Modelgo: A Tool for Machine Learning License Analysis

Moming Duan
National University of Singapore
Singapore
moming@nus.edu.sg

ABSTRACT

Productionizing machine learning projects is inherently complex, involving a multitude of interconnected components that are assembled like LEGO blocks and evolve throughout development lifecycle. These components encompass software, databases, and models, each subject to various licenses governing their reuse and redistribution. However, existing license analysis approaches for Open Source Software (OSS) are not well-suited for this context. For instance, some projects are licensed without explicitly granting sublicensing rights, or the granted rights can be revoked, potentially exposing their derivatives to legal risks. Indeed, the analysis of licenses in machine learning projects grows significantly more intricate as it involves interactions among diverse types of licenses and licensed materials. To the best of our knowledge, no prior research has delved into the exploration of license conflicts within this domain. In this paper, we introduce Modelgo, a practical tool for auditing potential legal risks in machine learning projects to enhance compliance and fairness. With Modelgo, we present license assessment reports based on 5 use cases with diverse model-reusing scenarios, rendered by XXX popular machine learning components. Finally, we summarize the reasons behind license conflicts and provide guidelines for minimizing them.

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

KEYWORDS

Software licenses, Software reuse, Open source software, Model mining

1 INTRODUCTION

Over the past decade, the advancement and productization of AI infrastructures have significantly accelerated the proliferation of machine learning (ML) components [11], including AI models [19, 24], software [9, 25], and big datasets [6, 22]. Concurrently, the reuse of these components has gained popularity, motivated by concerns about their significant demands on financial and energy resources [23], as well as the widespread recognition of the value advocated by the open-source movement [20]. Unlike code reuse in the OSS field, the reuse of AI models follow a distinct schema. A frequently employed approach for AI models reuse is fine-tuning Pre-Trained Models (PTMs) [8, 24], where PTMs are adapted on a domain-specific dataset, leveraging their robust generalization capabilities.

From a legal perspective, model reuse is generally uncontroversial when its developers or affiliated companies own the copyright for all components. However, data and models often have separate copyright holders in nowadays ML projects [17, 18, 21, 27]. For instance, GPT-2 [17], developed by OpenAI, was trained on 45 million web pages containing content from third-party platforms like WordPress, GitHub, and IMDb, none of which is owned by OpenAI. These crowdsourced content typically provides limited usage and distribution rights to users through pre-agreed licenses (e.g., Creative Commons Licenses[1]), which may restrict certain reuse methods like remixing, reproducing, and translating. To prevent legal risk, it is essential to ensure that the final ML projects remain compaliant with all license conditions associated with the reused components [4, 12, 14].

However, compared to assessing licensing compliance for OSS, ensuring license compliance in ML projects poses several unique challenges. First, a ML project is not only a combination of software like an OSS project but also composed of datasets and models [8], which may be under different types of licenses (e.g., Free Content Licenses and AI model licenses [3]). Second, ML components often follow more complicated coupling paradigms and nested workflows. For instance, Openjourney[2] is an image generation model derived from StableDiffusion [19], and fine-tuned on images generated by another commercial product, Midjourney[3]. This demonstrates that knowledge can be transferred between models without explicit code integration [26]. Another challenge is improper and ambiguity licensing in ML projects. For example, GPT-2 and BERT [5] are regarded as part of software and then licensed as OSS (e.g., MIT and Apache-2.0). However, ML projects like StableDiffusion and Llama2 [24] tend to apply responsible AI restriction terms for both model and code, using AI model licenses such as OpenRAIL-M [3] and Llama2 Community License[4]. Additionally, to circumvent the limitations of standard OSS licenses, some licensors adopt non-commercial content licenses or custom licenses to protect the Intellectual Property (IP) of their models by prohibiting commercial use [10], fine-tuning [13], and reverse engineering [7]. Such ambiguity and the diverse licensing practices within ML projects increase significant legal uncertainty in license compliance analysis. As a result, traditional OSS license analysis approaches [14, 15] only consider inclusion and linking relationships among software and lack support for AI model licenses, making them unsuitable for ML project license analysis.

