Remuneration for Copyright in AI

Limits and Implementation Challenges

reglab
center for strategy
& regulation

About Reglab

We are a private research center specializing in the media and technology sector, supporting companies, associations, and policymakers in making strategic decisions based on data and evidence.

Learn more at www.reglab.com.br

About the Policy Briefs Series

The Policy Briefs Series brings together studies that evaluate trends, existing public policies, or legislative proposals, using both qualitative and quantitative data to inform and guide decision-making. The goal is to present complex topics in an accessible way, highlighting key points of analysis, impacts, and possible recommendations.

Acknowledgements

Executive Director: Pedro Henrique Ramos

Research Coordinator: Marina Garrote

Authors: Pedro Henrique Ramos, Julia de Albuquerque Barreto, Marina Garrote

Researcher: Stephanie Mathias de Souza

Final Layout: Eliza Natsuko Shiroma

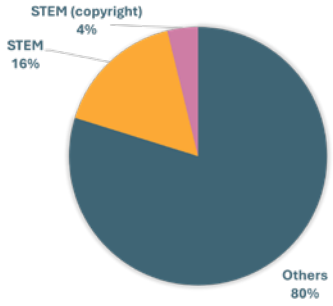
Suggested Citation: RAMOS, P. H.; BARRETO, J.; GARROTE, M. ***Remuneration for Copyright in AI: Limits and Implementation Challenges.*** *Reglab Policy Briefs, no. 3. São Paulo: Reglab, 2025.*

Executive Summary

The debate surrounding copyright and Generative Artificial Intelligence (Gen AI) is currently one of the most pressing issues in regulatory discussions, attracting a wide range of perspectives and opinions. Reglab has chosen to approach this topic from a **perspective that remains largely underrepresented in regulatory debates: the viewpoint of professionals from STEM fields** (Science, Technology, Engineering, and Mathematics).

To this end, we conducted interviews with **computer scientists, software engineers, machine learning experts, and university professors.** The objective was to understand how the training of Gen AI models involves the use of content protected by copyright, as well as to explore the technical challenges associated with implementing remuneration mechanisms linked to such usage.

This is a pioneering study in Brazil. To illustrate the relevance of this data, we analyzed the 24 public hearings held by CTIA – the Temporary Internal Commission on Artificial Intelligence in Brazil – which examined Draft Bill No. 2,338/23. Our findings revealed a notably limited participation of STEM professionals in discussions on Gen AI and copyright.



Participant profile and contributions of CTIA - Internal Temporary Committee on AI

Word cloud generated using Wordclouds.com, based on data automatically extracted from Atlas.ti using the "Concepts" tool

The **main findings of this study** are as follows:

- Data selection is a complex and critical process for AI models. **Not only the quantity but also the quality and diversity of the data** play a decisive role in model performance;

- Although there are technical approaches that allow for the tracking of data provenance and flow in training datasets, the experts interviewed indicated that **there are still no scalable and reliable solutions capable of measuring the specific contribution of each individual work within large-scale models.** This does not appear to be a strategic business decision but rather a structural limitation of current technology, particularly machine learning systems;

- This limitation arises from the fact that **machine learning models do not store data in a directly queryable database.** Instead, they encode information into statistical patterns — mathematical representations derived from probabilistic models — by breaking data down and converting it into numerical vectors. As a result, determining the precise impact of each copyrighted work on the final model is, in practice, impossible;

- **These technical challenges undermine traditional copyright remuneration mechanisms**, which fundamentally rely on quantifying the use of copyrighted works to establish payments. In the absence of accurate measurement, licensing agreements may disproportionately benefit major rights holders with substantial legal resources while disadvantaging independent creators, who would lack the means to track the use of their works.

When asked about the potential consequences of imposing strict restrictions on data availability — stemming from the application of licensing and copyright rules in Brazil — experts emphasized that, due to the global nature of the internet, **Gen AI training could easily be relocated to other jurisdictions. This would weaken the domestic AI ecosystem, rendering local regulations ineffective** and adversely affecting Brazil's regulatory credibility and competitiveness.

Additional impacts identified:

- **Decreased Model Quality:** Models trained with restricted datasets may present reduced accuracy and limited generalization capabilities;

- **Increased Development Costs:** The need to negotiate individual licenses for each data input would significantly raise costs, making AI development unfeasible for startups;

- **Market Concentration:** Companies with exclusive access to large proprietary datasets would Gen AIn a significant competitive advantage, hindering open innovation;

- **Economic Effects:** Should "usage" not be adopted as the primary metric for remuneration, alternative measurement frameworks may generate market distortions and reinforce existing structural inequalities within the sector;

- **Relocation of AI Hubs:** In the event of the implementation of overly restrictive rules in Brazil, there is a strong likelihood that Gen AI development hubs would relocate to more permissive jurisdictions.

The primary contribution of this study is to demonstrate that **a technically grounded and realistic understanding of how Gen AI models operate is essential for effective regulation.** Moreover, it highlights the urgent need to expand the participation of STEM professionals in the legislative process concerning artificial intelligence governance.

TABLE OF CONTENTS

Interactive Table of Contents

Executive Summary 3

1. Introduction 6

1.1. What is Generative Artificial Intelligence and Why Does It Matter? 6

1.2. Data Mining, Gen AI and Copyright 8

1.3. The Current Stage of the Debate in Brazil: Bill No. 2,338/23 10

1.4. The Methodological Approach of This Study 12

2. Results 14

2.1. The Quantity, Quality, and Diversity of Training Data Directly Impact Model Performance 14

2.2. Data Reduction May Lead to a Decline in the Quality of Gen AI Models 18

2.3. Individual Data Licensing Could Render the Development of Brazilian Models Unfeasible 19

2.4. Market Concentration: Exclusive Access to Datasets May Benefit Only Large Companies 20

2.5. Relocation of AI Hubs: Strict Regulations in Brazil Could Be Easily Circumvented—With Significant Economic and Social Impacts 20

3. Analysis and Commentary 22

3.1. The Lack of Technical Expertise in Public Policy Formulation on Gen AI 22

3.2. Economic Distortions: Criteria Beyond "Usage" Favor Companies With Large Proprietary Databases, and There Is No Evidence of How This Compensation Would Reach Creators 23

4. Conclusion 25

4.1. Direction for future studies 25

References 27

Methodology Annex 28

1. Introduction

Imagine an individual seeking information about the Brazilian economy during the 1990s. Instead of consulting a news article or a book, they pose the question to an artificial intelligence chatbot. Within seconds, the system delivers a clear, well-structured, and accurate response. While no content is reproduced verbatim, the model was trained on vast amounts of publicly available text from the internet, including news articles from that period, academic papers, and Wikipedia entries.

Does the way this chatbot used those texts infringe upon the rights of the authors whose works were used during the training process? This is a complex question that lies at the heart of a much broader debate: **the intricate relationship between generative artificial intelligence (Gen AI) and copyright law.**

Copyright law
Consists of rules designed to protect creators of intellectual works, such as music, texts, and images, granting them the ability to control how their creations are used and to receive compensation when others make use of these works. In Brazil, these rights are governed by Law No. 9,610/1998, which also establishes the applicable exceptions and limitations.

