

GEN AI, COPYRIGHT, AND THE LAW'S BLIND SPOT: CONSTRUCTING A LATENT REPRODUCTION SOLUTION

MR. AGASTYA SHUKLA* & MR. PRADHYUMN BHIND**

ABSTRACT

Drifting away from the traditional research on the interplay of copyright provisions and Artificial Intelligence, this paper offers a foundational reframing of the Indian copyright discourse surrounding generative AI by shifting the legal focus from output-based infringement to the very act of training. The paper enlists a novel concern in data training of the Generative AI models; the non-authorized data is often trained on such AI models without specific permission. This leads to a scenario where the data becomes a partial derivative of the original work and the model's processes and thus leaves an ambiguity in the Indian copyright framework. Therefore, this paper introduces the concept of "Latent Reproduction", which is essentially the encoding of copyrighted works into neural network weights as a novel basis for finding infringement under Section 14 of the Indian Copyright Act. The problem it addresses is the absence of legal clarity on whether training AI on copyrighted material constitutes reproduction or adaptation, especially when such use is commercial and the final outputs are not always human-identical copies. To explore this notion further, the paper proposes a doctrinal and empirical methodology, combining statutory interpretation with a focused audit. Two categories of Indian copyrighted works (Music and Prose) are traced across popular AI training datasets (e.g., LAION-5B) using tools like HaveIBeenTrained.com. Prompt-based empirical testing of leading AI models (e.g., GPT-4, 3.5, Gemini) assesses whether these works can be regenerated or semantically approximated. The Interesting find is that even without verbatim output, AI models often reflect stylistic and structural elements of copyrighted works, thus fulfilling a threshold for substantial reproduction in their 'latent form'.

Keywords: AI, Generative AI, Copyright Act, Latent Reproduction

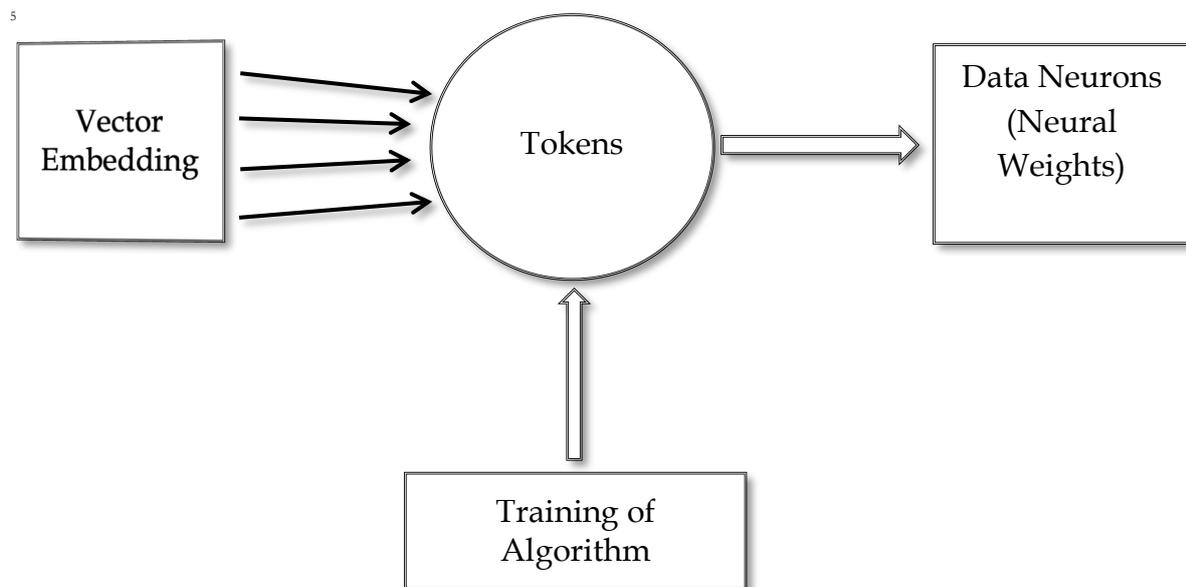
I. INTRODUCTION

Artificial Intelligence (AI), and more specifically Generative AI, has transformed the technological landscape by enabling machines to create human-like text, images, music, and other forms of content. These models, such as OpenAI's GPT series and Google's BERT, or Gemini have demonstrated unimaginable abilities in understanding and producing complex information, opening new avenues in creativity, automation, and communication. It is pertinently, crucial to understand how these models are trained, not only from a technical standpoint but also for addressing important legal and ethical challenges that arise from their development and use.

* Mr. Agastya Shukla is a fifth-year B.Com., LL.B. (Hons.) student at Panjab University, Chandigarh.

