"They've Stolen My GPL-Licensed Model!": Toward Standardized and Transparent Model Licensing

Anonymous Author(s)

Abstract

As model parameter sizes reach the billion-level range and their training consumes zettaFLOPs of computation, components reuse and collaborative development are become increasingly prevalent in the Machine Learning (ML) community. These components, including models, software, and datasets, may originate from various sources and be published under different licenses, which govern the use and distribution of licensed materials and their derivatives. However, commonly chosen licenses, such as GPL and Apache, are software-specific and may not be clearly defined and bounded in the context of model publishing. Meanwhile, the reused components may also have free-content licenses and model licenses, which pose a potential risk of license conflicts and rights infringement within the model production workflow. In this paper, we propose addressing the above challenges along two lines: 1) For license analysis, we have developed a new vocabulary for ML workflow management and encoded license rules to enable ontological reasoning for analyzing rights granting and license conflicts. 2) For standardized model publishing, we have drafted a set of model licenses that provide flexible options to meet the diverse needs of model publishing. Our analysis tool is built on Turtle language and the Notation3 reasoning engine, serving as a first step toward Linked Open Model Production Data. We have also encoded our proposed model license set into rules and demonstrated the effects of GPL and other commonly used licenses in model publishing, along with the flexibility advantages of our proposed model licenses, through experiments.

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, Automated Reasoning

ACM Reference Format:

Anonymous Author(s). 2018. "They've Stolen My GPL-Licensed Model!": Toward Standardized and Transparent Model Licensing. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06

https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In recent years, the compelling generalization capabilities provided by billion-parameter models [26], along with the high computational and data costs associated with their training [14], have motivated ML project developers to collaborate incrementally rather than train models from scratch. For example, a common approach is to download a Pre-Trained Model (PTM) [10] and fine-tune it for downstream task [8]. However, these paradigms may face potential legal risks if the use and redistribution practices violate the governing licenses of the reused components, akin to the GPL violation issues in the field of Open Source Software (OSS) [15]. Another risk arises from the choice of license used to republish the work. Some developers adhere to traditional software publishing practices and select OSS licenses for their models [6, 17], which often lack clear definitions and conditions regarding ML activities and do not effectively prevent undesirable use. For example, a licensee can close-source your published models, even if they are licensed under GPL, without violating any terms.

There are three possible ways to address above challenges. First, developers could avoid using any third-party materials. However, this is extremely difficult for individual developers, as training PTMs is expensive and requires vast amounts of data. For instance, the training dataset for GPT-2 [21] was collected from 45 million web pages, governed by various licenses and terms of use. Second, a new licensing framework for ML projects could be developed, which might include drafting specific licenses for models and datasets [2, 5], along with a compatibility table to guide their reuse policies. However, this approach also has limitations, as it does little to address existing conflicts in ML projects that rely on components released under traditional licenses. Furthermore, it is impractical to expect all publishers to relicense their previous works. Third, we can scan the reused components in ML projects and analyze existing license compliance risks to eliminate them. This is a common solution applied to OSS projects [9] but it cannot be directly extended to ML projects. The reason is that ML components can involve complex coupling mechanisms and different licensing frameworks that are interwoven within a project.

Take MixLoRA [12] as an example: it is licensed under Apache-2.0 (an OSS license) and is fine-tuned on Llama2 [24] (governed by Llama2 Community License [16], a model license) using the Cleaned Alpaca Dataset [23], which licensed under CC BY-NC-4.0 (a free content license from Creative Commons [4]). Previous OSS license analysis tools [15, 18] that only consider package reference dependencies and focus on software licenses will fall short in such ML scenarios, which involve implicit nested dependencies and various licensing frameworks. Therefore, to provide license analysis for ML projects, the key is to develop an interpretative solution that can cover all licensing frameworks and disambiguate their mapping rules related to ML activities. Moreover, the lack of consensus in standard model publishing practices and the inflexibility of available model licenses have led many developers to publish their models under OSS licenses or even free-content licenses [3, 7], further increasing the complexity of designing license analysis methods.

