

The "Fair Use" and "Fair Dealing" Dilemma in Large Language Model Pre-Training: A Comparative Analysis of US, EU, and UK Copyright Frameworks

Amaresh Patel¹, Dr. Jyoti Garg²

¹Research Scholar, Oriental University, Indore

ORCID: 0000-0002-6000-2839

Email: amareshpatelofficial@gmail.com

²Assistant Professor, Oriental University, Indore

ORCID: 0009-0003-0438-7054

Abstract

The deployment of large language models (LLMs) such as GPT-4, Claude, and Gemini has precipitated a global legal crisis in copyright law, centred on a deceptively simple question: does the wholesale ingestion of copyrighted text to train neural networks constitute permissible use or wholesale infringement? This paper provides a systematic comparative analysis of the three most consequential legal frameworks governing this question — the United States fair use doctrine under 17 U.S.C. § 107, the European Union's text and data mining (TDM) exceptions under Articles 3 and 4 of the Digital Single Market (DSM) Directive (Directive 2019/790/EU), and the United Kingdom's Copyright, Designs and Patents Act 1988 (CDPA), as amended and currently under consultation. Drawing on landmark judicial decisions — including Bartz v. Anthropic (N.D. Cal. 2025), Kadrey v. Meta (N.D. Cal. 2025), The New York Times Co. v. Microsoft Corp. (S.D.N.Y. 2023), Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc. (D. Del. 2025), and GEMA v. OpenAI (LG München I, 2025) — as well as legislative instruments including the EU AI Act (Regulation 2024/1689) and the UK Government's March 2026 Report on Copyright and Artificial Intelligence, this study maps the points of structural convergence and persistent divergence among the three regimes. The analysis reveals that while all three jurisdictions nominally balance innovation interests against creators' rights, their operational mechanisms are fundamentally different in design: the US adopts a post-hoc, litigation-driven four-factor balancing test; the EU employs an ex ante, legislatively structured opt-out framework; and the UK remains in a state of policy flux, unable to definitively settle between the two models. This article proposes that the emerging international standard should synthesise the EU's structural clarity with US jurisprudential flexibility to produce a regime that is both commercially workable and normatively defensible.

Keywords: large language models, fair use, fair dealing, text and data mining, DSM Directive, copyright, pre-training, comparative law, EU AI Act, CDPA

1. Introduction

The training of large language models (LLMs) is one of the most consequential and contested legal phenomena in contemporary intellectual property law. To build a model such as GPT-4, Claude, or Gemini, developers ingest billions of words of text — sourced from books, news archives, academic papers, websites, and code repositories — and use that text to calibrate the statistical weights of artificial neural networks. The process is, at the technical level, fundamentally different from human reading: the copyrighted expression is not reproduced in any output, but is instead dissolved

into billions of numerical parameters that shape the model's probabilistic responses. At the legal level, however, the threshold question is prior to and independent of that technical distinction: does the act of copying text into a training dataset, even for a purpose that does not reproduce the text in any output, constitute copyright infringement?

This question has generated a global wave of litigation. In the United States, more than fifty copyright actions have been filed against AI developers since 2022, including suits by the New York Times, music publishers, professional authors, and legal databases. In Europe, German courts have

handed down conflicting first-instance decisions on the applicability of the EU's text and data mining exceptions to commercial LLM training. In the United Kingdom, a consultation process initiated in December 2024 and followed by a government report in March 2026 has failed to produce legislative resolution, leaving commercial AI developers without a reliable safe harbour and creative industries without enforceable protection.

The stakes are enormous. The global AI content generation market is projected to exceed USD 1.3 trillion by 2032, and the legal framework that governs LLM pre-training will shape the competitive structure of that market and the economic interests of millions of rights holders (Geiger & Jütte, 2024). This paper examines the legal frameworks governing LLM pre-training across three major jurisdictions — the United States, the European Union, and the United Kingdom — through a comparative doctrinal methodology. Section 2 situates the technical problem in legal context. Section 3 analyses the US fair use framework. Section 4 examines the EU TDM regime. Section 5 surveys the UK position. Section 6 provides a comparative synthesis. Section 7 proposes a reform agenda, and Section 8 concludes.

