The KL3M Dataset: A Copyright Clean 80TB+ Dataset for Pretraining and Supervised Fine Tuning of Large Language Models

Michael J Bommarito II,[1,2] Jillian Bommarito[1] & Daniel Martin Katz[1,2,3,4]

[1] *Institute for the Advancement of Legal and Ethical AI (ALEA Institute)*
[2] *CodeX – The Stanford Center for Legal Informatics*
[3] *The Law Lab, Illinois Tech - Chicago Kent College of Law*
[4] *Center for Legal Technology & Data Science, Bucerius Law School*

March 13, 2025

Abstract

Practically all large language models have been pre-trained on data that is subject to global uncertainty related to copyright infringement, breach of contract, and privacy. In light of this reality, we present the 80TB+ sized Kelvin Legal Large Language Model (KL3M) dataset, the largest collections of pretraining and supervised fine-tuning data which is unencumbered by copyright risk. In total, KL3M contains over 125 million documents and more than 1.7 trillion tokens. This paper outlines our approach to creating a dataset free from copyright uncertainty including details on our data sources, and collection and preprocessing methods. While our data sources are limited to US and certain EU source materials, we open source not only the dataset itself but also the tokenizer, API's and other associated software in order to support further allied development in other jurisdictions. We believe that this represents a significant step towards developing A.I. systems that are legally and ethically sound, both now and in the face of future regulatory changes.

1 Introduction

Over the past decade, there has been significant progress on general-purpose language modeling driven by the application of neural based methods **??** to corpora of increasing larger scales.**???** Leading large language models (LLMs) have displayed significant progress on a range of challenging real-world tasks.**????**

While fewer and fewer question the technical progress of these models' capabilities, the training and use of large language models, however, is not without controversy. Indeed, there are an emerging range of questions associated with the development of generative A.I. technology. These objections vary and cover a range of topics including the 'openness' of the model creation process and the 'toxicity' of the subsequent outputs.**???** While these are certainly important issues, far less attention, by contrast, has been paid to use training data collected at scale without respect to the moral and legal rights of its creators.

Virtually all existing LLMs rely upon the large-scale collection and use of materials that are subject to copyright. Despite various clever efforts to ameliorate such issues at both training **?** and inference time **???** many leading models can still engage in relatively high levels of potentially infringing behavior.**?**

There is still the open question of whether notions “fair use" or “fair dealing" might serve as a legal defense to this otherwise problematic behavior.?? However, it is worth noting that while “[M]ore than forty countries with over one-third of the world’s population have fair use or fair dealing provisions in their copyright laws,"? the interpretation of such principles can vary significantly across jurisdictions.[1]

In light of this reality, we set out on an alternative path - one that is rooted in a rigorous data collection and curation processes designed to respect traditional legal and ethical frameworks. In this paper, we present the Kelvin Legal Large Language Model (KL3M) dataset, tokenizer, software, and APIs. These assets, open-sourced and maintained by the ALEA Institute,[2] represent one of the largest collections of pretraining and supervised fine-tuning data unencumbered by copyright risk. We outline a 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.?

2 Copyright and Generative A.I.

2.1 Brief Overview of Copyright

Copyright is sometimes described as a ‘bargain’ where creators receive, exclusive rights to their works for a specific period of time, in exchange for the public eventually gaining free access to those works after the copyright term expires.? Although all materials will eventually reach the public domain, creators may, during their period of exclusive ownership, choose to make their works available to others via licensing at a scope of their choosing. There is a wide continuum of popular copyright licenses including MIT, Apache, GNU GPL, AGPL as well as various flavors of Creative Commons licenses.? Such licenses provide a wide degree of latitude for creators to control the scope of how their respective works are used.

In the internet era, many individuals and organizations have chosen to make their otherwise copyrighted works available online for the scope delimited use of others. Through various platforms such as *Github* (computer code), *Getty Images* (image licensing), *YouTube* (video repositories) or directly through personal websites, the internet has arguably facilitate the most extensive period of information sharing in all of human history. At the same time, there have also been disputes surrounding the extent to which digitized information could be made available to users. Indeed, controversy has been ‘part and parcel’ of the internet era including major clashes over projects such as *Google Books*? and file sharing sites such as *Napster*.?

2.2 Copyright and A.I. Data Collection

Although disagreement over proper the scope of copyright is not new, the advent of LLMs has brought with it heightened concerns about the acquisition and use of otherwise copyrighted source materials.? Virtually all model providers and large scale datasets collected in furtherance of building LLMs have ignored both the website terms of service in the scraping of websites as well as licensing restrictions attached to the respective content collected.? As a result, the internet has seen a rise in restrictions upon sharing including a range of efforts to prevent data from being used in the training of A.I. systems.? Individual creators and content based organizations threatened by A.I. systems that arguably undermine creators future economic prospects have

[1] It is worth noting that although some model providers are offering “fair use” as a defense to their data collection practices, many such organizations are also inherently acknowledging the property rights of creators by entering into licensing deals.

[2] *See* Institute for the Advancement of Legal and Ethical AI (ALEA Institute) `https://aleainstitute.ai/`

begun to place additional restrictions on their works in an effort to limit reuse, redistribution, and commercial use of their copyrighted materials in the building of A.I systems.**?**

Since the advent of the large-scale public internet, there have been a variety of public and commercial efforts to track its development and growth.**??** Much of those efforts centered around "search" and helping route individuals to relevant webpages. However, indexing the internet is not the precisely akin to full collection of content. For example, in the early years of the current millennia, linguists were slow to include large-scale web content given anxiety over copyright in the underlying source materials.**?** While some of such materials did eventually make their way into important corpora,**?** the specter of legal restrictions caused some to limit their use of materials obtained from the internet.

Other groups, however, were far less motivated by such concerns and began to look at the internet as a premier source of data. Beyond mere tracking and indexing, internet data has been subjected to various large scale collection efforts including graph data, images, metadata and the underlying text.**???** *Common Crawl***?** one of longest standing efforts to collect web-scale data has served as a foundational dataset in many early LLMs. *Common Crawl* and subsequent efforts to build large-scale A.I. training datasets such as the *Colossal Cleaned Common Crawl* (C4)**?**, *The Pile***?** and *Dolma***?** are replete with copyrighted data.

As noted earlier, the collection and distribution of such materials relies upon "fair use" as a justification. "Fair use" is fact-specific inquiry meaning that whether a particular use of copyrighted material is considered "fair use" depends upon specific details that must be evaluated on a case-by-case basis. For example, when an academic institution or other non-profit type research organization collects data for research purposes that activity would likely be "fair use." Yet, if that same dataset were deployed for subsequent commercial use by an entity whose direct or indirect aim is to undercut the commercial viability of the original creator (*e.g.* coder, artist, author or musician) that activity might not be characterized as "fair use." Even if "fair use" were not deemed to cover a particular usage, it is possible that royalty system including perhaps compulsory licensing might be a vehicle for rewarding creators[3] while still allowing for innovation in A.I. model building to continue.[4]

Certain scholars working with datasets such as *Common Crawl*, *C4* and *The Pile* have recognized the looming copyright questions surrounding these efforts.**??** For example, authors of the recently released *Dolma* dataset stated "that the legal landscape of A.I. is changing rapidly, especially as it pertains to use of copyrighted materials for training models."**?** However, they still chose to distribute their dataset because the "sources were publicly available and already being usedin large-scale language model pretraining (both open and closed)."