In this paper, we propose a two-pronged approach to address these challenges. First, to resolve existing license conflicts in ML projects, we introduce *MG Analyzer*, a tool that constructs ML workflows as Resource Description Framework (RDF) graphs and assesses potential license compliance issues, improper license selection, rights grants, restrictions, and obligations within the projects. Second, to promote standardized model publishing in the future, we propose a new set of model-specific licenses, *MG Licenses*, offering Creative Commons-style licensing options for developers to choose from. To present potential risks of using traditional OSS, model, and free-content licenses in model publishing scenarios, we present XXXXXXXX use case. We also demonstrate the flexibility of *MG Licenses* in encompassing nearly all licensing conditions provided by other model licenses (by our analyzer?). The main contributions of our paper are:

- We identify the challenges of license compliance in ML projects and conduct a comprehensive analysis of the deficiencies in current model publishing practices.
- We developed MG Analyzer based on semantic technologies to enable automated reasoning for license analysis in ML projects. It incorporates a dedicated vocabulary capable of describing ML workflows with complex dependencies and the rules of OSS, data, and model licenses. We also provide a interface to convert user-input workflow descriptions into RDF graphs following this vocabulary. Based on these graphs, MG Analyzer constructs dependencies, performs reasoning, and detects potential license conflicts.
- We drafted a set of model licenses called MG Licenses to promote more standardized model publishing. These licenses are well-defined and cover a complete spectrum of model publishing scenarios. Furthermore, we have integrated support for MG Licenses within MG Analyzer.
- To the best of our knowledge, MG Analyzer and Licenses represent the first attempt at standardizing model publishing. Our code and license text are available at (link temporarily hidden due to double-blind policy).

The rest of the paper is organized as follows.

2 Background and Related Work

2.1 Model Licensing

2.2 License Compliance Analysis

3 MG Analyzer

This section aims to introduce the specific design of MG Analyzer by exploring three questions: *(i) How can we represent the workflows of ML projects? (ii) How do we establish a mapping from license text to reasonable rules? (iii) What types of license compliance issues can arise in ML projects, and how can we detect them?* Before delving into the detailed design that answers these questions, we first provide an overview that serves as a roadmap for this section. As illustrated in Figure 1, the process of MG Analyzer is divided into three main parts: Construction, Reasoning, and Analysis.

In the **Construction Stage**, user-input workflow descriptions (written in Python) are converted into an RDF graph (saved as *gen.ttl* in Turtle format) that contains the base information of the workflow. This conversion is achieved with the help of RDFLib [11] and the MG vocabulary, which is part of our analyzer. RDFLib provides an API for writing RDF graphs, while the MG vocabulary defines the specific semantics to represent the concepts and dependencies in ML projects. Then, we apply reasoning rules (written in Notation3 [1]) for the complete workflow construction using the EYE reasoner [25]. The reasoner concludes new properties that represent the input and output chains between components, enabling further reasoning about the *compositional dependencies* among them (reflecting Question (i); see Section 3.1 for details).

The main tasks in **Reasoning Stage** involve concluding the *definition dependencies* and *rights-using* dependencies. This is achieved through two substeps. First, a new property called *ruling* is created to record the definition of the output work in relation to the input work within the context of licensing. For example, if we merge GPL-licensed code into another software, the resulting work is considered a *derivative* of the original work. This relationship, which we refer to as *definition dependencies*, is crucial for determining the compliant license of the output work. We recursively identify such dependencies and ascertain the licenses of indeterminate works until all works in the workflow have a license. Based on the RDF workflow graph with complete license assignments, we can execute the second step of reasoning, termed *request*, which records *rights-using* dependencies that represent the rights required for the work according to practical reuse methods (reflecting Question (ii); see Section 3.2 for details).

So far, all necessary information for license analysis has been concluded before entering the final **Analysis Stage**. In this stage, MG Analyzer evaluates the correctness of the base workflow information, checks for the satisfaction of rights granting, and assesses license compliance and conflicts. *Reports* are genreated to present these results in the workflow graph (reflecting Question (iii); see Section 3.3 for details).

3.1 ML Workflow Representation

The representation of ML workflows, particularly when considering license analysis scenarios, differs significantly from common software workflows for the following three reasons. ① ML workflows often involve various components (e.g., code, datasets, images, model weights, services), each governed by licenses from different frameworks. Additionally, non-standard licensing practices are prevalent in current ML projects, for instance, the C4AI Command R+ model [3] is licensed under a free-content license: CC BY-NC-4.0. Therefore, the representation should be flexible enough to cover such situations.

② The component dependencies in ML workflows may be implicit and nested. For instance, Openjourney [20] is fine-tuned based on the StableDiffusion [22] model and the data generated by Midjourney [19]. In this case, knowledge from Midjourney is transferred to Openjourney without explicit compositional inclusion. Therefore, the representation should consider the multifarious dependencies present within ML projects.