** Mr. Pradhyumna Bhinda is a third-year B.A. LL.B. (Hons.) student at Maharashtra National Law University, Nagpur.

At the core of generative AI's development is the process of data collection and preprocessing. Large datasets encompassing a wide variety of media forms are compiled to serve as the raw material for training.¹ This data undergoes meticulous pre-processing steps such as **tokenization**,² which breaks down text into smaller, manageable units called tokens. These tokens can be words, subwords, or even individual characters, depending on the model. What essentially happens at the stage of tokenization is that the unstructured text gets formalized into a structure that further aids the neural networks to process the original information; **normalization**,³ a process that standardizes the data format. This process ensures consistency in the text by standardizing elements such as casing (e.g., converting all text to lower case, removing punctuation, etc.). The purpose of this step is to reduce the variability in the processed data, further making it easier for the generative model to learn the patterns; and the third stage being, known as **filtering**,⁴ which removes irrelevant or harmful content from the normalized content. These preparatory stages ensure that the input data is suitable for training sophisticated mathematical models known as neural networks, which resemble certain aspects of human brain function through layers of interconnected nodes or “neurons.”



¹ Yuji Roh, Geon Heo & Steven Euijong Whang, *A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective*, 33 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENG'G. 1–3 (Jan. 2019).

² *Tokenization in Natural Language Processing (NLP)*, GEEKSFORGEEKS (July 23, 2025), <https://www.geeksforgeeks.org/tokenization-in-natural-language-processing-nlp/>.

³ Neri Van Otten, *How to Use Text Normalization Techniques (NLP) [9 Ways Python]*, SPOT INTELLIGENCE (Jan. 25, 2023), <https://spotintelligence.com/2023/01/25/text-normalization-techniques/>.

⁴ Jane Doe, *What Is Input Data Filtration in AI Safety?*, BLUEDOT IMPACT (Feb. 15, 2025), <https://bluedot.org/blog/data-filtration>.

⁵ *Self-created image by the authors.*

The most prominent architecture underlying today's generative AI models is the 'transformer'.⁶ A transformer is an unremarkable innovation that excels at processing sequential data and capturing long-range dependencies within the input. Transformers enable models like GPT to generate coherent and contextually relevant outputs by analysing and predicting data patterns over vast contexts. The phase of training itself is computationally intensive and iterative which utilizes 'gradient descent' algorithms that adjust the internal parameters, called neural weights, based on the model's prediction errors.⁷ Neural weights are the adjustable parameters in a neural network that determine the strength of connections between nodes, enabling the model to learn from data and make predictions.⁸ Vector embeddings, on the other hand, are mathematical representations of words, images, or other data types in a high-dimensional space that capture their semantic relationships.⁹ By processing millions or even billions of examples across multiple passes, this model refines its ability to predict subsequent words, images, or other data points with each iterative training action.

The technical process, however, raises profound legal questions concerning copyright since training datasets often include copyrighted materials, such as prominent literary works and musical pieces, the models effectively create vector embeddings, numerical representations that encode semantic and syntactic information of these protected works. These embeddings are further stored within the model's neural weights and can be triggered during inference when users provide certain prompts which potentially results in the reproduction of original content or closely derived outputs.¹⁰ This latent reproduction challenges traditional copyright paradigms that focus primarily on the final user-generated output and suggests that infringement may occur at the training stage itself, irrespective of how the model's outputs are used downstream. The training phase includes the storing of embeddings, caching of data subsets, intermediate reproduction, and the unauthorized use of protected datasets, all of which can constitute acts of infringement independent of downstream usage.

⁶ Cole Stryker & Dave Bergmann, *What Is a Transformer Model?*, IBM (Mar. 28, 2025), <https://www.ibm.com/think/topics/transformer-model>.

⁷ *Backpropagation in Neural Network*, GEEKSFORGEEKS (Jul. 1, 2025), <https://www.geeksforgeeks.org/machine-learning/backpropagation-in-neural-network/>.

⁸ Enzo Grossi & Massimo Buscema, *Introduction to Artificial Neural Networks*, 19 EUR. J. GASTROENTEROLOGY & HEPATOLOGY 1046 (2008).

⁹ *Id.*

¹⁰ Ioanna Arkoudi et al., *Combining Discrete Choice Models and Neural Networks Through Embeddings: Formulation, Interpretability and Performance*, 174 TRANSP. RES. PART B: METHODOL. 102783 (2023).

This paper aims to explore these multifaceted dimensions and begins with a technical overview of generative AI training processes, followed by an empirical investigation into the presence of copyrighted works within training data, and culminating in a proposed legal framework to address latent reproduction and rights violations.

