ModelGo: A Tool for Machine Learning License Analysis

Authors

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 real-world 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

License analysis, AI licensing, 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 [18], including AI models [33, 38], software [14, 39], and big datasets [9, 36]. Concurrently, the reuse of these components has gained popularity, motivated by concerns about their significant demands on financial and energy resources [37], as well as the widespread recognition of the value advocated by the open-source movement [34]. 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) [12, 38], 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 [31, 32, 35, 41]. For instance, GPT-2 [31], 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 [6, 19, 25].

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 [12], which may be under different types of licenses (e.g., Free Content Licenses and AI model licenses [5]). Second, ML components often follow more complicated coupling paradigms and nested workflows. For instance, Openjourney[2] is an image generation model derived from StableDiffusion [33], 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 [40]. Another challenge is improper and ambiguity licensing in ML projects. For example, GPT-2 and BERT [7] 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 [38] tend to apply responsible AI restriction terms for both model and code, using AI model licenses such as OpenRAIL-M [5] 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 [16], fine-tuning [23], and reverse engineering [10]. 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 [25, 27] 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 5 use cases constructed using 15 datasets and 11 models from real-world, whose license types cover OSS, free content, and AI model. Our findings show that there exist potential legal risks when reusing components under copyleft, non-public, non-commercial licenses, and point out the need for attention to responsible 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.

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

- 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)

2 BACKGROUND AND RELATED WORK

In this section, we present the motivations for this work by introducing the background and prior related studies.

2.1 Machine Learning Project Licensing

Typically, a ML project is constructed with three components: data, software, and models, which are usually governed by different licensing frameworks. To profile current ML licensing, Table. 1 provides licensing details for 19 ML projects with over 1,000 likes available in Huggingface[5] model repository.

Data Licensing in ML. Based on our profile, more than half (53%) of ML projects claim their data is licensed in a mixture manner. Additionally, 26% of projects use a single dataset with a standard data license like Creative Commons (CC) licenses. The data source of remaining projects (21%) is unknown. Obviously, legal compliance cannot be guaranteed when using data from unknown sources. However, there is also potential risk associated with using data under a mixture of licenses or a single license based on follow reasons:

First, the mixture of data sources may involve content under copyleft, non-public, and non-commercial licenses. We investigated the sources of mixture and found that only one dataset, the Pile [9], explicitly removed non-permissive content. Common sources of risk include Wikipedia, arXiv, WordPress, Common Crawl, etc (Refer to Table. 4 for more examples). For instance, sharing derivatives based on non-public licensed content raises suspicion of a license violation, and integrating copyleft content also poses a risk of license incompatibility conflicts. Furthermore, some content sources like IMDb explicitly prohibit data mining in their *Conditions of Use*.[6]

Secondly, the single data license assigned by data collectors may be invalid. In our profile, all datasets with a single license contain risky data sources. Rajbahadur *et al.* [32] investigated the sources of six public datasets and shown their inherent incompatibility for commercial use. A real case is the copyright infringement lawsuit filed by Getty Images Inc., alleging that Stability AI Ltd. misused Getty Images photos to train its Stable Diffusion [33] generative model (1:23-cv-00135). However, the claimed license of training dataset [36] used for Stable Diffusion is CC-BY-4.0, which is a permissive license allowing for commercial use. This highlights that ML data licensing is currently irregular and has become a significant factor in legal non-compliance. Although Benjamin et al. [2] have proposed the Montreal Data License (MDL) to foster fair use of data in AI activities, unfortunately, none of the ML projects adopted this license as shown in our profile.

Software Licensing in ML. Distinct from OSS projects, less than half (47%) of ML projects release their code with standard OSS licenses. About one-third of ML projects do not declare the code license (but have a model license), which is much higher than in OSS projects [6]. Other projects switch to using AI model or custom licenses to insert additional disclaimers and restrictions related to AI activities, thereby increasing the diversity of licenses in this context. However, given that ML, especially Neural Networks (NNs), is still in its emerging stages, the license dependency chain is shorter compared to OSS projects, and most of them use the latest versions of OSS licenses like Apache-2.0 and MIT.

Model Licensing in ML.

