Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-8 |Aug: 2026
Original Researcher Article

Training Data and the Reproduction Right: Is Ingestion by Large Language Models "Copying" under Section 14 of the Copyright Act, 1957?.

Amaresh Patel[1], Dr. Jyoti Garg[2]
[1] Research Scholar, Oriental University, Indore, Madhya Pradesh, India.
Email: amareshpatelofficial@gmail.com
ORCID: 0000-0002-6000-2839
[2] Assistant Professor, Oriental University, Indore, Madhya Pradesh, India.
ORCID: 0009-0003-0438-7054
Corresponding Author:
Amaresh Patel
Research Scholar, Oriental University, Indore, Madhya Pradesh, India.
Email: amareshpatelofficial@gmail.com
ORCID: 0000-0002-6000-2839

.

ABSTRACT

The training of large language models (LLMs) proceeds by ingesting vast corpora of text, a substantial portion of which is protected by copyright. This article asks a narrow but foundational question of Indian law: does that act of ingestion amount to "reproduction" within the meaning of Section 14 of the Copyright Act, 1957? Adopting a doctrinal and comparative method, the study disaggregates the training pipeline into six technically distinct operations and tests each against the statutory language of Section 14(a)(i), the definition of "infringing copy" in Section 2(m), and the storage-oriented amendments introduced by the Copyright (Amendment) Act, 2012. It argues that at least three of those operations — corpus acquisition, corpus curation and durable retention — constitute prima facie reproduction, that a fourth (tokenised batching) is defensible as transient and incidental, and that model weights are ordinarily not copies save in the narrow case of demonstrable memorisation. The article then evaluates the Delhi High Court's interim ruling in ANI Media Pvt. Ltd. v. Open AI OpCo LLC (2026), which held that such storage prima facie falls within the "private or personal use, including research" limb of Section 52(1)(a). It contends that this characterisation, while pragmatically attractive, strains the statutory text, conflates commercial secrecy with private use, and imports an open-ended proportionality analysis into a closed-list exception regime. The study concludes that the reproduction question in India cannot responsibly be resolved through interpretive elasticity and proposes a calibrated statutory text and data mining exception coupled with a collective licensing mechanism..

Keywords: Artificial intelligence; copyright infringement; large language models; reproduction right; Section 14 Copyright Act 1957; fair dealing; text and data mining..

INTRODUCTION

The commercial deployment of conversational artificial intelligence has, within a remarkably short period, moved from a research curiosity to an object of sustained litigation across at least six jurisdictions. At the centre of that litigation lies a deceptively simple factual proposition: to build a large language model, a developer must first assemble and process an enormous body of text, and that text is overwhelmingly the work of human authors who hold copyright in it. Whether the assembly and processing of that text is a lawful act, an infringing act, or an act excused by statutory exception is now the single most consequential unresolved question in the law of copyright.

Public and academic debate has tended to collapse this question into the vocabulary of "fair use", a doctrine that has no direct analogue in Indian law. The Indian Copyright Act, 1957 does not confer on courts a general equitable power to excuse otherwise infringing conduct on grounds of transformative purpose or social utility. It instead establishes, in Section 14, a closed enumeration of exclusive rights, and, in Section 52, a closed enumeration of acts that shall not constitute infringement. The consequence is structural: in India the analysis must begin, and may well end, with the question whether the impugned act falls within the scope of an exclusive right at all. If ingestion is not reproduction, no exception is needed. If it is reproduction, the developer must locate its conduct within a specific statutory clause, not within a general appeal to innovation policy.

That structural point has been obscured by the tendency of Indian commentary to reason backwards from American outcomes. The rulings in Bartz v. Anthropic

and Kadrey v. Meta Platforms[1] were both decided under Section 107 of the United States Copyright Act, a provision whose four-factor calculus is expressly non-exhaustive and which invites precisely the kind of open-ended balancing that Section 52 forecloses. Importing their conclusions into Indian law without interrogating the statutory architecture that produced them is a category error, and one that this article seeks to correct.

The urgency of the enquiry has been sharpened by the Delhi High Court's decision of 24 July 2026 in ANI Media Pvt. Ltd. v. Open AI OpCo LLC.[2] Declining ANI's application for an interim injunction, Bansal J. held on a prima facie basis that OpenAI's storage of ANI's literary works for the purpose of training the models underlying ChatGPT falls within the "private or personal use, including research" limb of Section 52(1)(a), and therefore does not amount to infringement under Section 51. The judgment is the first substantive Indian judicial engagement with the question and is expressly interlocutory; its findings are stated not to bind the final adjudication of the suit. It nonetheless supplies the first authoritative Indian framework against which the present analysis must be tested.