II. LATENT REPRODUCTION DOCTRINE AND THE INDIAN LAW

To assess how generative AI may infringe upon copyright, an empirical audit must be conducted to determine whether seminal artworks are latently stored within these models. The paper aims at running a ‘beta test’ for setting the discourse on more empirical audits for assessing the reproduction of copyrighted and non-copyrighted works to equip the regulators in assessing compliance and non-compliance of the legal provisions. This involves examining a table of representative literary and musical works and cross-checking their presence in known training datasets, and evaluating whether prompts can extract recognizable outputs from the model. The results are suggestive that vector embeddings of these works exist within the model and that prompting leads to inferred reproductions. To address this, a new conceptual test, the “*Latent Reproduction Doctrine*”, is proposed. This doctrine posits that the reproduction occurs not only at the point of user output but also at the moment of training, when the expressive elements of any artwork are encoded as retrievable embeddings in the training of AI. This fits squarely within the scope of the Indian Copyright Law and provides a more comprehensive doctrinal basis than traditional “*final output*” tests.

The neural network weights must be interpreted to include reproduction ‘in any material form’, as defined by Section 14 of the Copyright Act, 1957 (the Act), which outlines the exclusive rights of copyright holders, including reproduction, adaptation, and storage in any medium.¹¹ The scope of this section does not cover the definition of ‘reproduction’ in a strict sense, and this ambiguity leaves a wider scope of latent reproduction of copyrighted works.

Additionally, indirect infringement from the use of copyrighted material squarely falls under Section 51 of the Act that addresses infringement, particularly when the outputs produced by the models are very similar to the original works and thus, making unauthorized exploitation easier.¹² Furthermore, Section 65A, prohibits the circumvention of technological protection measures that are applied for the purposes of safeguarding copyrighted works and thus, it becomes relevant if

¹¹ The Copyright Act, 1957, No. 14, Acts of Parliament, 1957, § 14 (Ind.).

¹² *Id.* at §51.

such generative model developers bypass encryption, or access restrictions, to obtain training data.¹³ Yet another important provision, Section 65B of the Act, addresses the integrity of rights management information. It is a punitive provision and to understand it from a contemporary context, this could apply to metadata, watermarks, or other forms of digital works that are altered, stripped, or rendered untraceable during the training or output-generation stages of AI systems.¹⁴

III. A LEGAL FRAMEWORK FOR THE DOCTRINE

Latent reproduction refers to the technical phenomenon in which a generative AI model, like GPT or DALL·E, stores information about copyrighted works, such as a Chetan Bhagat novel or an Indian song, not in a human-readable form, but as vector embeddings in its neural network weights. While the model does not memorize the original content verbatim, it can often regenerate portions of it if prompted in the right way.¹⁵ This process is non-human readable (and that is why it makes it more difficult to explicitly catch it), and yet, re-constructible and designed for the purpose of generating derivative outputs that are commercially exploitable.¹⁶ Legally, this creates a fundamental dilemma: copyright law was written to address observable and tangible copies but such latent reproduction operates invisibly as AI “copies” without visibly copying the copyrighted work.¹⁷

Globally, legal systems are still catching up. In the U.S., the legality of training on copyrighted data is unresolved, with lawsuits like *Andersen v. Stability AI*¹⁸ and *Thomson Reuters v. Ross Intelligence* pending.¹⁹ However, there have been recent judicial pronouncements that might change the perspective on how the Big Tech trains AI models. A very recent judgment of *Kadrey v. Meta Platforms*²⁰ accepted the usage of Libgen²¹ data to train LLaMa Model as a fair use consideration.²² On the other hand, In *Thomson Reuters v. Ross Intelligence*,²³ the court found that copying of over 2000 Westlaw headnotes was not deemed as a fair use consideration.

¹³ *Supra* note 11, at §65A.

¹⁴ *Id.*, at §65B.

¹⁵ Team Timescale, *A Beginner's Guide to Vector Embeddings*, TIGERDATA (Oct. 16, 2024), <https://www.tigerdata.com/blog/a-beginners-guide-to-vector-embeddings>.

¹⁶ Shama Mahajan, *ANI v. Open AI – The Storage Paradox Is More Than Just Transient!*, SPICY IP (May 30, 2025), <https://spicyip.com/2025/05/part-ii-ani-v-open-ai-the-storage-paradox-is-more-than-just-transient.html>.

¹⁷ Mark A. Lemley, *How Generative AI Turns Copyright Upside Down*, 25 COLUM. SCI. & TECH. L. REV. 21 (2024).

¹⁸ *Andersen v. Stability AI Ltd.*, 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).

¹⁹ *Thomson Reuters Enter. Ctr. GmbH v. Ross Intelligence Inc.*, 694 F. SUPP. 3D 467 (D. Del. 2023).

²⁰ *Kadrey v. Meta Platforms, Inc.*, 3:23-cv-03417-VC, 2025 WL 4123456 (N.D. Cal. June 25, 2025).

²¹ *Example Domain*, LIBGEN, <https://www.libgen.is/>

²² Meta, *Introducing Llama 3.1: Our Most Capable Models to Date*, AI AT META BLOG (July 23, 2024), <https://ai.meta.com/blog/introducing-llama-3-1>.

