Release of Pre-Trained Models for the Japanese Language

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui,
Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, Koh Mitsuda
rinna Co., Ltd., Tokyo, Japan
keisawada@rinna.co.jp

Abstract

AI democratization aims to create a world in which the average person can utilize AI techniques. To achieve this goal, numerous research institutes have attempted to make their results accessible to the public. In particular, large pre-trained models trained on large-scale data have shown unprecedented potential, and their release has had a significant impact. However, most of the released models specialize in the English language, and thus, AI democratization in non-English-speaking communities is lagging significantly. To reduce this gap in AI access, we released Generative Pre-trained Transformer (GPT), Contrastive Language and Image Pre-training (CLIP), Stable Diffusion, and Hidden-unit Bidirectional Encoder Representations from Transformers (HuBERT) pre-trained in Japanese. By providing these models, users can freely interface with AI that aligns with Japanese cultural values and ensures the identity of Japanese culture, thus enhancing the democratization of AI. Additionally, experiments showed that pre-trained models specialized for Japanese can efficiently achieve high performance in Japanese tasks.

Keywords: AI democratization, pre-trained model, Japanese language

1. Introduction

As AI technology advances, the idea of "AI democratization," which aims to create a world where everyone can easily use AI, has become widely popular. To contribute to AI democratization, many research institutions and companies are publicly releasing their latest methods, source codes, databases, and pre-trained models. Such steps are essential for supporting the rapid development of AI technology in the future.

Recently, methods using large-scale pre-trained models based on massive training data have achieved significant results and have become mainstream. The advent of self-supervised learning, which generates pseudo-ground-truth labels from training data, coupled with the introduction of the Transformer architecture (Vaswani et al., 2017), which enables efficient and accurate model training from massive data, has made large-scale modeling possible. The Generative Pre-trained Transformer (GPT, Radford et al. 2018) series engendered a breakthrough in text generation using self-supervised learning and Transformer architectures by discovering a scaling law suggesting that the performance improves as the model size, amount of training data, and computation increase (Kaplan et al., 2020). As a result, the size of the pre-trained models has escalated dramatically in the text domain as well as the image and speech domains. However, training high-performance pre-trained models incurs significant costs, such as creating training corpora and securing computational resources, making it infeasible for everyone to undertake pre-training easily. Fortunately, there is an active trend of releasing pre-trained models on platforms such as Hugging Face, and such models are now available.

While there is vibrant activity in the publishing of pre-trained models, many pre-trained models targeting languages are specialized for English. Consequently, AI democratization lags in non-English-speaking regions compared with English-speaking regions. Research is underway on multilingual models that support several languages. However, these multilingual models tend to have an increased number of parameters and often underperform compared with models specialized for a particular language given a fixed compute budget (Lin et al., 2022).

To address this issue in Japanese, we released pre-trained models optimized for Japanese on Hugging Face. By providing pre-trained models specialized for Japanese, we hope that users can freely access a model that aligns with Japanese cultural values but also ensures the identity of Japanese culture, leading to a more inclusive AI democratization that does not solely lean towards English-centric perspectives.

2. Japanese Pre-Trained Models

We built pre-trained models appropriate for the Japanese language and culture and released them in Hugging Face[1]. Table 1 presents an overview of the released pre-trained models we have released by September 2023. These models have fewer restrictive licenses, thereby allowing their wide use. In fact, between April 2021 and September 2023, these models were downloaded over four million times from Hugging Face. The de-

[1] https://huggingface.co/rinna

LREC-COLING 2024, pages 13898–13905
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Pre-trained model	Model size	License	Release date
Language model			
rinna/japanese-gpt2-xsmall	37M	MIT	August 2021
rinna/japanese-gpt2-small	110M	MIT	August 2021
rinna/japanese-gpt2-medium	336M	MIT	April 2021
rinna/japanese-gpt-1b	1.3B	MIT	January 2022
rinna/japanese-gpt-neox-small	110M	MIT	September 2022
rinna/japanese-gpt-neox-3.6b	3.6B	MIT	May 2023
rinna/bilingual-gpt-neox-4b	4B	MIT	July 2023
Language-image model			
rinna/japanese-clip-vit-b-16	197M	Apache 2.0	May 2022
rinna/japanese-cloob-vit-b-16	197M	Apache 2.0	May 2022
rinna/japanese-stable-diffusion	1.1B	CreativeML OpenRAIL M	September 2022
Speech model			
rinna/japanese-hubert-base	95M	Apache 2.0	April 2023

Table 1: Released pre-trained models in the Japanese language.

tails and specifics of these models are discussed in Sections 3 to 5.

