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

مجموعه دادهٔ Books3 در مدل LLaMA2

زینب پاپی *
دکتری علم اطلاعات و دانش‌شناسی، استادیار
گروه پژوهشی مدیریت اطلاعات و سازماندهی دانش،
پژوهشکده اسناد، سازمان اسناد و کتابخانه ملی ج.ا.ا.تهران، ایران

دریافت: ۱۴۰۵/۰۳/۲۳ | **پذیرش:** ۱۴۰۵/۰۷/۰۷ | مقاله برای اصلاح به مدت ۲۲ روز نزد پدیدآوران بوده است.

چکیده: پژوهش حاضر با هدف بررسی وضعیت حقوقی استفاده از آثار دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ، با تمرکز بر مجموعه‌داده Books3 و مدل LLaMA2، انجام شده است. پژوهش با تحلیل تطبیقی نظام‌های حقوقی ایران، ایالات متحده، اتحادیه اروپا، فرانسه و آلمان، در پی تبیین مبانی مسئولیت حقوقی بازیگران مختلف و ارزیابی کفایت قواعد موجود حقوق مؤلف در مواجهه با آموزش مدل‌های زبانی بزرگ است. پژوهش کنونی با رویکرد کیفی، مبتنی بر تحلیل محتوای اسناد و تحلیل تطبیقی نظام‌های حقوقی انجام شده است. داده‌ها از طریق بررسی قوانین و مقررات، اسناد بین‌المللی، آرای قضایی، مستندات فنی مربوط به Books3 و LLaMA2 و مطالعات علمی گردآوری و با روش تحلیل تطبیقی حقوقی مبتنی بر رویکرد کارکردی تحلیل شده‌اند. مجموعه‌داده Books3 و مدل LLaMA2 به‌عنوان نمونه‌ای تجربی برای آزمون تحلیل‌های حقوقی مورد استفاده قرار گرفته‌اند. یافته‌های پژوهش نشان داد که چالش اصلی در این حوزه صرفاً مشروعیت استفاده از داده‌های آموزشی نیست، بلکه احراز منشأ داده‌ها، قابلیت اثبات استفاده از منابع دارای حق مؤلف و تعیین مسئولیت بازیگران مختلف فرایند آموزش و بهره‌برداری از مدل‌های زبانی اهمیت اساسی دارد. بررسی تطبیقی نشان داد که در ایالات متحده مشروعیت استفاده عمدتاً بر پایه دکترین استفاده منصفانه و ارزیابی موردی انجام می‌شود، در حالی که اتحادیه اروپا در قالب دستورالعمل ۷۹۰/۲۰۱۹ چارچوب استثنای متن‌کاوی و داده‌کاوی را تعیین کرده و فرانسه و آلمان، به‌عنوان کشورهای عضو، این چارچوب را با تفاوت‌هایی در قوانین ملی خود اجرا کرده‌اند؛ شرایطی مانند دسترسی قانونی، رعایت الزامات قانونی و در برخی موارد امکان اعمال حق انصراف در هر دو سطح اتحادیه‌ای و ملی مطرح است. در مقابل، در حقوق ایران چارچوب

نشریهٔ علمی (رتبه ین المللی)
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳- ۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱- ۲۲۵۱
نمایه در LISTA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره xx | شمارهٔ x | ص ص xx-xx
۱۳xx x

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:
درون متن:
(پاپی، زودآیند)
در فهرست منابع:
پاپی، زی‌نب زودآیند. ملاحظات حقوقی بهرمبرداری از داده‌های آموزشی دارای حق مؤلف در مدل‌های زبانی بزرگ: نمونه‌پژوهی مجموعه دادهٔ Books3 در مدل LLaMA2 . *پژوهشنامه پردازش و مدیریت اطلاعات.*
http://Jipm.irandoc.ac.ir (دسترسی در روز/ماه/سال)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

x شمارهٔ | xx دورهٔ | ۱۳xx x

مستقلی برای استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ پیش‌بینی نشده و این موضوع موجب ابهام در تعیین مسئولیت توسعه‌دهندگان و سایر بازیگران شده است.

نتایج پژوهش بیانگر آن است که توسعه مسئولانه مدل‌های زبانی بزرگ مستلزم ایجاد چارچوبی شفاف برای استفاده از داده‌های آموزشی است. بر این اساس، بازنگری در مقررات مالکیت ادبی و هنری، پیش‌بینی قواعد مشخص درباره آموزش مدل‌های هوش مصنوعی، ایجاد الزامات قانونی برای شناسایی و مستندسازی داده‌های آموزشی، توسعه سازوکارهای مجوزدهی جمعی برای بهره‌برداری از منابع و تعیین حدود قانونی استفاده از داده‌های دارای حق مؤلف می‌تواند زمینه ایجاد تعادل میان حمایت از حقوق پدیدآورندگان و توسعه فناوری‌های هوش مصنوعی را فراهم سازد.

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

* z-papi@nlai.ir

۱. مقدمه

پیشرفت‌های سریع در حوزه هوش مصنوعی، به ویژه توسعه مدل‌های زبانی بزرگ[1]، تحولات چشمگیری در پردازش زبان طبیعی ایجاد کرده است. این مدل‌ها قادر به تولید متن‌های طبیعی، ترجمه، خلاصه‌سازی و پاسخ به پرسش‌ها هستند و کاربردهای گسترده‌ای در حوزه‌های مختلف ایجاد کرده‌اند (Raffel et al., 2020). برای آموزش این مدل‌ها، حجم وسیعی از داده‌های متنی به کار گرفته می‌شود که بخش قابل توجهی از آن‌ها تحت حمایت قانون حق مولف قرار دارند. داده‌های آموزشی نقش اساسی در توسعه مدل‌های زبانی بزرگ ایفا می‌کنند، زیرا کیفیت و گستردگی این داده‌ها مستقیماً بر عملکرد و دقت مدل تأثیرگذار است. استفاده از داده‌های دارای حق مولف برای آموزش مدل‌های زبانی در قالب پیش‌آموزش[2] چالش حقوقی ایجاد می‌کند.

---

1. Large Language Models (LLMs)
2. Pre-training

در حالی که جامعه یادگیری ماشین به تازگی بحث درباره قانونی‌بودن آموزش مدل‌ها با استفاده از داده‌های دارای حق مولف را آغاز کرده است، ملاحظات مهمی در خصوص پردازش و توزیع داده‌های متعلق به دیگران نیز مطرح می‌شود. توجه کافی به این نکته نمی‌شود که پردازش و توزیع این داده‌ها نیز می‌تواند نقض قانون حق مولف محسوب شود (Gao, 2020). چارچوب‌های حقوقی موجود، علی‌رغم تعدد در حوزه‌های قضایی مختلف، هنوز پاسخگویی کامل به چالش‌های حقوقی ناشی از استفاده گسترده از داده‌های دارای حق مولف در آموزش مدل‌های زبانی بزرگ ندارند و بسیاری از جنبه‌های استثنائات و محدودیت‌های قانونی در این زمینه مبهم باقی مانده است (Gervais, 2021). این ابهام، نیاز به بررسی دقیق‌تر حقوقی و تدوین راهکارهای حقوقی و فنی برای کاهش ریسک‌های ناشی از استفاده از این داده‌ها را ضروری می‌سازد. داده‌های آموزشی که در مدل‌های زبانی بزرگ مورد استفاده قرار می‌گیرند، همواره در معرض خطر نقض حق مولف هستند؛ به ویژه زمانی که داده‌ها بدون اخذ مجوز از پدیدآورندگان جمع‌آوری و پردازش شده باشند. استفاده منصفانه و سایر استثنائات و محدودیت‌های مشابه در برخی نظام‌های حقوقی به عنوان ابزارهایی میان حقوق انحصاری مولفان و منابع عمومی شناخته می‌شوند؛ با این‌حال، چنین سازوکارهایی هنوز در نظام حقوقی ایران توسعه نیافته است.

از جمله مجموعه داده‌هایی که در این زمینه چالش‌ها و مناقشات حقوقی جدی را ایجاد کرده، Books3 است که داده‌های آن در برخی از مدل‌های برجسته مانند LLaMA2 به کاررفته است. مجموعه دادهٔ Books3 شامل حدود ۱۹۶۰۰۰ عنوان کتاب داستانی و غیرداستان به زبان انگلیسی است که از پایگاه خصوصی و محدود Bibliotik استخراج شده است؛ نسخه‌ای از این پایگاه توسط پژوهشگری به‌نام شاون پرسر[1] در سال ۲۰۲۰ منتشر شد. پایگاه Bibliotik ترکیبی از کتاب‌های داستانی و غیرداستانی را دربرمی‌گیرد و تقریباً ده برابر بزرگ‌تر از دومین مجموعه دادهٔ بزرگ در این حوزه یعنی BookCorpus2، است. کتاب‌ها به عنوان منابعی با ساختار متنی منسجم و دامنه واژگانی غنی، می‌توانند مدل‌های زبانی را در تولید متن‌های بلند، یکپارچه و معنادار یاری کنند؛ از این‌رو، استفاده از منابعی چون Bibliotik به دلیل حجم بالای کتاب‌ها برای ارتقاء عملکرد مدل‌ها و بهبود کیفیت تولید متن به کار گرفته شده است (Gao, 2020). اما توجه به منشا

[1]. Shawn Presser

غیرقانونی داده‌ها، بهره‌برداری از آن‌ها را با مخاطرات جدی حقوقی مواجه کرده است. برپایهٔ آنچه بیان شد، مسئله اصلی پژوهش کنونی این است که بهره‌برداری از داده‌های دارای حق مولف، به‌ویژه مجموعه Books3، در آموزش مدل‌های زبانی بزرگ مانند LLaMA2، در چارچوب‌های قانونی کشورها – از جمله ایران – مجاز است یا مصداق نقض حقوق مؤلف تلقی می‌شود.

در تبیین این مسئله، مقاله به‌طور خاص می‌کوشد بررسی کند:

نخست، تا چه اندازه می‌توان بهره‌برداری از مجموعه‌داده Books3 در فرایند آموزش مدل‌های زبانی بزرگ (با تأکید بر LLaMA2) را در چارچوب مفاهیمی چون «استفاده منصفانه» یا سایر استثنائات قانونی مشابه توجیه کرد؟

دوم، مسئولیت حقوقی تولیدکنندگان و کاربران مدل‌های زبانی بزرگ در قبال نقض احتمالی حقوق مؤلف چگونه قابل تبیین است؟؛

سوم، با توجه به آراء، رویه‌های قضایی و دیدگاه‌های حقوقی موجود در نظام‌های حقوقی پیشرو، چه اصول یا ملاحظات کلیدی می‌توان برای سیاست‌های حقوقی در ایران استخراج کرد؟؛

و در پایان، چه الزامات یا راهکارهای حقوقی می‌تواند برای کشورهایی نظیر ایران پیشنهاد شود تا استفاده از داده‌های آموزشی به‌گونه‌ای مشروع و منصفانه قانون‌مند شوند؟

۲. پیشینه پژوهش

در سال‌های اخیر، پژوهش‌های داخلی نیز به تدریج به ابعاد مختلف حقوقی هوش مصنوعی پرداخته‌اند. بخشی از این مطالعات بر مسئله شخصیت حقوقی و اصالت منابع تولیدشده توسط هوش مصنوعی متمرکز بوده‌اند. در این زمینه، پاپی (۱۴۰۳) با بررسی تطبیقی آرای قضایی آمریکا، بریتانیا و چین نشان داده است که در نظام‌های حقوقی مورد مطالعه، هوش مصنوعی به عنوان پدیدآورنده یا مخترع شناخته نمی‌شود و حقوق ناشی از منابع تولیدشده، حسب مورد، به اشخاص حقیقی یا حقوقی دخیل در فرایند تولید اثر تعلق می‌گیرد. همچنین لطیف‌زاده (۱۴۰۳) با تفکیک میان خلاقیت انسانی و فرایندهای ترکیبی هوش مصنوعی، بر این باور است که هوش مصنوعی فاقد خلاقیت اصیل

انسانی است و ازاین‌رو اعطای شخصیت حقوقی به آن، دست‌کم در وضعیت کنونی، قابل توجیه نیست.

دسته دیگری از پژوهش‌ها به مسئولیت مدنی ناشی از فعالیت سامانه‌های هوش مصنوعی اختصاص یافته است. ذاکری‌نیا (۱۴۰۲) با تحلیل تطبیقی مقررات اتحادیه اروپا، الگوی دوگانه مسئولیت را برای سامانه‌های هوش مصنوعی تبیین کرده و در حقوق ایران، مسئولیت را بر مبنای قواعد فعل غیر قابل تحلیل دانسته است. تخشید (۱۴۰۰) نیز با تأکید بر پیچیدگی الگوریتم‌های یادگیری عمیق و دشواری اثبات ارکان مسئولیت مدنی، راهکارهایی مانند مسئولیت محض و بیمه اجباری را پیشنهاد کرده است. در همین راستا، احسانگر و یزدانیان (۱۴۰۴) و باقری (۱۴۰۴) نیز بر ضرورت بازنگری در مبانی مسئولیت مدنی و تدوین چارچوب‌های متناسب با فناوری‌های هوش مصنوعی تأکید کرده‌اند.