This perspective is emblematic of much of the broader literature on ethics in A.I. where there has been much greater focus on questions of model 'openness' and 'A.I. alignment' than respect for the scope of the moral and legal rights of creators. Although issues of transparency and model toxicity are important, they are far from the only consideration worthy of attention.

Most recently, the *Common Corpus* dataset**?** was released on the *Hugging Face* platform. The dataset

[3]A market based licensing and royalty system including perhaps a compulsory licensing would be more ethical than allowing individuals and organizations to seize the creative works of others without *any* compensation. Such ideas are explored in a recent report released by the U.S. Copyright Office.**?**

[4]In a letter sent to the White House Office of Science and Technology (OSTP), OpenAI argued that "[A]pplying the fair use doctrine to AI is not only a matter of American competitiveness — it's a matter of national security ... If the PRC's developers have unfettered access to data and American companies are left without fair use access, the race for AI is effectively over."**?** While clarity regarding the legal treatment of this question would be helpful, it is far from clear that the requirement of royalty payments to creators would materially impair the rate of innovation.

was promising as the authors claimed that the compilation "contains only data that either is uncopyrighted or permissively licensed." Unfortunately, the rhetoric surrounding the dataset does not match its reality. Although the authors recognize the ethical and potential legal issues associated with scraping data without the consent of the data creator, the authors provide very little description of their copyright audit process. It turns out that even a cursory inspection of the dataset reveals a significant volume of copyright materials contained therein.

3 Copyright Clean Collection Process

We believe that the continued development and use of AI systems should be predicated on such systems being both legally and ethically compliant, both in the current environment and in the wake of future regulatory and legislative changes. In this section, we introduce the Kelvin Legal Large Language Model (KL3M) dataset, the largest dataset developed to date that respects the moral and legal rights of copyright holders. 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.

3.1 Copyright Filtration Process

To build KL3M, we developed a multi-part filtration process designed to determine whether a given data source may be used without significant 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 tests is not included in the KL3M dataset. We describe the process below but also highlight the steps in Figure 1 below.

3.1.1 Test 1 - Free from Copyright Protection

Our first test is to determine whether the content is free from copyright **at the time of its creation**. A substantial percentage of content contained within the KL3M dataset meets this test and was thus eligible for inclusion.

Works of the United States government, for example, are not eligible for copyright protection under 17 U.S.C. § 105 ("Copyright protection under this title is not available for any work of the United States Government"). Specifically, work created by a federal government employee or officer is in the public domain, provided that the work was created in that person's official capacity.

A separate but related concept is the "government edict doctrine." This doctrine, developed through the common law, denies copyright protection to official "government edicts."[5] The doctrine exists to effectuate the principle that citizens must have unrestrained access to the laws that govern them. The "government edict doctrine" allows for the publication of various legislative and judicial pronouncements including judicial opinions. As noted by the U.S. Copyright Office, the doctrine extends to "all legislative enactments, judicial decisions, administrative rulings, public ordinances, or similar types of official legal materials."[6]

Most but not all countries follow this doctrine. For example, we had hoped to include certain UK legal sources within the KL3M dataset but limitations embedded in the Open Justice License (OJL) surrounding the

[5]The doctrine is long standing and dates back to *Wheaton v. Peters*, 33 U.S. (Pet. 8) 591 (1834) and has most recently been addressed in *Georgia v. Public.Resource.Org, Inc.*, 590 U.S. 255 (2020).

[6]U.S. Copyright Office, Compendium of U.S. Copyright Office Practices, §313.6(C)(2) (3d ed. 2017)

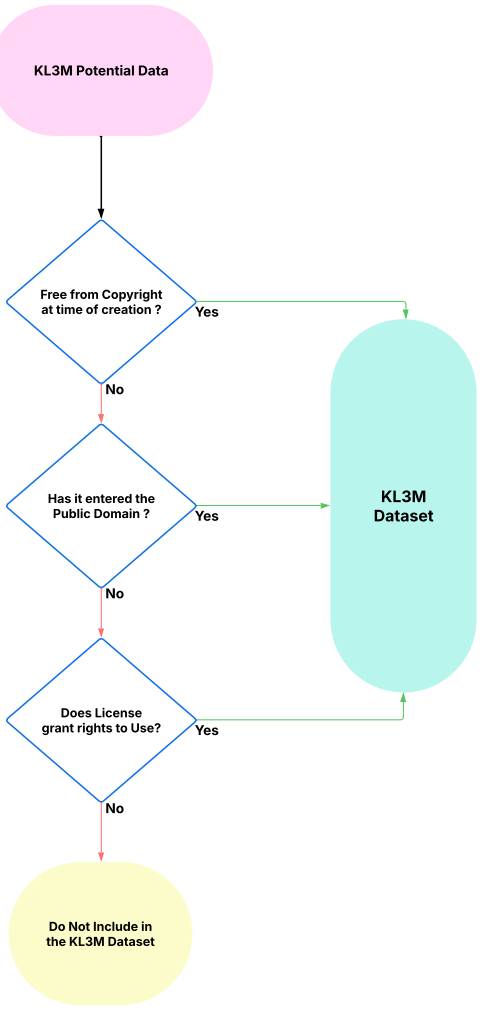

Figure 1: Overview of the Copyright Filtration Process

"computational analysis of the information" prevented us from including such UK data at this time.[7]

3.1.2 Test 2 - Public Domain Materials

In our second test, we determine whether content has **entered into the public domain** or its equivalent, such as a Creative Commons - No Rights Reserved (CC0) license where no rights are reserved. There are several ways that content, once subject to copyright, could thereafter enter the public domain. Most notably, although the amount of time has changed over the years, copyright is always a temporary grant of exclusive rights for a defined period of time. Therefore, once that designed time has lapsed, the work automatically enters the

[7] *See* Open Justice License (OJL) `https://caselaw.nationalarchives.gov.uk/open-justice-licence`

public domain. For example, while we would prefer to include the most recent Twelfth Edition**?** published in 2024, we instead include in the KL3M data the Second Edition of Black's Law Dictionary originally published in 1910.[8]

Other information can enter the public domain as part of a particular legal process. For example, we include granted patents in KL3M as patents are typically not subject to copyright restrictions. As noted by the USPTO, "[P]atents are published as part of the terms of granting the patent to the inventor." Absent a limited set of circumstances, "the text and drawings of a patent are typically not subject to copyright restrictions."[9]

One additionally important vehicle for adding information to the public domain is the US Federal Depository Library Program (FDLP). The Federal Depository Library Program (44 U.S.C. § 19), administered by the U.S. Government Publishing Office, was established to ensure that the American public has access to Government information. 44 U.S.C. § 1911 states that "[D]epository libraries shall make Government publications available for the free use of the general public." Although many of the documents required extensive pre-processing in order to be usable, the FDLP is a very large source of useful information for KL3M.

3.1.3 Test 3 - Minimally Encumbered Content with Clear Rights to Copy, Modify, and Redistribute

If the use of content is not otherwise permissible following the two prior tests, the final test is to determine whether the **license attached to the content grants a user the right to copy, modify, and redistribute the content without significant restriction**. As highlighted in Figure 1, if content fails this final test, we did not include it in the KL3M dataset.