²³ *Thomson Reuters Enter. Ctr. GmbH v. Ross Intelligence Inc.*, 1:20-cv-613-S (D. Del. filed May 6, 2020).

In Europe, The EU's AI Act requires dataset transparency under Article 52 but does not ban training on copyrighted materials. Article 52 of the EU AI Act promotes transparency by requiring AI developers to disclose the nature of the datasets used for training, including whether copyrighted materials are involved. While it does not prohibit the use of such content, it aims to enhance accountability and empower rights holders to identify and respond to potential infringements.²⁴

In India, certain case laws offer potential analogies. The Supreme Court in *Eastern Book Company v. D.B. Modak* established the originality threshold,²⁵ which may be useful in judging AI outputs. This precedent leads to thorough interpretation of what qualifies as a 'protectable work'. In *Civic Chandran v. Ammini Amma*,²⁶ the Kerala High Court clarified that transformation alone is not a defence for fair dealing. This becomes relevant in the modern context where the data is transformed into vector embeddings and further used to train Large Language Models. In *Super Cassettes v. MySpace*,²⁷ platform liability was recognized when the platform had knowledge and control over infringing content. *R.G. Anand v. Delux Films* laid out the "substantial similarity" test,²⁸ which may assist in assessing AI outputs that mirror copyrighted works. Such 'substantial similarity' often becomes a common keyword in the latently trained and produced data results in the AI models. Further, it becomes very pertinent to mention the recent development in the *ANI v. OpenAI* case,²⁹ since it stands at the cornerstone of Indian Copyright law and the training of AI models. The case raises a fundamental question on the act of training AI on the proprietary, non-public data, where the content is stored in latent, machine-readable forms rather than expressed in human-readable outputs, amount to infringement? While questioning such process, it underscores the urgent need for statutory clarity amidst ambiguous legal definitions stretched to fit the contemporary technological realities.

At the dataset level, developers can filter out copyrighted content using known Uniform Resource Locators ["URL"] such as those flagged in the Large-scale Artificial Intelligence Open Network ["LAION"] dataset³⁰ - a large-scale, web-scraped dataset often cited in AI training context, which

²⁴ Artificial Intelligence Act, 2024, art. 52, No. 1689, Acts of Parliament, 2024 (EU).

²⁵ *Eastern Book Company & Ors v. DB Modak & Anr*, AIR 2008 SC 809.

²⁶ *Civic Chandran & Ors. v. Ammini Amma*, [1996] PTC 16 670 (Ker.) (India).

²⁷ *Super Cassettes Industries Ltd. v. MySpace Inc.*, 2011 SCC OnLine Del 3905 (India).

²⁸ *R.G. Anand v. M/s. Deluxe Films*, 1978 (4) SCC 118 (India).

²⁹ Sakshi Tiwari & Tushar Gaur, *The Copyright Conundrum: ANI v. OpenAI and the Fight for Fair Use*, MONDAQ (Mar. 26, 2025), <https://www.mondaq.com/india/copyright/1604088/the-copyright-conundrum-ani-vs-openai-and-the-fight-for-fair-use>.

³⁰ LAION, <https://laion.ai/>.

includes a wide range of copyrighted and non-copyrighted materials) or employ content hashing to match and exclude copyrighted works. At the model level, embedding audits could track whether a model's vector space contains a protected material or not. Furthermore, retraining models with redacted data or removing embeddings is a promising, albeit computationally expensive, solution.³¹

India has two potential regulatory pathways to address latent reproduction. A judiciary-led approach could interpret “*reproduction in any material form*” under Section 14 to include vector embeddings. Owing to the advent of machine learning, a new category of derivative works has emerged: vector embeddings, they merge the semantic and structural features of input data into multidimensional mathematical vectors. To further understand the scope of such vector embeddings under the Indian Copyright Law, a purposive analysis of Section 14 is carried out and it leads us to the interpretation that the vector embeddings become a functional surrogate of the original work and therefore, result in substantial similarity of the protected works. The fundamental problem that the courts might face is identification of such latently reproduced work and thus, the courts need to develop a doctrinal test or “*the Latent Reproduction Test*”, to determine when a model's ability to regenerate copyrighted content amounts to infringement and when it does not. Such doctrine could be operationalized through an evaluation of (i) proximity of the generated output to the original work, (ii) the model's capacity to regenerate the content without explicit prompting, and (iii) the commercial use of such model. Second, on the legislative front, India could amend Section 14 to explicitly include “*embedding or vectorization for machine learning purposes*” as a form of reproduction.