در حوزه مالکیت فکری و فناوری‌های نوین نیز برخی پژوهش‌ها به چالش‌های حقوقی ناشی از بهره‌برداری دیجیتال از آثار پرداخته‌اند. صادقی و ملک‌زاده (۱۴۰۴) رویکرد تلفیقی حمایت از حقوق پدیدآورندگان همراه با پیش‌بینی استثنائات و محدودیت‌های قانونی را مناسب‌ترین شیوه تنظیم روابط در فضای دیجیتال دانسته‌اند. همچنین ذاکری‌نیا و همکاران (۱۴۰۴) با بررسی حقوق مؤلف در آثار تولیدشده توسط هوش مصنوعی، بر ضرورت تفکیک مرحله آموزش مدل از مرحله بهره‌برداری از خروجی و توجه مستقل به جایگاه حقوقی داده‌های آموزشی تأکید کرده‌اند.

نزدیک‌ترین پژوهش به موضوع حاضر، مطالعه قاسمیان (۱۴۰۲) است که با رویکرد تطبیقی، وضعیت حقوقی استفاده از داده‌های آموزشی در یادگیری ماشین را در حقوق ایران، ایالات متحده و اتحادیه اروپا بررسی کرده است. وی ضمن تحلیل چالش‌های مربوط به استفاده از آثار دارای حق مؤلف، پایگاه‌های داده، اسرار تجاری و داده‌های شخصی، به این نتیجه رسیده است که حقوق ایران فاقد چارچوب روشن برای استفاده از داده‌های آموزشی در توسعه سامانه‌های هوش مصنوعی است و اصلاح قوانین موجود و پیش‌بینی استثنائات مشخص برای اهداف پژوهشی و آموزشی ضروری به نظر می‌رسد.

در سال‌های اخیر، هم‌زمان با توسعه مدل‌های زبانی بزرگ، مطالعات متعددی به بررسی ابعاد حقوقی استفاده از آثار دارای حق مؤلف در آموزش این مدل‌ها پرداخته‌اند. Sobel (2017) با تحلیل دکترین استفاده منصفانه در حقوق آمریکا نشان داد که استفاده از منابع دارای حق مؤلف در فرایند

آموزش سامانه‌های هوش مصنوعی می‌تواند تعادل میان حمایت از حقوق مؤلف و توسعه فناوری را با چالش مواجه کند. Gervais و همکاران (2021) نیز با بررسی تطبیقی حقوق ایالات متحده، اتحادیه اروپا و بریتانیا به این نتیجه رسیدند که چارچوب‌های سنتی حقوق مؤلف پاسخ روشنی برای استفاده از داده‌های آموزشی در مدل‌های زبانی بزرگ ارائه نمی‌کنند و بازنگری در مقررات موجود ضروری است.

در ادامه، Lucchi (2023) با تمرکز بر سامانه‌های هوش مصنوعی مولد، به‌ویژه ChatGPT، نشان داد که مفاهیمی مانند اصالت اثر، اثر اشتقاقی و حدود استفاده از داده‌های آموزشی در پرتو قواعد سنتی حقوق مؤلف با ابهام‌های جدی روبه‌رو هستند. Manteghi (2024) نیز با بررسی مقررات اتحادیه اروپا، ظرفیت و محدودیت‌های استثنائات متن‌کاوی و داده‌کاوی و همچنین جایگاه داده‌های مصنوعی را تحلیل و بر ضرورت اصلاح مقررات مرتبط با استفاده از داده‌های آموزشی تأکید کرده است. در همین راستا، Schirru و همکاران (2024) با مطالعه تطبیقی مقررات متن‌کاوی و داده‌کاوی در کشورهای آمریکای لاتین نشان دادند که نبود استثنائات قانونی شفاف، یکی از موانع توسعه پژوهش‌های داده‌محور و آموزش سامانه‌های هوش مصنوعی است. همچنین Rosati (2025) با بررسی حقوق اتحادیه اروپا و بریتانیا نتیجه گرفت که استثنائات موجود پاسخگوی آموزش مدل‌های هوش مصنوعی نیستند و بهره‌گیری از منابع دارای حق مؤلف، در اغلب موارد، نیازمند اخذ مجوز از صاحبان حقوق خواهد بود.

در کنار این مطالعات حقوقی، برخی پژوهش‌ها ابعاد فنی و تجربی موضوع را بررسی کرده‌اند. Karamolegkou و همکاران (2023) نشان دادند که مدل‌های زبانی بزرگ در برخی شرایط قادر به بازتولید بخش‌هایی از منابع دارای حق مؤلف هستند و این موضوع می‌تواند پیامدهای مهمی از منظر نقض حقوق مؤلف ایجاد کند. Feretzakis و همکاران (2025) نیز با بررسی چالش‌های حقوقی و فنی مدل‌های زبانی در پرتو مقررات حفاظت از داده‌ها، بر ضرورت شفافیت، قابلیت ردیابی داده‌های آموزشی و طراحی سازوکارهای حاکمیتی مناسب تأکید کردند. همچنین Zhang و همکاران (2025) با مرور نظام‌مند مخاطرات ایمنی، امنیت و حریم خصوصی مدل‌های زبانی بزرگ، استفاده از داده‌های آموزشی فاقد منشأ و مجوزدهی مشخص را یکی از مهم‌ترین چالش‌های حقوقی این فناوری دانسته و بر ضرورت تدوین چارچوب‌های حقوقی و فنی برای مدیریت داده‌های آموزشی تأکید کردند.

جدول ۱. ماتریس پژوهش‌های پیشین

پژوهشگر/تاریخ	هدف	روش	نتایج	خلأهای نظری و مفهومی برای پژوهش‌های آینده	راهبردها و اقدامات عملی پیشنهادی
قاسمیان (۱۴۰۲)	شناسایی چالش‌های حقوقی در استفاده از داده‌های آموزشی برای یادگیری ماشین با نظام‌های حقوقی ایران، آمریکا، اتحادیه اروپا	تطبیقی - تحلیلی	-نبود چارچوب قانونی جامع در ایران برای استفاده از داده‌های آموزشی -تردید در مشروعیت استفاده از داده‌ها برای آموزش مدل‌های هوش مصنوعی -تعارضات میان توسعه فناوری و نظام حقوق مالکیت فکری	-نبود ادبیات بومی و نظام‌مند درباره استثنائات مجاز در استفاده از داده‌ها برای آموزش الگوریتم‌ها در ایران -فقدان ابزارهای تحلیلی برای ارزیابی ریسک حقوقی استفاده از داده‌ها	- اصلاح قوانین ایران با تمرکز بر استثنائات آموزشی و پژوهشی -طراحی ابزارهای ارزیابی ریسک حقوقی برای توسعه هوش مصنوعی به‌صورت قانون‌مند و ایمن
Sobel (2017)	بررسی مسئولیت حقوقی و چارچوب‌های قانونی مرتبط با بهره‌برداری از داده‌های دارای حق مولف در آموزش مدل‌های	تحلیل تطبیقی و مطالعه موردی	-مسئولیت اصلی متوجه تولیدکنندگان است؛ کاربران در استفاده آگاهانه یا عمدی مسئولیت تبعی دارند؛ چارچوب‌های قراردادی و مجوزدهی نقش کلیدی دارند	-فقدان مطالعات منسجم درباره مسئولیت حقوقی تولیدکنندگان و کاربران مدل‌های زبانی بزرگ در استفاده از منابع دارای حق مولف	-تحلیل حقوقی استفاده از داده‌های آموزشی دارای حق مولف در مدل‌های هوش مصنوعی و ارائه چارچوب ارزیابی ریسک حقوقی برای تولیدکنندگان و پژوهشگران

	هوش مصنوعی				
Gervais et al. (2021)	بررسی مدیریت استثنائات آموزشی در قوانین حق مولف کشورهای مختلف و شناسایی محدودیت‌ها و ظرفیت‌های قانونی موجود	مطالعه تطبیقی – تحلیلی	-چارچوب‌های حقوقی موجود نتوانسته‌اند پاسخگوی کامل چالش‌های حقوقی باشند - استثنائاتی مانند استفاده منصفانه و داده‌کاوی کافی نیستند	-ابهامات و خلأهای قانونی در حوزه مالکیت داده، حقوق استفاده و مسئولیت‌های حقوقی استفاده از داده‌های دارای حق مولف در آموزش مدل‌های زبانی بزرگ	-بازنگری قوانین مالکیت فکری با هدف شفاف‌سازی حدود استفاده از داده‌های دارای حق مولف در آموزش LLM ها -تدوین مقررات منعطف متناسب با ویژگی‌های فناوری و داده‌های آموزشی
Karamolegkou and et al. (2023)	بررسی تکثیر عینی محتوای دارای حق مولف در مدل‌های زبانی بزرگ	مطالعه تجربی و تطبیقی	مدل‌های بزرگ‌تر احتمال بیشتری برای تکثیر عینی محتوای دارای حق مولف دارند.	شناسایی مرزهای استفاده منصفانه و نقض حقوق مؤلف در مدل‌های زبانی	نیاز به تنظیم مقررات شفاف برای آموزش مدل‌ها با داده‌های دارای حق مولف
Lucchi (2024)	بررسی چالش‌های حق مؤلف در سامانه‌های	تحلیل حقوقی با مطالعه موردی	چالش‌های حق مؤلف به ورودی‌ها و خروجی‌های هوش مصنوعی	ناکارآمدی چارچوب‌های سنتی حق مؤلف در پوشش	پیشنهاد سازوکارهای شفاف مجوزدهی، مدل‌های جبران خسارت و دستورالعمل‌های

	هوش مصنوعی مولد با تمرکز بر داده‌های آموزشی		محدود نمی‌شود و داده‌های آموزشی نقش محوری دارند	پیچیدگی‌های هوش مصنوعی مولد	اخلاقی برای استفاده از داده‌های آموزشی
Manteghi (2024)	بررسی چالش‌های حقوقی استفاده از منابع دارای حق مولف در آموزش مدل‌های هوش مصنوعی با رویکرد تطبیقی	تحلیل تطبیقی مقررات اتحادیه اروپا	مقررات فعلی محدودیت‌هایی دارند؛ داده‌های مصنوعی به دلیل نبود استانداردهای قانونی بهره‌برداری محدود دارند.	تدوین مقررات مشخص برای داده‌های مصنوعی و بهبود چارچوب‌های حقوقی	استفاده از سازوکارهای جبران مالی برای حفظ تعادل حقوق مالکیت و دسترسی به اطلاعات
Schirru and et al.(2024)	مطالعه تطبیقی استثنائات متن‌کاوی و داده‌کاوی در کشورهای آمریکای لاتین و	مطالعه تطبیقی قوانین ملی و چارچوب‌های حقوقی	اکثر کشورهای آمریکای لاتین فاقد قوانین مشخص هستند؛ نیاز به اصلاح قوانین وجود دارد.	پژوهش در راهکارهای تطبیق قوانین با استانداردهای بین‌المللی	آموزش قانون‌گذاران و پژوهشگران برای تسهیل پژوهش‌های داده‌محور

	کشورهای توسعه‌یافته				
Feretzakis and et al. (2025)	بررسی موانع فنی و حقوقی به‌کارگیری مدل‌های زبانی بزرگ در چارچوب مقررات عمومی حفاظت از داده‌ها	تحلیل انتقادی ادبیات و چارچوب‌های قانونی و فنی	ناسازگاری حقوق داده‌محور با ویژگی‌های مدل‌های عمیق، ضرورت شفافیت و مسئولیت‌پذیری	نبود چارچوب‌های قانونی مشخص برای استفاده از داده‌های دارای حق مولف در آموزش مدل‌های زبانی بزرگ	پیشنهاد چارچوب حاکمیتی چهارلایه، استفاده از حذف هدفمند داده‌ها، یادگیری فدرال، حفظ حریم خصوصی
Zhang et al. (2025)	بررسی نظام‌مند مخاطرات ایمنی، امنیتی و حریم خصوصی در مدل‌های زبانی بزرگ	مرور نظام‌مند	شناسایی تهدیداتی مانند افشای داده‌های خصوصی، تکثیر محتوای دارای حق مولف، تضعیف اعتماد عمومی به مدل‌های زبانی بزرگ	فقدان چارچوب‌های جامع برای تمایز میان داده‌های مجاز و غیرمجاز، و نبود سیاست‌های اجرایی قابل تطبیق با نظام‌های حقوقی مختلف	طراحی سیاست‌های اجرایی برای استفاده مسئولانه از داده‌های آموزشی، تدوین استانداردهای اخلاقی و فنی منطبق با چارچوب‌های بین‌المللی

Rosati (2025)	بررسی محدودیت‌ها و دفاعیات قانونی استفاده از منابع دارای حق مولف در آموزش مدل‌های هوش مصنوعی در قوانین اتحادیه اروپا و بریتانیا	تحلیل تطبیقی	هیچ یک از استثنائات یا دفاعیات موجود آموزش AI بدون مجوز را به‌طور کامل پوشش نمی‌دهد و استفاده بدون مجوز نقض حقوق مؤلفان است.	نیاز به روشن‌شدن چارچوب‌های قانونی برای آموزش مدل‌های هوش مصنوعی	رویکرد مبتنی بر مجوزدهی؛ استفاده از منابع فقط با کسب اجازه مستقیم یا جمعی از صاحبان حقوق

بررسی پژوهش‌های داخلی و خارجی نشان می‌دهد که اگرچه در سال‌های اخیر توجه پژوهشگران به ابعاد حقوقی آموزش مدل‌های زبانی بزرگ افزایش یافته است، مطالعات موجود عمدتاً بر یکی از جنبه‌های موضوع، مانند استثنائات حق مؤلف، داده‌کاوی، مسئولیت مدنی یا منابع تولیدشده توسط هوش مصنوعی، متمرکز بوده‌اند. همچنین بخش عمده پژوهش‌های خارجی در چارچوب نظام‌های حقوقی خاص انجام شده و پژوهش‌های داخلی نیز هنوز به بررسی موردی استفاده از مجموعه‌داده‌های دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ نپرداخته‌اند. ازاین‌رو، خلأ اصلی ادبیات موجود، نبود تحلیلی تطبیقی و یکپارچه درباره وضعیت حقوقی استفاده از داده‌های آموزشی دارای حق مؤلف، با تکیه بر یک نمونه واقعی مانند Books3 در آموزش LLaMA2 و ارزیابی آن در پرتو نظام حقوقی ایران و نظام‌های حقوقی منتخب است؛ خلأیی که پژوهش حاضر درصدد پاسخ‌گویی به آن است. با توجه به تنوع مطالعات پیشین، در جدول ۱ فقط پژوهش‌هایی منعکس شده‌اند که به‌طور مستقیم به ابعاد حقوقی استفاده از داده‌های آموزشی دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ پرداخته‌اند و از این حیث بیشترین ارتباط را با موضوع پژوهش حاضر دارند.