While CC0 or No Rights Reserved is, of course, the most unencumbered form of license, there are a range of other content licenses that might theoretically be considered for inclusion. Overall, we evaluated content with the following license types using the reasoning below:

- CC0: included given content is shared with No Rights Reserved
- CC BY: included as the attribution is not overly burdensome (particularly in the case of an institution as author)
- CC BY-SA: excluded due to the uncertainty around meeting share-alike obligations in the context of a Large Language Model[10]
- 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

[8] Black's Law Dictionary contains definitions of specialized legal terms. The substantive meaning of many such terms has not changed for many years.

[9] *See* USPTO Terms of Service `https://www.uspto.gov/terms-use-uspto-websites`

[10] While we plan to meet the Share-alike requirements in the full Open Source release of KL3M, we recognize that others might consider this to be major encumbrance upon utilization.

In general, in the context of content created by an institution, we believe an attribution requirement is only a very modest restriction upon downstream use. For example, content released by the European Commission under 2011/833/EU is only minimally encumbered as it merely imposes an "obligation for the reuser to acknowledge the source of the [Commission's] documents."[11] Similarly, in most cases, compliance with the attribution requirement set forth in the Creative Commons License (CC BY) is similarly straightforward.

However, there are arguably some nuances and complexity in the context of non-institutional authors. Indeed, some of the most fruitful data that we might consider for inclusion could not meet our goal of building a dataset that was relatively unencumbered. For example, despite its inclusion in other training datasets such as *Colossal Cleaned Common Crawl* (C4) **?**, *The Pile* **?**, *Dolma* **?** and *Common Corpus* **?**, *Wikipedia* and many other knowledge commons are arguably encumbered by the "copyleft" licenses that the community has chosen.

From a historical perspective, *Wikipedia* content was originally licensed under the GNU Free Documentation License (GFDL).**?** Today, it is arguably licensed under CC BY SA which carries with it not only a share alike (SA) requirement but also an attribution requirement (BY). We contacted the *Wikimedia Foundation* to ascertain their perspective regarding the scope of attribution that they believe would be required to use the content on *Wikipedia*. Specifically, given the millions of total contributors who at some point have authored content on *Wikipedia*, we wanted to determine whether they believe that a general attribution statement or a specific attribution statement was required.

In the context of building or fine-tuning large language models, a general attribution statement highlighting the respective input sources to a given dataset or model is relatively easy to provide. However, specific attribution to the specific work or works that gave rise to a *specific model output* is a difficult if not impossible technical challenge. It was the Wikimedia Foundation's position that "providing a general notice to customers would not be an adequate solution to compliance." While Wikimedia's interpretation of the CC BY SA requirement is not the final word on this important legal question, we did feel comfortable including this content given it would significantly encumber downstream usage.[12]

In addition to the Wikimedia Foundation's interpretation of the (BY) requirement, *Wikipedia* has a share alike (SA) requirement which like other "copyleft" licenses carries with it the requirement that downstream users make any new works they create with the original content available on the same terms as the original content.**??** Specifically, the human-readable summary of the *Wikipedia Creative Commons Attribution-ShareAlike 4.0 International License* states "if you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar or a compatible license."[13] Given these restrictions, it is unclear how *any* model creator could use *Wikipedia* data without making their model fully available under similar CC BY SA / GFDL terms.[14]

[11] *See* On the reuse of Commission documents `https://eur-lex.europa.eu/eli/dec/2011/833/oj/eng`

[12] It is not clear how *any* model creator could comply with a *specific attribution* requirement given current technical limitations. At best, one could construct a system to assign *statistical attribution* by assigning attribution through some sort of potential *n-gram* based matching or other higher-order statistical style inference. However, from an attribution perspective, this would undoubtedly produce false positives as well as false negatives.

[13] *See* Wikipedia Creative Commons Attribution-ShareAlike 4.0 International License `https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_4.0_International_License`

[14] Relying solely on the "fair use," virtually *all* model providers have used *Wikipedia* data in constructing their models. To our knowledge, however, none of them have followed the attribution requirement (BY) as interpreted by the *Wikimedia Foundation* and most do not come close to complying with *Share Alike* (SA) requirement.

3.2 Fairly Trained Certification

In 2024, the KL3M dataset was audited by the independent non-profit *Fairly Trained*.[15] *Fairly Trained* is a non-profit certification and auditing organization supported by many creators which is devoted to the certification of models that uphold the highest standards of respect for copyright.

In certifying a model or dataset as *Fairly Trained*, we were required to provide the provenance of each source contained within the dataset including a detailed review of our KL3M Copyright filtration process. In addition, we needed to demonstrate any models or libraries we leveraged in the data collection, curation, processing steps were also not built upon the unauthorized use of third party content. After an extensive audit process, the KL3M dataset was the first large language model dataset ever to be certified as *Fairly Trained*.

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 Board of Directors of *The Free Law Project*.[16] in managing the tension between privacy and the public interest. Documents are only removed from the database under explicit court order.

3.4 Current Jurisdictional Coverage of KL3M

We chose to focus on content regulated by US 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 languages, but we hope that the process and tests that we have outlined in this paper as well as the associated infrastructure we have open sourced will enable others to create similar datasets for additional jurisdictions.

4 Overview of KL3M Collection Process

In developing the Kelvin Legal Large Langugage Model (KL3M), we sought to collect and curate a large body of information which was free from the specter of copyright infringement. As opposed to broad data collected from the internet writ large, we sought to identify sources of high quality information, free from copyright concerns that were reasonably available online. Governmental data, thus currently represents the vast majority of data within the Kelvin Legal Large Language Model (KL3M) dataset. As an output of the growth of the modern regulatory state,**??** various components of government produce very large bodies of relatively high-quality information content. In addition, various legal and regulatory processes require individuals and organizations to draft and submit various forms of information to government officials, often with guidance and oversight by lawyers, accountants and other professionals.

While the substantive quality of many forms of governmental data is quite high, the organization and accessibility of such information is not always nearly as strong. Although governmental and other related data has slowly made it way into the digital world, the mere digitization of governmental information or

[15] *See* Fairly Trained `https://www.fairlytrained.org/`

[16] *See* CourtListener `https://www.courtlistener.com/` and the Free Law Project `https://free.law/`

work product has not always allowed the public to obtain the crucial information they need. Poorly designed websites, a lack user-friendly navigation and outdated and inconsistent digital platforms are just some of the challenges faced by individuals engaging with public sector information systems. We faced these very challenges in collecting and curating the KL3M dataset. Although much of this data is theoretically accessible, it is often stored in various inconsistent formats which make cross-functional access very difficult even in discrete amounts (let alone at scale).

In this section, we begin by discussing trends in modeling building while providing some exemplars of the wide-ranging content contained within the KL3M dataset. Next, we detail our efforts to collect, pre-process and organize this vast and diverse corpora. Finally, we provide a high level overview of the distribution of sources contained within the current version of the Kelvin Legal Large Language Model (KL3M) dataset.

4.1 Breadth and Quality of Tokens Might Be What You Need

One challenge in building modern language models is to have both scale and a diversity of pre-training information content to cover the broad conceptual space of potential user queries. Users might want to ask a model to explain a scientific question, determine how best to cook a particular recipe, look to draft some long-form prose or perhaps even ask questions about how to sublease their apartment. The prevailing approach to cover the broad conceptual space of potential user queries is simply to scale models to increasing large scales.**?** "Chinchilla" and other related scaling laws have encouraged model builders to pre-train on the maximal amount of available pretraining data.**??**. As such, most developments in LLMs have been focused on the use of various collections of internet and so called "publicly available" data to build larger and larger language models.