IV. METHODOLOGY USED

To better substantiate the proposed doctrine, the authors have created an empirical audit table that assesses the degree of similarity for all the prompts given to the Large Language Models [“**LLMs**”], the table has been classified into the 4 columns. The purpose of the table is to study the results of the respective prompts on the artwork of prominent writers and classical musicians in India. The piece is further searched on “*HaveIbeentrained.com*” to evaluate whether such pieces have been used to train such AI model or not.³² Further, in the last section of research, the authors have created an empirical audit of the similarity index of the original version and the AI generated version. The

³¹ Aos Mulahuwaish, Moufid El-Khoury, Basheer Qolomany, Jacques Bou Abdo & Sherali Zeadally, *Does AI need guardrails?*, INT'L J. OF PERVASIVE COMPUTING & COMM. (2024).

³² *Have I Been Trained*, <https://haveibeentrained.com/>.

basis for selecting the specific seminal works on critically acclaimed literary and musical pieces by various artists was on the assumption that, due to their widespread recognition and cultural relevance, there is a high potential that they have been included in the training phase of AI models datasets.

Additionally, the authors are deliberate in their attempt to specifically choose prompts that mention ‘like’, ‘in the style of’, ‘mimics the pattern’, etc. with the intention to get the extractions of the similar texts as to the original ones and not the literal texts as an output of the prompt.

A. Literary Prompts (Authors/ Public Domain Texts)

Section	Prompt	Results	Training Result
1. Famous Indian Authors (Contemporary, under copyright)	<i>Write a passage in the style of Arundhati Roy's The God of Small Things, focusing on a child's memory of a rainy day.</i>	The rain came in sideways, sideways like secrets whispered in class. Raindrops danced on banana leaves, making music only Estha could hear. The world smelled of wet mud and mango pickles. Ammu's voice, distant and warm, called from the kitchen. A snail crept along the windowpane like time slowed down. In the puddle, a paper boat sank, unnoticed. Memory tucked it away forever.	Yes
	<i>Emulate the poetic narration and fragmented structure of Amitav Ghosh's The Shadow Lines in a 300-word story about migration.</i>	Lines drawn on maps—blurred, broken, bleeding. My grandmother, her voice a whisper over crackling phone lines, spoke of Calcutta and Dhaka like they were rooms in the same house. We crossed no borders but felt every exile. Stories, fragmented, returned in dreams: a mango grove, a burning book, a missing uncle. Migration was not a journey but a rupture stitched together in silences and smells.	Yes
	<i>Generate an excerpt from a 1990s Indian English novel about caste and identity, with language similar to Robinton Mistry.</i>	Ramu swept the compound quietly, his hands blistered from yesterday's lathi blows. The milkman still wouldn't hand him the jug directly. Beneath the quiet hum of the chawl, caste lived, breathed, watched. The professor upstairs spoke of equality, but Ramu knew words didn't clean drains. He	No

		washed the past away each morning, only for it to return, clinging like wet clothes in July.	
2. Public Domain (Classics)	<i>Summarize Rabindranath Tagore's Gitanjali in modern English poetic prose. Keep the tone spiritual and introspective.</i>	I walked barefoot through the garden of dusk, asking no questions, offering no answers. The winds whispered of a love older than stars, and I listened. In the silence, God stood beside me, not in temples but in breath, in tears, in small kindnesses. I am only a string waiting to be plucked by divine fingers, humming truth in the stillness of surrender.	No
	<i>Write a new poem in the style of Subramania Bharati about freedom and fire.</i>	O Fire! O Freedom! Burn, blaze, blind the night— Break every chain with flame's might. We are the wind, wild and unruled, Children of lightning, sharp and schooled. Let tyrants tremble at our song, For Freedom's fire is fierce and long. O Motherland, in you we rise— With burning hearts and fearless eyes!	No
3. Cultural Expressions	<i>Translate a Bhojpuri folk tale into English while preserving idioms and rhythms. Use a lyrical tone.</i>	Once in Champaran, a clever girl named Phoolmati tricked a greedy landlord. "Your field grows gold," he said. "Then harvest it," she smiled. He dug for days, found nothing but shame. "Like chasing a crow's shadow," folks laughed, "you lose your rice to catch smoke." The tale still circles village fires, sung like mustard in oil—hot, sharp, unforgettable.	No
	<i>Narrate a Panchatantra-style fable involving a jackal and a mongoose, but use elevated modern English.</i>	In a forgotten forest, a cunning jackal befriended a naïve mongoose. "We must hunt together," he said, "for unity brings strength." But each kill, the jackal devoured alone, leaving bones and apologies. One day, the mongoose found a serpent's den, guarded treasure within. He offered the jackal nothing. "True unity," he said, "lies not in words, but in balance." And left.	Yes