2. The Copyright Problem in LLM Pre-Training

2.1 The Technical Basis of the Legal Dispute

LLM pre-training involves the ingestion of large corpora of textual data into a computational process — specifically, the iterative adjustment of the model's neural network weights through exposure to sequences of tokens drawn from the training corpus. The training process is transformative in a technical sense: no human-readable copy of any specific training text is preserved in the deployed model, and the model cannot, under normal operating conditions, reproduce the verbatim content of its training data with any reliability. Research by Carlini et al. (2021) and subsequent work by Somepalli et al. (2023) has demonstrated, however, that LLMs are capable of "memorising" training data — producing near-verbatim reproductions of specific texts in response to carefully crafted prompts — with the frequency and extent of memorisation increasing with model size and the

density of a given work's representation in the training corpus.

The legal dispute begins before any question of output memorisation arises. At the input stage, the developer must copy the text into its training infrastructure. This copying is, *prima facie*, a reproduction within the exclusive rights of the copyright owner under the applicable statute. The legal question in each jurisdiction is whether that copying is excused by an applicable exception — fair use in the United States, TDM exception in the EU, or some equivalent in the UK — or whether it requires a licence from the rights holder.

2.2 The Commercial Significance of the Threshold Question

The commercial implications of the threshold question are asymmetric and substantial. If pre-training copying requires a licence, AI developers face a negotiation problem of essentially intractable scale: the training corpora of major LLMs span hundreds of millions of documents produced by tens of millions of rights holders across dozens of jurisdictions. The transaction costs of obtaining licences from individual rights holders would be prohibitive, and the alternative — collective licensing through a rights management organisation — requires the creation of new institutional infrastructure that does not currently exist at the scale required. Conversely, if pre-training copying is categorically exempt, creators and publishers lose the economic leverage that copyright ordinarily provides, and receive no remuneration for the use of their works as training material by commercially valuable AI systems (Henderson et al., 2023). The major LLM developers have characterised the legal uncertainty as a threat to innovation. The creative industries have characterised the absence of licensing requirements as an existential threat to the economic model of professional content creation. Neither characterisation is wrong, and it is the resolution of this tension that the three legal frameworks examined in this paper must navigate.

3. The United States: Fair Use and the Four-Factor Test

3.1 The Doctrinal Framework

The United States regulates the use of copyrighted works in LLM pre-training primarily through the fair use doctrine codified at 17 U.S.C. § 107. Fair use is a judicially determined exception that permits the use of copyrighted material without the rights holder's permission if, on balance, four statutory factors weigh in favour of the user: (1) the purpose and character of the use, including whether it is commercial or for non-profit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use on the potential market for or value of the copyrighted work. The Supreme Court has consistently held that the four factors are to be weighed holistically rather than mechanically, with no single factor dispositive, and that the doctrine is an "equitable rule of reason" requiring case-by-case evaluation (*Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 1994).

The practical consequence of this structure is that fair use analysis is inherently *ex post* and litigation-dependent. There is no bright-line rule that specifies in advance whether a particular form of training-data use is or is not fair use. AI developers have operated in this legal uncertainty since the emergence of large-scale LLM training, relying on the judgment — not yet judicially confirmed at the time of writing — that transformative use arguments would carry the day.

3.2 The Transformativeness Argument

The most significant application of fair use to LLM pre-training concerns the first statutory factor, and specifically the concept of "transformative use" — a judicially developed doctrine holding that a use that adds new expression, meaning, or message, or that serves a new and different function, is more likely to be fair use than a use that simply supersedes the original work. Lemley and Casey (2021) were among the first scholars to articulate the case for treating LLM training as transformative: the copying occurs not for the purpose of consuming the

expressive content of the copied works, but to extract statistical patterns — abstract, non-expressive relationships among tokens — that enable the model to generate new text. On this analysis, the trained model does not substitute for the original works; it is a categorically different product serving a categorically different function.