۳. روش پژوهش

پژوهش حاضر با رویکرد کیفی و بر پایه تحلیل اسنادی و تحلیل تطبیقی حقوقی انجام شد. هدف پژوهش، بررسی چالش‌های حقوقی ناشی از استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ و تحلیل نحوه مواجهه نظام‌های حقوقی با این پدیده است. با توجه به ماهیت میان‌رشته‌ای موضوع، تحلیل قواعد حقوقی با بهره‌گیری از شواهد فنی مرتبط با فرایند آموزش مدل‌های زبانی انجام شد تا امکان بررسی انطباق مقررات حقوق مؤلف با واقعیت‌های فنی این فناوری فراهم شود.

در این پژوهش، مجموعه‌داده Books3 و مدل زبانی LLaMA2 بر اساس نمونه‌گیری هدفمند و با توجه به ارتباط مستقیم آن‌ها با پرسش‌های پژوهش انتخاب شدند. درباره مجموعه‌داده Books3 اطلاعات نسبتاً مستندی درباره ساختار، ترکیب، نحوه گردآوری داده‌ها و ارتباط آن با آموزش مدل‌های زبانی در منابع فنی و حقوقی منتشر شده بود. همچنین این مجموعه‌داده در برخی مباحث و دعاوی مرتبط با استفاده از منابع دارای حق مؤلف در آموزش سامانه‌های هوش مصنوعی، از جمله دعاوی مطرح‌شده علیه شرکت Meta، مورد توجه قرارگرفت. درباره مدل LLaMA2 نیز مستندات فنی و گزارش‌های عمومی مرتبط با داده‌های آموزشی در دسترس بود. افزون بر این، انتشار پژوهش‌هایی درباره امکان شناسایی منابع دارای حق مؤلف در داده‌های آموزشی و بررسی ارتباط آن‌ها با خروجی مدل‌ها، امکان تحلیل هم‌زمان ابعاد فنی و حقوقی موضوع را فراهم کرده بود (Cooper et al., 2025). این انتخاب صرفاً بر پایه در دسترس‌بودن اطلاعات نبود. Books۳ و LLaMA۲ در زمان انجام این پژوهش دارای ترکیبی از سه ویژگی هم‌زمان بودند: نخست، بیان رسمی توسعه‌دهنده به استفاده از این مجموعه‌داده در پرونده Kadrey v. Meta Platforms, Inc., No. ۳:۲۳-cv-۰۳۴۱۷ N.D. Cal. ، پرونده‌ای که دادگاه در حکم بیست‌وپنجم ژوئن ۲۰۲۵ خود نیز به موضوع استفاده منصفانه در آن پرداخته بود؛ دوم، شواهد تجربی مستقل درباره امکان بازتولید محتوای آموزشی؛ و سوم، مستندات شناخته‌شده درباره منشأ داده. جزئیات کامل این سه مستند، از جمله نتیجه حکم یادشده، در جدول ۲ (بخش یافته‌ها) گزارش شد. این ترکیب هم‌زمان برای نمونه‌های دیگر گزارش نشده بود. با این حال، Books3 و LLaMA2 در این پژوهش فقط به عنوان یک نمونه واقعی انتخاب شدند تا امکان بررسی و تحلیل چالش‌های حقوقی استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ فراهم شود؛ از این رو، هدف پژوهش کنونی، تحلیل حقوقی یک مجموعه‌داده یا یک مدل خاص نیست، بلکه بررسی قواعد و چالش‌های حقوقی

مرتبط با استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ است.

برای استخراج منابع مختلف و مرتبط با موضوع، کلیدواژه‌های متن‌کاوی و داده‌کاوی، استثناهای پژوهشی، تحلیل داده، هوش مصنوعی و مالکیت فکری، حق مولف، مدل‌های زبانی بزرگ، مجموعه دادهٔ Books3 و مدل LLaMA2 در پایگاه‌های داخلی مانند گنج ایرانداک، اپک کتابخانه ملی، پارسه (پایان‌نامه‌های دانشگاه تربیت مدرس)، پایان‌نامه‌های دانشگاه تهران، پورتال جامع علوم انسانی، پورتال علمی جهاد دانشگاهی و سیویلیکا به جستجو پرداخت. همین‌طور برای جستجوی منابع خارجی از پایگاه‌های اطلاعاتی نظیر اشپرینگر[۱]، آرشیو اینترنت[۲]، گوگل اسکالر[۳]، امرالد[۴]، ریسرچ‌گیت[۵]، آرکایو[۶] و سیج‌پاب[۷] استفاده شد. کلیدواژه‌های مورد استفاده عبارت بودند از: Text mining and Data Mining, Copyright, Data Analysis, Exceptions and Limitations, Research Exceptions, artificial intelligence and intellectual property, Training Data, Large Language Models, Books3 Dataset, LLaMA2 Model. برای جستجو دقیق قوانین از پایگاه اطلاعاتی وایپو نیز بهره گرفته شد. در تحلیل تطبیقی، نظام‌های حقوقی ایران، ایالات متحده آمریکا، اتحادیه اروپا، فرانسه و آلمان بررسی شدند. ایالات متحده آمریکا به‌عنوان نماینده نظام حقوق عرفی[۸] انتخاب شد. در کنار آن، اتحادیه اروپا نه به‌عنوان یک نظام حقوقی مستقل و موازی با فرانسه و آلمان، بلکه به‌عنوان چارچوب هماهنگ‌کننده فراملی بررسی شد که استثنای متن‌کاوی و داده‌کاوی را در مواد ۳ و ۴ دستورالعمل ۷۹۰/۲۰۱۹ تعیین کرده است؛ فرانسه و آلمان نیز به‌عنوان دو نمونه از تبدیل این چارچوب اتحادیه‌ای به قانون ملی انتخاب شدند تا هم سطح حداقلی مقرر در سطح اتحادیه اروپا و هم تفاوت‌های ناشی از اختیار عمل کشورهای عضو در پیاده‌سازی آن قابل بررسی باشد. این طراحی امکان تحلیل هم‌زمان دو سطح حقوقی- چارچوب اتحادیه‌ای و اجرای ملی آن- را در کنار نظام حقوق عرفی ایالات متحده که فاقد چنین لایه فراملی است، فراهم کرد. اتحادیه اروپا همچنین به دلیل پیشگام‌بودن در تنظیم‌گری هوش مصنوعی و تصویب مقررات هوش مصنوعی، نظام حقوقی مهمی برای بررسی رویکردهای نوین به داده‌های آموزشی به شمار می‌رود. ایران نیز به‌عنوان نظام

---

1. Springer
2. Internet Archive
2. Google Scholar
3. Emerald
5. Research Gate
6. arxiv
7. Sage pub
8 Common law

حقوقی هدف این پژوهش انتخاب شد تا کفایت مقررات موجود آن در پرتو تجربه سایر کشورها سنجیده شود. در کنار مقررات قانونی، آرای شاخصی مانند Authors Guild v. Google Inc.، The New York Times v. OpenAI and Microsoft و Getty Images v. Stability AI نیز به دلیل تبیین معیارهای حقوقی مرتبط با استفاده از منابع دارای حق مؤلف در محیط‌های داده‌محور مورد مطالعه قرار گرفت. داده‌های گردآوری‌شده با روش تحلیل تطبیقی حقوقی بررسی شدند. تحلیل تطبیقی بر پایه رویکرد کارکردی انجام شد؛ به این معنا که به‌جای مقایسه صرف واژگان قانونی، نحوه پاسخ هر نظام حقوقی به یک مسئله مشترک بررسی و مقایسه شد. معیار مقایسه در سه محور مشخص بود: نخست، وجود یا نبود استثنای صریح متن‌کاوی و داده‌کاوی و شرایط اعمال آن؛ دوم، دامنه و معیارهای استفاده منصفانه یا نهادهای مشابه آن در سایر نظام‌های حقوقی؛ و سوم، رویه قضایی موجود درباره استفاده از داده‌های دارای حق مؤلف در آموزش سامانه‌های هوش مصنوعی. برای هر نظام حقوقی، نخست متن قانونی مرتبط استخراج شد، سپس محتوای آن بر پایه این سه محور بررسی و در نهایت میان نظام‌های حقوقی منتخب مقایسه شد؛ نتیجه این مقایسه، شامل هر سه محور، در جدول ۴ ارائه شده است. به منظور افزایش اعتبار یافته‌ها، از راهبرد مثلث‌سازی داده‌ها استفاده شد. در این فرایند، سه دسته منبع شامل: (۱) اسناد قانونی و حقوقی، (۲) مستندات فنی مرتبط با Books3 و LLaMA2 و (۳) مطالعات علمی و تجربی درباره مدل‌های زبانی بزرگ و امکان بازتولید داده‌های آموزشی، با یکدیگر مقایسه شدند. این مقایسه امکان ارزیابی هم‌خوانی یا تفاوت میان الزامات حقوقی، واقعیت‌های فنی و شواهد پژوهشی را فراهم کرد و مبنای تحلیل‌های ارائه‌شده در بخش یافته‌ها و نتیجه‌گیری قرار گرفت.

٤. یافته‌ها

تا چه اندازه می‌توان بهره‌برداری از مجموعه‌داده Books۳ در فرایند آموزش مدل‌های زبانی بزرگ، با تأکید بر LLaMA۲، را در چارچوب مفاهیمی چون «استفاده منصفانه» یا سایر استثنائات قانونی مشابه توجیه کرد؟

٤-١-١. احراز استفاده از منابع دارای حق مؤلف در فرایند آموزش مدل‌های زبانی بزرگ

پیش از بررسی مشروعیت حقوقی استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ، نخست باید احراز شود که اساساً چنین استفاده‌ای رخ داده است یا خیر. این موضوع از آن جهت اهمیت دارد که صرف امکان نظری استفاده از منابع دارای حق مؤلف در آموزش سامانه‌های هوش مصنوعی، برای انتساب مسئولیت یا ارزیابی مشروعیت حقوقی کافی نیست. توسعه‌دهندگان مدل‌های زبانی ممکن است ادعا کنند که داده‌های آموزشی از منابع عمومی، داده‌های دارای مجوز، داده‌های تولیدشده توسط کاربران یا سایر منابع مجاز تأمین شده‌اند. بنابراین، نخست باید امکان اثبات استفاده از منابع دارای حق مؤلف در فرآیند آموزش مدل بررسی شود (جدول ۲).

جدول ۲. شواهد فنی و حقوقی احراز ورود منابع دارای حق مؤلف به آموزش مدل‌های زبانی بزرگ (تحلیل منابع)

منابع انتخاب‌شده	یافته مستخرج	تحلیل حقوقی
Gao et al. (2020)	Books3 یکی از زیرمجموعه‌های اصلی The Pileاست و مشتمل بر حجم گسترده‌ای از کتاب‌های دیجیتال است.	ورود منابع کتابی دارای حق مؤلف به داده‌های آموزشی قابل احراز است.
Touvron et al. (2023)	گزارش فنی LLaMA2 به استفاده از داده‌های متنی متنوع از جمله منابع کتابی اشاره دارد.	استفاده از متون کتابی در آموزش مدل صرفاً فرضیه نظری نیست.
Karamolegkou et al. (2023)	مدل‌های زبانی در مواردی قادر به بازتولید یا شباهت قابل توجه به داده‌های آموزشی هستند.	رابطه میان داده آموزشی و خروجی مدل قابل بررسی تجربی است.
Kadrey et al. v. Meta	استفاده از Books3 در آموزش مدل‌های خانواده LLaMAموضوع دعوای قضایی قرار گرفته است؛ دادگاه در حکم Kadrey v. Meta Platforms, Inc., No. ۲۳-cv-۰۳۴۱۷-VC	Books3 یکی از نمونه‌های مستند اختلافات حقوقی درباره

	(N.D. Cal. June ۲۵, ۲۰۲۵) استفاده از این آثار برای آموزش مدل را، به دلیل ناتوانی شاکیان در اثبات ضرر بازار، مصداق استفاده منصفانه دانست؛ قاضی تصریح کرد این حکم رویه‌ای عام ایجاد نمی‌کند و به واقعیت‌های همین پرونده محدود است.	استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی است.
Cooper et al. (2025)	پژوهش با استفاده از روش استخراج احتمالاتی حضور کتاب *Harry Potter and the Sorcerer's Stone* موجود در Books3 را در نسخه ۷۰ میلیارد پارامتری مدل (Llma3.1 70B) از خانواده LLaMA بررسی و امکان استخراج متن آن را در شرایط آزمایشی نشان داد.	حضور برخی منابع مشخص در داده‌های آموزشی و ارتباط آن با خروجی مدل، در موارد خاص با روش‌های فنی قابل ارزیابی است؛ در همین مطالعه، LLaMA۲ نیز بررسی شد و میزان محدودتری از بازتولید محتوا را نسبت به Llma3.1 70B نشان داد.