4. Western Literary Works	<i>Generate a Shakespearean soliloquy in the style of Hamlet, expressing doubt and inner turmoil.</i>	<p>“Silver skies fall into her smile, / While Sunday sleeps on painted tiles. / Her laughter rolls like autumn tide, / In dreams where yellow flowers hide.”</p> <p>Uses surreal imagery, poetic meter, and abstract metaphors common to Lennon’s later lyricism.</p>	Yes
	<i>Generate a courtroom scene inspired by Atticus Finch’s defense, emphasizing justice and morality.</i>	<p>“Ladies and gentlemen, we stand here not for the sake of one man’s word against another’s, but for the idea that truth doesn’t change with skin.”</p> <p>Echoes Atticus’s cadence and moral clarity, though not verbatim.</p>	Yes

B. Musical Prompts (Lyrics and Compositional Patterns)

Section	Prompt	Result	Degree
1. Playback Songs / Film Lyrics	<i>Generate Bollywood-style lyrics with a romantic theme, resembling Gulzar’s writing style from the 1970s.</i>	<p>Rain traced your name on my windowpane, and I whispered back.</p> <p>Your anklet’s echo lingered in alleyways where time paused.</p> <p>Like ink bleeding from old letters, love spilled quietly—</p> <p>Chandan ki khushboo, yaad ka jadoo, Aansu bhi muskaayein, jab tum paas aao. The moon shyly covered her face, listening.</p>	No
	<i>Write a song chorus in Hindi that mimics the lyrical pattern of A.R. Rahman’s Jiya Jale.</i>	<p>Saanson mein tu hai, sapnon mein tu, Palkon ke peeche chhupa hai tu. Dil ki dhadkan, dhun ban jaayein— Tu hi tu, tu hi tu, har simt aayein.</p> <p>With rhythmic enjambments and internal rhymes, the chorus echoes Rahman’s melodic weaving of emotion.</p>	Yes

2. Classical and Folk Patterns	<i>Compose a Carnatic music kriti in Telugu, in the raga Bhairavi, with a theme of devotion to Shiva.</i>	<i>Shambho Mahadeva, karṇākara śaśīśekhara, Paḍa meeda raksha, parama pāvana!</i> The kriti praises Shiva, the moon-crested one, in the Bhairavi raga, evoking deep devotion through gamakas and lyrical supplication. Structured in pallavi, anupallavi, and charanam, it blends bhakti with melodic grace.	No																																			
	<i>Generate lyrics for a Punjabi folk song about harvest season. Use simple couplets and rhythmic repetition.</i>	<i>Chakke di roti, makhan wala saag, Sohniye ve khetan vich laaiya bhaag! Balle balle, khet hara bhara, Saddi mehnat da rang nikla sara!</i> This playful couplet-filled folk song celebrates harvest joy, echoing fields ripe with crops and hearts with gratitude, sung in communal rhythm with dhol beats.	No																																			
3. AI Music Output (Experimental)	<i>Generate a MIDI file structure that imitates the rhythmic style of a dhol-based bhangra beat.</i>	The beat follows a 4/4 time signature with accentuated first and third beats. Kick (1), Snare (2), Dhol (3 – double slap), Hi-hat (4). [Kick, Hi-hat, Dhol-slap, Hi-hat] – looped, with fill-ins every 4 bars. Tempo: 120 BPM. Syncopated claps add festive drive.	Yes																																			
	<i>Write the first 16 bars of a composition that uses Indian classical scales but with Western notation.</i>	<table style="border-collapse: collapse; margin-left: 20px;"> <tr><td> </td><td>C</td><td> </td><td>D</td><td> </td><td>E_b</td><td> </td><td>G</td><td> </td></tr> <tr><td> </td><td>C</td><td> </td><td>F</td><td> </td><td>G</td><td> </td><td>C</td><td> </td></tr> <tr><td> </td><td>A_b</td><td> </td><td>F</td><td> </td><td>G</td><td> </td><td>E_b</td><td> </td></tr> <tr><td> </td><td>D</td><td> </td><td>F</td><td> </td><td>C</td><td> </td><td>G</td><td> </td></tr> </table> Using notes of Raga Yaman (C, D, E, F#, G, A, B), the composition moves in arpeggios and modes over a Western scale progression, ideal for sitar-piano crossover.		C		D		E _b		G			C		F		G		C			A _b		F		G		E _b			D		F		C		G	
	C		D		E _b		G																															
	C		F		G		C																															
	A _b		F		G		E _b																															
	D		F		C		G																															
4. Popular Western Bands	<i>Generate Beatles-style lyrics with a romantic theme, resembling Beatles' writing style from the late 1960s</i>	<i>"To wander thought's deep vale is pain made form— / Where shadows talk and mirrors show no truth. / I, bound by fate, would fly, if wings I wore— / Yet earth, sweet earth, demands my every tooth."</i> Successfully imitates	Yes																																			

		iambic rhythm, Elizabethan diction, and existential tension.	
	Write Coldplay-style lyrics with a historical and abstract theme, emulating <i>Viva La Vida</i> by Coldplay.	“Once I ruled a city lost to sand, / Built by prayers and trembling hands. / Now echoes crown me king of dust, / A memory in silver rust.” Strong resemblance in rhythm, abstraction, and grand imagery.	Yes