This argument received its most significant judicial endorsement in *Bartz v. Anthropic PBC* (No. 3:24-cv-05417, N.D. Cal., June 2025), in which Judge William Alsup issued a landmark summary judgment ruling holding that Anthropic's use of millions of copyrighted books to train its Claude LLM was "quintessentially transformative" and therefore fair use. The court held that authors cannot exclude others from using their works to learn, and that the training process — which extracted patterns rather than reproducing expression — served a fundamentally different purpose from the original works. Critically, the court distinguished between lawfully acquired and pirated copies: training on pirated books obtained from shadow libraries was found not to constitute fair use, because the illegality of the acquisition negated the transformativeness argument as applied to those works. Within days, Judge Vince Chhabria reached a similar conclusion in *Kadrey v. Meta Platforms* (No. 23-CV-03417, N.D. Cal., June 2025), applying a somewhat different fair use analysis but reaching the same bottom line: training on copyrighted books was fair use.

The significance of these rulings is substantial but bounded. Both decisions are first-instance rulings subject to appeal. Both were issued in the same district and may not reflect the position of other circuits. The *New York Times Co. v. Microsoft Corp.* (No. 1:23-cv-11195, S.D.N.Y., filed December 2023) — the highest-profile pending AI copyright action — is proceeding in the Second Circuit, whose fair use jurisprudence has historically been less deferential to transformative use claims, and where early judicial rulings have been more favourable to the plaintiff than the California decisions (AI Lawsuit Tracker, 2026). A circuit split is a realistic prospect that may ultimately require Supreme Court resolution.

3.3 The Market Effect Factor

The fourth statutory factor — market effect — is widely regarded by copyright scholars as the "most important" element of fair use analysis (Harper & Row, Publishers, Inc. v. Nation Enterprises, 471 U.S. 539, 1985). Its application to LLM pre-training is deeply contested. AI developers argue that the training market is not a market that copyright has traditionally protected, that the training use does not substitute for the original works in their primary consumer market, and that the broader social and economic benefits of AI development outweigh any market harm. Rights holders argue that the emergence of AI training has created a new licensing market — many major publishers have entered licensing agreements with AI developers, and the New York Times, the Associated Press, and Axel Springer have all taken positions (litigation or licensing) based on the premise that the training-data market is real and economically significant (Li, 2024).

The Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc. (D. Del., February 2025) decision cuts against the broad transformative use argument. The Delaware court rejected ROSS's fair use defence at summary judgment, finding that ROSS had used Westlaw's copyrighted legal headnotes to train a directly competing AI legal research tool. The court emphasised the commercial and competitive character of the use and the creative and economic value of the curated headnotes. While the factual pattern — copying from a competitor to build a directly competing product — is more egregious than the general-purpose LLM training scenario, the decision illustrates the limits of the transformative use defence when the training use directly substitutes for an existing or foreseeable licensing market (Peukert, 2024).

3.4 The US Copyright Office Position

The US Copyright Office has maintained a cautious and non-prescriptive position on AI training. Its 2023 Registration Guidance confirmed that AI-generated works without human authorship are not eligible for copyright protection, but did not address the training-data question directly. Its 2025 Report on Copyright and Artificial Intelligence, Part Two:

Copyrightability addressed subsistence issues but deferred substantive comment on the fair use question pending the outcome of ongoing litigation. The Copyright Office's institutional reluctance to weigh in legislatively reflects the traditional US preference for allowing litigation — rather than regulation — to develop copyright doctrine in the context of new technologies. The consequence for AI developers is an environment of managed uncertainty in which sophisticated legal risk assessment, rather than regulatory compliance, drives decision-making.

4. The European Union: Structured TDM Exceptions and the AI Act

4.1 The DSM Directive Framework

The European Union's approach to LLM pre-training is structurally distinct from the US approach in two fundamental respects: it is *ex ante* (specifying the rules in advance through legislation rather than resolving them retrospectively through litigation), and it is structured through a tiered exception framework rather than an open-ended balancing test. The governing instrument is Directive (EU) 2019/790 on copyright in the Digital Single Market (the DSM Directive), which was required to be implemented by EU member states by June 2021. Articles 3 and 4 of the DSM Directive establish the TDM exceptions that directly govern LLM pre-training in the EU.