Undoubtably, the sheer power of scale has delivered some fairly remarkable results. Engineering, however, is not only about performance, it is also about cost.**?** Thus, an alternative strain of work has focused upon how to cost effectively train models using distillation and other related techniques.**???** The scaling laws upon which the field was once fixated have arguably given way to a more nuanced perspective where token diversity, token quality and test time inference scaling are also an important part of the overall calculus.**??**

4.2 The Incredible Expanse of Government Work Product

Legal and regulatory processes implicate a wide variety of pursuits and fields of human endeavour. Thus, taken as a whole, the work product of governments and governmental processes covers a significant amount of intellectual territory. While certainly not covering every topic that a user might find interesting, information from governments covers a surprising amount of the broader conceptual space. Laws, regulations, scientific reports, food safety bulletins, environmental impact assessments, public health guidelines, statistical reports, press releases, disaster preparedness plans, government contracts and certain private contracts, judicial opinions, public commentary, military directives, food and drug recalls, transportation safety reports, economic forecasts, patent filings, congressional testimony, securities filings, census data reports are just some examples of the outputs associated with the governmental work product and governmental processes.

KL3M features a wide variety of question and answer pairs, professional dialogues, formal definitions of key terminology and parallel assessment documents in multiple languages. It is quite difficult to fully characterize the incredible expanse of topics and materials contained therein. Appendix I highlights the KL3M Data Gallery, an online exploration tool which allows users to review millions of sample documents drawn from the broader KL3M dataset. However, consider Figure 2 which offers just a few selected examples that are

exemplars of the broader set of data contained in KL3M. Across the four examples, we observe a wide range of topics from turducken food handling, micronutrient testing in vitamins and carotenoids, mineral resources of the Owyhee River Canyon in Idaho and an analysis of the change in input impedance for electrically short dipole antennas. Agencies represented such as the United States Department of Agriculture, NIST, Department of Commerce and the Department of Interior are just a small subset of the total agencies producing government work product on a daily basis.

Figure 2: Samples of Content from the KL3M Dataset

4.3 The Collection & Pre-Processing Data Pipeline for KL3M

Working with the vast array of document types such as those displayed in Figure 2 is a challenging proposition. To do so effectively at scale, we developed a pre-processing pipeline designed to engage with the real-life challenges associated with the unstructured, inconsistent and complex forms of documents across the various information systems with which we engaged. Thus, in addition to the KL3M dataset, we are also releasing all of the tooling and libraries leveraged across the KL3M Pre-Processing pipeline as it is our hope that others

might adapt this work in order to build parallel corpora in other jurisdictions.

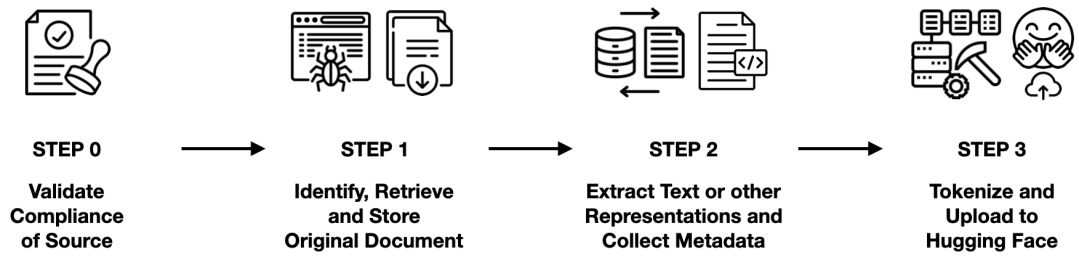

Figure 3: Overview of the Pre-Processing Pipeline - (icons via Flaticon)

Figure 3 provides a high-level overview of our pre-processing pipeline. Building from our copyright filtration process described in Figure 1, we begin by identifying a potential source of useful content. Next, we must determine whether that candidate source complies with our requirements. Thus, *Step 0* is high level recitation of the process described in Figure 1. Having then identified and validated compliance of the source material, we next proceed to *Step 1* of Figure 3 where we retrieve and retain the original source material for provenance purposes.[17]

In *Step 2*, we develop an alternative representation of the documents which varies depending upon the nature of the original source. Our ideal path is to build an alternative representation of all source materials in Markdown **?** while also retaining the original source material in parallel. However, this is not always possible given the nature of the original source. Finally, in *Step 2* were also collect and store *Dublin Core Metadata* for all source material.

In *Step 3*, we tokenize all objects using a custom tokenizer developed specifically for this task. The KL3M tokenizer has several specific elements that makes it unique **[MIKE ADD HERE 2-3 sentences]** Finally, we upload the tokenized document to their respective folder on *Hugging Face*.[18]

5 Dataset Characteristics

5.1 KL3M Components and Summary Statistics

While currently limited mostly to governmental sources, the KL3M dataset features a relatively large and diverse set of content such as the content highlighted in Figure 2. In Table 1, we present summary statistics for the overall dataset. In total, the KL3M dataset features more than 125 million documents, 1.7 trillion tokens and more than 80 terabyte of total information content.

Given the range of sources as the vast interconnectedness of ideas, concepts, there is almost certainly duplicate content contained herein. A document may quote elements of other documents or otherwise incorporate

[17]This ability to demonstrate original source provenance was critical in obtaining the *Fairly Trained* Certification described in Section 3.2.

[18]The KL3M Data can be access here `https://huggingface.co/collections/alea-institute/kl3m-data-679f9db9b6fd93f91c3c633e`

KL3M Features	
Total Documents in KL3M	126 Million Documents
Total Tokens in KL3M	1.7 Trillion Tokens
Total File Size	80 Terabytes

Table 1: Summary Statistics for KL3M Dataset

concepts from other documents. Thus, the overall token count is somewhat larger than if one were to consider something such as the total number unique n-gram combinations. However, we did not deduplicate the underlying content as we envision a wide range of potential uses for this overall source material. Downstream users can thus decide how to mix, match, segment or further pre-process the content in order to support their specific objective or use case.

