

چکیده توسعه یافته = Extended Abstract

عنوان: ملاحظات حقوقی بهره‌برداری از داده‌های آموزشی دارای حق مؤلف در مدل‌های زبانی بزرگ:

نمونه پژوهی مجموعه داده Books3 در مدل LLaMA2

نام نویسنده: زینب پاپی، استادیار گروه پژوهشی مدیریت اطلاعات و سازماندهی دانش، پژوهشکده اسناد،

سازمان اسناد و کتابخانه ملی؛ z-papi@nlai.ir

مقدمه: مدل‌های زبانی بزرگ برای آموزش به حجم عظیمی از داده‌های متنی نیاز دارند که بخش قابل توجهی از آن‌ها تحت حمایت قانون حق مؤلف است. مجموعه داده Books3، شامل حدود ۱۹۶ هزار کتاب و به کاررفته در آموزش LLaMA2، نمونه‌ای برجسته از این چالش است. چارچوب‌های حقوقی موجود پاسخ کاملی به این مسئله نمی‌دهند و در نظام حقوقی ایران سازوکارهایی مانند استفاده منصفانه توسعه نیافته است. این پژوهش می‌پرسد آیا بهره‌برداری از داده‌های دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ مجاز است یا نقض حق مؤلف محسوب می‌شود. هدف، بررسی امکان توجیه این استفاده، تبیین مسئولیت تولیدکنندگان و کاربران، و پیشنهاد راهکارهای حقوقی برای ایران است.

روش‌شناسی: پژوهش با رویکرد کیفی و بر پایه تحلیل اسنادی و تحلیل تطبیقی حقوقی انجام شد. به دلیل ماهیت میان‌رشته‌ای موضوع، قواعد حقوقی در پرتو شواهد فنی مربوط به فرایند آموزش مدل‌ها بررسی شد. مجموعه داده Books3 و مدل LLaMA2 با نمونه‌گیری هدفمند و به عنوان یک نمونه واقعی انتخاب شدند؛ زیرا هم‌زمان دارای سه ویژگی بودند: بیان رسمی توسعه‌دهنده درباره استفاده از داده در پرونده *Kadrey v. Meta*، شواهد تجربی مستقل درباره امکان بازتولید محتوای آموزشی، و مستندات شناخته‌شده درباره منشأ داده. منابع از پایگاه‌های داخلی و خارجی و با کلیدواژه‌های تخصصی گردآوری شد و قوانین از طریق پایگاه وایپو بررسی شد. در تحلیل تطبیقی، نظام‌های حقوقی ایران، ایالات متحده، اتحادیه اروپا، فرانسه و آلمان با رویکرد کارکردی و در سه محور مقایسه شدند: وجود استثنای متن کاوی و داده کاوی، دامنه استفاده منصفانه یا نهادهای مشابه، و رویه قضایی مرتبط. آرای شاخصی مانند *Authors Guild v. Google* نیز مطالعه شد. برای افزایش اعتبار یافته‌ها، از مثلث‌سازی داده‌ها میان اسناد حقوقی، مستندات فنی و مطالعات تجربی استفاده شد.

یافته‌های اصلی: یافته‌ها نشان داد استفاده از Books3 در آموزش LLaMA2 را نمی‌توان به طور مطلق مشروع یا نامشروع دانست. ورود منابع دارای حق مؤلف به فرایند آموزش قابل احراز بود و گزارش فنی LLaMA2 و پرونده *Kadrey v. Meta* آن را تأیید کرد. فرایند توسعه مدل چندمرحله‌ای بود (گردآوری، پاک‌سازی، توکن‌سازی، آموزش و استنتاج) و هر مرحله مسئله حقوقی متفاوتی ایجاد کرد؛ مرز «استخراج الگو»

و «بازتولید اثر» نیز همواره روشن نبود. هیچ‌یک از نظام‌های بررسی شده مجوز عامی برای آموزش مدل‌ها پیش‌بینی نکرده بود. در اتحادیه اروپا، فرانسه و آلمان، استثنائات متن‌کاوی و داده‌کاوی مشروط به دسترسی قانونی بود؛ در آمریکا، استفاده منصفانه به تحلیل موردی وابسته بود؛ و در ایران مقرر اختصاصی وجود نداشت و قواعد عمومی حق مؤلف حاکم بود. رویه قضایی عمدتاً در آمریکا شکل گرفته بود و نتیجه یکسانی نداشت. در بحث مسئولیت، توسعه‌دهندگان بیشترین سطح مسئولیت را داشتند، کاربران معمولاً مسئولیت تبعی محدودی می‌پذیرفتند و مجوزهای قراردادی ابزاری برای مدیریت خطر بودند.