2.2 FOSS License Assessment

SPDX Automating the license compatibility process in open source software with SPDX

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 [4], it raises another question: *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 [2], but this proposal is rarely implemented in practice (life would be easier if it were widely accepted). Making things more complex, some projects release their models under free content licenses, like LayoutLMv3 model [16], 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

[5] https://huggingface.co/. Projects in same series but different versions are omitted.

[6] You may not use data mining, robots, screen scraping, or similar data gathering and extraction tools on this site ...

Table 1: Summary of licensing details for machine learning projects with over 1K likes on 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.

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 licensing language into these four categories based on their outcome forms.

We leverage Figure 1 to illustrate this idea. The left side consists of a list of AI activities, many of which pertain to model reusing methods, categorized based on the forms of their results. The middle part is our taxonomy that can classify these AI activities. Following this rule, we can also identify the corresponding terms in natural language license text shown on the right side. For example, Mixture of Experts (MoE) leverages a gating network to ensemble a batch of weak learners [17], which leads to a combination with strong separation and aligns with licensing terms like link, portion, collection, etc. Unlike combination, the results of amalgamation are difficult (or impossible) to separate, corresponding to AI activities such as modification, fine-tuning, model fusion, etc [7]. These unrecoverable revision of original works are corresponding license text like adapt, alter, remix, etc. Distillation and generation are derivatives of original works, which means the results will not contain any portion of the original works. These two AI activities are mostly defined in AI model licenses but are not covered by traditional OSS licenses and free content licenses.

By now, we can ascertain the suitable permissions, limitations, and responsibilities for each AI activity based on the language of the license, even when the license type is not an exact match. However, its necessarily to emphasize three points. First, our proposed method only applies in cases where ambiguities exist in the definition. If the conditions of certain AI activities are explicitly defined in the license, then we should directly follow that. Second, due to the various definitions adopted in different licenses, the bridging rules depend on each specific case and may differ from Figure 1. Lastly, one AI activity may trigger multiple license conditions. For example, a fine-tuned model can be seen as a combination with weak separation of the original model, while it can also be viewed as a derivative from fine-tuning data. Therefore, we should design a mechanism to trace these multi-source dependency structures in ML projects, which we will detail in the next section.

3.2 Structure of ML Projects

ML projects have unique dependency relationships compared to OSS projects, like the dependencies between generated content and generation model, as well as between training data and trained model. We can summarize these dependencies in ML projects into three categories:

- **Mix-works** be embeded in the new work, either verbatim or in part, in a tangible form. They usually result from direct copying of original components or reusing them through AI activities like combination and amalgamation.

[7] Whether embeddings constitute a combination with weak separation depends on the specific case. In ModelGo, we classify embeddings as amalgamation if they are created under a content license that treats translation as a form of modification.

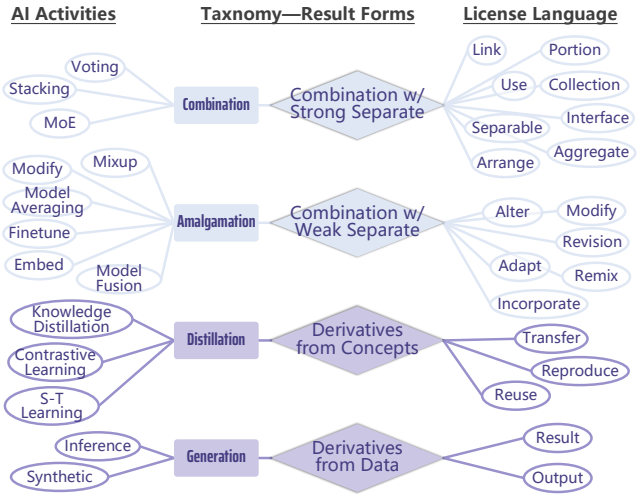

Figure 1: Our proposed taxonomy bridging AI activities and license terms based on their result forms.