Article 3 provides a mandatory, non-overridable exception for text and data mining by research organisations and cultural heritage institutions for the purposes of scientific research. Rights holders cannot opt out of Article 3 through contractual terms or technical restrictions. This exception covers academic research institutions and non-profit research organisations conducting scientific TDM, including, on its terms, the training of research models. However, its scope is confined to non-commercial research: the exception does not extend to commercial LLM training by AI companies, which is the commercially significant activity.

Article 4 provides the applicable exception for commercial TDM, including commercial LLM pre-training. It permits reproductions and extractions

from lawfully accessible works and other subject matter for the purposes of text and data mining, subject to one critical qualification: rights holders may opt out by expressly reserving their rights in a machine-readable manner, such as through robots.txt files, metadata, or terms and conditions of a website or service. Article 4(3) specifies that a reservation of rights expressed through "machine-readable means" displaces the exception, restoring the normal position of requiring a licence. The default position under Article 4 is therefore permission — commercial TDM of publicly accessible works is permitted unless the rights holder has actively opted out.

4.2 The AI Act and Article 53

The interaction between the DSM Directive's TDM exceptions and the EU AI Act (Regulation 2024/1689, applicable to general-purpose AI models from August 2025) is significant. Article 53(1)(c) of the AI Act requires providers of general-purpose AI models — a category that encompasses the major commercial LLMs — to "put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790." Recital 105 of the AI Act makes the legislative intent unambiguous: "Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights." The EU legislature has confirmed that the Article 4 TDM exception applies to AI training, including commercial LLM pre-training (Quintais, 2024).

The practical consequence is a compliance obligation, rather than a liability risk, for commercial LLM developers operating in the EU market: they must implement technical mechanisms to detect and honour machine-readable opt-out reservations. The GPAI Code of Practice, published in July 2025 to operationalise Article 53, requires providers to use legal data sources, avoid known piracy sites, and establish complaint mechanisms for rights holders. This framework is broadly understood to make commercial LLM pre-training

of web-crawled content legal in the EU, subject to opt-out compliance — a significantly clearer and more permissive position than the US litigation-dependent uncertainty.

4.3 European Litigation and the Judicial Landscape

Despite the DSM Directive's structural clarity, the EU legal landscape has generated judicial controversy. The Hamburg Regional Court decision in *Kneschke v. LAION* (310 O 227/23, 2024) addressed the application of the Article 4 TDM exception to an AI training dataset (the LAION-5B image dataset), finding that the exception applied. However, in *GEMA v. OpenAI* (Landgericht München I, 2025), the Munich Regional Court found against OpenAI in a case involving the use of musical works to train GPT models, holding that OpenAI had failed to comply with applicable opt-out reservations — a finding of non-compliance with Article 4(3), rather than a finding that Article 4 did not apply at all (Technollama, 2025).

The most significant forthcoming European decision is the CJEU Grand Chamber's pending reference in *Like Company v. Google Ireland Ltd* (C-250/25), which constitutes the first-ever judicial hearing before the EU's highest court on whether the DSM Directive's TDM exception extends to commercial LLM training at scale. A Grand Chamber ruling is expected to provide authoritative guidance on the scope of Articles 3 and 4, with potentially significant implications for the global AI industry's reliance on the EU TDM framework as a permissive baseline (European Copyright Society, 2026).

4.4 The Opt-Out Architecture: Structural Advantages and Limitations

The EU's opt-out architecture is, in theoretical terms, a sophisticated solution to the collective action problem that makes individual licensing of LLM training data practically impossible. It establishes permission as the default, avoiding the transaction cost problem, while preserving the rights of rights holders to exclude their works from training datasets through a machine-readable signal. In practice, however, the architecture has significant operational

limitations. Research by Peukert (2024) and Geiger and Jütte (2024) has identified three structural weaknesses. First, many rights holders — particularly individual authors and small publishers — are unaware of the opt-out mechanism and do not implement it, effectively surrendering control over their works without any economic compensation. Second, the technical implementation of opt-out mechanisms is not standardised, creating uncertainty about whether a given signal will be detected and respected by a given LLM developer. Third, the interaction between the opt-out mechanism and multilingual, multi-jurisdictional web crawling is complex: a robots.txt file configured for European users may not be detected or respected by crawlers operating from non-EU jurisdictions.