This article advances three claims. First, that the question "is ingestion copying?" cannot be answered at the level of the training pipeline as a whole, because that pipeline comprises technically and legally heterogeneous operations, some of which are plainly reproductive and some of which are plainly not. Second, that when those operations are disaggregated, the acquisition, curation and durable retention of a training corpus satisfy the statutory test of reproduction in a material form under Section 14(a)(i), while the encoding of statistical parameters into model weights ordinarily does not. Third, that the exceptional pathway identified in ANI Media — the private use and research limb of Section 52(1)(a) — is the least persuasive of the available routes to the same practical result, and that the doctrinal work it is being asked to perform is better discharged by legislative amendment than by judicial construction.

The article proceeds in seven further parts. Part 2 sets out the objectives, research questions and method. Part 3 disaggregates the training pipeline and characterises each stage. Part 4 undertakes the core statutory analysis of Section 14. Part 5 examines the exceptions in Section 52 and the reasoning in ANI Media. Part 6 places the Indian position in comparative perspective. Part 7 offers a critical evaluation. Part 8 sets out findings and reform proposals, and Part 9 concludes.

2. Objectives, Research Questions and Method

The principal objective of this study is to determine the scope of the reproduction right under Section 14 of the Copyright Act, 1957 as applied to the acts constituting the training of a large language model, and to assess the adequacy of the existing Indian exception regime to accommodate those acts. Four subsidiary objectives follow: to establish a technically accurate taxonomy of training operations suitable for legal characterisation; to test each operation against the statutory language of the Act; to evaluate the reasoning of the Delhi High Court in ANI Media against that statutory language; and to identify a reform pathway consistent with India's obligations under the Berne Convention and the TRIPS Agreement.

The study is guided by five research questions. (i) Does the acquisition and durable storage of copyrighted works for the purpose of model training constitute reproduction in a material form under Section 14(a)(i)? (ii) Does the transient tokenisation and batching of those works during a training run attract the reproduction right, or is it excluded as incidental? (iii) Do the trained parameters of a model constitute an "infringing copy" within Section 2(m)? (iv) Can such acts be accommodated within Section 52, and specifically within clauses 52(1)(a), 52(1)(b) and 52(1)(c)? (v) What legislative response would best reconcile the interests of rightholders and developers within India's closed-list framework?

The method is doctrinal and comparative, supplemented by a technical–legal mapping exercise. The doctrinal component proceeds from primary sources: the Copyright Act, 1957 as amended in 2012, the Copyright Rules, 2013, and the decisions of Indian courts on reproduction and substantiality, notably R.G. Anand v. Delux Films[3] and the line of authority following it. The comparative component examines the fair use jurisprudence of the United States, the text and data mining exceptions of the European Union, Japan and Singapore, and the contrasting German position in GEMA v. OpenAI. The technical component draws on the published architecture and training methodology literature of transformer-based language models in order to ensure that legal characterisation attaches to operations as they actually occur rather than to metaphor. No empirical or human-subject data were collected; the study raises no ethical clearance issues.

Two limitations should be acknowledged at the outset. First, the ANI Media ruling is interlocutory, and the analysis offered here is necessarily provisional pending final adjudication and any appellate consideration. Second, the technical description of training pipelines is generalised across model families, since the precise data-handling practices of individual developers are typically confidential and, in litigation, disclosed only under seal.

3. The Anatomy of Ingestion: A Technical–Legal Taxonomy

Legal analysis of machine learning has been persistently handicapped by the imprecision of the word "training". The term is used indiscriminately to describe operations ranging from the bulk downloading of a web archive to the gradient update of a parameter matrix. These are not merely different steps in a single process; they are

[1] Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal., 23 June 2025); Kadrey v. Meta Platforms Inc., No. 3:23-cv-03417 (N.D. Cal., 25 June 2025).

[2] ANI Media Pvt. Ltd. v. Open AI OpCo LLC, CS(COMM) 1028/2024, Neutral Citation 2026:DHC:5900 (Del. HC, 24 July 2026) (Bansal, J.) (interim order).

[3] R.G. Anand v. M/s Delux Films, (1978) 4 SCC 118.

different kinds of acts, with different relationships to the copyright work, and the reproduction right does not attach to them uniformly.

It is therefore necessary to disaggregate. The construction of a modern large language model may be decomposed into six operations, each of which is capable of independent legal characterisation.

The first is corpus acquisition. Text is obtained by crawling the public web, by downloading pre-assembled datasets, by licensing from publishers, or — as the American litigation has revealed — by downloading from unauthorised repositories. In every case the operation involves the making of a complete digital copy of the source work on the developer's storage infrastructure. The copy is verbatim, permanent until deleted, and functionally indistinguishable from any other digital reproduction.

The second is corpus curation. The acquired text is deduplicated, filtered for quality, stripped of boilerplate, language-classified, and frequently rewritten or reformatted. Curation typically generates further copies: a filtered corpus, a deduplicated corpus, and one or more derived datasets, each retained separately for reproducibility.