بررسی منابع فنی و حقوقی نشان داد که Books3 در شماری از تحلیل‌ها و منازعات مرتبط با استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ مورد استناد قرار گرفته است. همچنین مدل‌های خانوادهLLaMA ، از جملهLLaMA2 ، در بخشی از این مباحث و منازعات حقوقی مطرح شده‌اند. از این رو، مطالعه موردی Books3 و LLaMA2 امکان بررسی حدود قانونی استفاده از منابع دارای حق مؤلف را در یک مصداق عینی فراهم می‌کند.

٤-١-٢ ماهیت استفاده از اثر در فرایند آموزش مدل: استخراج الگو یا تکثیر اثر؟

این بخش بر پایه بررسی مستندات فنی گردآوری‌شده در همین پژوهش، به‌ویژه گزارش فنی LLaMA۲، تدوین شده است. پس از احراز ورود منابع دارای حق مؤلف به داده‌های آموزشی، گام بعدی تعیین ماهیت حقوقی استفاده از این منابع در مراحل مختلف آموزش مدل است. بررسی مستندات فنی مربوط به معماری مدل‌های زبانی و گزارش فنی LLaMA2 نشان می‌دهد که

آموزش مدل از طریق پردازش حجم گسترده‌ای از داده‌های متنی و تنظیم پارامترهای مدل برای پیش‌بینی توکن بعدی انجام می‌شود (Vaswani et al., 2017; Touvron et al., 2023). در نتیجه این فرایند، آنچه در مدل ذخیره می‌شود متن منابع نیست، بلکه پارامترهایی است که از پردازش داده‌های آموزشی حاصل شده‌اند. با این حال، این توصیف فنی به تنهایی برای ارزیابی حقوقی کافی نیست؛ زیرا در حقوق مؤلف، معیار اصلی ارزیابی، نوع بهره‌برداری از اثر و اعمال انجام‌شده بر آن است، نه صرف هدف یا سازوکار فنی فرایند. ازاین‌رو، در این پژوهش مراحل فنی آموزش مدل بر پایه مستندات فنی، گزارش فنی LLaMA2 و مطالعات تجربی مرتبط بازسازی و سپس از منظر حقوق مؤلف دسته‌بندی و تحلیل شده‌اند. حاصل این دسته‌بندی در جدول ۳ ارائه شده است.

جدول ۳. مراحل فنی آموزش مدل‌های زبانی از منظر حقوق مؤلف (بر پایه مستندات فنی Cooper et al., 2025 و تحلیل پژوهش)

مرحله فرایند	عملیات فنی انجام‌شده	وضعیت منبع دارای حق مؤلف	مسئله حقوقی قابل طرح
گردآوری داده[1]	انتخاب و دریافت داده از پایگاه‌ها و مخازن مختلف	اثر برای نخستین بار وارد چرخه آموزش می‌شود.	مشروعیت منشأ داده و قانونی بودن دسترسی
آماده‌سازی و پاک‌سازی داده[2]	تبدیل قالب، حذف داده‌های تکراری، یکسان‌سازی متن	نسخه‌ای پردازش‌شده از اثر ایجاد می‌شود.	ارتباط با حق تکثیر و ایجاد نسخه دیجیتال
نشانه‌گذاری (توکن‌سازی)[3] و نمایش محاسباتی	تبدیل متن به واحدهای قابل پردازش برای مدل	متن اثر به توکن‌ها و نمایش عددی قابل پردازش برای مدل تبدیل می‌شود.	آیا این تبدیل همچنان مشمول حمایت حق مؤلف است؟

[1] Data Collection
[2] Pre-processing
[3] Tokenization

آموزش مدل[1]	استخراج روابط آماری میان توکن‌ها و به‌روزرسانی وزن‌های مدل	محتوای اثر در فرایند یادگیری آماری استفاده می‌شود.	ماهیت حقوقی استفاده از اثر در آموزش مدل
استنتاج یا تولید خروجی[2]	تولید متن جدید براساس وزن‌های آموخته‌شده	امکان شباهت یا بازتولید بخش‌هایی از منابع وجود دارد.	احتمال نقض حق مؤلف از طریق خروجی مدل

دسته‌بندی مراحل فوق نشان می‌دهد که آموزش مدل‌های زبانی یک فرایند چندمرحله‌ای است و هر مرحله می‌تواند مسئله حقوقی متفاوتی ایجاد کند. بنابراین، مشروعیت آموزش مدل را نمی‌توان فقط بر مبنای نتیجه نهایی یا فرایند یادگیری به طور کلی ارزیابی کرد، بلکه وضعیت حقوقی هر مرحله باید به‌طور مستقل و با توجه به قواعد حق تکثیر، استثنائات متن‌کاوی و داده‌کاوی و سایر محدودیت‌های قانونی بررسی شود. در کنار این تحلیل، شواهد تجربی نیز نشان می‌دهد که مرز میان «استخراج الگو» و «بازتولید اثر» همواره مشخص نیست. Karamolegkou et al. (2023) گزارش کردند که با افزایش اندازه مدل‌های زبانی، احتمال بازتولید بخش‌هایی از داده‌های آموزشی نیز افزایش می‌یابد. همچنین Cooper et al. (2025) نشان دادند که مدل‌های خانواده LLaMAدر برخی موارد قادرند با دریافت محرک‌های بسیار کوتاه، بخش‌های قابل‌توجهی از منابع موجود در داده‌های آموزشی را بازتولید کنند. هرچند این یافته‌ها قابل تعمیم به همه مدل‌ها و همه منابع نیست، اما نشان می‌دهد که تفکیک میان یادگیری آماری و بازتولید لفظی در عمل همواره روشن نیست. از سوی دیگر، بررسی منابع حقوقی نشان می‌دهد که قیاس میان یادگیری انسان و آموزش مدل‌های زبانی نیز محل اتفاق نظر نیست. لطیف‌زاده (۱۴۰۳) با تفکیک میان خلاقیت تحولی و خلاقیت ترکیبی، تفاوت ماهوی میان یادگیری انسان و پردازش آماری مدل‌های هوش مصنوعی را مطرح می‌کند. بر این اساس، استناد به شباهت ظاهری میان یادگیری انسان و مدل‌های زبانی، به‌تنهایی نمی‌تواند مبنای ارزیابی حقوقی استفاده از منابع دارای حق مؤلف قرار گیرد (ص۵-۶).

[1] Training
[2] Inference

٤-١-٣ شناسایی و دسته‌بندی مقررات ناظر بر استفاده از منابع دارای حق مؤلف

برای پاسخ به این پرسش، ابتدا مقررات ملی و بین‌المللی مرتبط با استفاده از منابع دارای حق مؤلف در فرایند آموزش مدل‌های زبانی بزرگ استخراج و دسته‌بندی شد. تحلیل اسناد نشان داد که قواعد قابل اعمال را می‌توان در سه دسته اصلی دسته‌بندی کرد: ۱) قوانین مختص متن‌کاوی و داده‌کاوی؛ ۲) مقررات مربوط به استفاده منصفانه و سایر استثنائات حق مؤلف، و ۳) قواعد عمومی ناظر بر حق تکثیر و بهره‌برداری از اثر.

جدول ٤. دسته‌بندی مقررات مرتبط با استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ

نظام حقوقی / سند	منبع قانونی/ مقرره مرتبط	محدوده مجاز استفاده	شرط اعمال قاعده	برآیند حقوقی برای آموزش مدل‌های زبانی	رویه قضایی مرتبط
اتحادیه اروپا	مواد ۳ و ۴ دستورالعمل ۷۹۰/۲۰۱۹	قانون‌گذار اصل استفاده از منابع برای متن‌کاوی و داده‌کاوی را به عنوان استثنا پذیرفته است.	دسترسی قانونی به اثر؛ امکان حق انصراف صاحب حق در ماده ۴	مشروعیت آموزش مدل وابسته به قانونی‌بودن منشأ داده و رعایت شرایط TDM است.	در منابع بررسی‌شده در این پژوهش، رأی قضایی مشخصی در سطح اتحادیه اروپا شناسایی نشد؛ رویه قضایی موجود درباره این موضوع عمدتاً در نظام حقوقی ایالات متحده شکل گرفته است.
فرانسه	ماده-L122 5-3قانون	استخراج متن و داده در چارچوب استثنای	استفاده پژوهشی و	استفاده از منابع برای آموزش مدل تنها در	در منابع بررسی‌شده در این پژوهش، رأی قضایی

	مالکیت فکری	پژوهشی مجاز شناخته شده است.	دسترسی قانونی	محدوده استثنای مقرر قابل بررسی است.	مشخصی در نظام حقوقی فرانسه شناسایی نشد.
آلمان	مواد ۴۴b و d ۶۰ قانون حق مؤلف آلمان (UrhG)	پردازش خودکار منابع برای داده‌کاوی به‌عنوان استثنای محدود پذیرفته شده است.	دسترسی قانونی و رعایت الزامات نگهداری داده	استفاده از منابع برای آموزش مدل مستلزم احراز شرایط قانونی داده‌کاوی است.	در منابع بررسی‌شده در این پژوهش، رأی قضایی مشخصی در نظام حقوقی آلمان شناسایی نشد.
ایالات متحده	ماده ۱۰۷ قانون حق مؤلف ایالات متحده (U.S.C. §107 17)	استفاده از منابع بدون مجوز به‌طور مطلق مجاز یا ممنوع نیست و براساس آزمون استفاده منصفانه ارزیابی می‌شود.	ارزیابی چهار عامل قانونی	مشروعیت آموزش مدل به نتیجه تحلیل موردی استفاده منصفانه وابسته است.	Authors Guild v. Google Inc.، Kadrey v. Meta Platforms (حکم ۲۵ ژوئن ۲۰۲۵ به نفع Meta در موضوع استفاده منصفانه) و The New York Times v. OpenAI and Microsoft (در جریان)؛ رویه قضایی موجود در این پژوهش عمدتاً در همین نظام حقوقی متمرکز است. (دعوای Getty Images

					v. Stability AI که در این پژوهش نیز بررسی شد، در نظام حقوقی بریتانیا مطرح است و از این‌رو در ردیف ایالات متحده گنجانده نشده است.)
کنوانسیون برن	ماده (۲)۹	اصل بر حق انحصاری تکثیر مؤلف است و استثناها باید محدود تفسیر شوند.	رعایت آزمون سه‌گام	آموزش مدل تنها در صورت انطباق با معیارهای سه‌گانه قابل توجیه است.	در منابع بررسی‌شده در این پژوهش، رأی قضایی مشخصی که مستقیماً به تفسیر آزمون سه‌گام در پرتو آموزش هوش مصنوعی پرداخته باشد شناسایی نشد.
موافقت‌نامه تریپس	ماده ۱۳	استثناهای حقوق مؤلف فقط در موارد خاص قابل پذیرش هستند.	رعایت آزمون سه‌گام	مشروعیت استثناهای مورد استناد باید با معیارهای بین‌المللی سازگار باشد.	مشابه کنوانسیون برن، رأی قضایی مستقل و مشخصی در سطح بین‌المللی درباره این موضوع در منابع بررسی‌شده این پژوهش شناسایی نشد.
ایران	مواد ۳، ۵ و ۱۲ قانون حمایت حقوق مؤلفان،	حق تکثیر، نشر و بهره‌برداری از اثر در اختیار	اخذ اجازه یا وجود مجوز قانونی	مقرره خاصی برای آموزش هوش مصنوعی وجود ندارد و	چنین رویه قضایی در نظام حقوقی ایران هنوز شکل نگرفته است.

	مصنفان و هنرمندان ۱۳۴۸	صاحب حق قرار دارد.		ارزیابی براساس قواعد عمومی حق مؤلف انجام می‌شود.	
ایران	ماده ۱ قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی ۱۳۵۲	حق انحصاری ترجمه، تکثیر، نشر و بهره‌برداری از ترجمه را برای مترجم یا قائم‌مقام قانونی او شناسایی می‌کند.	هرگونه بهره‌برداری مستلزم رعایت حقوق صاحب حق است.	استفاده از کتاب‌های ترجمه‌شده در داده‌های آموزشی بدون مجوز، نیازمند بررسی در پرتو حقوق انحصاری مترجم است.	چنین رویه قضایی در نظام حقوقی ایران هنوز شکل نگرفته است.