5. The United Kingdom: Policy Flux and Unsettled Law

5.1 The Existing CDPA Framework

The United Kingdom's copyright law is governed by the Copyright, Designs and Patents Act 1988 (CDPA). The CDPA provides a narrower set of defined exceptions than either the US fair use doctrine or the EU TDM framework. The relevant exception for TDM is Section 29A CDPA, introduced in 2014 following the Hargreaves Review of 2011. Section 29A permits text and data mining by any person for the purposes of "non-commercial research", subject to lawful access requirements. The exception is absolute within its scope — there is no opt-out mechanism — but its commercial scope is minimal: commercial AI developers training LLMs for deployment in commercial products cannot rely on Section 29A. The absence of a commercial TDM exception under existing UK law means that commercial LLM pre-training of CDPA-protected works requires either a licence from each rights holder or the application of some other CDPA exception, none of which straightforwardly applies to commercial model training at scale (Recording Law, 2026).

The UK did not implement the DSM Directive following its withdrawal from the European Union. The commercial TDM exception introduced by Article 4 of the DSM Directive is therefore not

available to LLM developers operating under UK law. This places the UK in a position that is arguably more restrictive than the EU in relation to commercial TDM, while also being without the judicial mechanisms — such as the four-factor fair use balancing test — that provide US developers with a viable (if litigation-dependent) legal argument. The UK position is, in practical terms, the worst of both worlds for AI developers: no legislative exception and no established doctrine of fair dealing broad enough to encompass commercial LLM training.

5.2 The Consultation and the March 2026 Government Report

The UK Government initiated a public consultation on Copyright and Artificial Intelligence on 17 December 2024, which closed on 25 February 2025. The consultation attracted over 11,000 responses — an extraordinary volume reflecting the intensity of the competing interests — and proposed four options for reform, including maintaining the status quo, introducing a broad commercial TDM exception with rights-holder opt-out, and various intermediate positions. The consultation's preferred option was an EU-style commercial TDM exception with machine-readable opt-out, underpinned by transparency obligations on AI developers (UK Government, 2025).

The government's Report on Copyright and Artificial Intelligence, published on 18 March 2026, stepped back significantly from the consultation's preferred option. The Report declined to commit to any specific legislative proposal, instead announcing further engagement and evidence-gathering. The firmest conclusion in the Report was a proposal to remove the protection for "computer-generated works" under Section 9(3) CDPA — a relatively minor reform unrelated to the training-data question. The government's original preferred solution — a TDM exception with opt-out — was described as no longer the government's preferred approach, without a clear alternative being substituted. As of mid-2026, no commercial TDM exception has been enacted and no legislative timeline has been announced (Two Birds, 2026).

The Data (Use and Access) Act 2025, enacted in the intervening period, introduced transparency obligations on AI developers concerning their training data (Sections 135 to 137) but did not create any copyright exception or otherwise alter the CDPA's treatment of commercial TDM. The transparency obligations require AI developers to publish information about their training datasets — a measure that may assist rights holders in identifying uses of their works but does not resolve the underlying liability question.

5.3 The UK's Structural Dilemma

The UK's inability to resolve the training-data copyright question reflects a structural tension between two powerful competing interests. On the one hand, the UK government has explicitly identified AI development as a strategic priority and has expressed concern that over-protective copyright regulation could disadvantage UK-based AI companies relative to their US and EU counterparts. On the other hand, the UK creative industries — music, publishing, broadcasting, and film — are among the most economically significant and politically organised sectors of the UK economy, and they have mounted an intense campaign against any TDM exception that does not include robust licensing and compensation mechanisms. The government's March 2026 decision to retreat from its preferred option without substituting an alternative is best understood as a reflection of the political impossibility of satisfying both constituencies simultaneously, rather than a substantive resolution of the underlying legal question (Kluwer Copyright Blog, 2025).