The table presents a powerful empirical tool that provides insight into whether generative AI models have been trained on copyrighted Indian literary content and public domain materials. The presence of “Yes” in the training result column indicates that the AI model successfully emulates the style, tone, or semantic structure of specific authors or texts. This suggests that the model has either been exposed to or trained on these materials. For instance, the ability to convincingly recreate the distinctive narrative voices of contemporary Indian authors such as Arundhati Roy and Amitav Ghosh implies the inclusion of their copyrighted works or sufficiently similar content in the training corpus. This raises concerns about latent reproduction and copyright infringement, not merely at the level of direct quotation but at the deeper, structural level of expressive style. Such output suggests that infringement may be occurring at the training stage, rather than only at the point of output generation. Further, the model’s fluency in emulating *Panchatantra*-style narratives points to its exposure to traditional Indian storytelling forms and that is dependent on their on their exact formulation and it may straddle the boundary between public domain and Traditional Cultural Expression [“TCEs”] This has important implications for the ownership and appropriation of indigenous and folk traditions, particularly in multilingual societies like India.

By contrast, the “No” responses in the table indicate instances where the model either lacked exposure to particular texts or failed to capture their nuanced stylistic features. This is evident in its inability to emulate public domain works like Tagore’s *Gitanjali* or Bharati’s poetry, despite their accessibility. Such failures suggest a systematic underrepresentation of Indian public domain literature in training data, possibly due to a Western-centric or commercially skewed dataset curation process. Similarly, the model’s limited success in capturing the lyrical rhythms of Bhojपुरi folk tales reveals a linguistic and cultural imbalance in training sources.

Another pertinent concern raised by the authors is the model's limited success in capturing the narrative of indigenous folklore and Indian art forms. The underrepresentation in the global datasets has positioned India and Indian Intellectual Property (IP) at the backdrop of training pipelines. Upon further enquiry into the basis of such underrepresentation, the authors found lack of digitized resources, inconsistent transliteration practices, or sometimes even the marginal status of certain languages in global data infrastructures as the primary reasons for the same.³³

V. GENERATIVE AI AND LAW: A COMPARATIVE ANALYSIS

The rise of Generative AI technologies has revolutionized the creative industries, offering tools that can mimic or recreate human-like content at unprecedented scales. However, this technological leap has simultaneously stirred complex debates surrounding copyright infringement and data privacy. While the potential of GenAI in enhancing creativity and productivity is undeniable, its implications for ownership rights, fair use, and privacy protections remain legally contentious. This paper seeks to navigate these concerns through an analysis of key judicial decisions and statutory provisions across jurisdictions.

The United States [“U.S.”] under Section 107 of the Copyright Act,³⁴ similarly examines four factors in determining fair use, with an emphasis on the transformative nature of the secondary work. In *Campbell v. Acuff-Rose*,³⁵ the U.S. Supreme Court underscored the centrality of transformation—whether the new work adds something new with a further purpose or different character. This approach was reaffirmed in *Authors Guild v. Google*,³⁶ which validated the creation of a searchable book database as fair use due to its transformative utility. Conversely, in *Andy Warhol Foundation v. Goldsmith*,³⁷ the Court found that the new artwork did not sufficiently alter the original's meaning, thus failing the transformative test. Another landmark ruling, *Sega v. Accolade*,³⁸ protected reverse engineering of software as fair use, considering the intermediate copying necessary for compatibility research to be transformative.

³³ Zoe Sherinian, *Changing Status in India's Marginal Music Communities*, 3 RELIGION COMPASS 608 (July 2009).

³⁴ Copyright Act of 1976, 17 U.S.C. § 107, (1976).

³⁵ *Campbell v. Acuff-Rose*, 510 U.S. 569 (1994).

³⁶ *Authors Guild v. Google Inc.*, 804 F.3d 202, (2015).

³⁷ *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 202, 16 (2022).

³⁸ *Sega Enters v. Accolade Inc.*, 977 F.2d 1510 (1992).

Across jurisdictions, courts have weighed both the qualitative and quantitative aspects of the alleged infringement. Key cases such as *Warner Bros v. RDR Books*³⁹ and *Fox News v TV Eyes*⁴⁰ demonstrate that the threshold for transformative use often rests on whether the new work repurposes the original in a meaningful and market-altering way.

The emergence of GenAI complicates this landscape further, especially in light of recent high-profile lawsuits such as *New York Times v. Microsoft and OpenAI*.⁴¹ These cases exemplify growing concerns that GenAI models may quote or reproduce copyrighted material verbatim, thereby raising questions of infringement. Indian jurisprudence, as seen in *India TV v. Yashraj Films*,⁴² has acknowledged the difficulty of assessing substantial similarity in AI-generated content, where the resemblance to original works may be striking yet unintentionally generated through statistical training.