یافته‌های حاصل از دسته‌بندی مقررات نشان می‌دهد که هیچ‌یک از نظام‌های حقوقی مورد بررسی، مجوز عام و نامحدودی برای استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ پیش‌بینی نکرده‌اند. در همه نظام‌ها، مشروعیت این فرایند به احراز شرایط مقرر در قواعد حاکم بر همان نظام حقوقی وابسته است. بررسی رویه قضایی مرتبط (جدول ۴) نیز نشان می‌دهد که این رویه تقریباً به‌طور کامل در نظام حقوقی ایالات متحده شکل گرفته و در سایر نظام‌های مورد بررسی این پژوهش، رأی قضایی مشخصی درباره آموزش مدل‌های زبانی شناسایی نشد؛ حتی در همین نظام نیز نتیجه یکسانی برای همه پرونده‌ها گزارش نشده است، چنان‌که رأی Kadrey v. Meta به سود توسعه‌دهنده صادر شد در حالی که پرونده‌های دیگری مانند The New York Times v. OpenAI هنوز در جریان است. همچنین بررسی منابع نشان داد که بهره‌برداری از داده‌های

آموزشی در برخی موارد با حوزه‌هایی مانند حقوق رقابت، حقوق داده و تنظیم‌گری بازارهای دیجیتال نیز تلاقی پیدا می‌کند، هرچند بررسی تفصیلی این حوزه‌ها خارج از دامنه پژوهش حاضر است. بر اساس مقررات استخراج‌شده، وضعیت استفاده از مجموعه‌داده Books3 در آموزش مدل‌های خانواده LLaMA از منظر معیارهای شناسایی‌شده در نظام‌های حقوقی منتخب مورد بررسی قرار گرفت.

٤-١-٤ تطبیق قواعد استخراج‌شده بر وضعیت Books3

پس از استخراج قواعد حقوقی، این قواعد بر وضعیت مجموعه‌داده Books3 بر مبنای همین معیارها مورد ارزیابی قرار گرفت. هدف از این مرحله، تعیین مشروعیت یا عدم مشروعیت آموزش مدل نبود، بلکه بررسی این موضوع بود که مطالعه موردی حاضر در چارچوب هر یک از معیارهای حقوقی استخراج‌شده چگونه قابل ارزیابی است. نتایج این تطبیق نشان می‌دهد که در نظام حقوقی اتحادیه اروپا، قانونی‌بودن دسترسی به منابع مهم‌ترین متغیر در امکان استناد به استثنائات متن‌کاوی و داده‌کاوی است؛ بنابراین، پیش از بررسی سایر شرایط، وضعیت حقوقی منبع تأمین داده‌ها اهمیت پیدا می‌کند. در مقابل، در حقوق آمریکا، ارزیابی بر پایه یک استثنای خاص انجام نمی‌شود، بلکه نتیجه به تحلیل موردی عوامل استفاده منصفانه بستگی دارد. در اسناد بین‌المللی نیز آزمون سه‌گام کنوانسیون برن و موافقت‌نامه تریپس چارچوب کلی سنجش مشروعیت استثنائات را فراهم می‌کند و در حقوق ایران نیز، به دلیل فقدان مقرره اختصاصی درباره آموزش مدل‌های زبانی، ارزیابی ناگزیر بر قواعد عمومی حق مؤلف متکی است.

این یافته‌ها نشان می‌دهد که اختلاف اصلی میان نظام‌های حقوقی، نه در اصل حمایت از حق مؤلف، بلکه در مبنای ارزیابی مشروعیت استفاده از منابع در فرایند آموزش مدل است. به همین دلیل، پاسخ حقوقی نسبت به استفاده از داده‌های آموزشی، پیش از آنکه به فناوری وابسته باشد، به معیارهای حقوقی حاکم بر هر نظام بستگی پیدا می‌کند.

بررسی مطالعه موردی بر پایه معیارهای استخراج‌شده نشان می‌دهد که استناد به استثنائات قانونی در خصوص Books3 و LLaMA2 با ابهام‌هایی روبه‌رو است. در چارچوب آزمون سه‌گام[1]، استفاده از مجموعه‌ای متشکل از منابع کامل، پرسش‌هایی درباره محدود بودن استثنا، تعارض با بهره‌برداری عادی و احتمال ورود لطمه به منافع مشروع صاحبان منابع ایجاد می‌کند. همچنین در حقوق آمریکا، ارزیابی عوامل استفاده منصفانه نیز به دلیل ماهیت تجاری مدل، استفاده از آثار خلاقانه، دامنه بهره‌برداری و آثار احتمالی بر بازار، پاسخ قطعی و از پیش تعیین‌شده‌ای ارائه نمی‌کند و همچنان به بررسی موردی وابسته است.

تحلیل یافته‌ها نشان می‌دهد که در هیچ‌یک از نظام‌های حقوقی بررسی‌شده، آموزش مدل‌های زبانی با استفاده از منابع دارای حق مؤلف، مشروع یا نامشروع مطلق تلقی نشده است. نتیجه حقوقی این فرایند به عواملی مانند منشأ داده، مبنای دسترسی، نوع بهره‌برداری از اثر و شرایط اعمال استثنائات قانونی وابسته است. همچنین، مقررات موجود عمدتاً برای الگوهای سنتی بهره‌برداری از منابع تدوین شده‌اند و در مواجهه با فرایندهای داده‌محور آموزش مدل‌های زبانی، با چالش‌های تفسیری قابل توجهی روبه‌رو هستند.

پاسخ به پرسش دوم. مسئولیت حقوقی تولیدکنندگان و کاربران مدل‌های زبانی بزرگ در قبال نقض احتمالی حقوق مؤلف چگونه قابل تبیین است؟

پس از بررسی حدود قانونی استفاده از منابع دارای حق مؤلف در فرایند آموزش مدل‌های زبانی بزرگ، پرسش دوم به تعیین مسئولیت حقوقی بازیگران دخیل در این فرایند اختصاص دارد. برای پاسخ به این پرسش، ابتدا مقررات مرتبط در نظام‌های حقوقی مورد مطالعه استخراج و دسته‌بندی شد تا مبنای قانونی مسئولیت هر یک از بازیگران مشخص شود. سپس این مقررات بر فرایند توسعه و بهره‌برداری از مدل‌های زبانی بزرگ تطبیق داده شد.

۱. مسئولیت تولیدکنندگان مدل‌های زبانی

[1]. Three-step test

در بررسی مقررات، قواعد مرتبط با مسئولیت در سه دسته قابل دسته‌بندی است: ۱) مسئولیت ناشی از تکثیر و بهره‌برداری از منابع دارای حق مؤلف؛ ۲) مسئولیت ناشی از استفاده از خروجی‌های ناقض و ۳) مسئولیت ناشی از نقض تعهدات قراردادی و شرایط مجوزها.

جدول ۵. ضوابط قانونی مرتبط با مسئولیت در استفاده از منابع دارای حق مؤلف در مدل‌های زبانی

نظام حقوقی	منبع قانونی	مفاد قانونی مرتبط	شخص مسئول
آمریکا	ماده ۱۰۶ قانون حق مؤلف ایالات متحده (U.S.C. §106 17)	حقوق انحصاری صاحب اثر در تکثیر، تهیه منابع اشتقاقی و توزیع اثر	توسعه‌دهندگان و استفاده‌کنندگان از منابع
آمریکا	ماده ۱۰۷ قانون حق مؤلف ایالات متحده (U.S.C. §107 17)	امکان استناد به استفاده منصفانه با احراز شرایط قانونی	توسعه‌دهنده/استفاده‌کننده
اتحادیه اروپا	مواد ۳ و ۴ دستورالعمل ۷۹۰/۲۰۱۹	داده‌کاوی تنها در صورت دسترسی قانونی و رعایت شرایط مقرر مجاز است.	توسعه‌دهندگان/مؤسسات پژوهشی
فرانسه	ماده L122-5-3 قانون مالکیت فکری	داده‌کاوی پژوهشی با دسترسی قانونی به اثر	مؤسسات پژوهشی / توسعه‌دهندگان
آلمان	مواد ۴۴b و ۶۰ d قانون حق مؤلف آلمان(UrhG)	استخراج متن و داده در حدود مقرر قانون	اشخاص و نهادهای مجاز
ایران	مواد ۳، ۵ و ۱۲ قانون حمایت حقوق مؤلفان ۱۳۴۸	شناسایی حقوق انحصاری پدیدآورنده در نشر، تکثیر و بهره‌برداری از اثر	استفاده‌کنندگان از منابع بدون مجوز
ایران	ماده ۲۳ قانون حمایت حقوق مؤلفان ۱۳۴۸	ضمانت اجرای نقض حقوق انحصاری پدیدآورنده	ناقضان حقوق مؤلف

بررسی مقررات جدول ۵ نشان داد که اگرچه مبنای مسئولیت در نظام‌های حقوقی مورد مطالعه متفاوت است، اما در همه آنها نقطه آغاز مسئولیت، انجام اعمالی است که در قلمرو حقوق انحصاری

مؤلف قرار می‌گیرد. این اعمال عمدتاً شامل تکثیر، ذخیره‌سازی، پردازش و بهره‌برداری از منابع دارای حق مؤلف است. تفاوت اصلی نظام‌های حقوقی در حدود استثنائات قانونی و شرایط بهره‌مندی از آنها مشاهده می‌شود. در حقوق ایالات متحده، امکان استناد به استفاده منصفانه براساس ماده ۱۰۷ وجود دارد، در حالی که در اتحادیه اروپا مشروعیت داده‌کاوی به رعایت شرایط مقرر در مواد ۳ و ۴ دستورالعمل ۷۹۰/۲۰۱۹ وابسته است. در حقوق ایران نیز اصل بر لزوم اخذ اجازه از صاحب حق است و استثناء خاصی برای آموزش مدل‌های زبانی با توجه به زمان تصویب قانون پیش‌بینی نشده است.

اعمال مقررات فوق بر فرایند توسعه مدل‌های زبانی نشان داد که بیشترین سطح مسئولیت متوجه توسعه‌کنندگان مدل است؛ زیرا انتخاب داده‌های آموزشی، گردآوری، ذخیره‌سازی و آموزش مدل تحت کنترل آنان انجام می‌شود.

در حقوق ایالات متحده، استفاده از منابع دارای حق مؤلف در آموزش مدل در صورتی که خارج از قلمرو استفاده منصفانه باشد، می‌تواند مشمول حقوق انحصاری موضوع ماده ۱۰۶ قانون کپی‌رایت قرار گیرد. رویه‌های قضایی اخیر نیز نشان می‌دهد که منشأ داده‌های آموزشی در ارزیابی مسئولیت اهمیت اساسی دارد؛ به گونه‌ای که استفاده از داده‌های گردآوری‌شده از منابع فاقد مجوز، می‌تواند استناد به دفاع استفاده منصفانه را با دشواری بیشتری مواجه سازد. در اتحادیه اروپا نیز استثنائات متن‌کاوی و داده‌کاوی تنها در صورتی قابل اعمال است که دسترسی به منابع قانونی بوده و شرایط مقرر در دستورالعمل ۷۹۰/۲۰۱۹ رعایت شده باشد. بنابراین، استفاده از مجموعه‌داده‌هایی که بر پایه منابعی با منشأ قانونی نامشخص تهیه شده است، نمی‌تواند مشمول این استثنائات قرار گیرد. در حقوق ایران نیز، هرچند مقرره‌ای اختصاصی درباره داده‌کاوی و آموزش مدل‌های زبانی بزرگ وجود ندارد، اما مواد ۳، ۵ و ۱۲ قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان ۱۳۴۸، حقوق انحصاری نشر، تکثیر و بهره‌برداری مادی از اثر را برای پدیدآورنده شناسایی کرده است و ماده ۲۳ همان قانون، نقض این حقوق را مستوجب مسئولیت می‌داند. ازاین‌رو، در فرض استفاده از منابع دارای حق مؤلف برای آموزش مدل بدون اخذ مجوز، توسعه‌دهنده مدل بیشترین مواجهه را با مسئولیت حقوقی خواهد داشت. در مطالعه موردی Books3 نیز کانون اصلی اختلاف حقوقی دقیقاً به همین مرحله مربوط

است؛ زیرا ادعاها ناظر بر استفاده از کتاب‌هایی است که بدون رضایت صاحبان حق در مجموعه‌داده آموزشی قرار گرفته‌اند.

۲. مسئولیت کاربران مدل‌های زبانی

بررسی تطبیقی مقررات نشان می‌دهد که جایگاه کاربران مدل‌های زبانی به‌طور ماهوی از توسعه‌دهندگان این ابزارها متمایز است؛ زیرا کاربران، برخلاف توسعه‌دهندگان، نقشی در گردآوری داده‌ها یا فرایند آموزش مدل ایفا نمی‌کنند و ارتباط آن‌ها عمدتاً به سطح بهره‌برداری از خروجی‌های تولیدشده محدود می‌شود. بر این اساس، در نظام‌های حقوقی مورد مطالعه، صرف استفاده از مدل زبانی به‌تنهایی مبنای انتساب مسئولیت تلقی نمی‌شود، بلکه تحقق مسئولیت زمانی قابل طرح است که رفتار کاربر، خود به تنهایی مصداق نقض حق مؤلف باشد؛ از جمله در مواردی که خروجی مدل منتشر شده یا مورد بهره‌برداری قرار گیرد و در نتیجه، حقوق اشخاص ثالث نقض شود.

در همین چارچوب، Rosati (2024) تصریح می‌کند که کاربران نهایی مدل‌های زبانی، در صورت استفاده از خروجی‌هایی که حقوق مؤلفان شخص ثالث را نقض می‌کنند، ممکن است با مسئولیت حقوقی مواجه شوند. با این حال، این مسئولیت در نظام‌های حقوقی بین‌المللی، به‌ویژه در اتحادیه اروپا، غالباً به‌عنوان مسئولیت تبعی یا ثانویه تفسیر می‌شود و هدف آن محدودسازی دامنه مسئولیت کاربران به مواردی است که آنان در مقام استفاده‌کننده از منابع ناقض ظاهر می‌شوند، نه به‌عنوان عاملان اصلی تولید نقض.