6. Comparative Analysis: Convergences and Divergences

6.1 Structural Divergences

The three frameworks examined in this paper differ in four dimensions that are analytically fundamental. First, they differ in their legal mechanism: the US employs an open-ended, post-hoc judicial balancing test; the EU employs a structured, ex ante legislative exception framework; and the UK currently employs a narrow statutory exception that does not cover commercial use, without any compensating

flexibility mechanism. Second, they differ in the default position: the US default is that copying requires permission unless the copier can demonstrate fair use; the EU default under Article 4 is that TDM of accessible works is permitted unless the rights holder has opted out; the UK default is that commercial copying requires permission with no general TDM exception to displace it. Third, they differ in the locus of decision-making: the US relies on courts, producing expensive, slow, circuit-specific, and retrospective guidance; the EU relies on legislative instruments, producing prospective general guidance at the cost of political compromise and rigidity; and the UK is currently relying on governmental consultation, producing no legal certainty at all. Fourth, they differ in their treatment of rights-holder compensation: all three frameworks, as currently configured, provide for no mandatory compensation to rights holders whose works are used in LLM pre-training — permission (or exception) does not entail remuneration.

6.2 Points of Convergence

Despite these structural divergences, the three frameworks share several important features. All three agree, at the level of principle, that copyright protection for expressive works extends to the reproduction involved in LLM pre-training — none of the three frameworks provides a categorical exemption for all LLM training uses. All three frameworks are wrestling with the same underlying policy tension: the public interest in AI innovation and the private rights of creators whose works provide the raw material for that innovation. All three are converging, through different mechanisms, on the view that the legality of pre-training depends significantly on the source of the training data — whether lawfully acquired or obtained through piracy is critical in the US (Bartz), whether the rights holder has opted out is critical in the EU (Article 4(3)), and whether the developer has lawful access is a threshold requirement in the UK (Section 29A). The "lawful access" requirement has emerged, in different institutional forms, as the central criterion across all three frameworks.

6.3 The Opt-Out Model as an Emerging Global Standard

A significant pattern in the comparative analysis is the gravitational pull of the EU's opt-out model. The December 2024 UK consultation proposed an EU-style opt-out mechanism as its preferred option. The US litigation landscape — in which courts have begun to distinguish between training on lawfully acquired and pirated copies — implicitly incorporates an access-based criterion analogous to the EU's lawful-access requirement. The GPAI Code of Practice, which applies to general-purpose AI models placed on the EU market, operationalises the opt-out requirement in ways that de facto apply to any LLM with European users. The opt-out model is emerging as a candidate global standard not because of its theoretical elegance but because of its practical workability: it avoids the transaction cost problem of individual licensing while preserving a meaningful mechanism for rights holders to assert control (Geiger & Jütte, 2024; Quintais, 2024).

The opt-out model has, however, a fundamental distributional weakness: it benefits large publishers and technology companies — who have the resources to implement and monitor opt-out mechanisms — while disadvantaging individual creators who lack the technical sophistication or institutional capacity to configure machine-readable reservations effectively. This distributional concern is not adequately addressed by any of the three frameworks in their current form.

7. Towards a Workable International Framework: Reform Proposals

7.1 A Hybrid Fair Use and TDM Opt-Out Framework

This paper proposes that the international baseline for LLM pre-training copyright should synthesise the structural clarity of the EU TDM opt-out framework with the normative flexibility of US fair use doctrine. The core of the proposal is as follows. Pre-training on lawfully accessed works should be prima facie permissible, subject to a standardised machine-readable opt-out mechanism that is technically accessible to all rights holders regardless of their technological sophistication. This is more

permissive than the current US position (which requires the developer to justify fair use) and structurally equivalent to the EU's Article 4, but with enhanced accessibility requirements for the opt-out mechanism. The lawful access requirement — which all three jurisdictions effectively impose, though through different instruments — should be codified as a hard threshold condition, with training on unlawfully obtained copies (including from shadow libraries and piracy sites) categorically outside the exception.