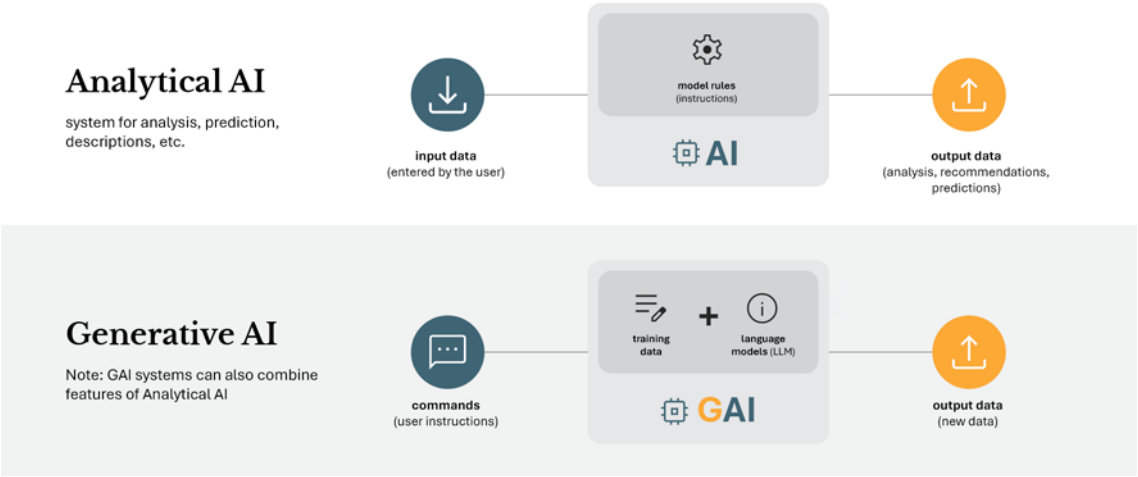
However, these discussions often take place without a thorough analysis of the technical foundations of Gen AI — an analytical gap that this study seeks to address. Our aim is **to translate the technical and operational aspects of Gen AI into accessible insights that can inform the regulatory debate surrounding Gen AI and copyright.** Our objective is not to prioritize political, economic, or legal dimensions — indeed, it is vital that the debate remains guided by a plurality of perspectives. Nevertheless, we contend that the technical dimension is essential for ensuring that the discussion progresses based on concrete evidence and feasible solutions.

1.1. What is Generative Artificial Intelligence and Why Does It Matter?

In this study, we define Generative Artificial Intelligence (Gen AI) **technologies as systems that employ statistical methods and machine learning techniques to generate new texts, images, or other types of content** (Daase et al., 2024). Unlike analytical models, which interpret, classify, and make decisions based on data (Amorim, 2025), Gen AI systems are capable of producing new data — such as texts and images — by leveraging patterns extracted from large datasets known as training data.

Datasets, or training data, are organized collections of data — such as texts, images, or videos — used to train Gen AI systems. These datasets enable the system to “learn” patterns and improve its ability to generate coherent and relevant outputs.

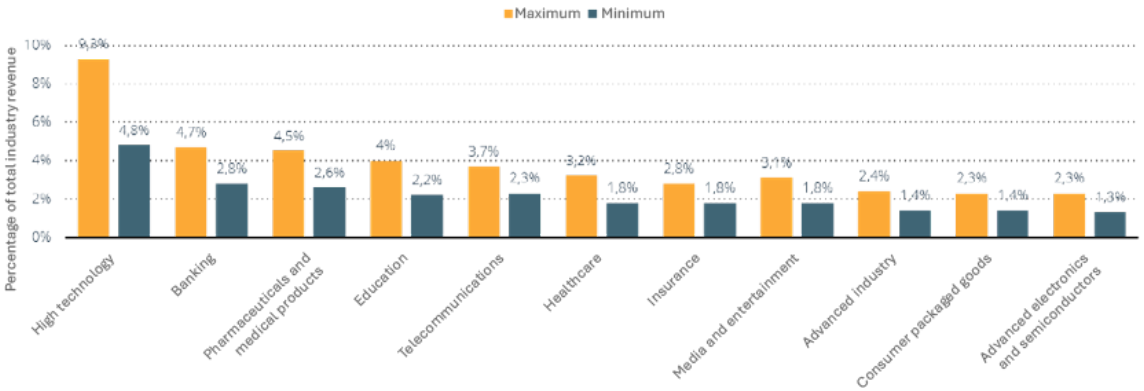
Figure 1. Differences between Analytical and Generative AI models



Source: Reglab's elaboration, based on Ramos (2023).

Gen AI-based systems have experienced rapid adoption over the past three years, and their economic impact is significant. A study by McKinsey (Chui et al., 2023) projects that the use of these technologies could generate up to a 5% increase in global economic growth over the next five years, directly impacting sectors such as agribusiness, insurance, consumer goods, and the pharmaceutical industry.

Figure 2. Future economic impact of generative AI on organizations worldwide in 2023, by economic sector.



Source: Chui et al (2023).

The Gen AI economy is not composed of a single player but rather functions as a structured ecosystem, with different companies performing complementary and interdependent roles. This ecosystem can be summarized into three main layers[1]:

i. **Infrastructure:** This layer comprises hardware manufacturers responsible for producing high-performance chips and data centers, which are essential for processing large volumes of data and supporting computationally intensive tasks;

ii. **Models:** This layer includes companies that develop and license foundational models, such as Large Language Models (LLMs). These models are based on neural networks with billions of parameters, trained on vast amounts of data, and are primarily designed for text generation; and

1 Other studies suggest a different division of layers, into four or even six (Simmons, A., 2023; Epical, 2024). For didactic purposes, and explicitly based on the model proposed by Benkler (2006), we have chosen to simplify it into just three layers.

iii. **Applications:** This layer encompasses companies that develop and provide software systems which, leveraging both the models and the underlying infrastructure, deliver solutions and services to end users—chatbots being one of the most well-known examples.

When discussing **training data, our primary focus is on the model layer.** It is important to clarify that the way these datasets are processed by the models differs significantly from how data is handled in applications based on storage or content reproduction, such as music or video streaming services (this distinction will be further explored in the presentation of this study's findings).

1.2. Data Mining, Gen AI and Copyright

There are at least two distinct debates concerning copyright and Gen AI that must be clearly differentiated. The first concerns the protection of works generated by AI systems (for instance, *when an AI creates a piece of music, who is considered the author?*). The second relates to copyrighted works that are embedded within the training data used to develop AI models. **This paper focuses exclusively on the latter — a debate that actually predates the emergence of Gen AI itself** (Fill-Flynn et al., 2022).

This is because the practice of **data mining** — a process that employs statistical methods to identify patterns and correlations within data — began to Gen AIn traction as early as the 1990s (Coenen, 2004). Around the same time, the **technique of crawling** emerged, in which automated systems systematically scan websites, web pages, or databases to analyze content, index it, and subsequently incorporate it into processes such as data mining.

***Crawling*, data mining, and machine learning are foundational techniques not only for Gen AI but also for a wide range of applications, including internet search engines, price comparison tools, scientific article indexing services, and platforms that monitor open government data.**

RECAP:

Crawling: The automated collection of data, such as texts and images, to create databases that will serve as the foundation for further analysis.

Data Mining: The process of analyzing large volumes of data to identify patterns and correlations, either prior to or independently from its use in training Gen AI systems

AI training: The stage in which a Gen AI system learns from data, adjusting its parameters to better recognize patterns and generate outputs.

Machine Learning: The process through which a system continuously improves its outputs by identifying patterns within the data during training.

The significance of these processes is so substantial that, in recent years, several countries have incorporated exceptions for data mining activities into their copyright laws, particularly when related to research or innovation in both the public and private sectors:

In Europe, countries such as the United Kingdom and Germany have adopted broad exceptions for data mining and training. More recently, the European Union, through the AI Act, has also incorporated specific provisions addressing this issue (Rosati, 2024).

Japan has distinguished itself by adopting a proactive stance that encourages the use of data for research and development in both the public and private sectors (Ueno, 2025).