3. Language Models

3.1. GPT

3.1.1. Overview

The Generative Pre-trained Transformer (GPT, Radford et al. 2018) is an autoregressive language model composed of an input embedding layer, stacked Transformer layers (Vaswani et al., 2017), and an output classification layer. It models $p(\mathbf{x})$, the probability of a sequence of text tokens $\mathbf{x} = [x_1, \cdots, x_{|\mathbf{x}|}]$, as factorized token-level probabilities, and then pre-trains a GPT model to minimize the negative log-likelihood (NLL) $\mathcal{L}_{\text{NLL}}$.

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)\cdots p(x_{|\mathbf{x}|}|x_{:|\mathbf{x}|-1}), \quad (1)$$

$$\mathcal{L}_{\text{NLL}} = -\log p(\mathbf{x}). \quad (2)$$

GPT-NeoX (Black et al., 2022) is a GPT variant that uses a modified architecture for the Transformer layer and an alternative position encoding mechanism called rotary embedding (Su et al., 2021) as a substitute for original learnable position embeddings.

For the most capable models, we also released their instruction-following versions, which were trained using either Supervised Fine-Tuning (SFT) or Reinforcement Learning from Human Feedback (RLHF, Ouyang et al. 2022) via the Proximal Policy Optimization (PPO, Schulman et al. 2017) algorithm in addition to SFT.

3.1.2. Training Data

For Japanese-specific GPT models, we used Wikipedia, the CC-100 (Conneau et al., 2020), and the mC4 (Raffel et al., 2020) datasets for pre-training. For bilingual English-Japanese GPT models, we additionally used the Pile (Gao et al., 2020) and Redpajama (Computer, 2023) datasets. The instruction-following models were trained on the Japanese translation of the Anthropic HH (Bai et al., 2022), the SHP (Ethayarajh et al., 2022), and the FLAN (Wei et al., 2022) datasets.

Tokenizers of the GPT models are trained via SentencePiece (Kudo and Richardson, 2018). Their vocabulary sizes vary from 32000 to 65536. While the tokenizers for Japanese-only models are trained from Japanese corpora, the tokenizer of *bilingual-gpt-neox-4b* is trained from a mixture of Japanese and English corpora for a better coverage of English tokens.

3.1.3. Experiments

We conducted few-shot evaluations of the GPT models to assess their performance on Japanese tasks. We used the JP Language Model Evaluation Harness[2] benchmark for evaluation. We conducted a comparison with *meta/llama-7b* (Touvron et al., 2023a), *meta/llama2-7b*, and *meta/llama2-7b-chat* (Touvron et al., 2023b), which were primarily trained using English data.

Table 2 lists the average scores for the jcommonsenseqa, jnli, marc-ja, jsquad (Kurihara et al., 2022), xwinograd (Muennighoff et al., 2023), jaqket-v2[3], xlsum-ja (Hasan et al., 2021), and mgsm (Shi et al., 2023) tasks. The few-shot numbers were 3, 3, 3, 2, 0, 1, 1, and 5. Our *rinna/japanese-gpt-neox-3.6b* and *rinna/bilingual-*

[2] https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable

[3] https://sites.google.com/view/project-aio/competition2

Pre-trained model	Average score
rinna/japanese-gpt2-xsmall	26.63
rinna/japanese-gpt2-small	27.33
rinna/japanese-gpt2-medium	28.33
rinna/japanese-gpt-1b	32.21
rinna/japanese-gpt-neox-small	30.11
rinna/japanese-gpt-neox-3.6b	36.60
rinna/bilingual-gpt-neox-4b	38.29
meta/llama-7b	33.28
meta/llama2-7b	42.97
meta/llama2-7b-chat	41.31
rinna/japanese-gpt-neox-3.6b-sft	45.24
rinna/japanese-gpt-neox-3.6b-ppo	46.37
rinna/bilingual-gpt-neox-4b-sft	47.65
rinna/bilingual-gpt-neox-4b-ppo	47.33

Table 2: Language model evaluation on the JP Language Model Evaluation Harness.

gpt-neox-4b pre-trained models outperformed *meta/llama-7b*, and instruction tuning via SFT or PPO significantly improved their capability. By specializing in Japanese, good performance was achieved while keeping the number of parameters low. We refer the readers to rinna's language model benchmark[4] for detailed benchmark results.

4. Language-Image Models

4.1. CLIP

4.1.1. Overview

Contrastive Language-Image Pre-training (CLIP, Radford et al. 2021) connects visual concepts with natural language in the embedding space. It comprises a pair of text and image encoders and is trained by minimizing contrastive loss. The Contrastive Leave One Out Boost (CLOOB, Fürst et al. 2022) demonstrated a better zero-shot performance than the original CLIP by introducing a novel loss function termed InfoLOOB.

