ALEA-D0: Minimally Encumbered Pretraining and SFT Dataset

Jillian Bommarito, Michael J Bommarito II, Daniel Martin Katz

Institute for the Advancement of Legal and Ethical AI (ALEA)

October 1, 2024

Abstract

Practically all large language models have been pretrained on data that is subject to global uncertainty related to copyright, breach of contract, and privacy. In light of this reality, we present the KL3M dataset, one of the largest collections of unencumbered pretraining and supervised fine-tuning data available for large language models. This paper outlines our approach to creating a dataset free from copyright uncertainty, details our data sources, and describes our collection and preprocessing methods. The KL3M dataset represents a significant step towards developing AI systems that are legally and ethically sound, both now and in the face of future regulatory changes.

1 Introduction

The training and use of large language models has generated substantial controversy. While fewer and fewer question the technical progress of these models' capabilities, more and more are questioning the legal and ethical implications of the technology, especially as it relates to training data. Practically all existing models are anchored in a foundation of "fair use" or "fair dealing" related to "publicly-available" data. This critical concept, loosely supported by precedent and limited statute in a small number of jurisdictions, is not guaranteed to survive - nor is it universally accepted by many within society that it should.

With over 30 copyright lawsuits against AI companies currently in progress in the US alone, the need for legal precedent on the matter of copyright infringement and model training is clear. The outcome of these cases will likely establish whether the argument for "fair use" in model training is a viable one. However, regardless of what one believes about the legal and ethical questions underlying this uncertainty, there is no denying the existence of the many lawsuits and investigations ongoing in major jurisdictions.

In light of this reality, we set out on an alternative research agenda - one that is rooted in legal and ethical practices that are well-accepted and free of

doubt. Our primary contribution to a space already saturated with datasets is the research into and development of a path that is free of reliance on the "fair use" argument that is often relied upon for model training.

In this paper, we present the KL3M dataset, tokenizer, software, and APIs. These assets, open-sourced and maintained by our 501(c)(3), represent one of the largest collections of unencumbered pretraining and supervised fine-tuning data available. We established and are publishing the framework for determining permissibility of content usage that can be employed by anyone who wants to gather or audit training data for model training or fine-tuning. In addition, we enhance the data provenance visibility of the KL3M dataset by providing Dublin Core metadata for all data in the dataset.

The continued development and use of AI systems is predicated on these systems being legal, both in the current environment and in the wake of future regulatory and legislative changes. We believe that the development of datasets and models that are transparent, freely available without legal restrictions, and high quality will enable downstream use that is free of the infringement concerns often present.

It's worth noting that "More than 40 countries with over one-third of the world's population have fair use or fair dealing provisions in their copyright laws." However, these provisions are not universally accepted and their interpretation can vary significantly across jurisdictions. Our approach aims to circumvent the need for reliance on fair use arguments by ensuring all data in the KL3M dataset is either in the public domain or explicitly licensed for unrestricted use.

In the following sections, we will detail our data sources, collection process, preprocessing methods, and the characteristics of the resulting dataset. We will also discuss the implications of our work for future AI development and the broader legal and ethical landscape of machine learning.

2 Data Sources

3 Collection Process

3.1 Permissibility of Use

As stated in the introduction, the novel nature of this dataset is the fact that it does not rely on fair use as a basis for establishing permissibility; instead we rely on a multi-part test to determine whether a given data source may be used without restriction. The test is based on a series of conditional assessments: if the data passes a test, we include it in the KL3M dataset; if the data does not pass the test, we move on to the next test. Data that does not pass any of our four tests is not included in the KL3M dataset.

3.1.1 Test 1 - Free from Copyright Protection

Our first test is whether the content is free from copyright **at the time of its creation**. Works of the United States government, for example, are not eligible for copyright protection under 17 USC § 105 ("Copyright protection under this title is not available for any work of the United States Government"). Content that meets this test is eligible for inclusion in the KL3M dataset.

3.1.2 Test 2 - Public Domain

The second test is whether content has been entered into the public domain or an equivalent, such as a CC0 license where no rights are reserved. Content that has entered the public domain as a result of the lapse of copyright protection falls into this category.

3.1.3 Test 3 - Right to Copy, Modify, and Redistribute

If content has not passed the prior two tests, the final test is whether the license grants the right to copy, modify, and redistribute the content without restriction. Licenses that meet this test include CC BY and the United Kingdom's Open Government License (OGL v3.0). If content failed this final test, we did not include it in the KL3M dataset.

3.1.4 Excluded Licenses

We excluded the following license types based on the reasoning below:

- CC BY-SA: excluded due to the uncertainty around meeting share-alike obligations with an LLM
- CC BY-NC: excluded due to non-commercial limitation
- CC BY-NC-SA: excluded due to non-commercial limitation and uncertainty around meeting share-alike obligations
- CC BY-ND: excluded due to limitation on derivative work
- CC BY-NC-ND: excluded due to non-commercial and derivative work limitations

3.2 Technical Description

3.3 Personal Information Considerations

Personal information instances within the KL3M dataset generally arise from their inclusion in documents that are a matter of the public record or that are works of the government. As a result, the personal information that could theoretically be obtained through a review of the KL3M dataset is already publicly available.

We follow the approach of CourtListener and the FreeLaw Project's Board of Directors in managing the tension between privacy and the public interest. Documents are only removed from the database under explicit court order.

3.4 Regulatory Compliance Considerations

4 Pre-Processing

5 Dataset Characteristics

5.1 Jurisdictional Coverage

We chose to focus on content related to US, UK, and EU law due to our familiarity with these jurisdictions and the legal intricacies of intellectual property rights. We recognize that this limited coverage does not address much of the world's population and laws, but we hope that the process and tests that we have outlined in this paper will enable others to create similar datasets for additional jurisdictions.

6 Conclusion

The KL3M dataset represents a significant step towards creating large language model training data that is free from copyright uncertainty. By carefully selecting and vetting our sources, we have created a resource that can be used with confidence by AI researchers and developers. We hope that our methodology will serve as a template for future efforts in this direction, ultimately leading to more legally and ethically robust AI systems.

As the field of AI continues to evolve rapidly, it is crucial that we address the legal and ethical challenges head-on. The KL3M dataset is our contribution to this effort, providing a foundation for the development of AI systems that can withstand future regulatory scrutiny and contribute positively to society.

While our focus has been on US, UK, and EU jurisdictions due to our familiarity with their legal systems, we recognize the need for similar efforts in other parts of the world. We encourage researchers and legal experts from other jurisdictions to adapt and expand upon our methodology to create globally representative datasets that are free from copyright concerns.

In conclusion, the KL3M dataset not only provides valuable training data for large language models but also sets a new standard for transparency and legal compliance in AI development. As we move forward, it is our hope that this work will inspire further innovation in the responsible and ethical advancement of AI technology.