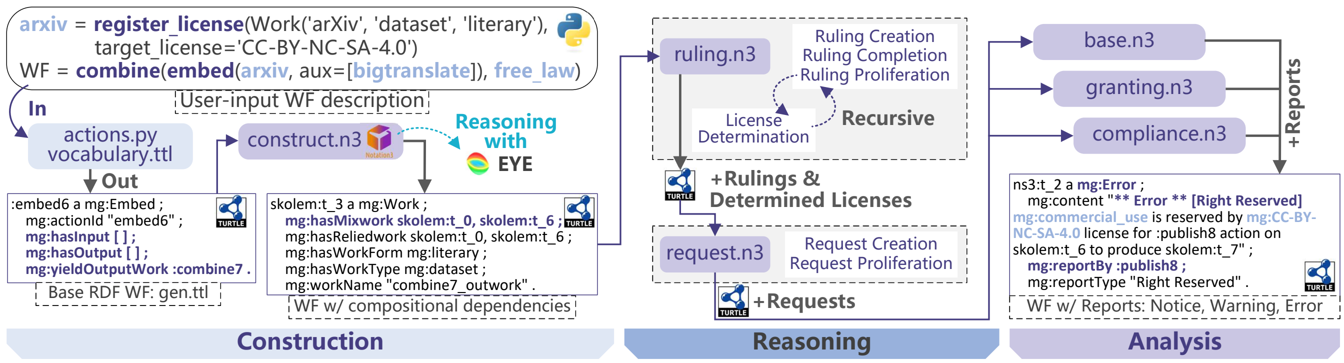

Figure 1: Overview of MG Analyzer ("mg" is the prefix of our proposed vocabulary).

③ The components' dependencies are also defined by the components' practical licenses and the ways they are reused. A common case in the OSS field is that republishing Software as a Service (SaaS) is considered to convey a *derivative* under AGPL-3.0 but has *no definition* under GPL-3.0. Therefore, terms like *derivative* and *independent* should be contextualized within specific licenses, and our representation should be capable of reflecting such meanings.

Therefore, we propose the MG Vocabulary to describe the properties and classes in an ML workflow. For the flexibility issue ①, we use the following terms to abstract key concepts in the workflow:

Work: Represents the components (e.g., models, datasets), each with a unique Type and Form. A work can have a license assigned through a Register License action, or its license can be determined through rules applied in the Reasoning Stage (see Figure 1).

Action: Represents operations performed on a Work, including Modify, Train and Combine, etc. In practice, we broaden the definition of these operations to make the vocabulary adaptable to different types of works. (See ref xxxxx for more details.)

Work Type: Includes software, dataset, model, and mixed-type. It is used to describe both the nature of the work and to identify the types of materials intended by a license. We use this information to detect any mismatch between the work type and the license.

Work Form: Divided into three subclasses: Raw, Binary, and Service to provide flexibility. For example, source code, model weights, and corpus fall under Raw; compiled programs are considered Binary; and services like SaaS or online chat LLMs are categorized as Service. Additionally, three general terms are offered: raw-form, binary-form, and service-form, which can work in conjunction with the work type. This approach helps represent situations that lack a formal ontology, such as "a dataset published as a service". We use mixed-form to represent the cases involving collections of works.

LicenseInfo: This contains the essential license information derived from conditions, including the license name, ID, intended types of works, whether it is copyleft or permissive, as well as granted and reserved rights, etc. While the license name is sufficient to describe the base workflow, to enable reasoning, LicenseInfo should bind rules, which we will discuss in the next section.

At this stage, we can describe a base ML workflow, as illustrated[1] schematically in Figure 2 (The RDF graph can be referred to in *gen.ttl* in Figure 1). In this base workflow, the derived input and output of each Action can be represented as blank nodes, serving as placeholders. These placeholders will be populated by reasoning the output yielded by the previous action and the input for the current action, respectively.

For dependency issues discussed in ② and ③, we identify three potential types of dependencies in an ML project: **compositional, definition, and rights-using**. These dependencies are visually represented by different colored dashed arrows in Figure 2. *Compositional dependencies* are categorized into four types: Mixwork, Subwork, Auxwork, and Provenance, each representing the containment relationships between input and output works. For example, when the output work includes the input work or a part of it, the input is considered the Mixwork of the output. This type of dependency is vital for license analysis because all actions performed on the output work, including any rights usage, will proliferate to the Mixwork components. For instance, when fine-tuning a MoE model, the fine-tuning operation will cascade to all submodels. Consequently, the license terms related to fine-tuning for each submodel are triggered, meaning that any constraints or rights imposed on the submodels must be honored. Additionally, if a work includes Mixworks in different forms, such as code and weights, we need to generalize the output's form to raw-form to accommodate these variations. A similar approach applies to the work type as well. Subwork and Auxwork are used to track works that are utilized by other works in the workflow, such as training datasets or distilled models. The key distinction is that Subwork is intended to be published alongside the output, which necessitates additional license analysis related to republishing. Provenance is specifically used for the Register License action to indicate that the output is simply the input itself, bound to a license. In such cases, any further proliferation of dependencies should cease.