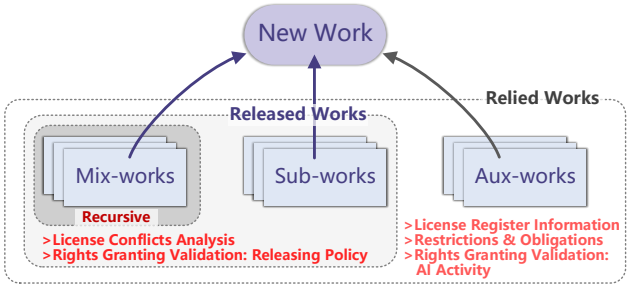

Figure 2: The proposed structure for capturing work dependencies in the context of ML projects with multiple reused components.

These components are embedded into ML projects and must be released with the new work. For example, if we release a new work utilizing Mixture of Experts (MoE), it is equivalent to releasing all weak learners.

- **Sub-works** are similar to mix-works, but the difference is that they are not embedded in the new work. For instance, if we manage to release MoE model along with the data used for training the gating network, then this data will be regarded as the sub-works of MoE model.
- **Aux-works** are components used to build the new work and are either included in it or released with it. For example, the original model used for knowledge distillation.

Figure 2 illustrates the structure of a work constructed by reusing multiple components in the context of ML projects. The final ML project may be constructed through iterative reuse of other works, resulting in a ternary dependencies tree for this project. The reason we need this specially-designed tree structure is that works with different dependency types have different license condition proliferation rules, which need to be handled separately during subsequent license analysis.

3.3 License Analysis in ML Projects

We have outlined all the necessary preparation steps for license analysis in previous sections. Their detailed implementations in ModelGo are as follows.

Preparation Step 1: Following our proposed taxonomy, we manually transcribed the terms in the license text to a standard machine-readable file in YAML format [8]. This file contain following informations for each license:

- Basic license descriptions, including its name, SPDX short ID, license version, license types (e.g., public domain, permissive, copyleft, proprietary), preferred work types (e.g., software, data, model), and supporting labels such as *disclose code required* or *auto-relicensing applied*.
- Rights granting information, including granted rights and reserved rights as defined by the license, along with the permitted reusing methods or result forms for redistribution. The prefix of such granting also be noted for cases where the granted rights can be revoked.
- Corresponding terms for each AI activity, which contain result forms and relicensability of the activity, corresponding restrictions, and obligations. This item will be marked as *No Defined* if both the activity and the result forms of this activity are not explicitly covered in the license text.

Preparation Step 2: To capture the dependency structure of works as shown in Figure 2, we encode the rules of dependencies construction for each AI activity. For example, if we generate embeddings of a corpus using an NN model, then the corpus is considered the sub-work of the generated embeddings, with the activity labeled as *embed*, and the NN model is categorized as the aux-work with the activity labeled as *use*. Furthermore, if the corpus is a collection of smaller corpora, then these smaller corpuses are categorized as the mix-works of the integrated corpus, with the activity labeled as *combine*. By recursively traversing this dependencies tree, we can gather all the dependent works and the activities used to build this ML project.

It is important to emphasize a concept in our license analysis approach called *activity proliferation*, which means that the activity performed by a work will recursively proliferate to all its mix-works. In the example of the corpus collection mentioned above, the *embed* calculation performed on the collection will be applied to all the smaller corpuses, triggering their license conditions related to *embed* as well. Similarly, as shown in Figure 2, all rights granting validation and license conflicts analysis of a work should be proliferated to all its mix-works. On the other hand, aux-works are not released with the project, so they are out of the scope of license conflict analysis and rights granting validation for release. In summary, mix-works, sub-works, and aux-works have different scopes in ML license analysis, which is why we need to distinguish between them.

Analysis Step: Given the license information and dependencies tree of ML projects, we are ready to analyze the license conflicts within it. ModelGo's license analysis consists of three phases:

Initial phase, where we register each component with clear license name, version, type, and format, and then construct their

[8] We attempted to use chatGPT to generate this content, but it often behaved unreliably in understanding our taxonomy and produced some stochastic answers [1].

Table 2: License warnings, errors, restrictions/obligations, and notices assessed by ModelGo in initial phase, license determination phase and license validation phase.