To train the Japanese-specific CLIP efficiently, we applied Locked-image Tuning (LiT, Zhai et al. 2022), in which both encoders were initialized using separate pre-trained models, and only the text encoder was trained. We used the pre-trained 12-layer 16×16-patch-size AugReg Vision Transformer (Dosovitskiy et al., 2021; Steiner et al., 2022) for the image encoder, and randomly initialized 12-layer Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2019) with a SentencePiece tokenizer (Kudo and Richardson, 2018) for the text encoder.

Pre-trained model	Accuracy
laion/clip-base	38.00
laion/clip-large	53.09
rinna/japanese-clip-vit-b-16	50.69
rinna/japanese-cloob-vit-b-16	54.64

Table 3: ImageNet image classification accuracy in a zero-shot setting.

4.1.2. Training Data

Owing to the absence of a large-scale dataset with Japanese captions, we used CC12M (Changpinyo et al., 2021). We translated all the English captions into Japanese. For data augmentation, we generated captions using Bootstrapping Language-Image Pre-training (BLIP, Li et al. 2022) trained on an English dataset.

4.1.3. Experiments

We evaluated CLIP for ImageNet (Deng et al., 2009) zero-shot image classification. We used open-sourced Japanese class names[5]. Additionally, we created 37 Japanese templates from 80 English templates by deduplicating captions that had the same meaning in Japanese. We compared our models with open-source multilingual CLIP models (Ilharco et al., 2021) trained on full LAION-5B (Schuhmann et al., 2022).

Table 3 shows the top-1 accuracy for each model. Our *rinna/japanese-cloob-vit-b-16* performed the best and achieved state-of-the-art accuracy. This is because, even with a limited amount of training data, the model can be efficiently trained by specializing in a specific language.

4.2. Stable Diffusion

4.2.1. Overview

Stable Diffusion (SD) facilitates high-quality image generation using simple text prompts. It is based on the Latent Diffusion Model (LDM, Rombach et al. 2022), which comprises three main components: a Variational AutoEncoder (VAE, Kingma and Welling 2014), a text encoder, and U-Net (Ronneberger et al., 2015).

[4] https://rinnakk.github.io/research/benchmarks/lm/index.html Due to the update of the evaluation code base, the latest benchmark adopts a different evaluation setting than that used in this paper. The results presented in this paper can be found in the benchmark spreadsheet on the *20231031* tab.

[5] https://gist.github.com/PonDad/4dcb4b242b9358e524b4ddecbee385e9

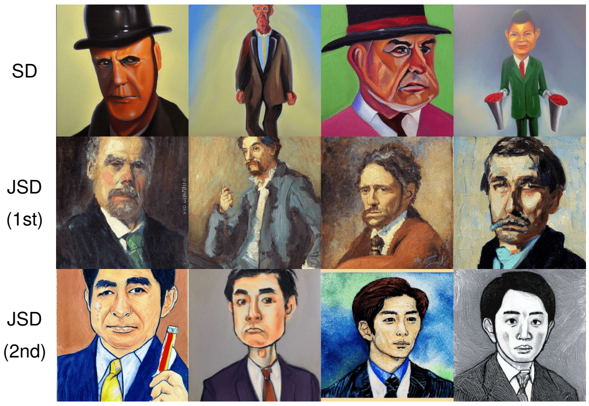

Figure 1: Outputs for the text prompt “salary man, oil painting”. For JSD, the translation in Japanese “サラリーマン 油絵” was used.

To train the Japanese-specific SD (JSD), we fine-tuned *CompVis/stable-diffusion-v1-4*[6] trained on the English dataset. We applied two training stages following the concept of Pretraining-based Image-To-Image translation (PITI, Wang et al. 2022); The text encoder was trained solely with U-Net fixed in the first stage and jointly trained in the second stage.

4.2.2. Training Data

We used approximately 100 million images with Japanese captions, including the Japanese subset LAION-5B (Schuhmann et al., 2022). To ensure data quality, we employed our *rinna/japanese-cloob-vit-b-16* introduced in Section 4.1 to calculate the similarity scores between images and their captions, and samples with scores below a certain threshold were removed.

4.2.3. Experiments

We used Japanglish “salary man”, which is commonly visualized as a man in a suit, as the text prompt for the evaluation. Figure 1 shows the results. The original SD failed to accurately interpret such distinctive Japanese terms. In the first stage, JSD understood the prompt’s meaning, but the generated images depicted businessmen with Western features because U-Net had not been updated. JSD at the second stage (*rinna/japanese-stable-diffusion*), JSD successfully generated images of businessmen with Japanese features. Using images reflecting Japanese culture as the training data, we were able to construct a model consistent with Japanese cultural identity.