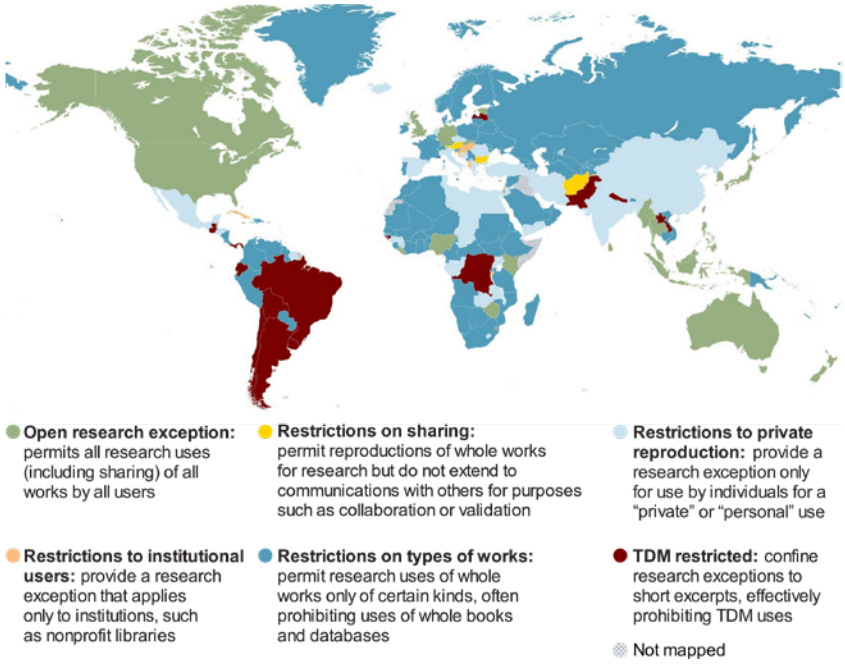
In the United States, the judicial doctrine of fair use has traditionally been interpreted as a valid exception for data mining and training. However, recent legal disputes have introduced significant legal uncertainty regarding how this concept is applied to Gen AI, raising concerns even among prominent civil society organizations (Noble, 2025).

In China, a similar degree of legal uncertainty exists, although the legislation appears somewhat clearer in supporting exceptions for data mining when compared to the U.S. framework (Karaganis, 2024).

In South America, the legal landscape is markedly different. Copyright laws across the region have not introduced specific exceptions for data mining and training, a gap that creates considerable legal uncertainty for investments in data centers in the region and poses significant barriers to the development of local technologies (Schirru et al., 2024).

The reliance on models trained in other jurisdictions may limit the ability of Latin American countries to develop technologies that are aligned with their cultural, linguistic, and social contexts. Applications in critical areas such as public health, justice, education, and local culture may be particularly impacted.

Figure 3. Copyright Exceptions for Research, Training, and Data Mining



Source: Fill-Flynn et al, 2022.

1.3. The Current Stage of the Debate in Brazil: Bill No. 2,338/23

The legislative debate on Gen AI in Brazil Gen AIned renewed momentum with the approval, by the Federal Senate, of Draft Bill No. 2,338/23. This bill originated from a draft prepared by a commission of legal experts and was introduced in 2023 by the President of the Senate, Rodrigo Pacheco. It consolidates provisions from seven other legislative proposals, including Draft Bill No. 21/2020, which had previously been approved by the Chamber of Deputies in 2021 but had since stalled in the Senate[2].

Inspired by the European Union's AI Act and normative frameworks in the field of personal data protection, the bill proposes a risk-based regulatory framework, coupled with a set of safeguards for individuals affected by AI systems. **Among the rights guaranteed are prior information regarding interactions with automated systems, the right to privacy and data protection, and the right to non-discrimination.** For AI systems classified as high-risk, the bill also establishes additional safeguards, including the right to explanation, the right to contest decisions, and the right to human review of automated decisions.

With respect to copyright, Draft Bill No. 2,338/2023 adopts a more restrictive approach compared to the European Union's AI Act and the legislative frameworks of other countries. Specifically, Articles 62 to 65 of the bill:

- Create exceptions for data mining and AI training exclusively for scientific and educational institutions, museums, archives, and libraries, provided that the use is non-commercial and that the data is lawfully accessed
- Establish transparency obligations for AI developers, including the public disclosure of the datasets used for training purposes; and
- Introduce remuneration mechanisms, allowing for either collective barGen AIning or direct negotiation with copyright holders, taking into account the size of the company and its economic impact.

The proposal generated immediate reactions. On one side, representatives from the cultural sector and creators emphasized the unprecedented nature of the measure and its commitment to protecting copyright in the digital age. On the other side, concerns emerged regarding the technical feasibility of the requirements and the potential negative impact on Brazil's competitiveness within the global AI development and innovation ecosystem.

Copyright remuneration mechanisms ensure payment for the use of creative works. Some examples include:

- **Music:** Rights holders receive payments for public performances or digital uses. In Brazil, ECAD (Central Office for Collection and Distribution) is responsible for collecting and distributing royalties to composers, performers, and publishers.
- **Audiovisual:** Screenwriters and directors are compensated for the exhibition of their works.

2 Source: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262 . Acessed on: May 12, 2025.

"In any type of economic activity, there is an essential input, and those who coordinate that activity must pay for it. In the case of artificial intelligence, the primary input is creativity—it is the result of what individuals have been able to create. This creativity is mined by the company developing the AI system, and that company must also pay for it, in recognition of the creative contributions embedded in musical, literary, or any other form of production."

Senator Humberto Costa (PT-PE)[3].

Draft Bill No. 2,338 does not align with international trends that seek to strike a balance between copyright protection and the development of AI. Countries such as Singapore and Japan broadly allow the training of AI models and systems. The European Union, in turn, has adopted more flexible rules that permit the computational analysis of publicly available works to enable AI training, recognizing the importance of fostering innovation in this field while simultaneously ensuring that rights holders can technically indicate their decision to opt out of having their works used for training purposes.

ABAG - Brazilian Agribusiness Association[4].

Figure 4. Comparative Table of Legislation Related to AI Model Training

How Does Draft Bill No. 2,338/23 Compare to the Legislation of Other Countries?

Question	USA	EU	China	Japan	Brazil
Can Models Be Trained Using Copyrighted Works That Are Publicly Available?	✓ **Yes,** if "transformative" *(under the fair use doctrine)*	✓ **Yes,** unless the rights holder has indicated their opt-out.	✓ **Yes,** there is a legal exception for AI training, although it is not explicitly stated.	✓ **Yes,** there is a legal exception for AI training.	✗ **No,** training is not an explicit legal exception if there is a commercial or for-profit purpose.
Can Copyright Holders Prevent Their Works From Being Used for Model Training?	✗ **No,** unless it is judicially proven that it does not qualify as fair use.	✓ **Yes,** it is possible through a technical opt-out *(via metadata)* or through licensing.	✗ **No,** the right to opt-out is not provided for under the law.	✗ **No,** the right to opt-out is not recognized by law.	✓ **Yes** — under an opt-in rule, companies must negotiate the use of works before training.
Can Copyright Holders Claim Compensation From Companies That Use Their Works for Model Training?	? **Uncertain** — it depends on a judicial decision.	! **Partial** — only if the opt-out is disregarded.	? **Uncertain** — there is an ongoing legal dispute on the matter.	! **Partial** — the law grants this right in cases of misuse or plagiarism in the output.	! **Partial** — the right to compensation exists, but there are no clear criteria for its calculation.

Source: Reglab's elaboration.

3 Source: https://www12.senado.leg.br/noticias/materias/2024/12/10/senado-aprova-regulamentacao-da-inteligencia-artificial-texto-vai-a-camara. Acessed on: May 12, 2025.

4 Source: https://abag.com.br/regras-equilibradas-de-direitos-autorais-e-a-competitividade-do-brasil-em-inteligencia-artificial/. Acessed on: May 12, 2025.