بحث و نتیجه‌گیری: چالش اصلی فقط مشروعیت استفاده از منابع دارای حق مؤلف نیست، بلکه احراز منشأ داده‌ها، اثبات نحوه استفاده و تعیین مسئولیت بازیگران نیز هست. هیچ نظام مطالعه‌شده استفاده نامحدود را مجاز نمی‌داند: آمریکا با معیار موردی استفاده منصفانه، اتحادیه اروپا با استثنای هماهنگ دستورالعمل ۷۹۰/۲۰۱۹ و شرط دسترسی قانونی، و فرانسه و آلمان با اجرای همین چارچوب. در ایران، قانون ۱۳۴۸ درباره داده‌کاوی و آموزش هوش مصنوعی ساکت است و رویه قضایی منتشرشده‌ای وجود ندارد. از آنجا که دسترسی به داده، آموزش و تولید خروجی آثار حقوقی متفاوتی دارند، قانون‌گذاری باید این مراحل را تفکیک کند. در این پژوهش با توجه به یافته‌های پژوهش، پیشنهادهایی به شرح زیر ارائه شد:

الزام به شناسایی، مستندسازی و ردیابی داده‌های آموزشی؛ بازنگری لایحه حمایت از حقوق مالکیت ادبی و هنری و پیش‌بینی مقرر مستقل برای داده‌کاوی بر پایه دسترسی قانونی و تفکیک استفاده پژوهشی از تجاری؛ طراحی سازوکارهای مجوزدهی به‌ویژه از طریق نهادهای مدیریت جمعی؛ و تعیین معیارهای قانونی برای حدود بهره‌برداری. این راهکارها ضمن حفظ حقوق مؤلفان، ابهام حقوقی را کاهش می‌دهند و توسعه مسئولانه فناوری را ممکن می‌سازند.

کلیدواژه‌ها: داده‌های آموزشی، حق مؤلف، مدل‌های زبانی بزرگ، استفاده منصفانه، LLaMA2، Books3،

نظام حقوقی ایران

Title: Legal Considerations of Exploiting Copyrighted Training Data in Large Language Models: A Case Study of the Books3 Dataset in the LLaMA2 Model

Name of Author: Zeinab Papi, Assistant Professor, Department of Information Management and Knowledge Organization, Research Center for Documents, National Library and Archives of the Islamic Republic of Iran, Tehran, Iran; z-papi@nlai.ir

Introduction: Large language models require vast volumes of textual data for training, a considerable portion of which is protected by copyright law. The Books3 dataset, which comprises approximately 196,000 books and was used in training LLaMA2, is a prominent example of this challenge. Existing legal frameworks do not offer a complete answer to this problem, and in Iran's legal system, mechanisms such as fair use have not been developed. This study asks whether the exploitation of copyrighted data in training large language models is permissible or constitutes copyright infringement. Its aims are to examine whether such use can be justified, to clarify the liability of developers and users, and to propose legal solutions for Iran.

Methodology: The study was conducted using a qualitative approach based on documentary analysis and comparative legal analysis. Given the interdisciplinary nature of the subject, legal rules were examined in light of technical evidence concerning the model training process. The Books3 dataset and the LLaMA2 model were selected through purposive sampling as a real-world case, because they simultaneously exhibited three characteristics: the developer's official statement regarding the use of the data in *Kadrey v. Meta*, independent empirical evidence on the possibility of reproducing training content, and well-established documentation on the data's provenance. Sources were collected from domestic and international databases using specialized keywords, and legislation was examined through the WIPO database. In the comparative analysis, the legal systems of Iran, the United States, the European Union, France, and Germany were compared using a functional approach along three axes: the existence of a text and data mining exception, the scope of fair use or comparable doctrines, and the relevant case law. Landmark rulings such as *Authors Guild v. Google* were also studied. To enhance the validity of the findings, data triangulation was employed across legal documents, technical documentation, and empirical studies.

Main Findings: The findings showed that the use of Books3 in training LLaMA2 cannot be considered categorically legitimate or illegitimate. The entry of copyrighted sources into the training process could be established, and both the LLaMA2 technical report and the *Kadrey v. Meta* case confirmed it. The model development process was multi-stage (collection, cleaning, tokenization,

training, and inference), and each stage raised a distinct legal issue; the boundary between "pattern extraction" and "reproduction of a work" was also not always clear. None of the legal systems examined had provided a general license for training models. In the European Union, France, and Germany, text and data mining exceptions were conditional on lawful access; in the United States, fair use depended on case-by-case analysis; and in Iran, no specific provision existed and the general rules of copyright applied. Case law had developed mainly in the United States and had not produced uniform outcomes. With respect to liability, developers bore the highest level of responsibility, users generally bore limited secondary liability, and contractual licenses served as a risk management tool.

Discussion and Conclusion: The main challenge is not only the legitimacy of using copyrighted sources, but also the establishment of data provenance, the proof of how the data were used, and the allocation of responsibility among the actors involved. None of the legal systems studied permits unrestricted use: the United States applies a case-by-case fair use standard; the European Union applies the harmonized exception of Directive 2019/790, conditional on lawful access; and France and Germany implement the same framework. In Iran, the 1969 Copyright Act is silent on data mining and AI training, and no published case law exists. Since data access, training, and output generation carry different legal consequences, legislation should distinguish between these stages. In light of the findings, the following recommendations were presented: requiring the identification, documentation, and traceability of training data; revising the Bill on the Protection of Literary and Artistic Property Rights and introducing an independent provision for data mining based on lawful access and a distinction between research and commercial use; designing licensing mechanisms, particularly through collective management organizations; and establishing statutory criteria for the limits of exploitation. These measures would safeguard authors' rights while reducing legal uncertainty and enabling the responsible development of technology.