KL3M Component	File Size	Document Count	Token Count	Avg Token Per Doc
Securities & Exchange Commission Filings	00.0GiB	0.0	0.0	0.0
Congressional Documents	00.0GiB	0.0	0.0	0.0
Congressional Bills	00.0GiB	0.0	0.0	0.0
Code of Federal Regulations	00.0GiB	0.0	0.0	0.0
Electronic Code of Federal Regulations	00.0GiB	0.0	0.0	0.0
Federal Depository Library Program	00.0GiB	0.0	0.0	0.0
Federal Register	00.0GiB	0.0	0.0	0.0
Federal Judicial Center	00.0GiB	0.0	0.0	0.0
CIA World Factbook	00.0GiB	0.0	0.0	0.0
Congressional Research Service	00.0GiB	0.0	0.0	0.0
United States Government Manual	00.0GiB	0.0	0.0	0.0
Library of Congress - Country Profiles	00.0GiB	0.0	0.0	0.0
Statutes at Large	00.0GiB	0.0	0.0	0.0
Regulatory Submissions	00.0GiB	0.0	0.0	0.0
United States Code	00.0GiB	0.0	0.0	0.0
Court Documents - Opinions	00.0GiB	0.0	0.0	0.0
Court Documents - Motions, Orders, etc.	00.0GiB	0.0	0.0	0.0
Court Documents - Dockets	00.0GiB	0.0	0.0	0.0
Black's Law Dictionary, 2nd Edition	00.0GiB	0.0	0.0	0.0
U.S. Federal Government Websites	00.0GiB	0.0	0.0	0.0
US Patent Grant Full Text Data	00.0GiB	0.0	0.0	0.0
Official Journal of the European Union	00.0GiB	0.0	0.0	0.0
Totals	**0.0**	**0.0**	**0.0**	**0.0**

Table 2: Summary Statistics of KL3M Components

Table 1 displays the document counts, token counts and average tokens per document for each of the KL3M components. While *Securities & Exchange Commission Filings* are the largest component of the dataset from a total documents and token perspective, there are some other large sources of input data. Other large subset of data include [FILL IN NEXT 3-4 items]

Among some of the smaller KL3M components, there are interesting elements worthy of exploration including millions of conversational messages extracted from Congressional hearings, nearly 20 billion tokens worth of docket entiries

Appendix II offers a more detailed description of each of the KL3M Components.

5.2 KL3M Additional Compositional Statistics

DISTRIBUTION OF DOCUMENT SIZE

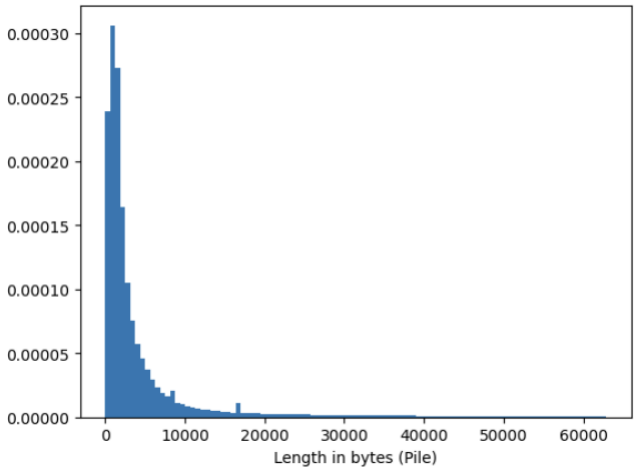

Figure 4: Overview of the Pre-Processing Pipeline - (icons via Flaticon)

DISTRIBUTION OF PERPLEXITY SCORES

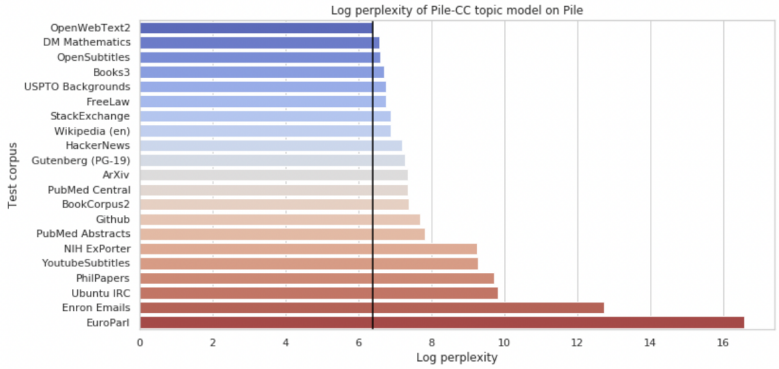

Figure 7: Log perplexity of 16-topic LDA trained on Pile-CC, on other Pile components. Dotted line indicates log perplexity of the topic model on OpenWebText2. Higher indicates a larger topical divergence from Pile-CC.

Figure 5: Overview of the Pre-Processing Pipeline - (icons via Flaticon)

TOXICITY ANALYSIS

Trying to copy some of the stuff reported in THE PILE paper

TOTAL TOKENS

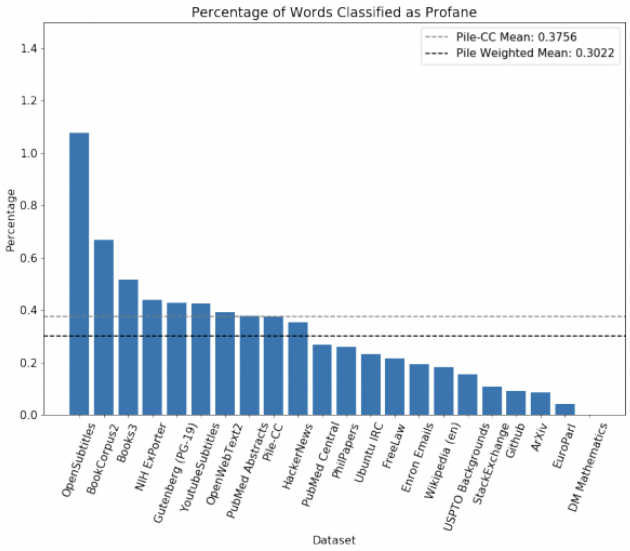

Figure 8: Percentage of words classified as profane in the Pile. The percentage of the CC component and the weighted mean of the Pile as a whole are shown as horizontal lines.

Figure 6: Overview of the Pre-Processing Pipeline - (icons via Flaticon)

TOTAL DOCS (with cavaet and discussion that doc is a tricky idea)

DISTRIBUTION OF TOKENS by Docs

SOME SORT OF TOKEN DIVERSITY MEASURE

Why US other countries – data availability ..

6 A Living Dataset and Infrastructure for the Collecting and Distributing Copyright Clean Data

In this paper, we introduced the Kelvin Legal Large Language Model (KL3M) dataset a large and diverse corpora of more than 125 million documents and more than 1.7 trillion tokens. We describe our copyright filtration process designed to identify only source materials with clear provenance from a copyright perspective. We then provided an overview of the pre-processing pipeline designed to

The KL3M can be used in several ways. First, it can serve as a baseline for model pretraining and could be combined with other appropriately license datasets. Alternatively, it could be used to fine tune an existing model. We realize that this dataset alone would likely be insufficent to allow for models to be built which cover the boundless set of possible use cases and user queries. However, we believe this large body of tokens could be combined with selected forms of licensed content to develop LLMs which can deliver strong performance on certain tasks. However, 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.

The paper reflects the current version of the KL3M dataset as of the time of this publication. Yet, rather than being a static snapshot, we hope that KL3M will persist as "living dataset" which we seek to update, maintain and extend as time moves forward. In addition, we hope that KL3M will become a federated project where others leverage or retrofit some of our underlying infrastructure to expand the set of copyright clean data available from a global perspective.

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 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.

Appendices

A Caselaw Access Project (CAP)

This appendix provides details about the Caselaw Access Project (CAP) dataset, including its collection methodology, data statistics, and examples of the data. The Caselaw Access Project is a collaborative effort led by Harvard Law School's Library Innovation Lab to digitize and freely share all U.S. case law.

A.1 Dataset Overview

The CAP dataset contains court opinions from U.S. state and federal courts, providing a comprehensive repository of legal precedents. The corpus includes published opinions from all federal courts and state appellate courts, covering nearly 6.92 million cases from the 1700s to the present day. This represents one of the most comprehensive collections of historical and contemporary U.S. case law available for research and analysis.