1.4. The Methodological Approach of This Study

This context underscores the relevance of this research: at a time when Brazil is actively discussing its legal framework for AI, **policymakers must develop a deep understanding of the technical dimensions involved.** Critical questions — such as the feasibility of tracking and quantifying the use of copyrighted content, attributing its individual contribution to a model's output, estimating the costs of a potential compensation system, and assessing who would actually benefit from such measures — **must be answered based on empirical evidence before any legislative solutions are adopted.**

> **This study aims to understand how the training of Gen AI models involves the use of copyrighted content, and to identify the technical challenges related to the feasibility of remuneration mechanisms associated with such use**.

In this study, we combine two methodological approaches. The first is evidence translation, a method still underexplored in the field of digital governance in Brazil, which aims to **produce robust and accessible evidence to support public decision-making** (Ingold, 2025).

> Whenever we use pink-highlighted boxes, charts, or featured examples in the layout, we do so intentionally. We are aware that this approach carries the risk of some technical imprecision; however, we believe that, **within the framework of translating complex evidence into applied knowledge, enhancing clarity and accessibility is a necessary methodological choice** — and a position we embrace with full transparency.

The second approach is qualitative research. Rather than relying on traditional literature reviews and desk research, we conducted **semi-structured interviews** to capture the perceptions and experiences of a group that is **often absent from regulatory discussions: STEM professionals** (Science, Technology, Engineering, and Mathematics).

Inspired by reception studies, we sought to understand how these professionals interpret the technical challenges related to the intersection of AI and copyright. Over the course of one month, we conducted eight interviews with experts in the field, focusing on senior-level professionals with both academic training and practical experience in STEM disciplines. The interviews followed predefined scripts and confidentiality protocols, and the transcripts and field notes were analyzed using Atlas.ti software, applying the thematic analysis technique to identify key patterns and insights.

Figure 5. Descriptive Table of Interview Participants

Interviewee	Description
1	Female, PhD, and data scientist at a large Brazilian company in the software sector.
2	Male, PhD, data scientist and university professor in the field of technology.
3	Male, data scientist at a large Brazilian company in the financial sector.
4	Male, PhD, data scientist and university professor in the fields of technology and business administration.
5	Male, software engineer and executive at a Brazilian startup.
6	Male, MSc, electrical engineer, and AI solutions architect at a big tech company.
7 *	Female, artificial intelligence professional at a big tech company.
8 *	Male, machine learning consultant at a Brazilian startup.

* Preliminary interviews

Source: Reglab's elaboration.

The complete methodology, including detailed information about the procedures adopted, is provided at the end of this study.

2. Results

2.1. The Quantity, Quality, and Diversity of Training Data Directly Impact Model Performance

The interviewees explained that Gen AI models are highly dependent on the quality, diversity, and quantity of data, and that there is no strict hierarchy among these factors — their importance depends on the specific objective of each model.

- **Data quality** is critical. Content containing errors, biases, or incomplete information compromises the model's inferences and may lead to the reproduction of distortions or omissions.
- Similarly, **data diversity**, in terms of languages, cultures, styles, and contexts, is essential to ensure inclusive and generalizable outputs.
- **Data quantity** is also relevant, particularly because of the mathematical models employed. As one interviewee noted, **"For neural models, the more data, the better."**[5].

However, none of these factors alone guarantees strong performance. For instance, models trained on large but homogeneous datasets may still reproduce biases and present significant limitations in applicability. As one interviewee stated:

> *"You have to be careful, because quantity is not the same as the number of images. There's no point in feeding a trillion images if they're all similar."*

Some interviewees noted that smaller, more specialized models may actually be more effective for certain applications, in addition to being more cost-efficient. This observation aligns with recent academic experiments aimed at developing smaller datasets whose diversity and quality compensate for their limited quantity (Gao et al., 2020; Leffer, 2025).

It is also noteworthy that several interviewees pointed out that some of the most popular models have already exhausted the publicly available data on the internet, gathered through crawling. This means that future differentiation will likely depend on either **(i) the technical performance of the models** — such as increased processing capacity, innovative computational methods, or enhanced personalization — or **(ii) the incorporation of datasets that cannot be captured via crawling,** but whose quality can become a key competitive advantage. This explains why many companies are now seeking licenses to access historical newspaper archives, which are typically private and not publicly available online (Barcott, 2025).

5 In order to preserve the anonymity and confidentiality of the research participants, minor modifications were made to the quotations presented in this study. In certain cases, specific linguistic adjustments were applied to ensure that the interviewees' original intent was accurately reflected in the textual transcription. The integrity of the discursive record was preserved whenever possible, in accordance with the established methodological principles.

Not all data can be collected through crawling. This is because many datasets are protected by technical barriers (such as paywalls), require login credentials for access, or are subject to legal restrictions, including sensitive personal data. Additionally, some content exists in formats that are not automatically accessible, such as offline files or private collections. These factors limit the reach of crawling and require alternative methods of access or authorization, which may involve financial agreements between companies.

These technical findings offer direct insights for the ongoing debate on copyright and Gen AI, highlighting that remuneration models based solely on the volume of works used **may fail to capture the actual impact of each contribution on a system's performance.** A more balanced approach would need to consider not only the quantity but also the quality and contextual relevance of the works used in training — a task that poses significant technical challenges, as will be further discussed later in this study.

The Technical Infeasibility of Measuring the Contribution of Works in Gen AI

The interviewees explained that large-scale models do not operate through direct data indexing (as a library does), but rather through statistical patterns extracted from the data. Each work is broken down into words, which are then transformed into billions of vector representations with no direct links to the original files — and, in fact, the original data is not even stored.

Thus, any attempt to identify how much an individual work contributed to a specific output is technically unfeasible.

This is because, during training, a model analyzes large volumes of data within each dataset to adjust its own mathematical representations — unique to that model — **without necessarily copying or storing the data itself.** In other words, while a music app processes information to reproduce content, Gen AI systems process information to generalize it.

It is precisely because of this generalization that a Gen AI system trained on millions of images can produce a new visual composition that replicates common characteristics of 16th-century artworks, without reproducing any specific work — merely incorporating recurring elements from various references.

EXAMPLE: How does ai transform numbers into new content?

During training, AI analyzes millions of data points and converts each of them into a **mathematical representation** — also known as a **vector.** These vectors represent the characteristics of what the system has learned. Let's look at an example to understand how a word can be transformed into a vector consisting of hundreds (sometimes thousands) of numbers:

Dog → [0,2, -1,3, 7,8, **-0,4**, ...]

These numbers have no isolated meaning. **What matters is how they relate to other vectors.** For instance, the word "hot" might not seem directly related to "dog", but the system could assign some correlation — something detectable by the repetition or similarity of certain values in the vector:

Hot → [65,1, -1,12, 32,8, **-0,4**, ...]

Over time, the model learns more and more correlations — that is, how different vectors can be connected to one another. This is an intensive process — we are talking about trillions of vectors — and it requires enormous data processing power and highly complex calculations. These processes are referred to as **neural networks** because of their resemblance to the functioning of biological nervous systems.

Now, imagine typing the following prompt into a chatbot:

"Complete the sentence: For lunch, I ordered a hot ____"

The first step the Gen AI model takes is to convert this prompt into **vectors**, turning words into sequences of numbers. Then, the model searches for correlations: **which of these numbers are related to others** it already knows. **This is a probability calculation** —the model does not choose words randomly but instead **selects the most statistically probable next word.** It's as if the model is asking itself: **"Based on this prompt, what is the most likely word to come next?"** Let's simplify and revisit our example vectors:

Lunch → [0,2, -1,3, 7,8, **-0,4**, ...]

Dog → [65,1, -1,12, 32,8, **-0,4**, ...]

Hot → [0,7, -8,3, 7,1, -**0,4**, ...]

