211.05100**, 2022.

[4] Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mgpt: Few-shot learners go multilingual, 2023.

[5] Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. Second language acquisition of neural language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13557–13572, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Tatsuya Aoyama and Nathan Schneider. Modeling nonnative sentence processing with L2 language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 4927–4940, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Boning Lyu, Chun Lai, and Jianing Guo. Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. **International Journal of Applied Linguistics**, 11 2024.

[8] Areti Vasmatzoglou and Neasa Ní Chiaráin. The development of an online game-based simulation for the training of english language teachers in virtual environments. In Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, and Sylvie Thouësny, editors, **CALL for Widening Participation: Short Papers from EUROCALL 2020**, pp. 334–341. Research-publishing.net, Voillans, France, December 2020.

[9] James Stanlaw. **Japanese English: Language and Culture Contact**. Hong Kong University Press, 2004.

[10] LLM-jp, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[11] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[12] 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, 2020.

[13] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[14] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[15] Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs, 2025.

[16] Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. A systematic assessment of language models with linguistic minimal pairs in chinese, 2025.

Table 1: Statistics of the Japanese corpus after filtering.

Subcategory	by whitelist	by English	by Chinese
CommonCrawl	6.89%	23.00%	0.17%
Kaken	13.13%	54.67%	0.00%
WARP HTML	2.61%	13.03%	0.07%
WARP PDF	14.10%	8.61%	0.18%
Wikipedia	7.35%	23.82%	0.22%
Overall	7.67%	21.65%	0.17%

Subcategory	Total Dropped	# docs remain
CommonCrawl	29.96%	295,309,943
Kaken	67.80%	2,228,758
WARP HTML	15.70%	1,659,736
WARP PDF	22.84%	35,528,902
Wikipedia	31.32%	1,364,529
Overall	29.39%	336,091,868

A Pretraining Details

We reproduced the customized Megatron-LM pretraining framework developed by llm-jp.[15] Most hyperparameters, such as the model architecture and optimization settings, were kept identical to those of the official llm-jp-3 models,[16] except that we increased the global batch size from 512 to 1024 and micro batch size from 4 to 8. The learning-rate schedule was set to a maximum of 376B tokens to match the size of the filtered corpus, but trained for only 100B tokens. In contrast, the final released llm-jp-3 model is scheduled for 2.1T tokens in total. Thus, our model and the corresponding intermediate llm-jp-3 checkpoint represent partial progress along their respective full training trajectories. Pretraining was conducted on 32 NVIDIA H200 GPUs (4 nodes × 8 GPUs) for approximately 26 hours.

B Filtered Documents: Case Study

We found severe English contamination in the Japanese corpus, including document-level (Figure 5) and phrase-level parallel data (Figure 6). In these examples, lines (separated by "\n") shown in red are flagged by the filter.

In contrast, we retain English that a typical Japanese speaker may encounter or use in everyday life, as shown in Figure 7. For example, common short tokens such as "hit" and "ok," proper nouns such as "Access" and "Excel," and short phrases (e.g., "stay at home order," "52 Reasons I Love You") do not trigger the English filter.

```
(.....)
若年女性の流行
Trends
Japan is the land of trends. Nowhere else do trends arise, spread and \n
die with such speed. The reasons for this are simple: affluent youth, merciless \n
advertising, high population density and an insatiable appetite for 'the \n
(.....)
流行
日本は流行り廃りの激しい国だ。こんなに速いスピードで流行が生まれ、広がり、そして廃れていく国は他にない。そうなった理由は単純だ。小金を持った若年層、容赦ない宣伝、高い人口密度、そして新し物好きの国民性。 \n
(.....)
```

Figure 5: Contamination: document-level parallel data.

```
(.....)
左ひだり left (side) \n
折おれて turn (toward) \n
血塔けっとう Bloody Tower \n
門もん gate \n
入はいる enter \n
今いま now; the present \n
昔むかし long ago; ancient times \n
薔薇しょうびの乱らん War of the Roses (15th century struggle for the throne of England) \n
(.....)
```

Figure 6: Contamination: word/phrase-level parallel data.

```
アクセス数:0 hit (累計:15524 hit) 逆アクセス数:0 hit (累計:6371 hit)

応募資格 IT 関連実務経験が少しでもあれば ok！ Access、Excel マクロ経験ある方は尚歓迎です！

ハワイでは、3 月 22 日に州政府より緊急事態宣言が発令され、翌日午後 4 時より 4 月 30 日まで不要不急の外出を禁止する “stay at home order” が発令されました。

(.....)52 枚のトランプに好きな人の好きなところを書いてプレゼントするラブレター"52 Reasons I Love You" に着想を得て生まれた曲だそう。
```

Figure 7: Occurrences of English that are allowed by the filtering pipeline.

15) https://github.com/llm-jp/scripts/tree/main/pretrain/scripts/v4-high-quality-abci

16) The full set of hyperparameters is available at https://github.com/llm-jp/scripts/blob/main/pretrain/scripts/v4-high-quality-abci/pretrain/params/v3-1.8b.sh.