Warning, Error, Restriction, Notice	Description
Copyleft / Revocable / No Public Notice	This license or its granted rights are **copyleft / revocable / no public**.
License Type Mismatch Warning	License preferred work type is **not compatible** with this work type.
License Disclose Self Warning	License requires this work (in binary or SaaS format) to remain **open source** or provide a **readable copy** of the source code.
Rights Not Granted Warning	License of this work does **not explicitly grant** you the right to do (...)
Rights Not Granted Error	License of this work **cannot grant** you the right to do (...)
License Incompatibility Error	Work has a license conflict as it involves **multiple incompatible** licenses.
Cannot Relicense Error	Work has a license conflict as it required **relicense** rights not be granted.
Cannot Share Error	License **prohibits sharing** of this work.
State Changes Restriction	This work must **state changes** according to related license(s).
Include License Restriction	This work must retain the **original license file** according to the related license(s).
Include Notice Restriction	This work must retain all **notice files** (may contain copyright, patent, trademark and attribution) according to the related license(s).
Use Behavioral Restriction	This work must comply with the **use restriction** terms according to related license(s).
Runtime Restriction	This work must comply with the **rumtime restriction** terms according to related license(s).

dependencies using our predefined reusing functions. The release policy should be preset here, and we support personal use, sharing, and selling. Normally, few conditions apply when you only use the work personally, and most licenses limit behaviors like redistribution and commercial use.

License determination phase, where we iteratively derive the appropriate new licenses for intermediate reused results. Copyleft proliferation occurs when there is a triggered copyleft license in the relied components. An error will raise if there are other copyleft licenses or if there are components that cannot be relicensed. To condense our analysis results, we prioritize using *Unlicense* for intermediate results once they are relicenseable. After this phase, all components and their derivatives should have a well-determined license name.

License validation phase, where we validate the required rights for construct and release this project whether can be granted. The validation also includes compliance with disclosure requirements, such as when a components is in binary format but subject to conditions that require source code disclosure. The releaseability of the final result will be validated upon its mix-works and sub-works, and then an assessment report will be generated.

Table 2 presents the warnings, errors, restrictions, obligations, and notices that can be detected using ModelGo. Table 3 lists the licenses supported by ModelGo, which collectively cover over 98% of licensed models and datasets on Huggingface [9]. In the next section, we will present five case studies based on real ML components.

[9] Licenses without clear names and versions are excluded from the calculation. Worth mentioning, our coverage represents only 24.8% and 5.2% of the models and datasets on the entire site due to the significant number of works without license information.

Table 3: List of licenses supported by ModelGo, covering over 98% of licensed models and datasets on Huggingface.

OSS License (99.8%)	**Content License** (99.2%)	**AI Model License** (98.2%)
Apache-2.0, Unlicense, MIT, AFL-3.0, GPL-3.0, AGPL-3.0, LGPL-3.0, LGPL-2.1, BSD-3-Clause, BSD-3-Clause-Clear, BSD-2-Clause, Artistic-2.0, WTFPL-2.0, OSL-3.0, ECL-2.0	CC0-1.0, CC-BY-4.0, CC-BY-SA-4.0, CC-BY-NC-4.0, CC-BY-ND-4.0, CC-BY-NC-ND-4.0, CC-BY-NC-SA-4.0, PDDL, C-UDA, LGPL-LR, GFDL	OpenRAIL++, CreativeML-OpenRAIL-M, BigScience-BLOOM-RAIL-1.0, Llama2, OPT-175B, SEER

4 CASE STUDY DETAILS

An ideal practice of ModelGo is to assess real-world ML projects and detect their potential license compliance issues. However, this can be challenging in practice due to three present situations:

(1) Prevalent Licensing Disorganization in ML Projects: Many ML projects lack organized licensing information, making it difficult to ascertain the licenses of individual components.

(2) Lack of Development Lifecycle Information for ML Reusing: ML reusing often occurs without a clear record, making it hard to trace the origins and licenses of components used.

(3) Non-compliance within Datasets: Crowdsourced datasets often suffer from license non-compliance issues [32], making the licenses (usually permissive) declared by dataset collectors invalid.

Consequently, directly analyzing real-world ML projects may result in uncertainty, over-optimistic results, and often fail to detect any license conflicts. Therefore, to validate ModelGo, we have designed five ML scenarios rendered using 15 common data sources and 11 models that cover 5 modalities and 7 tasks, respectively. Table 4 shows the specifications of the involved data sources and models, whose licenses include copyleft, permissive, public domain, and no public license [10]. Furthermore, our case studies can cover all events listed in Table 2, and the their details and findings are provided in the following section.