The model seems to have found a correlation! If this is the most statistically probable match, the model will then produce the following output:

"For lunch, I ordered a hot dog."

What's most interesting is that — even though the model has learned from thousands of phrases and pieces of content — **this output, as simple as it may seem, is a new**

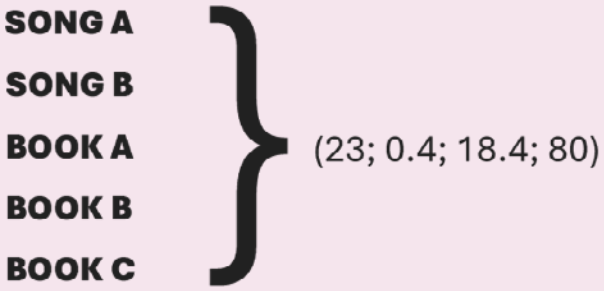
combination of words, generated from statistical patterns.

This characteristic distinguishes AI models from other applications, such as streaming services, which function more like digital libraries. In those cases, consumption can be directly linked to a discrete unit of content, **making it possible to attribute usage to a specific output.** In contrast, Gen AI techniques lack metadata structures or tracking systems capable of reconstructing the cause-and-effect relationships between input data and output results.

As one interviewee explained:

> *"I can confirm that author x was used to train a model, but the model won't be able to give a precise answer like: 'for this response, i used this specific text from author x.' (...) in other words, did it actually use author x's text that was part of the training, or did it, for example, use other texts written by people about author x?"*

Let's imagine that the model learns to correlate the word "rain" with "sadness" based on two song lyrics (**Song A** and **Song B**) and three books (**Book A, Book B,** and **Book C**). This correlation becomes so strong that it generates a specific vector.



When the model produces the sentence **"rain is sadness"**, it may be possible to audit and identify that the vector (23; 0.4; 18.4; 80) was used. However, it would not be possible to determine which of the five data sources contributed to this result, since there was no storage or indexing—only machine learning.

In other words, any attempt to isolate the influence of a single work or a narrow set of information becomes highly complex. The notion that it would be possible to calculate the "weight" of an individual work in a model's performance is fundamentally misaligned with the statistical functioning of machine learning systems, which learn through diffuse patterns and probabilistic recurrences. Another interviewee explained this issue by illustrating how models process different languages and transform concepts into vectors:

> *"In a model, when you write a question in either Portuguese or English, the first thing the model does is transform that sentence into a mathematical representation that is already language-agnostic. That's something truly remarkable. Imagine taking the word 'dog'— the model*

> *will convert it into a mathematical vector that represents 'dog' in any language. Across all languages, it will arrive at the same concept".*

The discussion on quality also touches upon the tension between cultural value and statistical value. A work may hold immense cultural significance (e.g., an excerpt from a literary classic), but for an AI model, its specific statistical contribution may be negligible — a reality that makes the development of a copyright remuneration system particularly complex. This conclusion is supported by a recent experiment conducted by De La Rosa et al. (2024), which demonstrates that fictional works are not particularly decisive in influencing model performance.

Let's imagine a model that has been trained with two different datasets: the first contains books written by Author X, and the second consists of academic articles analyzing Author X's work.

Now, consider someone asking the chatbot: "Write an original paragraph in the style of Author X." We know that the model used information from both datasets to generate this response. However, since the data was transformed into numbers and statistical patterns during training, it is impossible to determine which dataset contributed more to the final output.

Even though both datasets influenced the result, we cannot measure which was more significant. This leads to a critical question: **how can we fairly compensate those who contributed the most if it is impossible to identify the relative weight of each contribution?**

Finally, several interviewees emphasized that this issue does not appear to be a strategic business decision, but rather a limitation of the current state of the art in technology. In this regard, our research identified a recent surge of interest in this topic at academic conferences and in university-led experiments. For example, working papers by Wang et al. (2024) and Zhang et al. (2025) employ game theory techniques to attempt to estimate these weights. However, they acknowledge a range of methodological limitations, including computational complexity, the fragmentation of data across diverse sources, and the inherent difficulty of accurately identifying which works are protected by copyright.

2.2. Data Reduction May Lead to a Decline in the Quality of Gen AI Models

Among the interviewees, there was unanimous agreement that restricting data usage —whether due to regulations, costs, or legal risks — directly impacts the quality of AI models. **The smaller the available dataset, the more limited the universe the model is capable of representing,** resulting in outputs that tend to be poorer in nuance, accuracy, and applicability.

Some interviewees discussed the use of synthetic data — artificially generated to

compensate for a lack of diversity in training datasets — but were categorical in stating that:

> *"The accuracy is not as good as it would be if you were using real data."*

An issue that emerged as particularly relevant during the interviews is that **reducing the availability of data in Portuguese could lead to a significant decline in model quality when it comes to representing local cultural contexts.** The Portuguese language accounts for slightly less than 4% of the open content on the internet, whereas English represents nearly half (Statista, 2025). In other words, there is a real risk that models could become less relevant to local audiences. As one interviewee noted:

> *"[If the model] is trained only on data from other countries because Brazil restricts access, these models probably won't be able to handle typical Brazilian problems or certain things that are specific to Brazil. So, if you ask who won the Brazilian football championship, the model won't know — unless that information happens to be published in some external source that it is allowed to use".*

2.3. Individual Data Licensing Could Render the Development of Brazilian Models Unfeasible

When asked about the impacts of a regulatory framework that would require mandatory licensing of works for the training of Gen AI models, **the interviewees unanimously agreed that the consequences would be severe, particularly for Brazilian companies.** As one interviewee put it:

> *"The ones most affected would be Brazilian companies themselves, because to operate in Brazil and conduct this training locally, we simply wouldn't be able to—it would be unfeasible".*

The concerns raised during the interviews related both to the practical implementation and the cost of such licensing requirements. One interviewee emphasized that the diversity of creators and sources on the internet is so vast that it would be practically impossible to individually license every piece of content:

> *"The problem arises if I create a general law that demands everyone be compensated, you know? I even doubt how I would go about compensating everyone. I'm finding one photo here on the internet, another there, another elsewhere that's publicly available — and how am I supposed to compensate each of these people? I think the issue is exactly that: creating a law that, in practice, is unworkable".*

The financial impact was also highlighted as a major limitation, particularly for the emergence of new Brazilian startups, with several participants noting that this would make development **"unaffordable for startups".** This scenario would severely constrain the dynamism of the local innovation ecosystem and create significant barriers to the country's

competitiveness in the global landscape, especially since, as discussed in the Introduction, other countries are actively seeking ways to ease the use of training data. As one interviewee summarized, **"It's the kind of problem that could leave a country behind".**

2.4. Market Concentration: Exclusive Access to Datasets May Benefit Only Large Companies

In a scenario of strict regulation, c**ompanies with large proprietary datasets or exclusive access to data may consolidate even more dominant positions in the Gen AI market.** According to the interviewees, this dynamic could create distortions on both sides — among content owners and AI model developers.

As noted by Barcott (2025), major Gen AI companies are already pursuing exclusive agreements with organizations that hold large proprietary datasets. A concern raised repeatedly in the interviews is that **the diversity of content on the internet, combined with the technical challenges of attribution, would make it virtually impossible to compensate small creators individually** — even though their contributions might, statistically, be more relevant than those from large proprietary databases.