The third is tokenisation and batching. Curated text is segmented into sub-word tokens, converted into integer sequences, and loaded into memory in batches for presentation to the model. These representations are lossless and reversible — a token sequence can be decoded back into the original text — but they exist in volatile memory for the duration of a training step.

The fourth is the forward and backward pass. The model computes a prediction over the next token, compares it against the observed token, and propagates an error gradient backwards through the network. No copy of the work is created by this operation as such; what is generated is a set of numerical adjustments.

The fifth is parameter update. The computed gradients are applied to the model's weights. The resulting parameters encode statistical regularities abstracted across the whole corpus. In the ordinary case these parameters bear no retrievable relationship to any individual work; in the exceptional case of memorisation — typically associated with works that are heavily duplicated in the corpus — a sequence may be recoverable from the model with high fidelity.

The sixth is retention. Developers routinely preserve the acquired and curated corpora after training is complete, for the purposes of retraining, ablation studies, reproducibility and evaluation. It was this retention, rather than the training process itself, that proved decisive in the American litigation: Judge Alsup's ruling in Bartz distinguished sharply between training on lawfully acquired books, which he regarded as transformative, and the maintenance of a permanent library of unlawfully obtained copies, which he did not.[4]

Table 1 maps these operations against their principal legal characteristics under the Indian Act.

Table 1 A technical–legal taxonomy of operations in the large language model training pipeline.

Operation	Nature of the fixation	Duration	Characterisation under the Act
1. Corpus acquisition	Complete verbatim digital copy on developer-controlled storage	Durable	Reproduction: s. 14(a)(i) (storing in a medium by electronic means)
2. Corpus curation	Derived copies — deduplicated, filtered, reformatted datasets	Durable	Reproduction; adaptation under s. 14(a)(vi) may also arise
3. Tokenisation and batching	Lossless reversible token sequences in volatile memory	Transient	Not storage in a medium; outside s. 14 on the better view
4. Forward and backward pass	Numerical gradients; no fixation of expression	Transient	Outside s. 14

[4] Bartz v. Anthropic PBC, supra note 1. The retention holding formed the basis of the subsequent class settlement.

Operation	Nature of the fixation	Duration	Characterisation under the Act
5. Parameter update	Statistical abstraction across the corpus	Durable	Not an infringing copy under s. 2(m), save on proof of memorisation
6. Retention of corpora	Continued warehousing after the training run	Durable, open-ended	Reproduction; least connected to technical necessity

Note. Characterisation is the authors 'own, derived from the statutory analysis in Part 4.

Two conclusions follow from this taxonomy. The first is that the copyright-relevant conduct is heavily concentrated at the beginning and end of the pipeline — in acquisition, curation and retention — rather than in the mathematical operations that the word "training" tends to evoke. The second is that the developer's most defensible territory is the middle of the pipeline, and its most exposed territory is the part that most closely resembles conventional digital copying. A legal framework that analyses "AI training" as an undifferentiated whole will therefore either over-protect or under-protect, depending on which end of the pipeline it happens to fix upon.

4. Section 14 and the Reproduction Right

4.1 The statutory language

Section 14(a)(i) of the Copyright Act, 1957 confers upon the owner of copyright in a literary, dramatic or musical work the exclusive right to reproduce the work in any material form, and expressly extends that right to include "the storing of it in any medium by electronic or other means". The words in quotation were substituted by the Copyright (Amendment) Act, 1994 and retained in the 2012 amendments. Their significance for the present enquiry is difficult to overstate. Parliament did not leave the application of the reproduction right to digital storage to inference; it stated in terms that electronic storage is itself an exercise of the reproduction right.

This textual position distinguishes India from jurisdictions in which the reproduction right had to be extended to digital copies by judicial construction. It also means that arguments framed around the absence of a "human-readable" output, or around the fact that no person ever reads the stored copy, engage the wrong question. Section 14(a)(i) is concerned with the making of the copy, not with its consumption. A copy stored on a server and never opened is a reproduction; a copy stored on a server and used only as an input to a mathematical process is likewise a reproduction. The purpose to which the copy is put is a question for Section 52, not for Section 14.

Section 51(a)(i) provides that copyright is infringed when any person, without a licence, does anything the exclusive right to do which is conferred upon the owner. Section 2(m) defines an "infringing copy" in relation to a literary work as a reproduction otherwise than in the form of a cinematographic film. Read together, these provisions establish a low threshold for prima facie infringement: the making of an unauthorised electronic copy of a substantial part of a protected work, without more, engages Section 51.

4.2 Applying the right to acquisition, curation and retention

Against that statutory background, the characterisation of corpus acquisition is straightforward. The developer causes a complete and verbatim electronic copy of the work to be made on storage under its control. That is storage in a medium by electronic means. It is reproduction within Section 14(a)(i). The position is not altered by the fact that the source was publicly accessible: accessibility on the open web is not a licence, and Indian law contains no doctrine by which publication on a website exhausts the reproduction right. Nor is it altered by the presence or absence of a paywall, save as to the further question whether circumvention engages Section 65A.