It's worth noting that, as a license compliance analysis tool, ModelGo's goal is to report potential legal risks in ML projects related to licenses. It is not designed to address legal interpretation issues such as copyrightability of the final work, assessing copyright infringement, or establishing authorship, which typically require verification by a court of law in different regions [15, 24, 28].

4.1 CASE I : Corpus Combination

Our first case is corpus combination, which is very common in crowdsourced LLM datasets [9, 20, 29]. Additionally, we also consider scenarios where the corpus is extended with the help of translation LLM. As shown in Figure 3 (a), we first translate[11] *arXiv* and *Stack Exchange* using *Big Translate* model, then we combine these translated corpuses with *Deep-sequoia* and *FreeLaw*. This combined corpus is the final work, intended for commercial purposes. Figure 3 (b) depicts a variation in which the final work is a combination of translated corpus and the LLM. Note that, to simplify analysis, we treat these non-public licenses, such as CC-BY-ND-4.0 and CC-BY-NC-ND-4.0, as permissive licenses with limitations on sharing derivatives, as they do not include any copyleft terms. If not specified otherwise, the format of models and datasets is set to

[10] Some data sources contain crowdsourced content with multiple licenses, and we selected a non-public domain license among them. [11] In our cases, we treat translation as a specific form of embedding with a natural language output.

Table 4: Specifications of AI components used in case studies, which include Copyleft License, Permissive License, Public Domain Licens and Not-Public License.

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

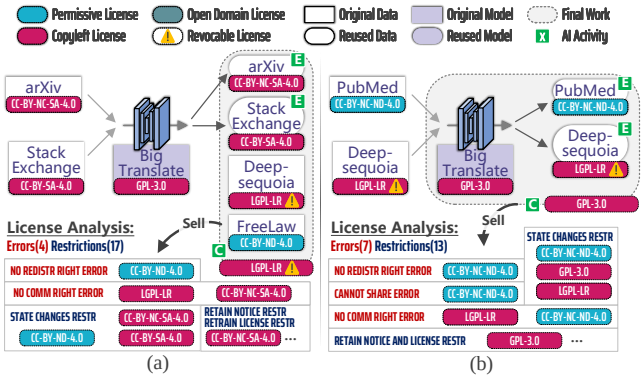

Figure 3: CASE I: Corpus Combination. (a) Example of copyleft proliferation rules ; (b) Example of LGPL-LR exemption and non-public license. AI Activities: E mbed, C ombine.

raw (i.e., modifiable), while the other supported formats are binary and SaaS. The interpretation of license analysis results is as follows:

Results of CASE I (a) The copyleft conditions about *translation* of the CC licenses were triggered, which means that the translated corpuses are also covered by the original licenses. As a result, the translated *arXiv* and *Stack Exchange* corpuses remain under the original copyleft CC ShareAlike licenses. However, combining these corpuses with another copyleft-licensed *Deep-sequoia* corpus did not result in the multiple copyleft licenses issue, as the combination with strong separate falls outside the proliferate coverage of LGPL-LR and CC ShareAlike licenses [4]. But, the proliferation extended to the final work and force it to be licensed under LGPL-LR as well. It is important to note that only the effort taken to combine the corpuses is under LGPL-LR, and the licensing action to the final work will not change the inherent licenses of its components.

There are two types of errors according to ModelGo's assessment. The first error arises from the CC-BY-NC-SA-4.0 license of the translated *arXiv*, which doesn't grant the right of commercial use[12]. The second error is caused by the fact that the redistribution rights of final work are not granted to comply with FreeLaw's CC-BY-ND-4.0 license. There are also many restrictions, such as the final work must state the changes compared to the original work and must retain the licenses and notice files of the original works. In addition, ModelGo also indicates that the granted rights of LGPL-LR are revocable, which poses a potential risk for further redistribution.

Results of CASE I (b) Different from CASE I (a), the final work in CASE I (b) is licensed under another copyleft license GPL-3.0 from *Big Translate*. This is because LGPL-LR has a license proliferation exemption for reused results that are no longer classified as linguistic resources. Consequently, the license of final work is proliferated by GPL-3.0. Additionally, besides the rights not granted error arising from CC-BY-NC-ND-4.0, this non-public license also explicitly prohibits any form of sharing derivatives, resulting in a cannot share error.