> *"We're not just talking about niche players like media companies or book publishers. I feel like it's becoming a scenario where almost every website on the internet now has the right to claim copyright compensation—if we assume that the model was trained on their content".*

On the side of Gen AI development, this market concentration is already visible globally, with only a handful of companies dominating model development—companies that are far better positioned to bear the high costs associated with licensing. This comes in addition to the already enormous costs of training itself, as one interviewee emphasized:

> *"You can set up the training, but you also have to provide the funding the money—you need PhDs working on it, and there are a lot of indirect costs involved".*

2.5. Relocation of AI Hubs: Strict Regulations in Brazil Could Be Easily Circumvented—With Significant Economic and Social Impacts

The imposition of overly restrictive rules on data usage for AI training may lead to a phenomenon referred to as **Relocation of AI Hubs—the relocation of innovation hubs and investment to countries with more flexible regulatory environments.**

Interviewees noted that Gen AI companies could, from a technical standpoint, simply relocate their activities to jurisdictions where copyright remuneration obligations do not apply. **Given the global and open nature of the internet, executing such a strategy would be straightforward — and equally simple to circumvent.** As one interviewee explained:

> *"It's like banning it here, but you're not banning it anywhere else (...) so the question is: do you not want the technology here while all your neighboring countries have it?".*

This situation would directly contradict public policies aimed at promoting local data centers and would give a competitive advantage to larger companies with distributed cloud infrastructure, which can choose to train models in jurisdictions where the law is more favorable.

> *"Absolutely. No doubt about it. Think about it: where does training happen? It happens at different scales (...) Companies today are already training on clouds that exist in multiple countries. There are data centers spread across many nations. So if the question is whether the technology for training is already global, the answer is absolutely yes — and it will become even more so".*

Additionally, if the regulation applies only locally, **its effectiveness over a foreign entity that provides a model via the internet would be severely limited.** Unless the government resorts to extreme measures, such as total website or application bans, a foreign company could still offer its services to Brazilian users. This would undermine the credibility of the country's regulatory framework.

> *"What I could also do is this: if I face restrictions in Brazil but not in the United States, I can simply send my model to the U.S., train the parts I cannot process here, then bring it back and continue training in Brazil".*

3. Analysis and Commentary

This section analyzes the research findings, connecting them with academic literature and expert opinions, through the lens of the authors of this study.

3.1. The Lack of Technical Expertise in Public Policy Formulation on Gen AI

Following the completion of the interviews and a detailed review of the copyright provisions included in Draft Bill No. 2,338/23, it became evident to us that there is a significant gap between the bill's proposals and their technical feasibility. **What could have led to this discrepancy?**

This is a challenging question to answer empirically. However, our exploratory hypothesis is that **the legislative debate on AI and copyright in Brazil was conducted without a deep understanding of the technology involved.** Several factors support this argument.

Firstly, the issue of copyright did not feature prominently among the key topics discussed during the Senate Commission's proceedings. An analysis of the transcripts from the 24 sessions of the Federal Senate's Temporary Internal Commission on Artificial Intelligence (CTIA) reveals that the discussions primarily focused on topics such as personal data protection, risk classification, system definitions, and impacts on innovation. While copyright is certainly a relevant dimension in AI regulation, its discussion was significantly less prominent compared to other issues.

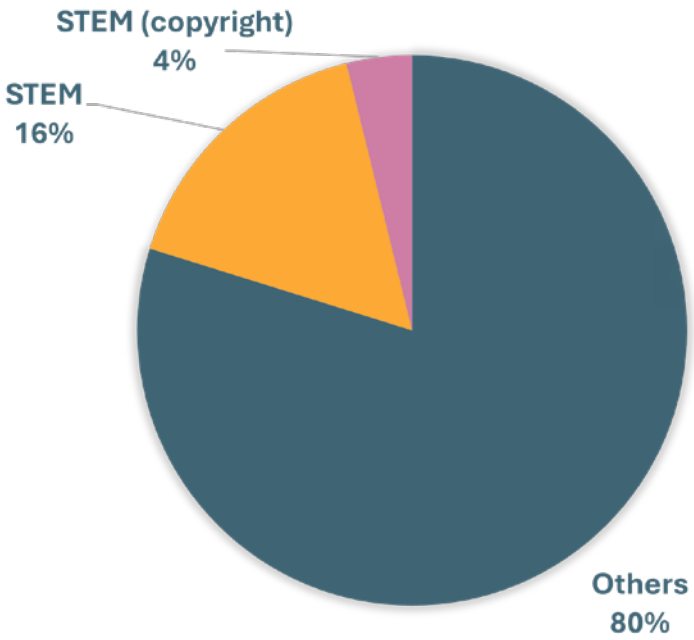
Figure 6. Word Cloud from the Sessions of the Temporary Internal Commission on Artificial Intelligence in Brazil (CTIA) of the Federal Senate, generated using Atlas.ti software (Concepts tool).

Source: Word cloud generated using Wordclouds.com, based on data automatically extracted from Atlas.ti using the "Concepts" tool[6].

6 Based on the automated mapping, a thematic clustering of concepts was performed, grouping similar expressions (e.g., legislation and regulation; privacy and personal data). The size of each word reflects the weighted frequency of the concepts after clustering.

Additionally, we observed a **low presence of professionals from STEM fields in the debate** — even lower when considering how many of them addressed copyright issues from a technical perspective in their statements:

Figure 7. Chart of Participant Profiles and Contributions in the CTIA



Source: Reglab's elaboration.

In other words, it appears that the absence of technical experts within the CTIA may have contributed to the proposal of **measures that do not align with the practical realities of how AI models function in relation to copyright.** Moreover, this disconnect between regulation and technology does not seem to be an issue unique to Brazil, but rather a global challenge. However, to avoid the creation of laws that are either unenforceable or detrimental to the country's competitiveness, it is essential to institutionalize mechanisms for qualified technical consultation, grounded in concrete scientific and economic evidence, thereby ensuring that regulatory frameworks are both feasible and effective.

3.2. Economic Distortions: Criteria Beyond "Usage" Favor Companies With Large Proprietary Databases, and There Is No Evidence of How This Compensation Would Reach Creators

The technical challenges in attributing the use of works within Gen AI systems undermine the economic rationale of copyright, which is fundamentally based on quantifying how protected content is reproduced, distributed, or transformed in order to allocate remuneration (Watt, 2009). However, this rationale — rewarding creators proportionally to the use of their works — collapses when we consider that Gen AI systems are incapable of reliably tracking or measuring the use of such works.

This breakdown has the potential to distort incentives: licensing frameworks that are not based on actual usage could exacerbate market concentration. Large rights holders with significant legal resources may be able to negotiate bulk licensing agreements, while independent creators — lacking the barGen AIning power to prove the use of their works by Gen AI systems — may be disadvantaged. This dynamic risks marginalizing smaller voices and reducing creative diversity (Martens, 2024).

> **In the current state of machine learning technology, copyright runs the risk of becoming a mechanism that both excessively hinders AI innovation and inadequately protects human creators.**

Addressing these challenges in future research may require a **reconceptualization of both the concept of copyright and the impact of Gen AI on creative industries** through a broader historical perspective. Previous technological disruptions — such as the shift from physical to digital distribution — initially generated controversy but ultimately led to industry transformation rather than decline, fostering new business models and revenue streams that leveraged emerging technologies to create additional value (Masnick & Beadon, 2024).

4. Conclusion

The advancement of Gen AI raises legitimate questions about how to ensure an ecosystem that balances innovation with social welfare, with regulation emerging as a key tool to promote this balance. However, the findings of this study suggest that, for such regulation to be effective, its **guidelines must be technically feasible.**

This research demonstrates that, although it is possible to identify the datasets used to train models, **there are still no scalable and reliable solutions to measure the specific contribution of individual works** in large-scale models. At present, this appears to be a **structural limitation of the technology,** particularly in machine learning.