The *definition dependencies* represent the relationships between works based on the definitions established by their licenses. For example, if a new work is created by modifying GPL-licensed code, that modification is considered a *derivative* of the original work. These dependencies should be understood in the context of the original license and can extend to subsequent actions if the same conditions are activated again. These dependencies are the main factor in determining the applicable license and restrictions for the output work. For example, under GPL-3.0, the republication of

[1] The "mg" prefix is omitted, and some properties are merged or filtered for presentation.

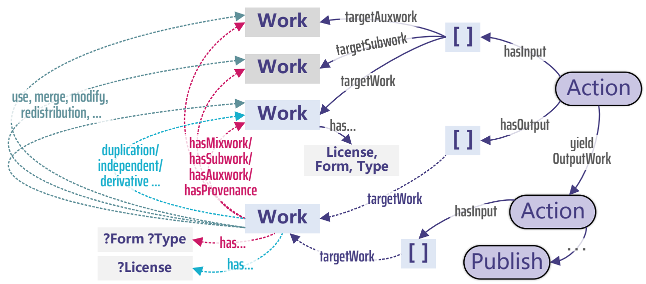

Figure 2: A Typical ML Workflow Represented by MG Analyzer. Dashed arrows with different colors indicate the properties related to three kinds of dependencies: compositional, definition, and rights-using dependencies.

derivatives must apply the same license. Additionally, a work may have multiple definition dependencies in a complex workflow, and these dependencies should be simultaneously satisfied (if possible; otherwise, an error should be reported) during license determination. The corresponding implementation in the MG Analyzer is illustrated in the Reasoning Stage of Figure 1. New instance nodes called *Ruling* are created to track the definition dependencies and triggered rules for each work, determining their applicable licenses in an alternating manner.

The *rights-using dependencies* describe the rights that must be granted for actions performed on works. For example, when executing a training action on a model, it requires the rights to *use* and *modify* (termed as Usage in MG vocabulary) from the model's license[2]. Similarly, the rights-using dependencies should proliferate according to compositional dependencies. For instance, the requirement to *modify* a model extends to all its submodels. In the MG Analyzer, we create new nodes called *Request* to represent this dependency. Additionally, it is insufficient to only check the granting rights; the reserved rights must also be verified. Depending on the clarity of the license text, some rights may either be explicitly granted or reserved. Furthermore, certain license clauses can waive the requirement for specific rights. For instance, both GPL-3.0 and CC licenses include automatic relicensing clauses for downstream recipients, which eliminate the need for a *sublicense* right.

By MG Vocabulary, we are able to describe complete ML workflows and represent the necessary dependencies for license analysis. The *compositional dependencies* are license-independent and can be reasoned from the base workflow. However, *definition dependencies* and *rights-using dependencies* are associated with specific rules expressed in natural language within each license. To facilitate automated reasoning for these dependencies, the next step is to develop a viable method for encoding license terms into formal logic rules.

3.2 License Rule Encoding and Reasoning

Typically, licenses are designed to govern the use and distribution of specific types of works. For example, the GPL-3.0 is tailored for source code and object code, while CC licenses focus on literary, musical, and artistic works. As a result, it is challenging to map their rules within a unified framework for logical reasoning. Meanwhile, many ML projects actually incorporate non-standard licensing components, as mentioned in Section 1. If we consider these claimed licenses to be invalid, then they would not pose any license compliance issues. However, the validity of these licenses depends on specific cases and the dispute resolution process by the jurisdictional courts in accordance with the applicable laws in different regions. As a license analyzer, we aim to maximize the detection of all potential legal risks under various interpretations, rather than merely granting a green light with low confidence. To this end, we perform three generalizations to encode license rules.

The first generalization is called *fuzz form matching*. We broaden the definitions related to a work's form to encompass its general form. For instance, we expand the license terms of GPL-3.0 concerning source code to include all forms of work in the Raw categories, such as model weights and corpus. In this way, we can extend the scope of interpretation of GPL-3.0 to cover models and datasets.

The second generalization is called *composition-based rule alignment*, which aims to resolve the pervasive ambiguities across different licensing frameworks. This ambiguity often arises in non-standard scenarios, for example, when licensing a model under GPL-3.0, it may be unclear whether *model aggregation* (a technology used in federated learning [13]) triggers the "Aggregate" clause in GPL-3.0. Therefore, we propose a composition-based method to align these rules. Specifically, we generalize the concept of action to represent the compositional relationships between input and output. For instance, the action *Combine* signifies that the input work has been entirely included in the output work without modification. This action corresponds to terms such as "Link" and "Aggregate" in software licenses, "Collection" in CC licenses, and "MoE" in model licenses. In the case of *model aggregation*, which produces an output that contains parts of the input works and is difficult to separate, it is not considered a *Combine*. Consequently, according to our rule alignment, it will not activate the "Aggregate" clause in GPL-3.0.