در حقوق ایران نیز معیار اصلی، تحقق رفتار ناقض است. به موجب ماده ۲۳ و سایر مقررات مرتبط، از جمله قوانین ناظر بر جرائم رایانه‌ای، حمایت از نرم‌افزارهای رایانه‌ای و تجارت الکترونیکی، هرگونه استفاده از منابع دارای حق مؤلف بدون رضایت مالک می‌تواند موجب مسئولیت مدنی و کیفری شود. بر این اساس، صرف استفاده از ابزارهای هوش مصنوعی فی‌نفسه موجب مسئولیت

کاربر نخواهد بود، مگر آنکه نحوه استفاده وی مصداق بهره‌برداری غیرمجاز از منابع حمایت‌شده تلقی شود. افزون بر این، حتی در فرض فقدان دخالت مستقیم در فرایند نقض، احراز آگاهی کاربر از ماهیت ناقض اثر و استمرار در استفاده بدون مجوز، می‌تواند مبنای انتساب مسئولیت قرار گیرد. بنابراین، مسئولیت کاربران، در قیاس با توسعه‌دهندگان مدل، دارای دامنه‌ای محدودتر و ماهیتی غالباً تبعی است و بیش از هر عامل دیگری، به نحوه استفاده از خروجی‌های مدل و سطح آگاهی کاربر از نقض حقوق مؤلف وابسته است؛ ازاین‌رو، صرف استفاده از مدل زبانی، بدون تحقق عناصر مادی و معنوی نقض، برای انتساب مسئولیت کفایت نمی‌کند.

۳. مسئولیت قراردادی و مجوزهای استفاده از داده‌ها

تحلیل مقررات نشان می‌دهد که در حوزه مدل‌های زبانی، مجوزهای قراردادی نه به‌عنوان مبنای اولیه ایجاد مسئولیت، بلکه به‌عنوان ابزاری برای تنظیم و تعدیل آن عمل می‌کنند. به بیان دیگر، برخلاف مسئولیت‌های مستقیم ناشی از نقض حق مؤلف که در بندهای پیشین بررسی شد، در اینجا قراردادها مشخص می‌کنند که آیا استفاده از داده‌ها اصولاً مشروع است و در صورت بروز نقض، مسئولیت چگونه میان اشخاص توزیع می‌شود.

در حقوق اتحادیه اروپا، دستورالعمل 2019/790 (EU)، به‌ویژه در مواد ۳ و ۴، استفاده از منابع برای «داده‌کاوی و متن‌کاوی» را مشروط به دسترسی قانونی و عدم منع از سوی دارنده حق کرده است. این امر نشان می‌دهد که اراده قراردادی مالک اثر می‌تواند دامنه استفاده مجاز را گسترش دهد یا محدود کند. در حقوق فرانسه ماده L122-5-3 و آلمان مواد 44b و ۶۰d، نیز، اگرچه استثنائات محدودی برای استفاده بدون مجوز پیش‌بینی شده، اما خارج از این موارد، اصل بر لزوم اخذ مجوز است. بنابراین، قرارداد در این نظام‌ها نقش کلیدی در تبدیل استفاده غیرمجاز به استفاده مشروع دارد.

در حقوق ایالات متحده، با وجود نهاد «استفاده منصفانه»، حقوق انحصاری مقرر در ماده ۱۰۶ قانون کپی‌رایت همچنان مبنای اصلی است. در عمل، موافقت‌نامه‌های کاربری و مجوزهای بهره‌برداری تعیین می‌کنند که داده‌ها تا چه حد قابل استفاده هستند. از این رو، قراردادها علاوه بر کاهش خطر نقض، چارچوبی برای تقسیم مسئولیت میان نقش‌آفرینان مختلف، از جمله ارائه‌دهندگان خدمات و

کاربران، فراهم می‌کنند. در حقوق ایران نیز، ماده ۲۳ بر لزوم اخذ اجازه از پدیدآورنده تأکید دارد، در حالی که ماده ۶ نقش قرارداد را در تعیین مالکیت و به تبع آن، مسئولیت‌ها به رسمیت می‌شناسد. همچنین، قواعد عمومی قراردادها در امکان تعیین حدود استفاده، پیش‌بینی مسئولیت و توزیع خطر را فراهم می‌آورد. در نتیجه، قرارداد می‌تواند استفاده از داده‌ها را از حالت غیرمجاز خارج کرده و آن را در چارچوبی قانونی قرار دهد.

بر این اساس، نقش اصلی مجوزهای قراردادی در این حوزه، «مدیریت خطر حقوقی» از طریق مشروع‌سازی استفاده و تعیین حدود مسئولیت است، نه ایجاد مستقیم مسئولیت. این رویکرد، جایگاه قرارداد را به‌عنوان یک ابزار تنظیمی مؤثر در نظام حقوقی داده‌ها و فناوری‌های مبتنی بر هوش مصنوعی نشان می‌دهد.

پاسخ به پرسش سوم. با توجه به آراء، رویه‌های قضایی و دیدگاه‌های حقوقی موجود در نظام‌های حقوقی پیشرو، چه اصول یا ملاحظات کلیدی می‌توان برای سیاست‌های حقوقی در ایران استخراج کرد؟

بررسی رویه‌های قضایی و دیدگاه‌های حقوقی نشان می‌دهد که تاکنون هیچ نظام حقوقی، قاعده‌ای واحد و قطعی درباره مشروعیت استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ ارائه نکرده است. اختلافات مطروحه در دادگاه‌ها نیز عمدتاً حول این پرسش شکل گرفته‌اند که آیا استفاده از منابع در مرحله آموزش مدل، صرفاً یک فرایند فنی و تحلیلی محسوب می‌شود یا آنکه بهره‌برداری اقتصادی از منابع دارای حق مؤلف است و مستلزم اخذ اجازه از صاحبان حقوق خواهد بود. به همین دلیل، مهم‌ترین معیارهایی که در رویه‌های قضایی مورد توجه قرار گرفته‌اند عبارت‌اند از: منشأ قانونی داده‌های آموزشی، هدف و ماهیت استفاده، آثار اقتصادی استفاده بر بازار اثر و میزان شباهت خروجی‌های مدل با آثار اصلی.

نخستین رأی اثرگذار در این زمینه، رأی Authors Guild v. Google Inc. است. در این پرونده، دادگاه تجدیدنظر حوزه دوم ایالات متحده، اسکن میلیون‌ها کتاب برای ایجاد موتور

جستجوی Google Books را مشمول دکترین استفاده منصفانه دانست. استدلال دادگاه بر این مبنا بود که استفاده گوگل، ماهیتی «تحول‌آفرین» دارد؛ زیرا هدف آن جایگزینی کتاب‌ها یا بهره‌برداری از محتوای آن‌ها نبود، بلکه ایجاد امکان جستجو و بازیابی اطلاعات محسوب می‌شد و همچنین لطمه قابل توجهی به بازار اصلی منابع وارد نمی‌کرد. هرچند موضوع این پرونده آموزش مدل‌های زبانی نبود، اما از آنجا که در هر دو مورد منابع دارای حق مؤلف برای پردازش ماشینی مورد استفاده قرار می‌گیرند، معیارهای مطرح‌شده درباره تحول‌آفرین بودن استفاده و تأثیر آن بر بازار اثر می‌تواند در تحلیل حقوقی استفاده از مجموعه‌داده Books3 نیز مورد توجه قرار گیرد. با این حال، تفاوت اساسی آن است که Google Books صرفاً امکان جست‌وجوی متن را فراهم می‌کرد، در حالی که مدل‌های زبانی مانند LLaMA2 قادر به تولید متن جدید هستند و همین ویژگی، ارزیابی حقوقی آن‌ها را پیچیده‌تر می‌کند.

با این حال، دعاوی جدید نشان می‌دهد که دادگاه‌ها دیگر صرف استفاده از منابع در مرحله آموزش را موضوع اختلاف نمی‌دانند، بلکه آثار خروجی مدل‌ها نیز به کانون توجه تبدیل شده است. در پرونده The New York Times v. OpenAI and Microsoft، روزنامه نیویورک تایمز مدعی شد که میلیون‌ها مقاله خبری بدون مجوز در آموزش مدل‌های زبانی استفاده شده و مدل‌ها در برخی موارد قادر به بازتولید بخش‌های قابل توجهی از این مقالات هستند؛ امری که می‌تواند جایگزین اقتصادی منابع اصلی شود و به بازار منابع و اعطای مجوز آن‌ها لطمه وارد کند. هرچند این دعوا هنوز به رأی نهایی منتهی نشده است، اما اهمیت آن در این است که برای نخستین بار رابطه میان فرایند آموزش مدل، خروجی تولیدشده و آثار اقتصادی آن بر بازار اثر را به صورت مستقیم در معرض رسیدگی قضایی قرار داده است. این پرونده نیز از جهت موضوع با مطالعه موردی Books3 شباهت قابل توجهی دارد؛ زیرا در هر دو مورد منشأ اختلاف، استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی است. تفاوت آن است که در پرونده نیویورک تایمز، علاوه بر استفاده از داده‌های آموزشی، قابلیت بازتولید بخش‌هایی از آثار در خروجی مدل نیز به عنوان مبنای ادعای نقض حق مؤلف مطرح شده است.

در بریتانیا نیز دعوای Getty Images v. Stability AI رویکرد متفاوتی را آشکار کرده است. در این پرونده، استفاده از میلیون‌ها تصویر دارای حق مؤلف برای آموزش مدل هوش مصنوعی بدون

اخذ مجوز مورد اعتراض قرار گرفت. برخلاف نظام حقوقی آمریکا که امکان استناد به استفاده منصفانه را پیش‌بینی کرده است، در حقوق بریتانیا چنین دفاع گسترده‌ای وجود ندارد و همین موضوع موجب شده است که مشروعیت استفاده از منابع دارای حق مؤلف برای آموزش مدل‌های هوش مصنوعی با محدودیت بیشتری مواجه باشد. اهمیت این پرونده برای پژوهش حاضر در آن است که همانند Books3، موضوع اختلاف به استفاده از منابع دارای حق مؤلف در مرحله آموزش سامانه هوش مصنوعی، بدون اخذ مجوز از صاحبان حقوق، مربوط می‌شود. هرچند موضوع پرونده حاضر تصاویر و مطالعه موردی این پژوهش کتاب‌هاست، اما پرسش حقوقی اصلی در هر دو یکسان است: آیا استفاده از منابع دارای حق مؤلف در مرحله آموزش مدل، بدون اخذ مجوز، با قواعد حقوق مؤلف سازگار است یا خیر.

در مجموع، آراء و دیدگاه‌های حقوقی بررسی‌شده نشان می‌دهد که هنوز معیار واحد و تثبیت‌شده‌ای درباره استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ شکل نگرفته است و ارزیابی هر پرونده همچنان به اوضاع و احوال خاص آن، از جمله منشأ داده‌های آموزشی، هدف استفاده و آثار اقتصادی آن وابسته است. این وضعیت با دیدگاه‌های مطرح‌شده در ادبیات پژوهش نیز همخوانی دارد؛ به‌گونه‌ای که Gervais، Lucchi و Rosati نیز ناکافی‌بودن قواعد سنتی حقوق مؤلف برای پاسخ‌گویی به چالش‌های مدل‌های زبانی بزرگ را مورد تأکید قرار داده و بر ضرورت تدوین چارچوب‌های حقوقی متناسب با این فناوری تأکید کرده‌اند.

در حقوق ایران نیز تاکنون رویه قضایی منتشرشده‌ای که به استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ پرداخته باشد، شناسایی نشد. ازاین‌رو، امکان استخراج معیارهای قضایی برای ارزیابی مشروعیت استفاده از داده‌های آموزشی یا تعیین حدود مسئولیت توسعه‌دهندگان وجود ندارد. در نتیجه، تحلیل حقوقی مطالعه موردی Books3 در این پژوهش بر پایه مقررات موجود حقوق مؤلف، اصول کلی حقوقی و بهره‌گیری از معیارها و استدلال‌های شکل‌گرفته در رویه‌های قضایی سایر نظام‌های حقوقی انجام شده است؛ زیرا در وضعیت فعلی، این رویه‌ها نزدیک‌ترین مبانی قابل استناد برای تحلیل این موضوع در حقوق ایران به شمار می‌آیند.

پاسخ به پرسش چهارم. چه الزامات یا راهکارهای حقوقی می‌تواند برای کشورهایی نظیر ایران پیشنهاد شود تا استفاده از داده‌های آموزشی به‌گونه‌ای مشروع و منصفانه قانون‌مند شوند؟

بررسی تطبیقی نظام‌های حقوقی مورد مطالعه نشان داد که سه عنصر در مواجهه با چالش‌های حقوقی آموزش مدل‌های زبانی بزرگ نقش کلیدی دارند و می‌توانند الگویی برای نظام‌های حقوقی فاقد مقرره اختصاصی، از جمله ایران، فراهم کنند: نخست، وجود یک استثنای صریح و شفاف برای متن‌کاوی و داده‌کاوی (مانند مواد ۳ و ۴ دستورالعمل ۷۹۰/۲۰۱۹ اتحادیه اروپا) که مرز مشروعیت استفاده از داده‌های آموزشی را از پیش تعیین می‌کند؛ دوم، الزامات شفافیت و مستندسازی منشأ داده‌های آموزشی، که در نبود آن، همان‌طور که بررسی موردی Books3 نشان داد، احراز مسئولیت حقوقی دشوار می‌شود؛ و سوم، وجود رویه قضایی که بتواند معیارهای کلی قانون را در پرونده‌های عینی تفسیر کند، امری که رویه قضایی ایران در حال حاضر فاقد آن است. جزئیات راهکارهای حقوقی پیشنهادی برپایه این سه عنصر، در بخش نتیجه‌گیری و پیشنهادها به‌تفصیل ارائه شده است.