Corpus curation stands in the same position, and arguably in a worse one, since curation generates copies that did not previously exist even in the developer's own systems. A deduplicated derivative corpus is a new fixation of the underlying works. Where curation involves reformatting or extraction, the further question of adaptation under Section 14(a)(vi) may also arise, though it will rarely add materially to the reproduction claim.

Retention is the operation most vulnerable to challenge, precisely because it is the operation least connected to any technical necessity of training. A corpus retained for two years after a training run has been completed is not being used to train anything; it is being warehoused. The purposive arguments available in respect of training — that copying is a necessary intermediate step towards a non-substitutive end — apply with markedly less force to a permanent archive. This is the distinction that proved decisive in Bartz, and there is no reason of Indian law why it should not be equally decisive here.

4.3 Tokenisation, transience and the incidental copy

Tokenisation presents the harder case. The token sequence is a lossless representation of the work, and if it were written to disk it would plainly be a reproduction. The developer's answer is that it exists only fleetingly in volatile memory as part of a technological process, and that the law has never treated such evanescent representations as reproductions.

Indian law offers partial support for that answer. Section 52(1)(b) exempts the transient or incidental storage of a work or performance purely in the technical process of electronic transmission or communication to the public, and Section 52(1)(c) exempts transient or incidental storage for the purpose of providing electronic links, access or integration. Both clauses were inserted by the 2012 amendment with the intermediary context principally in mind. Neither maps comfortably onto training. Clause (b) requires that the storage be part of a process of transmission or communication to the public; a training run transmits nothing to the public. Clause (c) requires a link, access or integration function; training provides none.

The better view is that tokenised batching escapes the reproduction right not because it falls within Section 52(1)(b) or (c), but because the transience and non-retrievability of the representation mean that it does not constitute "storing in a medium" in any meaningful sense. That conclusion is consistent with the reasoning of the Court of Justice of the European Union in Infopaq International A/S v. Danske Dagblades Forening and in Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd,[5] which treated the temporary reproduction exception as applying to acts of an evanescent character forming an integral part of a technological process. Indian law lacks an equivalent general exception, but the underlying reasoning informs the antecedent question of what "storage" means.

This conclusion is of limited comfort to developers, however, because tokenisation is downstream of acquisition. One cannot tokenise a corpus one has not first copied. The transient-copy defence therefore protects a step that is never reached without a prior act of reproduction that is not transient at all.

4.4 Model weights and the memorisation exception

Whether a trained model is itself an infringing copy is conceptually the most interesting question in this field, and practically the least important. Conceptually, it asks whether a lossy statistical abstraction derived from millions of works can be a reproduction of any one of them. Practically, it matters little, because a claimant who establishes infringement at the acquisition stage has no need to establish it at the parameter stage.

The orthodox answer, and in the present author's view the correct one, is that model weights are not copies. The parameters of a large language model encode conditional probability relationships across a vocabulary; they do not encode the works. The compression ratio alone makes the point: a model trained on corpora measured in terabytes may itself occupy a few hundred gigabytes, and no process of decompression will recover the corpus. What has been abstracted is information about expression, and information about expression is not expression. The idea–expression dichotomy, codified in Article 9(2) of TRIPS and reflected in Indian jurisprudence since R.G. Anand, points firmly in this direction.

The exception is memorisation. Where a work — or, more commonly, a passage — appears with unusual frequency in the training corpus, the model may encode it sufficiently precisely that it can be elicited verbatim through prompting. In such a case the claim that no copy resides in the parameters becomes empirically false as to that passage. The Delhi High Court in ANI Media treated memorisation as central, holding that ANI had failed to establish that ChatGPT had memorised, regurgitated or communicated exact copies of its reports to the public.[6] The allocation of the burden of proof is significant: it is the rightholder who must demonstrate memorisation, and the evidential means of doing so are largely within the developer's control.

4.5 Substantiality

Section 14 protects the work as a whole and any substantial part of it. In the training context the substantiality enquiry is unusual, because the copying is total at the acquisition stage and negligible at the output stage. Indian courts assess substantiality qualitatively rather than quantitatively, asking whether what has been taken constitutes a material and essential part of the plaintiff's work. Where an entire news report has been downloaded and stored, that enquiry is answered before it is asked. The defensive argument that any individual work contributes only an infinitesimal fraction of the model's learning is an argument about proportion of the defendant's product, not about proportion of the plaintiff's work, and Indian doctrine has never accepted the former as relevant.

5. Section 52 and the Reasoning in ANI Media

5.1 The closed-list structure

Section 52 lists some fifty-six categories of act that shall not constitute infringement. It is exhaustive. Indian courts have consistently declined to read into it a residual equitable discretion, and the Delhi High Court itself observed in Syndicate of the Press of the University of Cambridge v. B.D. Bhandari[7] that the fair dealing enquiry in India operates within, not beyond, the enumerated purposes. Any defence to a training claim must therefore be located in a specific clause.