The complete rule alignment method utilized in the MG Analyzer can be found in Table 1. It is worth mentioning that the meanings of these actions have been broadened and may differ from their original definitions. In some cases, multiple actions may align with the same license terms. For instance, the license term "Modify" can align with both the actions *Modify* and *Amalgamate*, as such licenses do not distinguish the extent of changes made or whether those changes can be reverted.

The final aspect is *applicable term generalization*, where we encode the triggering conditions of a license term into the following properties: range of input work forms, range of output work forms, and types of actions. Figure 3 illustrates an example of the GPL-3.0 derivatives rule[3], which represents the Combine action applied to input and output works in code format. This action triggers the "derivative" clause in GPL-3.0, indicating that the license of the output work must be compatible with GPL-3.0 (e.g., APGL-3.0). Additionally, five restrictions apply to the output work if it is to

[2]An example of such rights granting can be found in the Llama2 Community License [16], which states, "You are granted a ... to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Materials.

[3]The original license text reads: "You may convey a work based on the Program ... in the form of source code ... provided that you also meet all of these conditions: ... stating that you modified it ... it is released under this License ... keep intact all notices ... license the entire work, as a whole, under this License ..." .

Table 1: List of Supported Actions in MG Analyzer Following Rule Alignment. The symbols = and ≈ denote that the Types/Forms of output work and input work are the same and may differ, respectively. The corresponding license terms in OSS, Free-content, and Model for each action are listed.

Action	Type	Form	Composition	Terms
Copy	=	=	Output and input are exactly **same**.	Copy Duplicate
Combine	≈	≈	**Entire** input included in output.	Link Aggregate MoE Arrange Collect
Modify	=	=	Output includes a **significant** portion of input and **can be reverted**.	Modify Fine-tune
Amalgamate	=	=	Output includes portions of input but **cannot be reverted**.	Modify Remix Fusion
Train	=	=	Output has the **same structure** as input and may contain a **negligible** portion of it.	Alter Adapt Train
Generate	≈	≈	Output does not contain any portion of input and may **perform differently** from it.	Output Generate Synthetic
Distill	=	=	Output does not contain any portion of input but **performs similarly** to it.	Distill Transfer Extract
Embed	=	=	Output does not contain any portion of input, but there has a **mapping** that converts input to output.	Translate Transform
Publish	=	≈	Output is the **same** as the input but may have a **different form**.	Redistribute Perform Display Disseminate

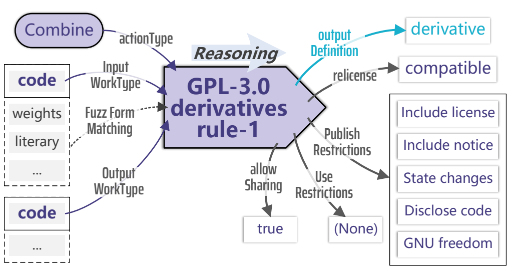

Figure 3: Example of a Generalized GPL-3.0 Derivatives Rule in MG Analyzer.

be republished, as dictated by this *definition dependency*. This rule does not include any *Use Restrictions*, which apply to output works regardless of whether they are republished. We found this subtle distinction to be crucial in ML license analysis, as most OSS license terms are triggered by distribution, and their definitions of "distribution" typically exclude publishing as a service. However, in the case of models, the common deployment method is through a web interface, leading to many OSS license restrictions being circumvented in such scenarios.

Furthermore, multiple rules may lead to the same output definition, and we also provide the option of *fuzz form matching* within the rules to enhance the interpretive capabilities of the MG Analyzer. It is worth mentioning that the reasoning behind the restrictions derived from the rules requires further analysis for validation. Taking the rule in Figure 3 as an example, whether the final work remains in a Raw form determines whether the warnings about disclosing code should be reported. We present snippets of encoded rules written in Turtle format in Appendix B. The list of supported licenses, whose terms have been encoded in MG Analyzer, is shown in Table 2, covering nearly all top-ranking licenses for published models on HuggingFace.

3.3 Compliance Analysis

4 MG Licenses

References

[1] Dörthe Arndt and Stephan Mennicke. 2023. Notation3 as an existential rule language. In *International Joint Conference on Rules and Reasoning (RuleML+RR)*. Springer, 70–85. https://doi.org/10.1007/978-3-031-45072-3_5

[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] CohereForAI. 2024. C4AI Command R+. Retrieved October 1, 2024 from https://huggingface.co/CohereForAI/c4ai-command-r-plus

[4] Creative Commons. 2024. Creative Commons Licenses List. Retrieved October 1, 2024 from https://creativecommons.org/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] 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

[7] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838* (2024).