In this paper, we introduce Modelgo, a tool designed to analyze potential license conflicts, improper license choices, use restrictions and obligations in ML projects that involve nested component reuse procedures. To demonstrate the usefulness of Modelgo, we present

[1] https://creativecommons.org/licenses/
[2] https://openjourney.art/
[3] https://www.midjourney.com/
[4] https://huggingface.co/meta-llama/Llama-2-7b

5 use cases constructed using 15 datasets and 11 models from real-world scenarios, whose license types cover OSS, free content, and AI model. Our findings show that there exist potential legal risks when reusing components under copyleft or non-commercial licenses, and point out the need for attention to AI model licenses. The main contributions of our paper are:

- We raise the challenge of license analysis for ML projects and propose Modelgo to assessing it. To the best of our knowledge, our work is the first attempt to deal with this challenge in the ML context.
- As part of our work, we introduce a new taxonomy based on the forms of reused components to identify the corresponding conditions for various ML reuse mechanisms. This method helps mitigate ambiguity in cases of mismatch between applied license type and actual component type, allowing Modelgo to analyze components under various license types, including OSS, free content and AI models.
- We provide legal compliance assessment reports based on 5 use cases to showcase the effectiveness of our approach. Through our use cases, we offer valuable insights and experiences in achieving legal compliance in ML projects. Additionally, we also provide license choosing recommendations to minize the risk of non-compliance.

The rest of the paper is organized as follows. (TBD)

Table 1

Table 1: Summary of xxxxx. Copyleft Permissive Public Domain No public

Work Name	License Name	Type	Modality/Usage
Wikipedia	CC-BY-SA-4.0	Data	Text
StackExchange	CC-BY-SA-4.0		
FreeLaw	CC-BY-ND-4.0		
arXiv	CC-BY-NC-SA-4.0		
PubMed	CC-BY-NC-SA-4.0		
Deep-sequoia	CC-BY-NC-ND-4.0		
Midjourney Gen	CC-BY-NC-ND-4.0		Image
Flickr	CC-BY-NC-SA-4.0		
StockSnap	CC0-1.0		
Wikimedia	CC-BY-SA-4.0		
OpenClipart	CC0-1.0		
ccMixter	CC-BY-NC-4.0		Voice
Jamendo	CC-BY-NC-ND-4.0		
Thingverse	CC-BY-NC-SA-4.0		3D model
Vimeo	CC-BY-NC-ND-4.0		Video
Baize	GPL-3.0	Model	Text Generation
BLOOM	BigScience-BLOOM-RAIL-1.0		
Llama2	Llama2		
BigTranslate	GPL-3.0		
BERT	Apache-2.0		Fill-Mask
Stable Diffusion	CreativeML-OpenRAIL-M		Text to Image
MaskFormer	CC-BY-NC-4.0		Image Segmentation
DETR	Apache-2.0		
Whisper	MIT		Voice to Text
X-Clip	MIT		Video to Text
I2VGen-XL	CC-BY-NC-ND-4.0		Image to Video

There is no consensus on whether the use of copyright works as input to train an AI system is an exercise of an exclusive right. There remains significant legal uncertainty about whether copyright applies to AI training, which means it may not always be clear whether a CC license applies. The larger model was trained on 256 cloud TPU v3 cores. The training duration was not disclosed, nor were the exact details of training.

Open source software license compliance [15]

The open source definition [16]

AFL [20]

Wudao2.0 1.75T MoE [FASTMOE: A FAST MIXTURE-OF-EXPERT TRAINING SYSTEM] [GLM-130B: AN OPEN BILINGUAL PRE-TRAINED MODEL]

Objectives and challenges associated with analyzing dataset license compliance? Getty Images (US), Inc. v. Stability AI, Inc. (1:23-cv-00135) Andersen et al v. Stability AI Ltd. et al (3:23-cv-00201) We are not aware of any copyright restrictions of the material

C4, Pile Common Crawl crowdsourced

COCO (CC-BY 4.0), CIFAR10 -> Flickr Unsplash License *Custom*: Compiling photos from Unsplash to replicate a similar or competing service. https://unsplash.com/license Pixabay License: Data mining, extraction, scraping and the use of programs or robots for automatic data collection and/or extraction of digital data on the Services and/or the content available therein is strictly prohibited for all purposes, including without limitation for machine learning purposes.

Google Street View (SVHN) https://about.google/brand-resource-center/products-and-services/geo-guidelines/

Software reuse is very simple from the legal point of view, if a company or an individual reuses software for which it has copyrights. However, things change dramatically if one wants to reuse software made by others, since software is protected by copyright and possibly by patents. Without explicit permission, no person other than the copyright holder is allowed to copy, distribute, or make derivative works from the original work.

2 BACKGROUND AND RELATED WORK

2.1 Machine Learning Project Licensing

Benjamoin *et al.* [1] propose Montreal Data License (MDL).

2.2 FOSS License Assessment

2.3 Machine Learning IP Protection

3 METHOD

This section is organized around three key questions in the context of ML license analysis: (i) How to determine the corresponding conditions in licenses for certain model reuse mechanisms? (ii) How to capture the dependency structure of a machine learning project? (iii) What types of non-compliance exist in ML projects and how to assess them? We will present our solutions to these questions in the following sections.