Figure 1 sets out the analytical sequence that follows from this structure.

[5] Case C-5/08 Infopaq International A/S v. Danske Dagblades Forening [2009] ECR I-6569; Case C-360/13 Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd, ECLI:EU:C:2014:1195.

[6] ANI Media Pvt. Ltd. v. Open AI OpCo LLC, supra note 2.

[7] Syndicate of the Press of the University of Cambridge v. B.D. Bhandari, 2011 (47) PTC 244 (Del.) (DB).

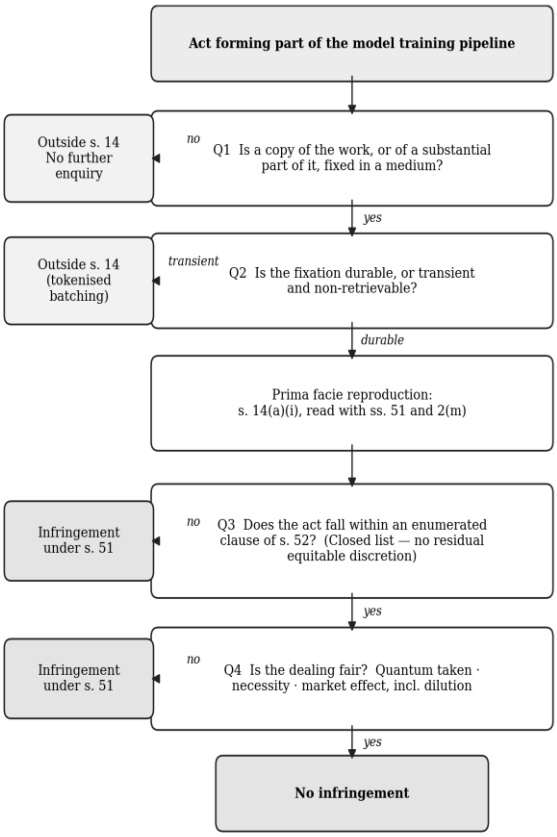

Figure 1 The sequence of enquiry under the Copyright Act, 1957 for an act forming part of the training pipeline. Note. Prepared by the authors.

5.2 The private use and research limb

Section 52(1)(a) provides that fair dealing with a literary work, not being a computer programme, for the purposes of private or personal use including research, criticism or review, or the reporting of current events, shall not constitute infringement. It was this limb that the Delhi High Court applied in ANI Media, holding that OpenAI's storage of ANI's works for the purpose of training the models underlying ChatGPT prima facie falls within "private or personal use, including research".[8]

The reasoning proceeded in three principal steps. First, the Court held that it possessed territorial jurisdiction, rejecting OpenAI's preliminary objection on maintainability, on the basis that ANI's principal and registered offices were in Delhi and that OpenAI deliberately offered paid and free services to users there. Second, the Court held that ANI had not established a prima facie case of infringement in respect of either training or outputs, having failed to demonstrate memorisation, regurgitation or communication of exact copies to the public, and having failed to establish market substitution. Third, the Court held that the balance of convenience favoured OpenAI, reasoning that an injunction would cause irreparable injury not only to the developer but to the very large Indian user base of the service, and that a requirement to license every data source before training would render the development of large language models economically unviable.

It should be emphasised that the Court expressly characterised its findings as prima facie and stated that they would not control the final adjudication of the suit. The judgment runs to some 135 pages following 32 hearings, and its analytical framework for the application of Section 52 to AI systems is likely to be influential irrespective of the eventual outcome.

5.3 Difficulties with the private use characterisation

The characterisation nonetheless raises four difficulties of construction.

The first concerns the meaning of "private". The natural reading of "private or personal use" in Section 52(1)(a) contemplates use by an individual for purposes not commercial and not public — the domestic copy, the student's note, the researcher's file. The clause has never been understood to describe the internal operations of a commercial enterprise merely because those operations are conducted confidentially. If confidentiality sufficed, the internal reproduction activities of any corporation would qualify, and the clause would swallow a large part of Section 51. To treat industrial-scale corpus assembly as "private" because it occurs behind a firewall is to conflate secrecy with privacy.

The second concerns "research". The word appears in Section 52(1)(a) as an illustration of private or personal use, not as a free-standing purpose. Whatever latitude is available for research is therefore bounded by the private and personal character of the use. Model training conducted as the input to a commercial product is research in the loose sense in which all product development is research, but it is not obviously research of the kind the clause contemplates. Indian authority on the educational and academic limbs of Section 52 has generally required a nexus with non-commercial ends.