Summary To minimize the license violation risk when collecting ML data, avoid using content under non-public or non-commercial licenses, and be cautious about the proliferation scope of GPL-like licenses. Based on our assessment, using CC-licensed content (including CC ShareAlike) carries less risk.

4.2 CASE II : Mixture of Experts

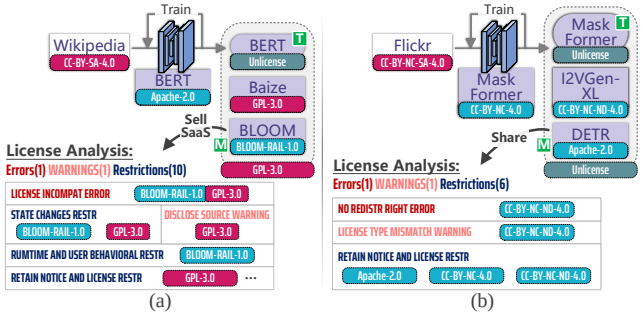

Figure 4: CASE II: Mixture of Experts. (a) Source code disclosure of GPLed work; (b) Sharing work under CC NoDerivs. AI Activities: T rain, M oE.

In this case study, we consider the MoE scenario, in which we combine two models with a newly trained model using a gating network. There are two variations in this case, each involving different models, training data, release policies (SaaS and sharing), as depicted in Figure 4 (a) and (b), respectively. A real-wrold counterpart could be Wu Dao 2.0, which is a LLM trained using MoE technology with input from tens of thousands of experts [14]. Additionally, releasing models as a service is commonly observed in commercial AI applications such as chatGPT and Midjourney.

[12] This error also arises from *Deep-sequoia* and *arXiv* (since it is a sub-work of the translated *arXiv*), we will omit this type of redundant in the rest of the case studies.

Results of CASE II (a) There is still significant legal uncertainty regarding whether CC-licensed works can be applied to AI training [4]. Since there is no explicit definition of AI training and corresponding restrictions for resulting models within the license text, we consider training as an undefined activity that falls outside the scope of CC agreements. Therefore, even though the copyleft CC-BY-SA-4.0 license is used for *Wikimedia*, the trained model *BERT* does not trigger the license proliferation conditions and can be relicensed to Unlicense. The final work's license is proliferated to GPL-3.0 from *Baize*, as in CASE I (b).

There is one error in the assessment: the copyleft-style user behavioral restriction claimed in BLOOM-RAIL-1.0 is consider as *non-permissive additional terms*, which can conflict with GPL-3.0. Therefore, an license incompatibility error is reported when we combine *Baize* and *BLOOM* using MoE. The warning is that the final work released as SaaS should remain open source or provide a readable copy of the source code to comply with GPL-3.0. Meanwhile, user behavioral restrictions also apply to the final work, as it is a derivative of *BLOOM* governed by responsible AI conditions [5].

Results of CASE II (b) In this case study, we replaced experts with CV models. The assessment reveals that the final work cannot be shared, whether modified or not, even for non-commercial purposes, if the project includes CC NoDerivs licenses, as these licenses do not grant redistribution rights to the licensee. This feature is helpful for licensors who intend to prohibit any derivation and commercialization of their models without the need to draft a custom proprietary license. However, this disorganization of ML projects' licensing has a negative effect on the entire ecosystem.

Summary Both OSS and CC licenses lack definitions and corresponding limitations related to model training, leaving freedom to use the trained results. However, responsible AI licenses provide comprehensive definitions for AI activities and copyleft-style restrictions, making their derivatives not GPL-compatible.

4.3 CASE III : Generation Pipeline

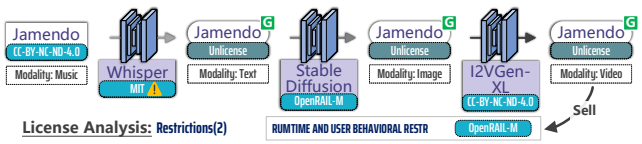

Figure 5: CASE III: Generation Pipeline. AI Activities: G eneration.