A.2 Data Processing Statistics

Based on the counts files in the KL3M project, the CAP dataset has been processed through multiple stages:

Processing Stage	Number of Documents
Documents (Initial Collection)	6,919,296
Representations (Processed)	6,919,272
Parquet (Final Format)	6,919,272

Table 3: CAP Dataset Document Counts by Processing Stage

As shown in Table 3, nearly all collected documents (over 99.99%) were successfully processed through each stage of the pipeline. The minimal difference between the initial collection count and the final processed count (24 documents) indicates the robustness of the processing pipeline for this dataset.

A.3 Collection Methodology

The CAP dataset was collected from the Harvard Law School's Case Law Access Project API and static archive. The collection process involved the following steps:

1. A list of ZIP file URLs was compiled from the static.case.law archive, stored in the `zip_urls.txt.gz` file within the KL3M codebase. Each URL points to a ZIP file containing multiple court opinions from a specific reporter.
2. For each ZIP file URL, the collection pipeline:
    - Downloads the ZIP archive from static.case.law
    - Extracts both the HTML files (containing the case text) and the corresponding JSON metadata files
    - Embeds each HTML fragment into a proper HTML document structure

- Creates a Document object with the following metadata:
  - Unique ID from the CAP dataset
  - Case name as the title
  - Format identifier (text/html)
  - Description (case name)
  - Source URL (https://static.case.law/)
  - License information (CC0 1.0 Universal)
  - Content hash for integrity verification
  - Case-specific metadata (court, jurisdiction, date, etc.)
- Uploads each document to the KL3M storage system

The collection methodology leverages the publicly available data from the Case Law Access Project, which digitized over 40 million pages of U.S. court decisions in collaboration with Ravel Law. The dataset is licensed under CC0 1.0 Universal, placing it in the public domain. This ensures that the entire corpus is freely available for research, analysis, and reuse without copyright restrictions.

A.4 Content Examples

B PACER Dockets

This appendix provides details about the PACER Dockets dataset, including its collection methodology, data statistics, and examples of the data. PACER (Public Access to Court Electronic Records) is the electronic system that provides access to case and docket information from federal appellate, district, and bankruptcy courts.

B.1 Dataset Overview

The PACER Dockets dataset contains docket sheets from federal courts across the United States. These docket sheets serve as the official records of court proceedings, containing chronological listings of all events and filings in a case. The dataset includes dockets from district courts, bankruptcy courts, and appellate courts, providing a comprehensive record of federal court activity and procedural history.

The dockets provide essential metadata about federal cases, including:

- Case numbers and titles
- Judge assignments
- Filing dates
- Party information (plaintiffs, defendants, attorneys)
- Chronological listing of all events and filings
- Case status and outcomes

B.2 Data Processing Statistics

Based on the counts files in the KL3M project, the PACER Dockets dataset has been processed through multiple stages:

Processing Stage	Number of Documents
Documents (Initial Collection)	641,964
Representations (Processed)	641,961
Parquet (Final Format)	641,945

Table 4: PACER Dockets Dataset Document Counts by Processing Stage

As shown in Table 4, the processing pipeline for the PACER Dockets dataset maintained high consistency across stages. From the initial collection to the representation stage, only 3 documents were lost (99.9995% retention). The final parquet conversion stage had a minimal additional loss of 16 documents. Overall, 99.997% of the originally collected documents were successfully processed through the entire pipeline, demonstrating the robustness of the processing methodology for this dataset.

B.3 Collection Methodology

The PACER Dockets dataset was collected from the CourtListener and Internet Archive's joint effort to make federal court records freely accessible. The collection process involved the following steps:

1. Obtaining the source data file: The dataset retrieves a compressed CSV file (`dockets-2024-08-31.csv.bz2`) from the CourtListener's S3 storage bucket (`com-courtlistener-storage`). This file contains metadata about hundreds of thousands of federal docket sheets.
2. Filtering and processing records: The source code filters the data to include only records with valid Internet Archive JSON URLs (records containing a `filepath_ia_json` field with a valid HTTP URL). Each record in the CSV contains extensive metadata about a court case, including:
    - Court identifiers (e.g., `flnd` for Florida Northern District)
    - Case numbers and PACER case IDs
    - Case names (e.g., `Salvador v. Morgan`)
    - Filing and termination dates
    - Nature of suit and cause of action
    - Judge assignments
    - Jurisdiction information
3. Downloading and processing JSON data: For each valid record, the system:
    - Downloads the complete docket sheet in JSON format from the Internet Archive URL
    - Creates a document record with the docket data
    - Assigns a unique ID based on the JSON filename
    - Preserves all original metadata in the document's `extra` field

- Calculates a cryptographic hash (blake2b) of the content for integrity verification
- Uploads the document to the KL3M storage system

The collection methodology leverages public domain federal court records made accessible through CourtListener and the Internet Archive. As noted in the source code, these docket entries are "Not subject to copyright under 17 U.S.C. 105 and provided under CC0 by CourtListener/IA." This status ensures that the entire dataset is freely available for research and analysis without copyright restrictions.

The dataset specifically focuses on the docket sheets themselves, which provide the procedural history and metadata of cases, rather than the full text of court documents (which are collected separately in the RECAP Documents dataset).

B.4 Content Examples

C Federal Websites

This appendix provides details about the Federal Websites dataset, including its collection methodology, data statistics, and examples of the data. The Federal Websites dataset consists of content from U.S. government websites in the .gov, .mil, and select other government-related domains. These websites contain official information from various federal agencies, departments, and institutions.

C.1 Domain Coverage

The KL3M dataset includes content from 343 federal websites across multiple government branches and agencies. These domains can be organized into several major categories:

- **Executive Branch Agencies** – Includes websites from cabinet departments (e.g., www.dhs.gov, www.treasury.gov), independent agencies (e.g., www.epa.gov, www.nasa.gov), and their sub-agencies
- **Legislative Branch** – Includes www.house.gov, www.senate.gov, www.gao.gov, www.cbo.gov
- **Judicial Branch** – Includes court websites such as www.uscourts.gov, cafc.uscourts.gov
- **Military Domains** – Various .mil domains including www.army.mil, www.navy.mil, www.af.mil

C.2 Selected Domains

The full list of 343 websites is available in the KL3M codebase (scripts/dotgov_datasets.txt). Some of the most notable websites include:

- **Executive Departments**
  - www.usda.gov – Department of Agriculture
  - www.commerce.gov – Department of Commerce
  - www.defense.gov – Department of Defense
  - www.ed.gov – Department of Education
  - www.energy.gov – Department of Energy

  - www.hhs.gov – Department of Health and Human Services
  - www.dhs.gov – Department of Homeland Security
  - www.hud.gov – Department of Housing and Urban Development
  - www.doi.gov – Department of the Interior
  - www.justice.gov – Department of Justice
  - www.dol.gov – Department of Labor
  - www.state.gov – Department of State
  - www.transportation.gov – Department of Transportation
  - www.treasury.gov – Department of the Treasury
  - www.va.gov – Department of Veterans Affairs
- **Independent Agencies**
  - www.epa.gov – Environmental Protection Agency
  - www.nasa.gov – National Aeronautics and Space Administration
  - www.nrc.gov – Nuclear Regulatory Commission
  - www.nsf.gov – National Science Foundation
  - www.ssa.gov – Social Security Administration
- **Federal Commissions and Boards**
  - www.fcc.gov – Federal Communications Commission
  - www.ftc.gov – Federal Trade Commission
  - www.sec.gov – Securities and Exchange Commission

C.3 Collection Methodology

C.4 Data Statistics

C.5 Content Examples

D Electronic Code of Federal Regulations (eCFR)

This appendix provides details about the Electronic Code of Federal Regulations (eCFR) dataset, including its collection methodology, data statistics, and examples of the data. The eCFR is a web version of the Code of Federal Regulations (CFR) that is maintained by the U.S. Government Publishing Office (GPO) and updated daily to better reflect its current status.