The third concerns the "fair dealing" chapeau. Even where a purpose is enumerated, the dealing must be fair. Fairness in Indian law has been assessed by reference to the quantum taken, the effect on the market for the work, and the necessity of the taking. Where the entirety of a corpus is taken, retained indefinitely, and used to build a product that competes for the attention of the same readers who would otherwise consult the source, the fairness enquiry is not obviously answered in the defendant's favour. The Court's conclusion that there would be no market substitution is defensible as to the direct substitution of a chatbot answer for a news article, but it engages less directly with what Chhabria J. in Kadrey termed market dilution — the indirect substitution effected by flooding the market with competing works generated at negligible marginal cost.[9]

The fourth concerns the role of policy. The Court's balance-of-convenience reasoning rested substantially on

[8] ANI Media Pvt. Ltd. v. Open AI OpCo LLC, supra note 2 (holding that storage of the plaintiff's literary works for training prima facie falls within s. 52(1)(a) and does not amount to infringement under s. 51).

[9] Kadrey v. Meta Platforms Inc., supra note 1 (articulating the market dilution theory under the fourth fair use factor).

the economic consequences of a licensing requirement for the Indian AI sector and on the interests of the user base. These are legitimate considerations at the interlocutory stage, where the question is the appropriateness of interim relief. They are less obviously legitimate as inputs into the construction of Section 52, which is a question of statutory meaning rather than of industrial policy. The risk is that reasoning appropriate to the grant of an injunction migrates, in later citation, into the interpretation of the exception itself.

5.4 The transient storage clauses

Clauses 52(1)(b) and (c) were considered but offer little. As noted in Part 4.3, clause (b) is tethered to transmission or communication to the public and clause (c) to linking, access or integration. Neither describes training. Their presence in the Act is nonetheless instructive: Parliament in 2012 addressed itself specifically to the problem of technologically necessary copying and legislated a narrow and carefully bounded answer. That it did so for intermediaries and not for data mining is a reason to doubt that the general clauses were intended to bear the weight now placed upon them.

6. Comparative Perspectives

6.1 The United States

Three district courts have now ruled on the fair use question, and they have not agreed. In Bartz v. Anthropic, Alsup J. held that training on lawfully acquired books was exceedingly transformative and fair, while the retention of a permanent library of pirated copies was not; the case subsequently settled for USD 1.5 billion covering approximately 482,000 works, at an implied rate in the region of USD 3,000 per work.[10] In Kadrey v. Meta Platforms, Chhabria J. reached a similar conclusion on training but on materially different reasoning, treating the downloaded copies as integral to the training effort and advancing the market dilution theory as the analytically correct route to the fourth fair use factor. In Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc., Bibas J. reached the opposite conclusion, finding against fair use in a context where the defendant's product competed directly with the plaintiff's.

Two observations follow. First, the American position is unsettled at the appellate level; the Third Circuit heard argument in the Ross appeal in June 2026 and its ruling remains awaited. Second, and more importantly for Indian purposes, the reasoning in all three cases is inseparable from the open-textured structure of Section 107. The concept of transformative purpose that carries the analysis in Bartz has no counterpart in Section 52. Indian courts that invoke these authorities must therefore do so as persuasive illustrations of policy, not as guides to construction.

6.2 The European Union and Germany

The European Union has legislated directly. Articles 3 and 4 of Directive (EU) 2019/790 on copyright in the Digital Single Market create two text and data mining exceptions: Article 3, a mandatory and non-overridable exception for research organisations and cultural heritage institutions carrying out TDM for scientific research on works to which they have lawful access; and Article 4, a general exception for any TDM on lawfully accessible works, subject to an express reservation of rights by the rightholder in an appropriate machine-readable manner. Article 4 thus adopts an opt-out architecture, coupled in Article 53 of the AI Act with transparency obligations concerning training data.

The German position illustrates that legislation does not guarantee uniform outcomes. In GEMA v. OpenAI the Munich Regional Court declined to accept a TDM justification in respect of song lyrics used to train ChatGPT, reasoning that the reproduction of protected lyrics in outputs fell outside the exception. The decision is a reminder that an input-side exception does not resolve output-side liability, and that the two questions must be kept analytically distinct.

6.3 Japan, Singapore and the United Kingdom

Article 30-4 of the Japanese Copyright Act permits the exploitation of a work where the purpose is not to enjoy the expression contained in it, a formulation broad enough to encompass machine learning and widely regarded as the most permissive in any major jurisdiction. It is nonetheless bounded by a proviso excluding uses that would unreasonably prejudice the interests of the copyright owner. Singapore's computational data analysis exception, introduced by Section 244 of the Copyright Act 2021, is narrower: it requires lawful access to the copy used and prohibits the supply of the copy for any other purpose.

The United Kingdom retains only the narrow non-commercial TDM exception in Section 29A of the Copyright, Designs and Patents Act 1988. The Government's proposal to broaden it to commercial mining, and its subsequent consultation on an opt-out model, generated substantial opposition from the creative sector, and the position at the time of writing remains unresolved. The British experience is instructive for India precisely because the two systems share a closed-list heritage: the difficulty of accommodating machine learning within enumerated exceptions is not an Indian peculiarity but a structural feature of the common law copyright tradition.