[8] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

[9] Michael C Jaeger, Oliver Fendt, Robert Gobeille, Maximilian Huber, Johannes Najjar, Kate Stewart, Steffen Weber, and Andreas Wurl. 2017. The FOSSology project: 10 years of license scanning. *International Free and Open Source Software Law Review* 9 (2017), 9.

[10] 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

[11] Daniel Krech, Gunnar AAstrand Grimnes, Graham Higgins, Jörn Hees, Iwan Aucamp, Niklas Lindström, Natanael Arndt, Ashley Sommer, Edmond Chuc, Ivan Herman, Alex Nelson, Jamie McCusker, Tom Gillespie, Thomas Kluyver, Florian Ludwig, Pierre-Antoine Champin, Mark Watts, Urs Holzer, Ed Summers, Whit Morriss, Donny Winston, Drew Perttula, Filip Kovacevic, Remi Chateauneu, Harold Solbrig, Benjamin Cogrel, and Veyndan Stuart. 2023. *RDFLib*. https://doi.org/10.5281/zenodo.6845245

[12] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. MixLoRa: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159* (2024).

[13] 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.

[14] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. *The AI Index 2024 Annual Report*. Stanford University, Stanford, CA.

[15] 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

[16] Inc. Meta Platforms. 2024. Llama2 Community License. Retrieved October 1, 2024 from https://ai.meta.com/llama/license/

[17] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 1–18. https://doi.org/10.1007/978-3-031-19772-7_1

[18] 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

[19] Midjourney platform. 2024. Midjourney's Terms of Service. Retrieved October 1, 2024 from https://docs.midjourney.com/docs/terms-of-service

[20] PromptHero. 2024. Openjourney v4. Retrieved October 1, 2024 from https://www.openjourney.art/

Table 2: List of MG Analyzer-Supported Licenses (or Aggrements) with Their Comparisons in Clarity and Freedom. Grouped by OSS, Free-Content (&Dataset), Model and Sorted First by Clarity Score, Then by Freedom Score.