D.1 Dataset Overview

The eCFR dataset contains the full text of all federal regulations organized into 50 titles, covering broad subject areas of federal regulation. The dataset provides a complete, up-to-date version of the U.S. Code of Federal Regulations, including all active regulations from federal agencies. This dataset is particularly valuable as it represents the official codification of the general and permanent rules published in the Federal Register by the executive departments and agencies of the Federal Government.

D.2 Data Processing Statistics

Based on the counts files in the KL3M project, the eCFR dataset has been processed through multiple stages:

Processing Stage	Number of Documents
Documents (Initial Collection)	262,243
Representations (Processed)	262,243
Parquet (Final Format)	262,243

Table 5: eCFR Dataset Document Counts by Processing Stage

As shown in Table 5, the eCFR dataset maintained perfect consistency across all processing stages, with 100% of documents successfully preserved through the entire pipeline. This demonstrates the exceptional quality and robustness of both the source data and the processing methodology for this dataset.

D.3 Collection Methodology

The eCFR dataset was collected using the official eCFR API provided by the U.S. Government Publishing Office. The collection process involved the following steps:

1. **Title Retrieval**: The system first retrieved metadata about all 50 CFR titles using the `/api/versioner/v1/titles.jso` endpoint.
2. **Version Discovery**: For each title, the system obtained the latest available version date using the `/api/versioner/v1/versions/title-{title}.json` endpoint.
3. **Structure Mapping**: For each title as of its latest version date, the system retrieved the complete hierarchical structure using the `/api/versioner/v1/structure/{date}/title-{title}.json` endpoint. This structure contains all levels of the regulatory hierarchy, including:
    - Titles (e.g., Title 1: General Provisions)
    - Chapters (e.g., Chapter I: Office of the Federal Register)
    - Subchapters (when applicable)
    - Parts (e.g., Part 1: Federal Register)
    - Subparts (when applicable)
    - Sections (e.g., §1.1 Definitions)

4. **Section Content Retrieval**: The system then traversed each title's structure to identify all section nodes, which represent the actual regulatory content. For each section, it retrieved the HTML content using the `/api/renderer/v1/content/enhanced/{date}/title-{title}?section={section}` endpoint.
5. **Document Creation**: For each section, a Document object was created with comprehensive metadata:
    - Unique ID in the format `{date}/{title}/{section}`
    - Title from the section label
    - Description combining the title, section, and date information
    - HTML content of the regulation
    - Publisher information (U.S. Government Publishing Office)
    - Date of the regulation version
    - Cryptographic hash (blake2b) for integrity verification
6. **Storage**: Each document was uploaded to the KL3M storage system.

The collection process incorporates a rate-limiting mechanism (with a delay of 0.1 seconds between requests) to ensure compliance with the eCFR API's usage policies and to avoid overwhelming the government servers.

D.4 Legal Status

As noted in the source code, the eCFR dataset is "Not subject to copyright under 17 U.S.C. 105," which means that the content is in the public domain as it was created by the federal government. This makes the dataset freely available for research, analysis, and redistribution without copyright restrictions.

D.5 Content Structure

The eCFR content is structured hierarchically, following the official organization of the Code of Federal Regulations:

- **Titles** (1-50): Broad subject areas (e.g., Title 10 - Energy, Title 26 - Internal Revenue)
- **Chapters**: Typically corresponding to the issuing agency
- **Subchapters**: Further divisions within chapters (when applicable)
- **Parts**: Major topical divisions within chapters or subchapters
- **Subparts**: Divisions within parts (when applicable)
- **Sections**: The basic unit of the CFR, containing the actual regulatory text
- **Appendices**: Supplementary material to parts or sections

This hierarchical structure is preserved in the KL3M dataset, allowing for comprehensive coverage and navigation of the entire federal regulatory framework.

E SEC EDGAR Database

This appendix provides details about the SEC EDGAR Database dataset, including its collection methodology, data statistics, and examples of the data. The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system is the primary system for companies and individuals to submit required forms and documents to the U.S. Securities and Exchange Commission (SEC).

E.1 Dataset Overview

The EDGAR dataset contains corporate filings submitted to the U.S. Securities and Exchange Commission (SEC) under various securities laws including the Securities Act of 1933, the Securities Exchange Act of 1934, the Trust Indenture Act of 1939, and the Investment Company Act of 1940. These filings include annual reports (10-K), quarterly reports (10-Q), registration statements, prospectuses, proxy statements, and various other corporate disclosures required by federal securities laws.

The dataset provides a comprehensive repository of public company disclosures dating back to 1996, serving as a critical resource for financial research, regulatory compliance analysis, and corporate transparency studies. The EDGAR archive is particularly valuable as it contains standardized corporate financial information across thousands of public companies over multiple decades.

E.2 Data Processing Statistics

Based on the counts files in the KL3M project, the EDGAR dataset has been processed through multiple stages:

Processing Stage	Number of Documents
Documents (Initial Collection)	74,063,501
Representations (Processed)	30,474,244
Parquet (Final Format)	44,768,118

Table 6: EDGAR Dataset Document Counts by Processing Stage

As shown in Table 6, the EDGAR dataset exhibits some interesting patterns through the processing pipeline. The initial collection contained over 74 million documents, of which approximately 41% were successfully converted to representations. However, the parquet format shows a higher document count than the representations stage, suggesting that some documents may have been processed through alternative pipelines or that the parquet conversion included additional fields or transformations that resulted in multiple parquet files per original document.

The large difference between the initial document count and the representations count likely reflects the complex nature of EDGAR filings, which often contain multiple document types, some of which may be difficult to process (like certain PDF formats, image-based documents, or proprietary formats) or may have been intentionally filtered during processing.

E.3 Collection Methodology

The EDGAR dataset was collected directly from the SEC's EDGAR system using their public API. The collection process involved several sophisticated steps:

1. **Daily Feed Collection**: The system downloads daily feed files from the SEC EDGAR archives, going back to 1996 (specified by the `EDGAR_MIN_DATE` constant). Each feed is a tar.gz archive containing multiple "nc" (News Condensed) files.
2. **Feed Processing**: For each daily feed archive, the system:
    - Extracts all member files from the tar.gz archive
    - Processes each .nc file containing multiple submissions
3. **Submission Parsing**: Within each .nc file, the system:
    - Identifies submission blocks using regex pattern matching
    - Extracts metadata from the submission header
    - Identifies individual document blocks within each submission
4. **Document Extraction**: For each document in a submission:
    - Extracts document metadata (type, sequence, filename, description)
    - Extracts the document content from between `<TEXT>` tags
    - Handles UUEncoded content by decoding when necessary
5. **Document Creation**: For each extracted document, a Document object is created with extensive metadata:
    - Unique ID in the format `cik/accession_number/sequence`
    - URL to the document on the SEC website
    - Document content and size
    - Content hash for integrity verification
    - MIME type based on file extension
    - Document title from the description field
    - Source name from filer/issuer/subject company information
    - SEC as the publisher
    - Filing date
    - Form types as subjects
    - Complete submission and document metadata in the extra field
6. **Storage**: Each document is uploaded to the KL3M storage system.