What emerges from the comparison is a spectrum. At one end lies Japan, where the purpose of non-enjoyment defines the exception. At the other lies Germany's application of the DSM framework, where output-side reproduction defeats it. India presently occupies no position on this spectrum at all, because it has legislated nothing. Its position is being determined instead by the construction of a clause drafted in 1957 for an entirely different purpose.

7. Critical Evaluation

[10] Bartz v. Anthropic PBC, supra note 1; settlement of approximately USD 1.5 billion covering some 482,000 works, final approval proceedings concluded in 2026.

The foregoing analysis supports a conclusion that is uncomfortable for both sides of the debate. Rightholders are correct that the acquisition, curation and retention of training corpora engage the reproduction right under Section 14(a)(i), and that the express extension of that right to electronic storage makes the Indian textual position stronger for them than the position in several comparator jurisdictions. Developers are correct that the resulting exposure is disproportionate to any identifiable harm, that no Indian exception was drafted with machine learning in mind, and that a strict application of Section 51 to every training run would place the domestic AI sector at a severe disadvantage relative to jurisdictions with express exceptions.

The response of the Delhi High Court has been to resolve that tension through construction, by reading "private or personal use, including research" broadly enough to accommodate commercial model training. The practical result is defensible; the route is not. Three consequences follow from taking that route.

First, it destabilises Section 52(1)(a) beyond the AI context. A construction of "private or personal use" capacious enough to include the internal operations of a multinational technology company is not readily confined to that company. It will be cited by every defendant whose reproduction occurred out of public view.

Second, it produces an all-or-nothing outcome where a calibrated one is required. A statutory exception can be conditioned upon lawful access, upon an opt-out mechanism, upon transparency as to training sources, upon retention limits, or upon remuneration. A judicial construction of an existing clause can be conditioned upon none of these. The court can only hold that the clause applies or that it does not.

Third, it leaves the output-side question entirely unresolved. ANI's allegations regarding hallucinated attributions were not determined, and the memorisation question was decided only on the burden of proof. A developer relying on ANI Media as authority that Indian law permits training would be reading the judgment for considerably more than it says.

There is also a question of institutional competence. The Court's reasoning that a per-source licensing requirement would render model development economically unviable is a proposition about market structure that was not, and in an interlocutory application could not have been, tested by evidence. The emergence of a substantial licensing market for training data — including agreements between major news organisations and AI developers — suggests that the proposition is at least contestable. Where a statutory exception is construed by reference to an untested economic premise, the construction inherits the fragility of the premise.

8. Findings and Recommendations

The study yields five findings. First, corpus acquisition, curation and retention constitute reproduction in a material form under Section 14(a)(i) of the Copyright Act, 1957, and the express inclusion of electronic storage within that provision forecloses any argument that digital ingestion falls outside the right. Second, tokenisation and batching do not constitute reproduction, not by virtue of Section 52(1)(b) or (c), which do not apply, but because transient non-retrievable representations do not amount to storage in a medium. Third, trained model parameters are not ordinarily infringing copies within Section 2(m), save where memorisation of a specific work or passage is demonstrated; the burden of that demonstration presently rests on the rightholder, who lacks the access necessary to discharge it. Fourth, none of the clauses of Section 52 was drafted to accommodate machine learning, and the private use and research limb accommodates it only at the cost of a construction difficult to reconcile with the clause's ordinary meaning and its place in the statutory scheme. Fifth, the resulting uncertainty is itself a harm, disadvantaging Indian developers relative to competitors in jurisdictions with express provision and denying rightholders any reliable basis on which to negotiate.

Four recommendations follow.

The first is the insertion of a dedicated text and data mining exception, as a new clause in Section 52(1), permitting the reproduction and storage of works for the purpose of computational analysis where the person making the reproduction has lawful access to the work. Lawful access should be the central condition. It preserves the distinction that proved decisive in Bartz between lawfully acquired and unlawfully obtained corpora, it aligns India with the European, Japanese and Singaporean models, and it leaves untouched the position of a developer who obtains works from unauthorised sources.

The second is a retention limit. The exception should extend to storage only for so long as is reasonably necessary for the analysis and for verification of its results, with a defined outer limit. This addresses the operation identified in Part 3 as least connected to technical necessity and most closely resembling conventional copying.

The third is a machine-readable reservation mechanism. Rightholders should be entitled to reserve their rights in an appropriate machine-readable form, on the Article 4 model, with the reservation operating prospectively. A registry maintained under the Copyright Rules would reduce the transaction costs that a purely distributed opt-out imposes on smaller rightholders.