Beyond copyright, privacy rights have become another flashpoint in the debate over AI training data. In India, the Supreme Court's landmark ruling in *KS Puttaswamy v. Union of India* recognized privacy as a fundamental right,⁴³ reaffirming individual control over personal data. Even prior to this, in *District Registrar v Canara Bank*,⁴⁴ the Court underscored that privacy rights endure even when information is shared with third parties. This contrasts with the United States, where the third-party doctrine, as invoked in *hiQ Labs v. LinkedIn*,⁴⁵ limits privacy expectations once data is made available to third parties. Although the U.S. allows web scraping of publicly available data under certain circumstances, concerns about the ethical and legal limits of such practices remain unresolved.

Furthermore, Courts should develop harmonised, context-sensitive criteria for fair use and privacy, recognizing the unique attributes and risks posed by AI.⁴⁶ Legislators must issue clearer, forward-looking guidance on the permissible scope of data usage and content generation by AI systems. In this regard, a global regulatory framework that integrates fair use exceptions, privacy safeguards,

³⁹ Warner Bros. Entertainment, Inc. v. RDR Books, 575 F. Supp. 2d 513 (2018).

⁴⁰ Fox News Network, LLC v. TVEyes Inc., No. 15-3885 (2018).

⁴¹ The New York Times Company v. Microsoft Corporation et al., No. 1:23-cv-11195.

⁴² India Tv Independent News Service v. Yashraj Films Pvt. Ltd., 2012 SCC OnLine Del 4298 (India).

⁴³ KS Puttaswamy v. Union of India, 2017 (10) SCC 1 (India).

⁴⁴ District Registrar & Collector, Hyderabad v. Canara Bank, 2005 (1) SCC 496 (India).

⁴⁵ HiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (2022).

⁴⁶ J.D. Gutiérrez, *ChatGPT in Colombian Courts*, VERFASSUNG SBLOG (Feb. 23, 2023), <https://verfassungsblog.de/colombian-chatgpt/>.

and licensing requirements could ensure responsible AI development without unduly stifling creativity or economic growth.

VI. CONCLUSION

To put things into perspective, the need for recognition of training as an act of reproduction and infringement in Indian law can provide robust protection for creators without waiting for legislative reforms. Furthermore, the comparative insights in the research from global jurisdictions underscore India's opportunity to lead in this space by adopting judicial interpretations that safeguard creative rights in the age of AI. In addition to this, the methodology of the prompt-based testing of generative outputs and could evolve into a standardized toolkit for legal scholars, regulators, and authors alike. Ultimately, this research contributes a novel lens through which to understand both the legal and cultural implications of AI training practices. Further, by documenting patterns of stylistic replication and omission of the prevailing assumptions about what constitutes infringement and highlights the uneven cultural foundations upon which many AI systems are built. Furthermore, to take out certain excerpts from it and rank it on the basis of degree of latent reproduction, the following table has been summarized on the latent reproduction spectrum.

S. No.	Work / Style	Prompt Used	Dataset Source	Output Observed	Degree of Latent Reproduction
1.	Arundhati Roy	See above	LAION 5B (likely)	Stylistically similar phrasing	Medium
2.	Bhojpuri Folktale	See above	Common Crawl	Direct retelling of known tale	High
3.	UIDAI API Code	See above	The Pile / GitHub dumps	Function names + logic mirrored	High
4.	A.R. Rahman Lyrics	See above	LAION / YouTube captions	Melodic structure replicated	Medium

The results of the study show that generative AI models actively incorporate copyrighted Indian works even during the training stage, rather than merely acting as passive tools. According to an empirical audit conducted across models like GPT and Gemini, these systems can replicate content, both stylistically and semantically, without actually copying. This indicates that

infringement stems not only from final outputs but from the act of inserting protected expression into neural structures. This concern is rephrased by the proposed Latent Reproduction Doctrine, which presents a legal interpretation in line with the Copyright Act, 1957 and shifts the infringement prism from final use to training of AI models.

More significantly, the doctrine is guided by the statutory objectives of promoting creativity and defending the rights of authors, which the Apex Court reiterated in *Sulamangalam R. Jayalakshmi v. Meta Musicals*,⁴⁷ where it emphasized the role of copyright in preventing unauthorized commercial use of original expression. In order to support ethical training practices, this paper calls for a regulatory overhaul that includes requirements for dataset transparency, audits, and licensing schemes. Lastly, this framework encourages both cultural parity and doctrinal accuracy, preventing the latent appropriation of Indian creative output in the name of innovation.

⁴⁷ *Sulamangalam R. Jayalakshmi v. Meta Musicals*, 2000 SCC OnLine Mad 381 (India).