The "market effect" criterion from US fair use doctrine should be incorporated as a limiting condition on the opt-out framework: where a commercial LLM developer is training on content from a rights holder who has actively sought to enter a licensing negotiation or who operates in a market where AI training is a commercially significant use, the developer should not be able to rely on the opt-out default but should be required to negotiate. This preserves the economic incentive for collective licensing in high-value content markets while maintaining the default permission that prevents the transaction cost problem from blocking AI development more generally.

7.2 Mandatory Transparency and Disclosure

A second reform, applicable in all three jurisdictions, concerns training data transparency. None of the current frameworks adequately enables rights holders to determine whether their works have been used in LLM pre-training, without which the opt-out mechanism and any future compensation claim are practically unenforceable. The UK Data (Use and Access) Act 2025 introduces transparency obligations, as does Article 53(1)(d) of the EU AI Act, which requires general-purpose AI model providers to publish "a sufficiently detailed summary" of training data content. These transparency requirements are important but underspecified: "sufficiently detailed" is not defined, and existing disclosures by major AI developers — citing datasets like "Common Crawl" or "The Pile" without granular content-level information — do not enable individual rights holders to identify whether their works are included. Standardised, granular training data disclosure

should be a mandatory component of any reform, implemented through technical standards bodies such as IETF or ISO rather than left to regulatory discretion (Henderson et al., 2023).

7.3 Collective Licensing and Remuneration Rights

The most substantial structural gap in all three frameworks is the absence of any remuneration mechanism for rights holders whose works are used in LLM pre-training. The EU's approach — allowing commercial TDM subject to opt-out — provides no compensation to rights holders who do not opt out. The US approach — subject to ongoing litigation — has not established any licensing or compensation entitlement even for rights holders whose fair use claims fail. Some scholars have proposed the introduction of extended collective licensing (ECL) mechanisms as a solution: a designated collecting society would negotiate a blanket licence on behalf of all rights holders in a given category, with developers paying a standard rate and rights holders receiving a proportionate share of the licence revenue (Geiger & Jütte, 2024). ECL mechanisms are already well-established in the Scandinavian copyright tradition and would, if applied to LLM training, simultaneously resolve the transaction cost problem for developers and the compensation deficit for creators. The EU AI Act's Article 53 framework could be extended to mandate ECL for commercial LLM training as a condition of accessing the Article 4 TDM exception — a reform that would transform the current "no compensation" default into a "fair remuneration" standard without requiring individual licensing transactions.

8. Conclusion

The legal question of whether LLM pre-training infringes copyright is one of the most commercially and normatively consequential questions in contemporary intellectual property law. This paper has examined that question across three major jurisdictions, reaching five principal conclusions.

First, the three frameworks — US fair use, EU TDM exception, and UK CDPA — are structurally divergent in their design logic, default positions, and decision-making loci, but are converging on a

common substantive principle: lawful access to the training data is a threshold requirement, and training on unlawfully obtained copies is not protected regardless of the doctrinal framework applied.

Second, the recent US decisions in *Bartz v. Anthropic* and *Kadrey v. Meta* represent significant pro-developer rulings that establish "transformative use" as a plausible fair use defence for commercial LLM training on lawfully acquired content, but these rulings are first-instance decisions in a single circuit, subject to appeal, and likely to be tested against different factual patterns — particularly in the New York Times litigation in the Second Circuit — before any stable national standard emerges.

Third, the EU's Article 4 TDM opt-out framework is the most structurally coherent existing response to the LLM training copyright problem, but it suffers from a distributional deficit: the opt-out mechanism is practically accessible only to large, technically sophisticated rights holders, leaving individual creators without effective protection.

Fourth, the UK's failure to resolve the training-data copyright question — despite extensive consultation and a government report — represents a significant governance failure with real commercial consequences for UK-based AI developers and UK-based creators alike. The absence of either a commercial TDM exception or a flexible fair dealing doctrine leaves UK law in a position that is more restrictive than either of its major trading partners.