The collection process includes extensive error handling to manage the variations in document formats and metadata structures found in the EDGAR archive. It also carefully maintains the SEC-specific metadata hierarchy, preserving the relationships between companies, submissions, and documents.

E.4 Legal Status

As noted in the source code, the EDGAR dataset is "Generally accepted to available for free use and distribution" under various securities laws including Sections 19 and 20 of the Securities Act of 1933, Section 21 of the Securities Exchange Act of 1934, Section 321 of the Trust Indenture Act of 1939, Section 42 of the Investment Company Act of 1940, Section 209 of the Investment Advisers Act of 1940, and Title 17 of the Code of Federal Regulations, Section 202.5.

The code does note the ISDA v. Socratek case as offering some potentially countervailing guidance, but the general understanding is that these materials are not subject to copyright and are available for public use as records of the U.S. government.

E.5 Content Types

The EDGAR dataset includes a wide variety of document types, reflecting the diverse nature of corporate filings:

- **Annual Reports (10-K, 10-KSB)** - Comprehensive reports on a company's financial performance
- **Quarterly Reports (10-Q, 10-QSB)** - Updates on a company's financial status for a fiscal quarter
- **Registration Statements (S-1, S-3, etc.)** - Filings for new security offerings
- **Beneficial Ownership Reports (Schedule 13D, 13G)** - Reports of ownership of more than 5% of a company
- **Insider Trading Reports (Form 3, 4, 5)** - Reports of insider transactions
- **Proxy Statements (DEF 14A)** - Information provided to shareholders before annual meetings
- **Current Reports (8-K)** - Reports of significant events between regular filings
- **Foreign Company Reports (20-F, 40-F)** - Annual reports for foreign companies

The dataset preserves both the structured metadata about these filings and the full content of the documents themselves, enabling comprehensive analysis of corporate disclosures across time and across different regulatory requirements.

F European Union Official Journal

This appendix provides details about the European Union Official Journal dataset, including its collection methodology, data statistics, and examples of the data.

G Federal Depository Library Program

This appendix provides details about the Federal Depository Library Program dataset, including its collection methodology, data statistics, and examples of the data.

H Federal Register

This appendix provides details about the Federal Register dataset, including its collection methodology, data statistics, and examples of the data.

I GovInfo

This appendix provides details about the GovInfo dataset, including its collection methodology, data statistics, and examples of the data. GovInfo (formerly known as the Federal Digital System or FDsys) is a service of the United States Government Publishing Office (GPO) that provides free public access to official publications from all three branches of the Federal Government.

I.1 Collections

The KL3M dataset includes content from multiple collections available in GovInfo. Some documents may appear in multiple collections, as indicated by the semicolon-delimited collection IDs in the source data. The primary collections are:

- **BILLS** – Congressional Bills
- **BUDGET** – Budget of the United States Government
- **CCAL** – Congressional Calendars
- **CDIR** – Congressional Directory
- **CDOC** – Congressional Documents
- **CHRG** – Congressional Hearings
- **CMR** – Commerce Business Daily
- **COMPS** – Compilations of Presidential Documents
- **CPD** – Congressional Pictorial Directory
- **CPRT** – Congressional Committee Prints
- **CREC** – Congressional Record
- **CRECB** – Congressional Record Bound
- **CRI** – Congressional Record Index
- **CRPT** – Congressional Reports
- **CZIC** – Coastal Zone Information Center
- **ECONI** – Economic Indicators
- **ERIC** – Education Resources Information Center
- **ERP** – Economic Report of the President
- **GAOREPORTS** – Government Accountability Office Reports

- **GOVMAN** – Government Manual
- **GOVPUB** – Government Publications
- **GPO** – Government Publishing Office Collections
- **HJOURNAL** – House Journal
- **HMAN** – House Manual
- **HOB** – History of Bills
- **LSA** – List of CFR Sections Affected
- **PAI** – Privacy Act Issuances
- **PLAW** – Public and Private Laws
- **PPP** – Public Papers of the Presidents
- **SERIALSET** – Serial Set
- **SMAN** – Senate Manual
- **SJOURNAL** – Senate Journal
- **STATUTE** – Statutes at Large
- **USCOURTS** – United States Courts Opinions

I.2 Cross-Collection Documents

Some documents in GovInfo appear in multiple collections. The major cross-collection occurrences include:

- Documents that appear in both their primary collection and the GPO collection (e.g., documents labeled as GPO;CDOC, GPO;CFR, GPO;CPRT, GPO;CRECB, GPO;CRPT, GPO;FR, GPO;SJOURNAL)
- Documents in the Serial Set that also belong to other collections (e.g., SERIALSET;CDOC, SERIALSET;CRPT, SERIALSET;HJOURNAL, SERIALSET;SJOURNAL)
- Congressional documents that appear in multiple collections (e.g., ERP;CDOC, HMAN;CDOC, SMAN;CDOC)
- Government publications in specific categories (e.g., GOVPUB;CHRG)
- Budget documents in multiple collections (SERIALSET;CDOC;BUDGET)
- Economic reports in multiple collections (SERIALSET;CRPT;ERP)

I.3 Collection Methodology

I.4 Data Statistics

I.5 Content Examples

J RECAP Archive

This appendix provides details about the RECAP Archive dataset, including its collection methodology, data statistics, and examples of the data.

K RECAP Documents

This appendix provides details about the RECAP Documents dataset, including its collection methodology, data statistics, and examples of the data.

L Regulations.gov Documents

This appendix provides details about the Regulations.gov Documents dataset, including its collection methodology, data statistics, and examples of the data.

M UK Legislation

This appendix provides details about the UK Legislation dataset, including its collection methodology, data statistics, and examples of the data. The UK Legislation dataset consists of primary and secondary legislation from the United Kingdom, published on the legislation.gov.uk website maintained by The National Archives.

M.1 Dataset Coverage

The dataset includes various types of UK legislation:

- **Acts of Parliament** – Primary legislation enacted by the UK Parliament
- **Statutory Instruments** – Secondary legislation made under powers delegated by Acts of Parliament
- **Statutory Rules and Orders** – Historical statutory instruments predating 1948
- **UK Statutory Rules** – Northern Ireland statutory rules
- **UK Church Instruments** – Measures and instruments related to the Church of England
- **UK Local Acts** – Acts that apply to specific localities or entities rather than the general public

The dataset covers legislation dating back to 1267, with comprehensive coverage of legislation enacted after 1988. Older legislation is added to the dataset as it is digitized by The National Archives.

M.2 Collection Methodology

M.3 Data Statistics

M.4 Content Examples

N United States Code

This appendix provides details about the United States Code dataset, including its collection methodology, data statistics, and examples of the data.

O USPTO Granted Patents

This appendix provides details about the USPTO Granted Patents dataset, including its collection methodology, data statistics, and examples of the data.