3.1 Taxonomy for ML License Analysis

Determining the corresponding conditions in licenses is a challenging task for ML projects due to the conceptual ambiguities in existing licensing language and the disorganization in current ML licensing practices. For example, CC-BY-ND prohibits the sharing of derivatives of licensed materials. However, its definition of making derivatives is unclear in the context of ML domain. For instance, should embeddings of a corpus be considered a derivative work upon that corpus? Unfortunately, even though Creative Commons provides a flow chart to illustrate the trigger conditions of CC licenses in the context of AI activity [2], it raises another question:

Table 2: Summary of machine learning projects in Huggingface.

ML Project	Task	Data License	Software License	Model License	Dataset	Risk Resource
Stable Diffusion v1-5	Text to Image	CC-BY-4.0	CreativeML-OpenRAIL-M	CreativeML-OpenRAIL-M	LAION-5B	Common Crawl
BLOOM	Text Generation	*Mixture*	*Unknown*	BigScience-BLOOM-RAIL-1.0	*Crowdsourced*	Common Crawl, Wikipedia, etc.
OrangeMixs	Text to Image	*Mixture*	*Unknown*	CreativeML-OpenRAIL-M	*Crowdsourced*	Danbooru
ControlNet	Text to Image	*Unknown*	Apache-2.0	OpenRAIL	*Unknown*	n/a
Openjourney	Text to Image	CC-BY-NC-4.0	*Unknown*	CreativeML-OpenRAIL-M	Midjourney Gen	Midjourney Gen
ChatGLM-6B	Text Generation	*Mixture*	Apache-2.0	*Custom*	the Pile, Wudao, *Crowdsourced*	PubMed, Wikipedia, arXiv, GitHub, etc.
Llama2	Text Generation	*Unknown*	Llama2 Community License	Llama2 Community License	*Unknown*	n/a
StarCoder	Text Generation	*Mixture*	Apache-2.0	BigCode-OpenRAIL-M	The Stack	none
Falcon-40B	Text Generation	ODC-By	Apache-2.0	Apache-2.0	RefinedWeb	Wikipedia, Reddit, StackOverflow, etc.
Waifu Diffusion	Text to Image	*Mixture*	*Unknown*	CreativeML-OpenRAIL-M	*Unknown*	n/a
Dolly-v2-12B	Text Generation	CC-BY-SA-3.0&4.0	MIT	MIT	databricks-dolly -15k, the Pile	PubMed, Wikipedia, arXiv, GitHub, etc.
Dreamlike Photoreal	Text to Image	*Unknown*	*Unknown*	*Modified* CreativeML-OpenRAIL-M	*Unknown*	n/a
Counterfeit	Text to Image	*Unknow*	*Unknown*	CreativeML-OpenRAIL-M	*Unknown*	n/a
GPT-2	Text Generation	*Mixture*	*Modified* MIT	*Modified* MIT	*Crowdsourced*	WordPress, GitHub, wikiHow, IMDb, etc.
GPT-J-6B	Text Generation	*Mixture*	Apache-2.0	Apache-2.0	the Pile	PubMed, Wikipedia, arXiv, GitHub, etc.
LLaMA-7B	Text Generation	*Mixture*	*Custom*	*Custom*	*Crowdsourced*	GitHub, arXiv, etc.
BERT	Fill Mask	*Mixture*	Apache-2.0	Apache-2.0	Book Corpus, Wikipedia (en)	Wikipedia (en)
Whisper	ASR	*Unknown*	MIT	MIT	*Unknown*	n/a
MPT	Text Generation	*Mixture*	Apache-2.0	Apache-2.0	*Crowdsourced*	Common Crawl, Wikipedia, etc.

Is the output considered protectable copyright subject matter? The answer depends on how the embedding activity is interpreted, for example, considering it as a translation of the original work can trigger the CC license.

MDL advocates the use of a "Top Sheet" to delineate what ML activities are allowed with data [1], but this proposal is rarely implemented in practice (things would be easier if it were widely accepted). Making things more complex, some projects release their models under free content licenses, like LayoutLMv3 model [10], which is licensed under CC-BY-NC-SA-4.0. This disorganization makes it unclear what kinds of ML activities can trigger licenses conditions in different contexts. An ideal and elegant solution would be to encourage licensors to make context-appropriate adaptations in their license agreements or terms of use to clarify the granted rights related to ML activities. However, some ML components may be composed of prior works that are shared under copyleft license templates, which may disallow such relicensing of their derivatives to a new license. Therefore, it is necessary to establish practical rules to bridge AI activities and existing licensing language.