License Name	Clarity of Definitions				Freedom of Verbatim Copy			Freedom of Derivative					Freedom of Use		Clarity	Freedom
	Prefixes	Rights	Rules	Remote	Share	Close	Non-excl	Share	Close	Non-excl	Sublicense	Attribute	Comm	Behav		
AGPL-3.0	✓	✓	≈	✓	≈	✗	≈	≈	✗	≈	≈	✗	✓	✓	3.5	4.5
AFL-3.0	≈	✓	≈	✓	≈	✓	✓	≈	≈	✓	✓	✗	✓	✓	3.0	7.5
OSL-3.0	≈	✓	≈	✓	≈	✓	✗	≈	✗	✗	✓	✗	✓	✓	3.0	5
Apache-2.0	✓	✓	≈	✗	≈	✓	✓	≈	≈	✓	✓	✗	✓	✓	2.5	7.5
LGPL-3.0	✓	✓	≈	✗	≈	✗	✓	≈	✓	✓	≈	✓	✓	✓	2.5	7.5
Artistic-2.0	≈	≈	≈	✓	≈	✗	✓	≈	✗	✓	✓	✗	✓	✓	2.5	6.0
GPL-3.0	✓	✓	≈	✗	≈	✗	≈	≈	✗	≈	≈	✗	✓	✓	2.5	4.5
ECL-2.0	≈	✓	≈	✗	≈	✓	✓	≈	≈	✓	✓	✗	✓	✓	2.0	7.5
Unlicensed	≈	≈	≈	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1.5	10
MIT	≈	≈	≈	✗	≈	✓	✓	✓	✓	✓	✓	✗	✓	✓	1.5	8.5
GPL-2.0	✗	✓	≈	✗	≈	✗	✓	≈	✗	✓	≈	✗	✓	✓	1.5	5.5
LGPL-2.1	✗	✓	≈	✗	≈	✗	✓	≈	✗	✓	≈	✗	✓	✓	1.5	5.5
BSD-3-Clause	✗	≈	≈	✗	≈	✓	✓	≈	✓	✓	✓	✗	≈	✓	1.0	7.5
BSD-3-Clause-Clear	✗	≈	≈	✗	≈	✓	✓	≈	✓	✓	✓	✗	≈	✓	1.0	7.5
BSD-2-Clause	✗	≈	≈	✗	≈	✓	✓	≈	✓	✓	✓	✗	≈	✓	1.0	7.5
WTFPL-2.0	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0	10
CC0-1.0	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓*	✓	✓	✓	3.0	10
ODC-By-1.0	≈	≈	✓	✓	≈	✓	✓	≈	✓	✓	≈	✗	✓	✓	3.0	7.5
PDDL-1.0	✓	≈	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.5	10
CC-BY-4.0	✓	✓	≈	✗	≈	✓	≈	≈	≈	≈	≈	✗	✓	✓	2.5	6.0
CC-BY-SA-4.0	✓	✓	≈	✗	≈	✓	≈	≈	≈	≈	≈	✗	✓	✓	2.5	6.0
CC-BY-NC-4.0	✓	✓	≈	✗	≈	✓	≈	≈	≈	≈	≈	✗	✗	✓	2.5	5.0
CC-BY-NC-SA-4.0	✓	✓	≈	✗	≈	✓	≈	≈	≈	≈	≈	✗	✗	✓	2.5	5.0
CC-BY-ND-4.0	✓	✓	≈	✗	≈	✓	≈	✗	n/a**	n/a	n/a	n/a	✓	✓	2.5	4.0
CC-BY-NC-ND-4.0	✓	✓	≈	✗	≈	✓	≈	✗	n/a**	n/a	n/a	n/a	✗	✓	2.5	3.0
GFDL	✓	≈	≈	✗	≈	✗	✗	≈	✗	✗	≈	✗	✓	✓	2.0	3.5
C-UDA	✓	✗	≈	✗	✓	✓	✓	≈	✓	✓	✗	✗	✗	✗	1.5	5.5
LGPLLR	✗	✓	≈	✗	≈	✓	✗	≈	≈	✗	≈	✗	≈	✓	1.5	4.5
MG0	✓	✓	✓	✓	≈	✓	✓	≈	✓	✓	✓	≈	✓	✓	4.0	8.5
MG-BY	✓	✓	✓	✓	≈	✓	✓	≈	≈	✓	✓	✗	✓	✓	4.0	7.5
MG-BY-NC	✓	✓	✓	✓	≈	✓	✓	≈	✓	✓	≈	✗	✗	✓	4.0	6.5
MG-BY-RAI	✓	✓	✓	✓	≈	✓	✓	≈	≈	✓	✓	✗	✓	✗	4.0	6.5
‡ OpenRAIL-M	✓	✓	✓	✓	≈	✓	✓	≈	≈	✓	✓	✗	✓	✗	4.0	6.5
MG-BY-OS	✓	✓	✓	✓	≈	✗	✓	≈	✗	✓	✓	✗	✓	✓	4.0	6.0
MG-BY-NC-RAI	✓	✓	✓	✓	≈	✓	✓	≈	✓	✓	≈	✗	✗	✗	4.0	5.5
MG-BY-NC-OS	✓	✓	✓	✓	≈	✗	✓	≈	✗	✓	≈	✗	✗	✓	4.0	4.5
MG-BY-ND	✓	✓	✓	✓	≈	✓	✓	✗	n/a	n/a	n/a	n/a	✓	✓	4.0	4.5
MG-BY-NC-ND	✓	✓	✓	✓	≈	✓	✓	✗	n/a	n/a	n/a	n/a	✗	✓	4.0	3.5
† OPT-175B	✓	≈	✓	✓	≈	✓	✓	≈	✓	✓	✗	✗	✗	✗	3.5	5.0
† Llama3	≈	✗	✓	✓	≈	✓	✓	≈	✓	✓	✗	✗	≈	✗	2.5	5.5
† Llama3.1	≈	✗	✓	✓	≈	✓	✓	≈	✓	✓	✗	✗	≈	✗	2.5	5.5
† • AI2-ImpACT-LR	≈	✗	✓	✓	≈	✓	✓	≈	✓	✓	✗	✗	≈	✗	2.5	5.5
† • AI2-ImpACT-MR	≈	✗	✓	✓	✗	n/a	n/a	≈	✓	✓	✗	✗	✗	✗	2.5	2.0
† • AI2-ImpACT-HR	≈	✗	✓	✓	✗	n/a	n/a	✗	n/a	n/a	n/a	n/a	✗	✗	2.5	0
† Gemma	✗	✗	✓	✓	≈	✓	✓	≈	≈	✓	✗	✗	≈	✗	2.0	4.5
† Llama2	≈	✗	✓	✗	≈	✓	✓	≈	✓	✓	✗	✗	≈	✗	1.5	5.5