As shown in Table 1, artifact generation has become the most popular application of ML. In this case study, we leverages generative models to produce data for different modalities in a pipeline fashion. The final generated content is released for commercial use.

Results of CASE III There is still an ambiguity in traditional OSS licenses and free content licenses when it comes to the use of licensed materials for generating artifacts. From the perspective of the license agreement, this AI activity is permitted as long as the right to use is granted, and there are also no further claims for the generated content. However, there is one restriction from OpenRAIL-M. The AI model license clearly defines the conditions for AI activities and applies copyleft-style restrictions to its licensed work. Therefore, once AI model licensed components are used in ML projects, all subsequent work should comply with these user behavioral restrictions, which can potentially lead to the final work becoming closed source [11].

Summary Leveraging generative models can bypass the no-sharing conditions of CC NoDerivs licenses and making the generated content almost ungoverned. However, if responsible AI licensed works are involved, the content should comply with their restrictions, potentially leading to further GPL-compatibility issues.

4.4 CASE IV : Knowledge Transfer and Fusion

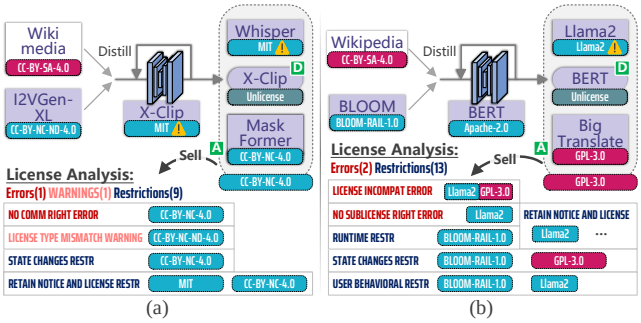

Figure 6: CASE IV: Knowledge Transfer and Fusion. AI Activities: D istillation, A malgamation.

The knowledge can be transferred or integrated from one model to another without the need for explicit code replication or linking. This is achieved through technologies such as Student-Teacher Learning [8], Contrastive Learning [22], Federated Learning [26], Model Fusion [21], etc. Traditional OSS licenses expose a loophole regarding these unique reusing methods from ML, and these methods also pose challenges for deep IP protection [30]. With the assistance of ModelGo, we further explore the compliance of these knowledge transfer methods within existing licensing framework.

Results of CASE IV (a) The distillation and amalgamation both yield a weak separation result from the original work, which can be interpreted as one form of modification. Therefore, the final work should be under a CC-BY-NC-4.0, the same as *Mask Former*. However, the CC licenses do not define the terms for the materials used for distillation, so there is no effect from the copyleft licenses of *Wikimedia* and *I2VGen-XL*.

There is one error in the assessment. Since the modification of a CC NonCommercial licensed work cannot be relicensed according to its conditions, the amalgamated result face a no commercial rights error when commercialized.

Results of CASE IV (b) This case study assess license compliance towards NLP models. There have two errors all detected from *Llama2*. The first error is the license incompatibility between its use limitations terms and the GPL-3.0. The second error is because the Llama2 license does not grant sublicense rights for further republication, conflicting with the releasing policy. Additionally, the rights granted by the Llama2 license are revocable, posing a potential risk in the final ML project. Furthermore, the final work should

also comply with the user behavioral restrictions demanded by BLOOM-RAIL-1.0 and Llama2.

Summary Knowledge transfer is a powerful method to bypass the reproduction prohibition of models. However, model fusion may trigger the terms like remix, incorporate, and adapt, necessitating the reusing procedures to remain in compliance. In addition, the rights may be revocable even if granted by a permissive license.

4.5 CASE V : Remix Data

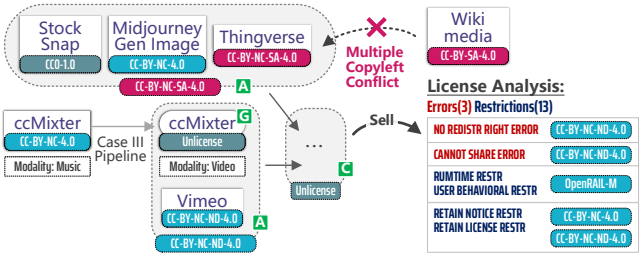

Figure 7: CASE V: Remix Data. AI Activities: G generation, A malgamation, C ombination.