To address the above challenge, we propose a result-based taxonomy that categorizes all AI activities into four categories based on the forms of their results. In our taxonomy, there are four categories of AI activities: Combination, Amalgamation, Distillation, and Generation, which are defined by four forms of their results, respectively: (1) Combination with strong separation; (2) Combination with weak separation; (3) Derivatives from concepts; and (4) Derivatives from data. Correspondingly, we can also categorize the usage behaviors in license language into these four categories based on their outcome forms. We leverage Figure 1 to illustrate this idea, and the details of the four categories are as follows:

Combination

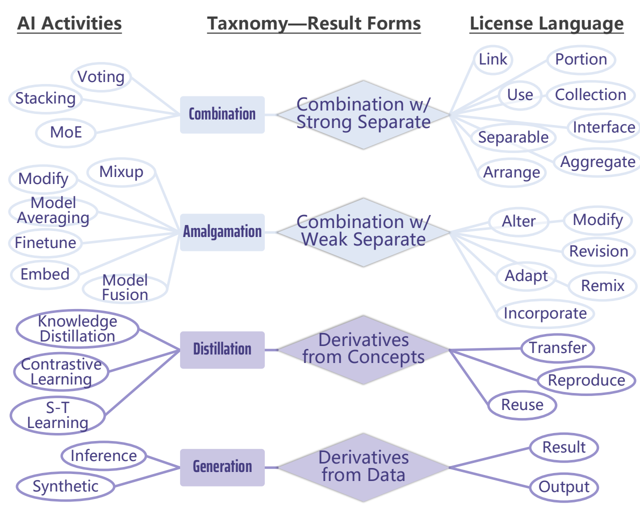

Figure 1: Our proposed taxonomy bridging AI activities and license terms based on their result forms.

translated, altered, arranged, transformed, or otherwise modified
Licensing Language Requires Standardization and to ML and AI the notion of derivative work is ill defined conceptual ambiguities in existing licensing language There is no consensus on whether the use

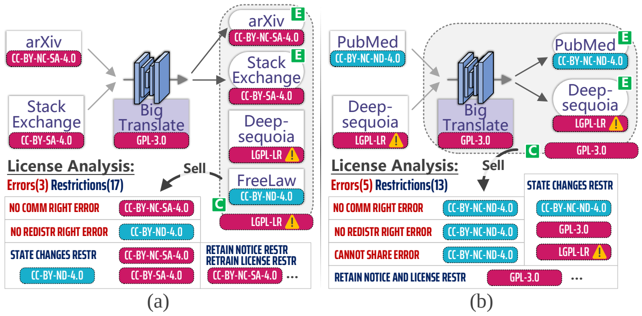

Figure 2: Case Study I: Combination of Corpus. (a) LGPL-LR proliferation, CC collection; (b) LGPL-LR no linguistic resource, CC No redistribution.

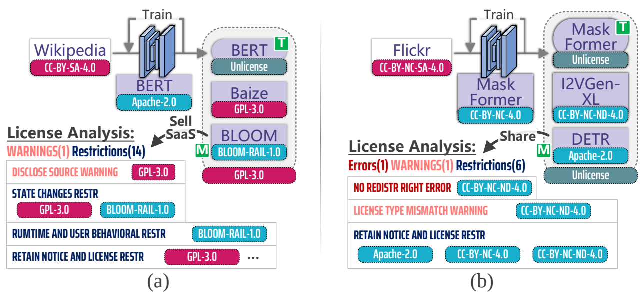

Figure 3: Case Study II: Mixture of Experts. (a) BLOOM-RAIL, binary of GPL; (b) Unlicense, CC-BY-NC no distribute derivative. GPL Automatic Licensing of Downstream Recipients

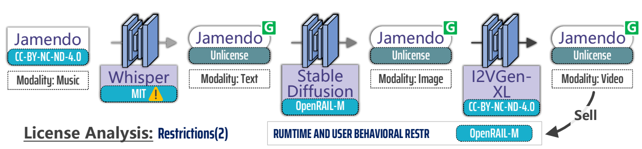

Figure 4: Case Study III: Pipeline.

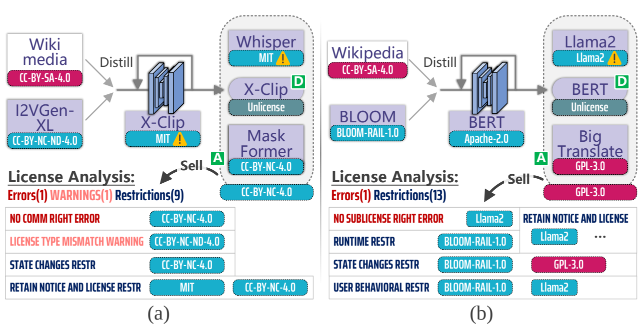

Figure 5: Case Study IV: distillation and model averaging.