۵. نتیجه‌گیری و پیشنهادها

این پژوهش با هدف بررسی وضعیت حقوقی بهره‌برداری از داده‌های دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ، با تمرکز بر مجموعه‌داده Books3 و مدل LLaMA2 ، انجام شد. بررسی تطبیقی قوانین ایران، ایالات متحده، اتحادیه اروپا، فرانسه و آلمان، همراه با تحلیل اسناد بین‌المللی، رویه‌های قضایی و مطالعات علمی، نشان داد که مهم‌ترین چالش حقوقی در این حوزه صرفاً به مشروعیت یا عدم مشروعیت استفاده از منابع دارای حق مؤلف محدود نمی‌شود، بلکه پیش از آن، به امکان احراز منشأ داده‌های آموزشی، قابلیت اثبات نحوه استفاده از منابع و تعیین مسئولیت اشخاص دخیل در فرایند آموزش و بهره‌برداری از مدل‌های زبانی مربوط است. در عمل، هرچه فرایند گردآوری و پردازش داده‌ها شفاف‌تر و قابل ردیابی باشد، ارزیابی حقوقی مشروعیت آموزش مدل و رسیدگی به دعاوی احتمالی نیز با دقت بیشتری امکان‌پذیر خواهد بود.

بررسی تطبیقی مقررات نشان داد که هیچ‌یک از نظام‌های حقوقی مورد مطالعه استفاده نامحدود از منابع دارای حق مؤلف برای آموزش مدل‌های زبانی را مجاز نمی‌دانند، اما حدود و شرایط

بهره‌برداری در آنها متفاوت است. در ایالات متحده، مشروعیت استفاده همچنان به ارزیابی موردی بر اساس دکترین استفاده منصفانه وابسته است، در حالی که در اتحادیه اروپا، مشروعیت داده‌کاوی و متن‌کاوی از طریق استثنای هماهنگ مقرر در دستورالعمل ۷۹۰/۲۰۱۹ تعیین شده و فرانسه و آلمان، به‌عنوان کشورهای عضو، همین چارچوب را با تفاوت‌هایی در قوانین ملی خود اجرا کرده‌اند؛ در هر دو سطح، مشروعیت به وجود دسترسی قانونی، رعایت شرایط مقرر در قانون و در برخی موارد امکان اعمال حق انصراف بستگی دارد. این یافته با تحلیل‌های Sobel (2017) ، Gervaisو همکاران(2021) ، Lucchi (2023) و Manteghi (2024) همخوانی دارد که نشان داده‌اند قواعد سنتی حق مؤلف، بدون بازنگری، پاسخگوی چالش‌های ناشی از آموزش مدل‌های زبانی بزرگ نیستند. در مقابل، بررسی حقوق ایران نشان داد که قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان مصوب ۱۳۴۸، با وجود حمایت از حقوق انحصاری پدیدآورندگان، مقرره مستقلی درباره داده‌کاوی، متن‌کاوی یا آموزش سامانه‌های هوش مصنوعی پیش‌بینی نکرده است. این نتیجه با یافته‌های قاسمیان (۱۴۰۲) نیز هماهنگ است که خلأ تقنینی حقوق ایران در این زمینه را مورد تأکید قرار داده است.

از سوی دیگر، بررسی انجام‌شده در سامانه ملی آرای قضایی و سایر منابع منتشرشده نشان داد که تاکنون آراء یا رویه قضایی منتشرشده‌ای که مستقیماً به استفاده از منابع دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ یا مجموعه‌داده‌های آموزشی پرداخته باشد، در دسترس نیست. آرای موجود عمدتاً به موضوعاتی مانند تکثیر، ترجمه، نشر یا بهره‌برداری غیرمجاز از منابع اختصاص دارند و هنوز اختلافی که آموزش مدل‌های زبانی را موضوع رسیدگی قرار داده باشد، مشاهده نشد. ازاین‌رو، تحلیل حاضر ناگزیر بر پایه قوانین موجود، اصول عمومی حقوق مؤلف و مطالعه تطبیقی رویه‌های خارجی متمرکز شد. این وضعیت نشان می‌دهد که ابهام موجود تنها ناشی از سکوت قانون نیست، بلکه در سطح رویه قضایی نیز هنوز تفسیر منسجمی درباره این فناوری شکل نگرفته است.

یکی دیگر از یافته‌های مهم پژوهش آن بود که در تحلیل مسئولیت حقوقی نمی‌توان فرایند توسعه مدل‌های زبانی را یک فعالیت واحد تلقی کرد. بررسی موردی Books3 نشان داد که دسترسی به داده، آموزش مدل و تولید خروجی، هر یک آثار حقوقی متفاوتی دارند و مسئولیت اشخاص نیز متناسب با نقش آنان در هر مرحله قابل ارزیابی است. این نتیجه با مطالعات Feretzakis و همکاران

(۲۰۲۵) و Zhang و همکاران (۲۰۲۵) که بر ضرورت شفافیت چرخه داده، قابلیت ردیابی منابع آموزشی و پاسخگویی الگوریتمی تأکید کرده‌اند، همسو است؛ با این تفاوت که پژوهش حاضر نشان می‌دهد منشأ بسیاری از اختلافات حقوقی، پیش از مرحله تولید خروجی و در مرحله گردآوری و استفاده از داده‌های آموزشی شکل می‌گیرد. همچنین یافته‌های تجربی Karamolegkou و همکاران (۲۰۲۳) درباره افزایش احتمال بازتولید منابع دارای حق مؤلف در مدل‌های بزرگ‌تر، مؤید این نتیجه است که نحوه انتخاب و استفاده از داده‌های آموزشی در ارزیابی مسئولیت حقوقی نقشی اساسی دارد.

بررسی پژوهش‌های داخلی نیز نشان داد که اگرچه پژوهش‌های اخیر ابعاد مختلف حقوقی هوش مصنوعی، از جمله مالکیت منابع، شخصیت حقوقی، مسئولیت مدنی و مبانی مالکیت دیجیتال را بررسی کرده‌اند، اما موضوع استفاده از داده‌های دارای حق مؤلف در آموزش مدل‌های زبانی بزرگ همچنان کمتر به صورت مستقیم مطالعه شده است. در این میان، پژوهش‌های پاپی (۱۴۰۳)، لطیف‌زاده (۱۴۰۳)، ذاکری‌نیا (۱۴۰۲)، تخشید (۱۴۰۰)، احسانگر و یزدانیان (۱۴۰۴)، باقری (۱۴۰۴) و صادقی و ملک‌زاده (۱۴۰۴) بیشتر بر مسائل مربوط به مالکیت، مسئولیت و جایگاه حقوقی هوش مصنوعی متمرکز بوده‌اند؛ در حالی که پژوهش حاضر با تمرکز بر وضعیت حقوقی داده‌های آموزشی، تلاش کرده است حلقه واسط میان قواعد سنتی حقوق مؤلف و چالش‌های عملی آموزش مدل‌های زبانی بزرگ را تبیین کند.

در مجموع، یافته‌های این پژوهش نشان می‌دهد که حل تعارض میان توسعه مدل‌های زبانی بزرگ و حمایت از حقوق مؤلفان،، مستلزم طراحی سازوکارهایی است که امکان بهره‌برداری مشروع، شفاف و قابل نظارت از داده‌های آموزشی را فراهم کند. بنابراین، بر پایه یافته‌های پژوهش کنونی، مجموعه‌ای از پیشنهادهای تقنینی و اجرایی برای کاهش ابهام‌های موجود و تنظیم بهره‌برداری از داده‌های آموزشی در توسعه مدل‌های زبانی بزرگ ارائه می‌شود:

۱. ایجاد الزامات قانونی برای شناسایی، مستندسازی و قابلیت ردیابی داده‌های آموزشی

یافته‌های پژوهش نشان داد که مهم‌ترین مانع رسیدگی حقوقی در پرونده‌هایی مانندBooks3، اثبات منشأ داده‌های آموزشی است. بنابراین، پیش از بحث درباره مسئولیت یا مجاز بودن آموزش مدل، باید امکان شناسایی داده‌های مورد استفاده وجود داشته باشد. به همین دلیل پیشنهاد می‌شود توسعه‌دهندگان مدل‌های زبانی مکلف شوند:

- مجموعه‌داده‌های مورد استفاده در آموزش مدل را ثبت کنند؛

- منشأ هر مجموعه‌داده را مشخص نمایند؛

- نحوه دسترسی به داده‌ها (مجوز، مالکیت عمومی یا سایر وضعیت‌های حقوقی) را مستند کنند؛

- این اطلاعات را برای مراجع صالح در صورت بروز اختلاف نگهداری کنند.

۲. بازنگری پیش‌نویس لایحه حمایت از حقوق مالکیت ادبی و هنری و حقوق مرتبط و اختصاص مواد قانونی برای داده‌کاوی و آموزش هوش مصنوعی

تحلیل انجام‌شده در این پژوهش نشان داد که قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان مصوب ۱۳۴۸ در زمانی تصویب شده که موضوعاتی مانند داده‌کاوی، متن‌کاوی و آموزش مدل‌های زبانی بزرگ اساساً وجود نداشته است؛ ازاین‌رو، بسیاری از مفاهیم و سازوکارهای مورد نیاز برای مواجهه با این فناوری‌ها در قانون پیش‌بینی نشده است. این شکاف فقط به استفاده از داده‌های آموزشی محدود نمی‌شود، بلکه در موضوعاتی مانند حدود استفاده آموزشی، اقتباس دیجیتال، بهره‌برداری از نسخه‌های الکترونیکی و استثناءهای پژوهشی مانند داده‌کاوی و متن‌کاوی نیز مشاهده می‌شود.

بررسی تطبیقی انجام‌شده نشان داد که چارچوب اتحادیه اروپا، که فرانسه و آلمان نیز آن را در قوانین ملی خود اجرا کرده‌اند، بدون خروج کامل منابع از قلمرو حمایت، شرایط مشخصی را برای متن‌کاوی و داده‌کاوی پیش‌بینی کرده است. وجه مشترک این مقررات، چه در سطح اتحادیه‌ای و چه در سطح ملی، آن است که استفاده از منابع را به دسترسی قانونی، رعایت شرایط مقرر و حفظ حقوق صاحبان آثار منوط کرده‌اند.

بر همین اساس، پیشنهاد می‌شود در جریان بررسی لایحه حمایت از حقوق مالکیت ادبی و هنری و حقوق مرتبط، مقرره‌ای مستقل درباره استفاده از منابع در آموزش سامانه‌های هوش مصنوعی پیش‌بینی شود. این مقرره می‌تواند حداقل بر چهار اصل متمرکز باشد: ۱) استفاده از منابع صرفاً در صورت دسترسی قانونی؛ ۲) تفکیک میان استفاده پژوهشی و بهره‌برداری تجاری؛ ۳) الزام به مستندسازی منابع داده و امکان ردیابی آن‌ها؛ ۴) پیش‌بینی حقوق و سازوکارهای لازم برای صاحبان آثار در صورت استفاده از منابع‌شان در آموزش مدل‌ها.

چنین رویکردی ضمن حفظ حقوق مؤلفان، می‌تواند ابهام موجود در وضعیت حقوقی آموزش مدل‌های زبانی را نیز کاهش دهد.

۳. طراحی سازوکارهای مجوزدهی برای استفاده از منابع در آموزش مدل‌های زبانی

یکی از یافته‌های این پژوهش آن بود که بخش مهمی از اختلافات حقوقی پیرامون مجموعه‌داده Books3ناشی از فقدان سازوکاری مشخص برای اخذ مجوز استفاده از منابع بوده است. در چنین شرایطی، توسعه‌دهندگان مدل‌های زبانی ناچار به استفاده از مجموعه‌داده‌هایی شده‌اند که وضعیت حقوقی آن‌ها محل اختلاف است.

بررسی تجربه اتحادیه اروپا و دو نمونه از اجرای ملی آن در فرانسه و آلمان نشان داد که توسعه فناوری هوش مصنوعی در عمل بیش از آنکه بر حذف حقوق مؤلف متکی باشد، مبتنی بر تنظیم روابط میان صاحبان آثار و استفاده‌کنندگان است. از همین رو، ایجاد نظام‌های مجوزدهی، به‌ویژه در قالب نهادهای مدیریت جمعی، می‌تواند راهکاری عملی برای کاهش تعارض میان توسعه فناوری و حمایت از حقوق مؤلف را فراهم کند. در حقوق ایران نیز می‌توان ایجاد نهادی برای مدیریت جمعی حقوق مربوط به بهره‌برداری دیجیتال از منابع را مورد توجه قرار داد تا توسعه‌دهندگان مدل‌های زبانی بتوانند از طریق یک مرجع مشخص، مجوز استفاده از منابع را دریافت کنند و صاحبان آثار نیز از منافع اقتصادی بهره‌برداری از منابع خود منتفع شوند. در همین راستا، شورای عالی انقلاب فرهنگی نیز با تصویب ماده‌واحده «تأسیس مؤسسات مدیریت جمعی حقوق مالکیت ادبی و هنری»

در سال ۱۴۰۲، گامی در جهت ایجاد بستر حقوقی این نهادها برداشته است که می‌تواند زمینه مناسبی برای توسعه سازوکارهای مجوزدهی در حوزه هوش مصنوعی را نیز فراهم سازد.