Mirroring the CASE IV, this case considers the scenario of data remix and integration, which can arise when using data augmentation methods such as *mixup*[42], SMOTE[3], ADASYN [13], etc. We reuse the generation pipeline depicted in Figure 5 to increase the complexity of the assessment.

Results of CASE V We first analysis the remix of *StockSnap*, *Midjourney Gen Image* and *Thingverse*. For content under public domain licenses like CC0-1.0, we can freely remix this content without worrying about any conflicts. However, conflicts may arise when remixing content under CC-BY-NC-4.0 and CC-BY-NC-SA-4.0 licenses. As shown in Figure 6 (a), CC-BY-NC-4.0 cannot be relicensed for its remixed result, while CC-BY-NC-SA-4.0 requires performing license proliferation. But the outcome is this remixed work can be relicensed to CC-BY-NC-SA-4.0 because there is a one-way compatibility between CC licenses, as indicated by a supplementary interpretation from Creative Commons[13]. A conflict due to multiple copyleft licenses will arise if we attempt to further remix with *Wikimedia*. Furthermore, there will be a *cannot relicense* issue if we attempt to augment *Wikimedia* and relicense it to a new permissive license to bypass the mentioned conflict.

On the other hand, remixing the generated *ccMixter* and *Vimeo* is governed by CC-BY-NC-ND-4.0, which is responsible for almost all errors and restrictions in the final product. However, we can get rid of these constraints by leveraging the loophole of generative content as shown in CASE III.

Summary Directly remixing raw data should ensure compatibility between licenses, which can be challenging in crowdsourced scenarios where multiple copyleft-licensed content exists. One feasible solution is to exclude all content under copyleft and non-public licenses. Alternatively, an irregular tactic is to exploit the current ambiguity in licensing frameworks regarding generated content.

[13] https://wiki.creativecommons.org/wiki/Wiki/cc_license_compatibility

5 CONCLUSION

6 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.

REFERENCES

[1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency (FAccT)*. 610–623. https://doi.org/10.1145/3442188.3445922

[2] 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).

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research (JAIR)* 16, 1 (2002), 321–357. https://doi.org/10.1613/jair.953

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

[5] 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

[6] 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

[7] 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

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, 1607–1616.

[9] 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).

[10] 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).

[11] Eli Greenbaum. 2016. The Non-Discrimination Principle in Open Source Licensing. *Cardozo Law Review* 37, 4 (2016), 1297–1344.

[12] 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

[13] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IJCNN)*. IEEE, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[14] 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

[15] Samantha Fink Hedrick. 2019. I Think, Therefore I Create: Claiming Copyright in the Outputs of Algorithms. *New York University Journal of Intellectual Property & Entertainment Law (JIPEL)* 8, 2 (2019), 324–375.

[16] 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

[17] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

[18] 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

[19] 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

[20] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. 2023. The Stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research (TMLR)* (2023).

[21] Thanh Chi Lam, Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. 2021. Model Fusion for Personalized Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 5948–5958.

[22] Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10713–10722.

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

[24] Thomas Margoni. 2018. Artificial Intelligence, Machine learning and EU copyright law: Who owns AI? *Machine Learning and EU Copyright Law: Who Owns AI* (2018). https://doi.org/10.2139/ssrn.3299523

[25] 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

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 1273–1282.

[27] 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

[28] National Commission on New Technological Uses of Copyrighted Works (US). 1979. Final Report of the National Commission on New Technological Uses of Copyrighted Works, July 31, 1978. Library of Congress.

[29] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023).

[30] Sen Peng, Yufei Chen, Jie Xu, Zizhuo Chen, Cong Wang, and Xiaohua Jia. 2022. Intellectual property protection of DNN models. *World Wide Web* (2022), 1–35. https://doi.org/10.1007/s11280-022-01113-3

[31] 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.

[32] 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).

[33] 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

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

[35] 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).

[36] 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.

[37] 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

[38] 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).

[39] 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

[40] 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

[41] 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)*.

[42] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.