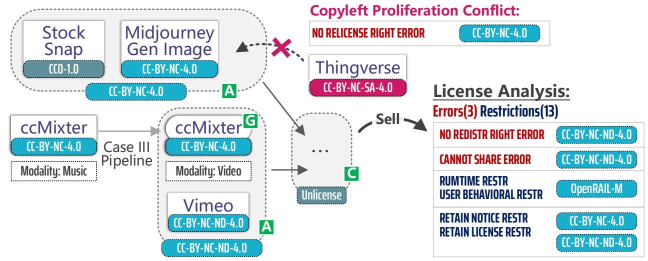

Figure 6: Case Study V: distillation and model averaging.

4 CASE STUDY DETAILS

5 DISCLAIMER

The content presented in this article is intended for general informational purposes only and should not be construed as legal advice. Any views, opinions, findings, conclusions, or recommendations expressed in this material are the sole responsibility of the author(s) and do not represent the perspectives of any organization or entity.

ACKNOWLEDGMENTS

Ack.

REFERENCES

[1] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. 2019. Towards standardization of data licenses: The montreal data license. *arXiv preprint arXiv:1903.12262* (2019).

[2] Creative Commons. 2023. Artificial intelligence and CC licenses. Retrieved September 25, 2023 from https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses

[3] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral use licensing for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 778–788. https://doi.org/10.1145/3531146.3533143

[4] Xing Cui, Jingzheng Wu, Yanjun Wu, Xu Wang, Tianyue Luo, Sheng Qu, Xiang Ling, and Mutian Yang. 2023. An Empirical Study of License Conflict in Free and Open Source Software. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 495–505. https://doi.org/10.1109/ICSE-SEIP58684.2023.00050

[5] 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 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] 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 preprint arXiv:2101.00027* (2020).

[7] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360* (2022).

[8] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open* 2 (2021), 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

[9] Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. 2022. FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. 120–134. https://doi.org/10.1145/3503221.3508418

[10] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*. 4083–4091. https://doi.org/10.1145/3503161.3548112

[11] Wenxin Jiang, Nicholas Synovic, Matt Hyatt, Taylor R Schorlemmer, Rohan Sethi, Yung-Hsiang Lu, George K Thiruvathukal, and James C Davis. 2023. An

empirical study of pre-trained model reuse in the hugging face deep learning model registry. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering (ICSE)*. 2463–2475. https://doi.org/10.1109/ICSE48619.2023.00206

[12] Georgia M Kapitsaki, Frederik Kramer, and Nikolaos D Tselikas. 2017. Automating the license compatibility process in open source software with SPDX. *Journal of Systems and Software (JSS)* 131 (2017), 386–401. https://doi.org/10.1016/j.jss.2016.06.064

[13] Dreamlike Tech Ltd. 2023. Dreamlike Photoreal 2.0. Retrieved September 25, 2023 from https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0

[14] Arunesh Mathur, Harshal Choudhary, Priyank Vashist, William Thies, and Santhi Thilagam. 2012. An empirical study of license violations in open source projects. In *2012 35th Annual IEEE Software Engineering Workshop (SEW)*. IEEE, 168–176. https://doi.org/10.1109/SEW.2012.24

[15] Philippe Ombredanne. 2020. Free and open source software license compliance: tools for software composition analysis. *Computer* 53, 10 (2020), 105–109. https://doi.org/10.1109/MC.2020.3011082

[16] Bruce Perens. 1999. The open source definition. *Open sources: voices from the open source revolution* 1 (1999), 171–188.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[18] Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin, Boyuan Chen, Zhen Ming, Daniel M German, et al. 2021. Can I use this publicly available dataset to build commercial AI software?–A Case Study on Publicly Available Image Datasets. *arXiv preprint arXiv:2111.02374* (2021).

[19] 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)*. 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

[20] Lawrence Rosen. 2005. *Open Source Licensing: Software Freedom and Intellectual Property Law*. Prentice Hall Professional Technical Reference, New Jersey.

[21] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 25278–25294.

[23] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. 3645–3650. https://doi.org/10.18653/v1/p19-1355

[24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[26] Shan You, Chang Xu, Fei Wang, and Changshui Zhang. 2021. Workshop on Model Mining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4177–4178. https://doi.org/10.1145/3447548.3469471

[27] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2023. GLM-130B: An Open Bilingual Pre-trained Model. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.