٤. تعیین معیارهای قانونی برای حدود بهره‌برداری از منابع در آموزش هوش مصنوعی

بررسی انجام‌شده نشان داد که یکی از عوامل اصلی اختلاف‌های حقوقی در پرونده‌هایی مانند Books3، نبود معیار مشخص برای تعیین حدود استفاده مجاز از منابع دارای حق مؤلف است. در نتیجه، تشخیص اینکه استفاده از منابع در چه سطحی همچنان در چارچوب بهره‌برداری مجاز قرار می‌گیرد، با ابهام روبه‌رو می‌شود.

تجربه برخی نظام‌های حقوقی، از جمله آلمان، نشان می‌دهد که تعیین حدود مشخص برای استفاده آموزشی و پژوهشی می‌تواند از بروز بسیاری از اختلافات جلوگیری کند. بنابراین، پیشنهاد می‌شود قانون‌گذار در صورت تنظیم مقررات مربوط به آموزش هوش مصنوعی، معیارهایی مشخص برای میزان مجاز استفاده از منابع، با تفکیک میان استفاده پژوهشی و استفاده تجاری، پیش‌بینی کند تا هم حقوق مؤلفان حفظ شود و هم امکان توسعه مسئولانه فناوری فراهم شود.

بخشی از پژوهش کنونی برگرفته از طرح پژوهشی مصوب شورای پژوهشی سازمان اسناد و کتابخانه ملی ایران با عنوان "جایگاه استثناهای پژوهشی داده‌کاوی و متن‌کاوی در نظام‌های حقوق مالکیت ادبی و هنری به منظور ارائه پیشنهادهایی برای اصلاح قوانین موضوعه ایران و بهره‌برداری در کتابخانه‌ها" است که در بهمن ۱۴۰۳ مصوب شده است.

فهرست منابع

احسانگر، نورا و علیرضا یزدانیان. ۱۴۰۴. تحلیل مسئولیت مدنی ناشی از عملکرد هوش مصنوعی با معیار مالکیت یا نگهداری با مطالعه تطبیقی در حقوق فرانسه. *فصلنامه علمی پژوهش حقوق خصوصی*، ۱۴(۵۳): ۵۳-۹۰.

باقری، پرویز. ۱۴۰۴. شخصیت حقوقی و مسئولیت مدنی هوش مصنوعی؛ چالش‌ها و راهکارهای حقوقی. *فصلنامه علمی پژوهش حقوق خصوصی*، ۱۳(۵۱): ۴۷-۸۲.

پاپی، زینب. ۱۴۰۳. چه کسی مالک آثار تولیدشده توسط هوش مصنوعی است؟ با تحلیلی بر آرای قضایی برخی از کشورها. هشتمین کنگره متخصصان علوم اطلاعات، تهران، ایران.

تخشید، زهرا. ۱۴۰۰. مقدمه‌ای بر چالش‌های هوش مصنوعی در حوزه مسئولیت مدنی. *حقوق خصوصی*، ۱۸(۱): ۲۲۷-۲۵۰.

ذاکری‌نیا، حانیه. ۱۴۰۲. ماهیت و مبنای مسئولیت مدنی ناشی از هوش مصنوعی در حقوق ایران و کشورهای اتحادیه اروپا. *حقوق خصوصی*، ۲۰(۱): ۱۳۵-۱۵۲ http://doi.org/10.22059/jolt.2023.356703.1007186

صادقی، محسن و امیر ملک‌زاده قلعه. ۱۴۰۴. بازاندیشی مبانی مالکیت خصوصی در بستر دارایی‌های دیجیتال. *دوفصلنامه تحقیق و توسعه در حقوق خصوصی*، زودآیند.
https://jpl.illrc.ac.ir/article_732397_aef545fbf8fe872592d6046e7033b1f6.pdf

قاسمیان، سعیده. ۱۴۰۲. ابعاد و چالش‌های مالکیت فکری استفاده از داده‌های آموزشی در یادگیری ماشین (پایان‌نامه کارشناسی ارشد، دانشکده حقوق، دانشگاه شهید بهشتی). تهران، ایران

قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی ۱۳۵۲. https://rc.majlis.ir/fa/law/show/97295

قانون جرائم رایانه‌ای ۱۳۸۸ https://rc.majlis.ir/fa/law/show/138825

قانون حمایت حقوق مولفان، مصنفان و هنرمندان ۱۳۴۸. https://rc.majlis.ir/fa/law/show/96427

لطیف‌زاده، مهدیه. ۱۴۰۳. پیامدهای حقوقی نقش‌آفرینی خلاقیت هوش مصنوعی مولد در خلق آثار ادبی و هنری. *حقوق خصوصی*، ۲۱(۱): ۱۳۹-۱۵۳ http://doi.org/10.22059/jolt.2024.374825.1007290

References

Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG) (2021).

https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/de/de236en.pdf (accessed Mar. 03, 2024)

Berne Convention for the Protection of Literary and Artistic Works (1979).
https://www.wipo.int/wipolex/en/treaties/textdetails/12807 (accessed Mar. 03, 2024)

Code de la propriété intellectuelle (2022).
https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414/2022-06-07
(accessed Mar. 03, 2024)

Cooper, A. F., Lemley, M. A., Casasola, A., Ahmed, A., Gokaslan, A., Cyphert, A. B., De Sa, C., Ho, D. E., & Liang, P. 2025. Extracting memorized pieces of (copyrighted) books from open-weight language models. arXiv preprint arXiv:2505.12546.

Copyright Law United States OF THEand Related Laws Contained in Title 17 of the United States Code, December 2022.

European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market. Official Journal of the European Union, L 130: 92–125. https://eur-lex.europa.eu/eli/dir/2019/790/oj (accessed Aug. 16, 2025)

Feretzakis, G., Vagena, E., Kalodanis, K., Peristera, P., Kalles, D., & Anastasiou, A. 2025. GDPR and Large Language Models: Technical and Legal Obstacles. Future Internet, 17(4): 151. https://doi.org/10.3390/fi17040151 (accessed Jun. 29, 2025)

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. 2020. The pile: An 800GB dataset of diverse text for language modeling. arXiv. https://arxiv.org/abs/2101.00027 (accessed Jun. 12, 2025)

Gervais, D. 2021. The Machine as Author. *Iowa Law Review*, *106*(5): 2053-2110. https://ilr.law.uiowa.edu/print/volume-106-issue-5/the-machine-as-author/ (accessed Jun. 8, 2025)

Getty Images (UK) Ltd. & Others v. Stability AI Ltd., [2023] EWHC 3090 (Ch). Available at: https://www.judiciary.uk/judgments/getty-images-and-others-v-stability-ai/ (accessed Jun. 8, 2025)

Ghasemian, S. 2023. Dimensions and Challenges of Intellectual Property in the Use of Training Data in Machine Learning (Master's Thesis, Faculty of Law, Shahid Beheshti University). Tehran, Iran. (In Persian).

Hugging Face, Inc. 2023. *meta-llama: LLaMA 2 models*. https://huggingface.co/meta-llama (accessed Apr. 16, 2025)

Karamolegkou, A., Li, J., Zhou, L., & Søgaard, A. 2023. Copyright Violations and Large Language Models. In Proceedings of EMNLP 2023, pp. 7403–7412.

The Law for the Protection of the Rights of Authors, Composers, and Artists (1969). Available at: https://rc.majlis.ir/fa/law/show/96427 (In Persian).

Lucchi, N. 2023. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. European Journal of Risk Regulation, 15(3): 1–23.

Manteghi, M. 2024. Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI?, Law, Innovation and Technology, 16 (2): 663-686. DOI: 10.1080/17579961.2024.2392928 (accessed May 27, 2025)

The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 .S.D.N.Y. Dec. 27, 2023. Available at: https://www.courthousenews.com/wp-content/uploads/2023/12/new-york-times-microsoft-open-ai-complaint.pdf (accessed Apr. 6, 2025)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. .2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*(140): 1-67. https://jmlr.org/papers/v21/20-074.html (accessed Jun. 8, 2025)

Rosati, E. (2024). Infringing AI? Liability for AI-generated outputs under international, EU and UK copyright law. Computer Law & Security Review, 60, 105938. https://doi.org/10.1016/j.clsr.2024.105938 (accessed Jan. 28, 2025)

Rosati, E. 2025. Copyright exceptions and fair use defences for AI training done for 'research' and 'learning', or the inescapable licensing horizon**.** European Journal of Risk Regulation**,** advance access, 1–24. https://doi.org/10.1017/err.2025.10035 (accessed Jan. 28, 2025)

Schirru, L., Rocha de Souza, A., Valente, M. G. & Lana, A. de P. 2024. Text and Data Mining Exceptions in Latin America. *International Review of Intellectual* Property and Competition Law (IIC), 55: 1624–1653. https://doi.org/10.1007/s40319-024-01511-2 (accessed Jan. 8, 2025)

Sobel, B. L. W. 2017. Artificial intelligence's fair use crisis**.** Columbia Journal of Law & the Arts, 41(1): 45–97. https://doi.org/10.7916/jla.v41i1.2036 (accessed Jan. 3, 2026)

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Xiang, J., Xu, K., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialo, T. 2023. *LLaMA 2: Open foundation and fine-tuned chat models*. arXiv. https://doi.org/10.48550/arXiv.2307.09288 (accessed May12, 2025)

U.S. Copyright Office. 2015, October 16. Authors Guild, Inc. v. Google Inc., No. 13-4829-cv (2d Cir.) [Court decision summary]. https://www.copyright.gov/fair-use/summaries/authorsguild-google-2dcir2015.pdf?utm_source=chatgpt.com (accessed Mar. 25, 2025)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

World Trade Organization (WTO). 1994. *Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS).* Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, Article 13. Geneva: WTO.

Zhang, R., Li, H., Qian, X., Jiang, W., Chen, H.-X., Zhang, H., Wang, J., & Li, C. 2025. On large language models safety, security, and privacy: A survey. Journal of Electronic Science and Technology, 100301. https://doi.org/10.1016/j.jnlest.2025.100301 (accessed Jun. 10, 2025)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

x xx ۱۳ | دورهٔ xx | شمارهٔ x

Legal Considerations in the Use of Copyright-Protected Training Data in Large Language Models: Insights from the Books3 Dataset and the LLaMA2 Model

Abstract

This study aims to examine the legal status of the use of copyrighted works in the training of large language models (LLMs), with particular emphasis on the Books3 dataset and the LLaMA2 model. Through a comparative analysis of the legal frameworks of Iran, the United States, the European Union, France, and Germany, the study seeks to identify the legal bases of liability for the various actors involved and to evaluate the adequacy of existing copyright rules in addressing the challenges posed by the training of LLMs. The research adopts a qualitative approach based on documentary and comparative legal analysis. The data were collected from legislation and regulations, international legal instruments, judicial decisions, technical documentation relating to Books3 and LLaMA2, and relevant academic literature. The materials were analysed using a functional comparative method, examining how each legal system addresses a common set of issues rather than comparing statutory wording alone. The Books3 dataset and the LLaMA2 model were employed as an empirical reference point for testing the legal analysis.

The findings indicate that the principal challenge extends beyond the mere legality of using training data. Rather, the identification of the provenance of training datasets, the ability to establish whether copyrighted works have been used, and the allocation of legal responsibility among the various actors involved in the development and deployment of LLMs are of fundamental importance. The comparative analysis demonstrates that, in the United States, the lawfulness of such use is primarily assessed under the doctrine of fair use on a case-by-case basis. By contrast, the European Union has established a harmonised text and data mining exception under Directive (EU) 2019/790, which France and Germany, as member states, have transposed into national law with certain variations; lawful use under this framework depends on conditions such as lawful access, compliance with statutory requirements, and, in certain cases, the right

of rightholders to opt out. In Iran, however, no specific legal framework governs the use of copyrighted works for training LLMs, resulting in significant uncertainty regarding the liability of developers and other actors involved in the AI lifecycle.

Conclusion: The findings suggest that the responsible development of LLMs requires a transparent legal framework governing the use of training data. Accordingly, the study recommends revising national copyright legislation, introducing explicit legal provisions concerning the training of AI models, establishing statutory obligations for the identification and documentation of training datasets, developing collective licensing mechanisms for the lawful use of copyrighted works, and defining clear legal limits on the use of protected materials for AI training. These measures would contribute to achieving an appropriate balance between the protection of copyright and the continued development of AI technologies.

Keywords: training data, copyright, large language models, fair use, LLaMA2, Books3, Iranian legal system

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

"زینب پاپی: متولد سال ۱۳۶۱ دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی از دانشگاه علوم و تحقیقات تهران است. ایشان هم‌اکنون استادیار گروه پژوهشی مدیریت اطلاعات و سازماندهی دانش سازمان اسناد و کتابخانه ملی است. سازماندهی اطلاعات، حق مؤلف، حقوق دیجیتال، مالکیت فکری در هوش مصنوعی و حقوق داده‌ها از جمله علایق پژوهشی وی است."