The fourth is a collective licensing pathway for reserved works, administered through registered copyright societies under Section 33. Where a rightholder has reserved, a developer should be able to obtain a licence on tariff terms rather than through individual negotiation. This directly answers the economic objection that animated the balance-of-convenience reasoning in ANI Media: the objection is not to paying, but to the transaction costs of clearing rights work by work. A collective mechanism removes the objection without removing the entitlement.

Two ancillary measures merit consideration. A transparency obligation requiring the maintenance and disclosure, on order, of a summary of training data sources would rebalance the evidential asymmetry identified in Part 4.4. And an express saving making clear that the exception operates only at the input stage, leaving output-side liability for memorised or substantially similar

generations to be determined under Section 51, would avoid the difficulty that arose in the German proceedings.

9. Conclusion

The question posed in the title admits of a clear answer. Ingestion by large language models is copying under Section 14 of the Copyright Act, 1957 — not in all of its aspects, but in those aspects that involve the acquisition, curation and durable retention of protected works. The statutory language is not ambiguous on this point; Parliament placed electronic storage expressly within the reproduction right, and no interpretive ingenuity removes bulk corpus assembly from its scope.

The genuinely difficult question is not whether the right is engaged but what should follow. India's closed-list exception regime, drafted for an analogue world and amended in 2012 with intermediaries rather than model developers in view, offers no clause that fits. The Delhi High Court's response in ANI Media — to widen the private use and research limb until it does fit — resolves the immediate dispute at the cost of doctrinal coherence, and does so through a mechanism incapable of the calibration the problem requires.

The better course is legislative. A text and data mining exception conditioned on lawful access, bounded by retention limits, subject to machine-readable reservation and backed by collective licensing would give developers the certainty they need and rightholders the control and remuneration they are entitled to expect. It would also restore the analysis to its proper place: the scope of the reproduction right is a question of statutory construction, but the accommodation of a new technology within that right is a question of legislative choice. Indian copyright law is presently asking its courts to make a choice that Parliament has not yet made.

Declarations

Conflict of interest: The authors declare that they have no financial or non-financial conflicts of interest in relation to the subject matter of this article.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Ethical approval: The study is doctrinal and comparative in method. No human or animal subjects were involved and no personal data were collected; accordingly, no ethical clearance or informed consent was required.

Author contributions: Both authors contributed to the conceptualisation, doctrinal analysis, drafting and revision of the manuscript, and both have approved the final version...

REFERENCES

1. Bansal, A. (2026). Judgment in ANI Media Pvt. Ltd. v. Open AI OpCo LLC, CS(COMM) 1028/2024, 2026:DHC:5900. High Court of Delhi.

2. Bently, L., Sherman, B., Gangjee, D., & Johnson, P. (2022). Intellectual property law (6th ed.). Oxford University Press.

3. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying memorization across neural language models. In Proceedings of the Eleventh International Conference on Learning Representations.

4. Copyright Act, No. 14 of 1957 (India), as amended by the Copyright (Amendment) Act, No. 27 of 2012.

5. Copyright, Designs and Patents Act 1988, c. 48, s. 29A (U.K.).

6. Copyright Act 2021, No. 22 of 2021, s. 244 (Sing.).

7. 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, 2019 O.J. (L 130) 92, arts. 3-4.

8. Geiger, C., Frosio, G., & Bulayenko, O. (2018). Text and data mining in the proposed Copyright Reform: Making the EU ready for an age of big data? IIC - International Review of Intellectual Property and Competition Law, 49(7), 814-844.

9. Ginsburg, J. C., & Budiardjo, L. A. (2019). Authors and machines. Berkeley Technology Law Journal, 34(2), 343-448.

10. Government of India, Ministry of Commerce and Industry. (2021). Report of the Parliamentary Standing Committee on Commerce: Review of the intellectual property rights regime in India (161st Report). Rajya Sabha Secretariat.

11. Lemley, M. A., & Casey, B. (2021). Fair learning. Texas Law Review, 99(4), 743-785.

12. Narayanan, A., & Kapoor, S. (2024). AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference. Princeton University Press.

13. Nazeer, M., & Reddy, T. P. (2024). Fair dealing in the digital age: Rethinking Section 52 of the Indian Copyright Act. Journal of Intellectual Property Rights, 29(3), 201-214.

14. Sag, M. (2019). The new legal landscape for text mining and machine learning. Journal of the Copyright Society of the USA, 66(2), 291-367.

15. Sag, M. (2023). Copyright safety for generative AI. Houston Law Review, 61(2), 295-347.

16. Samuelson, P. (2023). Generative AI meets copyright. Science, 381(6654), 158-161.

17. Scassa, T. (2023). Text and data mining exceptions in copyright law: A comparative perspective. Canadian Journal of Law and Technology, 21(1), 1-28.

18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

19. World Trade Organization. (1994). Agreement on Trade-Related Aspects of Intellectual Property Rights, art. 9(2), Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299....

.