5. Speech Models

5.1. HuBERT

5.1.1. Overview

The Hidden-unit BERT (HuBERT, Hsu et al. 2021) is a pre-trained model that can learn self-supervised speech representations. HuBERT comprises two main components: a convolutional waveform encoder and a BERT encoder (Devlin et al., 2019). HuBERT is trained with a BERT-like masked prediction objective: a portion of the encoded speech feature sequence is randomly masked, and a label corresponding to the masked portion is predicted from the unmasked portion. However, because speech signals, unlike text, are continuous-valued sequences, the model is trained by targeting discrete pseudo-labels obtained from speech using offline k-means clustering.

5.1.2. Training Data

We used the ReazonSpeech corpus (Yin et al., 2023), a 19,000-hour speech corpus collected from Japanese TV programs with 16 kHz sampling. To generate pseudo-labels, we ran k-means clustering with 100 clusters on 39-dimensional Mel-frequency cepstral coefficient features for the first iteration of HuBERT training and 500 clusters on the latent features extracted from the 6-th Transformer layers’ of the first iteration for the second iteration of HuBERT training.

5.1.3. Experiments

We evaluated the performance of the pre-trained HuBERT model for Japanese Automatic Speech Recognition (ASR). We used Corpus of Spontaneous Japanese (Maekawa et al., 2000). Two training subset sizes were prepared: core data only (approximately 32 h) and all data (approximately 552 h). ASR fine-tuning using the Connectionist Temporal Classification (CTC, Graves et al. 2006) loss was performed as described in (Hsu et al., 2021). The final projection layer was replaced with a softmax layer before ASR fine-tuning. The target vocabulary included 40 Japanese phonemes, a word boundary symbol, and a special CTC blank symbol. The public HuBERT model *meta/hubert-base-ls960*[7], pre-trained with 960 hours English speech from Librispeech (Panayotov et al., 2015), was used for comparison. In this study, we used the beam search with a beam size of 20 without a language model.

[6] https://huggingface.co/CompVis/stable-diffusion-v1-4

[7] https://huggingface.co/facebook/hubert-base-ls960

Pre-trained model	Eval1	Eval2	Eval3
meta/hubert-base			
32-hour labeled	13.12	10.33	10.66
552-hour labeled	7.88	5.66	6.48
rinna/japanese-hubert-base			
32-hour labeled	9.30	7.07	6.87
552-hour labeled	5.72	4.45	4.73

Table 4: The word error rates for the fine-tuned HuBERT models with the different sizes of data.

The results are presented in Table 4. For both sizes of labeled data, the *rinna/japanese-hubert-base* outperformed the *meta/hubert-base-ls960*. This result indicates that the pre-trained HuBERT model trained with a large Japanese speech corpus has the potential to provide better performance in Japanese speech-processing tasks.

6. Conclusion

Aiming to advance AI democratization, this paper discusses the released Japanese GPT, CLIP, Stable Diffusion, and HuBERT models. Experiments with GPT, CLIP, and HuBERT showed that pre-trained models specialized for Japanese can efficiently achieve high performance in Japanese tasks. Additionally, the Stable Diffusion results indicate that it handles Japanese input and produces output that reflects Japanese culture. Pre-trained models are continuously refined, and technically challenging tasks for improving them have now become achievable. We plan to continue releasing pre-trained models to further contribute to technological progress.

7. Bibliographical References

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, volume 1, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR 2021)*.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T. Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, and Sepp Hochreiter. 2022. CLOOB: Modern hopfield networks with InfoLOOB outperform CLIP. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 20450–20468.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pages 369–376.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP. *Zenodo*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv, arXiv:2001.08361*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR 2014)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, volume 162, pages 12888–12900.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 9019–9052.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pages 27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pages 8748–8763.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, pages 234–241.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv, arXiv:1707.06347*.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. 2022. How to train your ViT? Data, augmentation, and regularization in Vision Transformers. *Transactions on Machine Learning Research*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced Transformer with rotary position embedding. *arXiv, arXiv:2104.09864*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv, arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, pages 5998–6008.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. 2022. Pretraining is all you need for image-to-image translation. *arXiv, arXiv:2205.12952*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 18123–18133.

8. Language Resource References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv, arXiv:2204.05862*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, pages 3558–3568.

Together Computer. 2023. RedPajama: An open source recipe to reproduce llama training dataset. *GitHub*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8440–8451.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, volume 162, pages 5988–6008.

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

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 2957–2966.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC 2000)*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 15991–16111.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2015)*, pages 5206–5210.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations (ICLR 2022)*.

Yue Yin, Daijiro Mori, and Seiji Fujimoto. 2023. ReazonSpeech: A free and massive corpus for Japanese ASR. In *Proceedings of the Twenty-ninth Annual Meeting of the Association for Natural Language Processing*, pages 1134–1139.