Header Definitions:
Prefixes: ✓ The license explicitly includes sufficient prefixes that clearly describe scope and conditions of granting rights (e.g., revocable, sublicensable); ≈ Some important prefixes are indeterminate; ✗ No prefixes are declared.
Rights: ✓ The license explicitly declares whether a patent license or a copyright license is granted; ≈ Only the granting of a patent license or copyright license is stated; ✗ No explicit grant of either is provided.
Rules: ✓ The license terms cover all ML activities listed in (TBD); ≈ Some ML activities fall outside the definition of this license; ✗ Almost no rules are set forth.
Remote: ✓ The license considers remote access situations (e.g., via API, Web, SaaS); ✗ No definitions or rules regarding remote access behaviors are set forth.
Share: ✓ The license permits the sharing of verbatim copies/derivatives created by you without any restrictions; ≈ Some restrictions apply to sharing; ✗ Sharing verbatim copies/derivatives is prohibited.
Close: ✓ The license does not require you to disclose the source files of verbatim copies/derivatives created by you; ≈ Modification statements are required; ✗ You must disclose the source files of your created copies/derivatives.
Non-exclusive: ✓ The license does not restrict you from adding new terms when republishing; ≈ Certain types of terms are prohibited in republishing; ✗ All republishing must adhere to the original terms and conditions.
Sublicense: ✓ The license explicitly grants sublicensing rights; ≈ The license prohibits sublicensing but offers automatic licensing instead; ✗ Sublicensing is either prohibited or not explicitly permitted.
Attribute: ✓ The license does not require retaining the original attribution and licenses in redistributed derivatives; ≈ Attribution or license must be retained; ✗ Redistributed derivatives must retain the attributions and licenses.
Commercial: ✓ The license explicitly grants commercial rights; ≈ Commercial rights are not explicitly granted but not reserved either, or compromised commercial rights are granted; ✗ Commercial rights are reserved.
Behavioral: ✓ The license does not restrict user behaviors; ≈ Includes runtime controls (e.g., forced updates); ✗ Certain behaviors involving the licensed materials or derivatives are prohibited (e.g., harming, medical advice).
Clarity/Freedom Score: ✓ +1.0, ≈ +0.5, ✗ +0, n/a: +0 . Maximum Clarity Score: 4.0, Maximum Freedom Score: 10.
Explanations:
* Although CC0-1.0 explicitly states that sublicensing is not allowed, sublicensing becomes unnecessary due to the Waiver of Rights.
** Since CC-BY-ND-4.0 and CC-BY-NC-ND-4.0 prohibit the sharing of derivatives, judgments regarding redistributed derivatives are marked as "n/a" in the table.
† These licenses (or terms of use, or agreements) are specifically drafted for certain products and are not intended for general model publishing purposes.
‡ As there are no fundamental differences between CreativeML Open RAIL-M, OpenRAIL++-M, BigCode Open RAIL-M, BigScience RAIL, and BigScience Open RAIL-M, these licenses are grouped under OpenRAIL-M.
• We have used an archive of the AI2 ImpaACT license; the version is 2.0, with an effective date of January 8, 2024.

[21] 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.

[22] 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

[23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html* 3, 6 (2023), 7.

[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] Ruben Verborgh and Jos De Roo. 2015. Drawing conclusions from linked data on the web: The EYE reasoner. *IEEE Software* 32, 3 (2015), 23–27.

[26] Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. 2023. Inverse Scaling Can Become U-Shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 15580–15591. https://doi.org/10.18653/v1/2023.emnlp-main.963

A Disclaimers

The information in this article is for general informational purposes only and does not constitute legal advice. Views, opinions, and recommendations expressed are solely those of the author(s) and do not represent any organization. Do not rely on this material as a substitute for professional legal advice tailored to your specific circumstances.

B Encoded Rules

Here is a code snippet:

MGLicenseRule.ttl > GPL-3.0-derivative-rule-1

```
@prefix mg: <http://~/rdf/terms#> .
mg:GPL−3.0−derivative−rule−1 a mg:Rule ;
  mg:hasOutputDef mg:derivative ;
  mg:targetActionType mg:Combine ;
  mg:targetInputWorkForm mg:code ;
  mg:targetOutputWorkForm mg:code ;
  mg:relicense mg:compatible−license ;
  mg:hasPublishRestriction mg:include_license_restriction,
    mg:include_notice_restriction, mg:disclose_self_restriction,
    mg:state_changes_restriction, mg:gnu_freedom_restriction;
  mg:allowSharing true .
```

MGLicenseRule.ttl > Llama2-derivative-rule

```
@prefix mg: <http://~/rdf/terms#> .
mg:Llama2−derivative−rule a mg:Rule ;
  mg:hasOutputDef mg:derivative ;
  mg:targetActionType mg:Amalgamate, mg:Combine,
    mg:Modify, mg:Train, mg:Embed, mg:Distill ;
  mg:targetInputWorkForm mg:weights, mg:exe ;
  mg:targetOutputWorkForm mg:weights, mg:exe ;
  mg:relicense mg:any−license ;
  mg:hasPublishRestriction mg:include_license_restriction ;
  mg:hasUseRestriction mg:llama2_exclusive_use_restriction,
    mg:use_behavior_restriction ;
  mg:allowSharing true .
```