Fifth, the most viable international standard is a hybrid that incorporates the structural clarity and ex ante certainty of the EU opt-out model, the normative flexibility of US market-effect analysis, mandatory transparency requirements, and a collective licensing framework that provides remuneration to rights holders at scale. The design of that hybrid framework is a policy challenge that is, at present, being addressed more effectively by litigation in the United States than by regulation in either Europe or the United Kingdom. That priority ordering is itself a symptom of the governance deficit that comparative copyright scholarship must work to address.

References

1. AI Lawsuit Tracker. (2026). New York Times v. OpenAI & Microsoft. Retrieved from <https://ailawsuittracker.com/cases/new-york-times-v-openai/>
2. Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal. June 24, 2025).
3. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).
4. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Papernot, N. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium, 2633–2650.
5. European Copyright Society (CREATE). (2026). Comment on C-250/25 (Grand Chamber). IIC – International Review of Intellectual Property and Competition Law. <https://doi.org/10.1007/s40319-026-01717-6>
6. Geiger, C., & Jütte, B. J. (2024). Elaborating a human rights-friendly copyright framework for generative AI. International Review of Intellectual Property and Competition Law, 55, 1129–1165. <https://doi.org/10.1007/s40319-024-01486-2>
7. GEMA v. OpenAI, Landgericht München I (LG München I), 2025.
8. Harper & Row, Publishers, Inc. v. Nation Enterprises, 471 U.S. 539 (1985).
9. Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., & Liang, P. (2023). Foundation models and fair use. Journal of Machine Learning Research, 24(400), 1–64.
10. Kadrey v. Meta Platforms, Inc., No. 23-CV-03417-VC, 2025 WL 1752484 (N.D. Cal. June 25, 2025).
11. Kluwer Copyright Blog. (2025). Copyright and AI policy in the UK in 2025. Retrieved from <https://legalblogs.wolterskluwer.com/copyright-blog/copyright-and-ai-policy-in-the-uk-in-2025/>
12. Kneschke v. LAION e.V., Landgericht Hamburg, 310 O 227/23 (Hamburg Regional Court, 2024).
13. Lemley, M. A., & Casey, B. (2021). Fair learning. Texas Law Review, 99(4), 743–785.
14. Li, X. (2024). Copyright protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI. Computer Law and Security Review, 55, 106047. <https://doi.org/10.1016/j.clsr.2024.106047>
15. Like Company v. Google Ireland Ltd, CJEU Grand Chamber, Case C-250/25 (pending, 2026).
16. Peukert, A. (2024). Copyright in the Artificial Intelligence Act — a primer. GRUR International, 73(6), 497–511.
17. Quintais, J. P. (2024). Generative AI, copyright and the AI Act. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.4941701>
18. Recording Law. (2026). UK AI copyright laws: CDPA s.9(3) and what it covers. Retrieved from <https://www.recordinglaw.com/world-laws/world-ai-copyright-laws/united-kingdom-ai-copyright-laws/>
19. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? Investigating data replication in diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6048–6058.
20. Technollama. (2025). We need to talk about the EU TDM exception and AI training. Retrieved from <https://www.technollama.co.uk/we-need-to-talk-about-the-eu-tdm-exception-and-ai-training>
21. Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., No. 1:20-cv-00613 (D. Del. February 2025).
22. Two Birds. (2026, March 19). Copyright & AI in the UK: The debate rolls on. Retrieved from <https://www.twobirds.com/en/insights/2026/uk/copyright-,-a-,-ai-in-the-uk-the-debate-rolls-on>
23. UK Government. (2025). Copyright and Artificial Intelligence: Consultation. Department for Science, Innovation and Technology and Department for Culture, Media and Sport. Retrieved from <https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence>
24. UK Government. (2026). Report on Copyright and Artificial Intelligence (CP2602959). Retrieved from https://assets.publishing.service.gov.uk/media/69ba692226909a14239612e4/CP2602959_-_Report_on_Copyright_and_Artificial_Intelligence_web.pdf
25. US Copyright Office. (2023). Copyright registration guidance: Works containing material generated by artificial intelligence, 88 FR 16191 (March 16, 2023).
26. US Copyright Office. (2025). Copyright and artificial intelligence, Part 2: Copyrightability. Retrieved from <https://www.copyright.gov/ai/>