Consequently, regulatory proposals that fail to account for this reality may result in arbitrary assessments. Additionally, the interviews highlighted that **excessive restrictions on data usage** could create significant barriers to entry for startups, independent researchers, and public institutions, ultimately favoring market concentration. Another important consideration is the risk that companies may relocate their training processes to jurisdictions with more permissive regulations, which would undermine both the **credibility of local regulation and the country's global competitiveness.**

It is important to emphasize that the findings of this study should not be interpreted as an argument aGen AInst **regulation** or aGen AInst the protection of creators' rights. On the contrary, the conclusions point to the need for **evidence-based regulation that considers the realities of the sector and the technical limitations of current technology.**

As a final message, this study underscores the urgent need to broaden the participation of technical experts in the AI regulatory process. The current legislative discussions require an effective dialogue with the technical community — not to prioritize their perspective over others, but to ensure that public policies accurately reflect the complexity of the systems they aim to regulate.

4.1. Direction for future studies

This study analyzed the feasibility of copyright remuneration systems in AI, but several questions remain open. Below, we outline research avenues that could further advance the debate and inform public policy.

- **Economic Impact of Gen AI:** There is limited evidence on whether Gen AI generates losses or creates new opportunities for creators. Future research could analyze how different sectors are impacted, assess changes in income distribution, and explore alternative monetization models.

- **Creators' Perception of the Use of Their Data in AI Training:** Regulation often overlooks the perceptions of creators. Qualitative research from a media reception perspective could investigate how creators evaluate the use of their data, their levels of acceptance or rejection, and their views on the regulatory debate.

- **Dynamics of Interests in the Regulatory Debate:** Studies could map the key stakeholders involved in the legislative process, how they influence policymaking, what their agendas are, and whether there is balanced representation across different sectors.

- **Remuneration Models: Is There a Viable Path?:** Given the lack of traceability, research could assess the impacts of estimated compensation models, opt-out mechanisms, or copyright exceptions using econometric methods or cost-benefit analyses with a focus on social welfare.

- **The Effect of Data Restrictions on Innovation and Competitiveness:** Studies could measure how restrictions affect the quality of AI models, whether they favor large players, and how they incentivize the relocation of companies to jurisdictions with more flexible regulations.

References

AMORIM, P. **Analytical AI: A Better Way to Identify the Right AI Projects.** Available at: https://sloanreview. mit. edu/article/analytical-ai-a-better-way-to-identify-theright-ai-projects/. Accessed on: 10 may. 2025.

AUDENHOVE, L. V.; DONDERS, K. **Talking to People III: Expert Interviews and Elite Interviews.** In: VAN DEN BULCK, H.; PUPPIS, M.; DONDERS, K.; VAN AUDENHOVE, L. (Eds.). The Palgrave Handbook of Methods for Media Policy Research. Palgrave Macmillan, 2019.

BARCOTT, B. **How the Emerging Market for AI Training Data is Eroding Big Tech's "Fair Use" Copyright Defense,** 2025. Available at: <https://www.techpolicy. press/how-the-emerging-market-for-ai-training-data-iseroding-big-techs-fair-use-copyright-defense>. Accessed on: 12 may. 2025.

CHUI, M. et al. **Economic potential of generative AI.** McKinsey, 2023. Available at: https://www.mckinsey. com/ capabilities/mckinsey-digital/our-insights/ the-economic-potential-of-generative-ai-the-nextproductivity-frontier. Accessed on: 11 may. 2025.

COENEN, F. **Data Mining: Past, Present and Future.** The Knowledge Engineering Review, v. 00, p. 0–1, 2004.

DE LA ROSA, J., et al. **The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective.** arXiv preprint, 2024. Available at: <https://arxiv.org/ html/2412.09460v1>. Accessed on: 11 may. 2025.

DAASE, C. et al. **On the Current State of Generative Artificial Intelligence: A Conceptual Model of Potentials and Challenges.** 26th International Conference on Enterprise Information Systems, 2024.

FIIL-FLYNN, Sean M. et al. **Legal reform to enhance global text and data mining research.** Science, v. 378, p. 951-953, 2022.

GAO, L. et al. **The Pile: An 800GB Dataset of Diverse Text for Language Modeling.** arXiv preprint, 2020, disponível em: < https://arxiv.org/abs/2101.00027>. Accessed on: 10 may. 2025.

GUEST, G.; BUNCE, A.; JOHNSON, L. **How Many Interviews Are Enough? An Experiment with Data Saturation and Variability.** Field Methods, 18(1), 59-82, 2006.

HERZOG, C; HANDKE, C.; HITTERS, E. **Analyzing Talk and Text II: Thematic Analysis.** In: VAN DEN BULCK, H.; PUPPIS, M.; DONDERS, K.; VAN AUDENHOVE, L. (Eds.). The Palgrave Handbook of Methods for Media Policy Research. Palgrave Macmillan, 2019.

INGOLD, Jo; MONAGHAN, Mark. **Evidence translation: an exploration of policy makers' use of evidence.** Policy & Politics, v. 44, n. 2, p. 171-190, 2016.

KARAGANIS, J. **Emerging Copyright Governance Frameworks Across the US, China, and Europe.** AI, Media & Democracy, 2024. Available at: < https:// www. aim4dem.nl/is-ai-training-infringement/>. Acessed on: 12 may. 2025.

LEFFER, L. **When It Comes to AI Models, Bigger Isn't Always Better,** 2025. Available at: <https://www. scientificamerican.com/article/when-it-comes-to-aimodels-bigger-isnt-always-better/>. Accessed on: 12 may. 2025.

MARTENS, Bertin. **Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs.** Working Paper, Bruegel, 2024.

MASNICK, M.; BEADON, L. **The Sky Is Rising: A detailed look at the state of the entertainment industries, 2024 Edition.** Copia Institute & CCIA Research Center, 2024.

NOBLE, T. **AI and Copyright: Expanding Copyright Hurts Everyone — Here's What to Do Instead.** Electronic Frontier Foundation, 2025.

RAMOS, P. H. (coord). **Digital Governance in Focus: Strategies for the Use of Generative AI in Companies.** Gtech – Grupo de Estudos em Direito e Tecnologia. Relatório de Pesquisa. São Paulo: Ibmec SP, 2023.

ROSATI, E. **Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law.** European Journal of Risk Regulation, p. 1–25, 31 oct. 2024.

SALDAÑA, Johnny. **The Coding Manual for Qualitative Researchers.** 4. ed. Thousand Oaks: SAGE Publications, 2021.

STATISTA SEARCH DEPARTMENT. **Languages most frequently used for web content**, 2025. Available at: < https://www.statista.com/statistics/262946/mostcommon-languages-on-the-internet/>. Accessed on: 12 may. 2025.

SCHIRRU, L. et al. Text and Data Mining Exceptions in Latin America. **IIC - International Review of Intellectual Property and Competition Law**, 19 sep. 2024.

SOBEL, B. L. W. Artificial intelligence's fair use crisis. **Columbia Journal of Law & the Arts,** v. 41, p. 45-96, 2017.

UENO, H. J**apan's New Approach to Collaborative International R&D.** Issues in Science and Technology. Vol. XLI, Winter, 2025.

WANG, J. T. et al. **An Economic Solution to Copyright Challenges of Generative AI.** arXiv preprint, 2024. Available at: <https://arxiv.org/abs/2404.13964>. Accessed on: 12 may. 2025.

Zhang, L. et al. **Fairshare Data Pricing for Large Language Models.** arXiv preprint, 2025. Available at: <https://arxiv.org/ html/2502.00198v1>. Accessed on: 12 may, 2025.

Methodology Annex

FORMAT: POLICY BRIEF

Title	Remuneration for Copyright in AI: Limits and Implementation Challenges
Research Question	**In what ways does the training of generative AI models involve the use of copyright-protected content, and what are the technical challenges to the feasibility of remuneration mechanisms associated with such use?**
Methodology Summary	This research adopts a **qualitative** approach, combining primary data collection through **in-depth expert interviews** and secondary data analysis, including documents, academic literature, and practical cases. The methodological choice is grounded in the exploratory nature of the subject: as an emerging topic with few consolidated experiences of remuneration for AI training data, it is particularly valuable to capture the **perceptions, insights, and expertise** of key stakeholders and specialists.
Data Collection	Data collection followed the **expert interviews** methodology (Audenhove and Donders, 2019), using qualitative semi-structured interviews with an exploratory character. The choice of this method is justified by the technical nature of the subject and the lack of systematized data on the investigated problem, making the accumulated knowledge of experts actively working in the field essential. The sample was defined based on diversity and representativeness criteria, including: minimum participation of women; presence of representatives from academia or research centers; professionals from Brazilian companies; and experts from large technology companies. The selection process combined convenience sampling with a snowballing technique. A total of 16 individuals were contacted, of whom eight agreed to participate in the study; the others declined due to unavailability. The interviews were conducted between March 12 and March 31, 2025, in an online format (via Teams), with an average duration of 45 to 60 minutes. Each session included the presence of at least two Reglab researchers. The interview guide used is attached to this report. Two of the interviews were conducted in a preliminary manner to test the structure of the questionnaire and validate initial hypotheses. These preliminary interviews were not included in the coding process but contributed substantially to the final design of the data collection. The six interviews analyzed were considered sufficient to achieve theoretical saturation, since in qualitative approaches with semi-structured and in-depth interviews, thematic recurrence and analytical density tend to consolidate with a relatively small number of participants (Guest et al., 2006). All interviews were recorded with the participants' consent, fully transcribed, and accompanied by interviewer memos. The material was stored and coded using Atlas.ti software. The names and affiliations of the interviewees were anonymized.
Data Analysis	The data were analyzed using thematic analysis, as outlined by Herzog et al. (2019), employing two cycles of inductive coding. The first cycle consisted of open conceptual coding, while the second applied pattern coding to group and refine analytical categories (Saldaña, 2021). The process was conducted using Atlas.ti software. The choice of thematic analysis is justified by its suitability for exploratory studies aimed at structuring and interpreting technical information, allowing the identification of conceptual patterns in highly complex contexts. The research team maintained a reflexive stance throughout the analytical process, recording interpretative memos and systematically discussing potential analytical biases. Themes were defined based on recurrence, conceptual density, and relevance to the research objectives. The final categories included, among others: "technical impossibility of attribution," "remuneration models," "market concentration," "traceability limitations," and "regulatory impact." To support critical analysis and the triangulation of evidence, the visualization, mapping, and correlation tools provided by Atlas.ti were employed. The analysis was conducted between April 2 and April 15, 2025.

Bias Reduction Procedures

Established Theoretical and Methodological References: The data collection and analysis techniques adopted in this study followed practices widely recognized in the academic literature. The methodological approach was discussed internally both before and after the preliminary interviews, allowing the incorporation of feedback and suggestions into the final research design prior to the start of the analytical process.

Open Categorization: Data coding followed an inductive logic with no predefined categories, enabling codes and themes to emerge directly from the empirical material. This methodological choice aimed to minimize interpretive biases arising from the imposition of prior conceptual frameworks.

Methodological Triangulation: Empirical findings were cross-referenced with documentary analysis of secondary sources, with the aim of comparing, validating, and reinforcing the consistency of the interpretations developed from the interviews. These references were explicitly cited throughout the text.

Double Validation at Critical Stages: Coding was conducted and cross-reviewed by two researchers. The final definition of themes was the result of collective discussion among all three authors, ensuring the inclusion of multiple perspectives and mitigating individual biases in data interpretation.

Documentation and Methodological Transparency: All stages of the analytical process were thoroughly documented, including successive versions of coding files and decision logs. This practice ensures full traceability of the methodological pathway, in line with Reglab's guidelines for transparency and replicability.

Other Methodological Limitations

Qualitative Scope and Limits of Generalization: The limited number of interviews prioritized analytical depth but does not allow for statistical generalization.

Convenience and Network-Based Sampling: The sample selection may reflect availability bias and professional network limitations, despite the application of diversity criteria.

Technological and Regulatory Evolution: The findings reflect the state of the art at the time of the research and may be impacted by future developments in the technological and regulatory landscape.

Dependence on External Tools: Although Atlas.ti is one of the most widely recognized analytical tools in the academic sector, the analysis relied significantly on the functionalities and capabilities of this software.

Software Use

SOFTWARE	USE FOR RESEARCH PURPOSES
Suíte MS Office	Text editing, spreadsheet and chart development, and interviews (via Teams)
Suíte Adobe C	Layout design and finalization of charts and illustrations
Atlas.ti	Organization, coding, and analysis of qualitative data
Cockatoo	Transcription of interview audio into text
ChatGPT 4o	brainstorming, information systematization, grammatical review (spelling, grammar, synonym refinement), language adjustment, and alignment with the Reglab Style Guide
Notion AI	Text editing and proofreading (spelling and grammar, synonym refinement, language adjustment, translations); research organization and timeline development.
Wordclouds	Creation of word clouds
Lex.page	Advanced text review (conciseness, avoidance of clichés, readability, reduction of passive voice, elimination of unsupported claims, and removal of redundancies)

Ethical Guidelines	**Research Funding:** This publication is part of a series of publications sponsored by Google, Meta, and B/Luz, in which Reglab maintains full editorial control. Unlike commissioned research, Reglab independently defined the scope, objectives, and methodology of this study with complete autonomy. The authors retain full professional independence and are solely responsible for the content and conclusions of this work. **Processing of Personal Data:** The research involved the processing of personal data exclusively during the data collection and analysis phases, in a manner that was limited and proportionate to the study's objectives, in compliance with Law No. 13.709/2018 (Brazilian General Data Protection Law – LGPD). **Legal Basis:** All participants formally authorized their participation by signing an informed consent form, acknowledging the research objectives and data usage. **Purpose and Adequacy:** Data were used exclusively for the purposes of this study, in accordance with the consent provided, and were not employed for any other purposes. **Data Minimization and Anonymization:** Personally identifiable information that was not relevant to the research objectives was anonymized in the transcripts and removed from the active database. **Confidentiality and Privacy:** In the presentation of results, all data were kept confidential, and quotations were adjusted when necessary to ensure the anonymity of sources. Only a limited number of researchers directly involved in the project had access to personal data and original documents. **Information Security and Record Keeping:** All files were stored with password-protected access in accordance with Reglab's internal information security policies. **Retention and Disposal:** Data will be stored for up to 12 months solely for methodological audit purposes and potential replication, after which they will be permanently deleted. Responsible Use of Public Data: Although some of the data analyzed are publicly available, their use was carried out in a responsible and ethical manner, strictly for the purposes of independent research. **Methodological Transparency:** The research methodology has been fully detailed to ensure transparency and replicability, contributing to scientific integrity and enabling independent validation of the results. **Non-Discrimination and Respect for Diversity:** The research was conducted in a manner that respects diversity and avoids any form of discrimination.

ANNEX II - SEMI-STRUCTURED INTERVIEW GUIDE

#	QUESTION
0	In the research output, how would you prefer to be identified? What title should we use to refer to you?
1	To begin, could you share a bit about your professional background and your work in the field of Artificial Intelligence and Machine Learning?
2	How does the selection and use of data for training AI models take place?
3	Which types of data are most critical for model performance? Are there specific characteristics that make a dataset particularly valuable?
4	If there were limitations on data availability due to licensing rules, what would be the technical and practical impacts on model development and performance?
5	Considering the possibility of data being restricted or unavailable due to copyright constraints, what would be the implications for AI models?
6	Based on your experience, is it possible to trace and document which specific data were used in the training of an AI model?
7	How is this process carried out in practice?
8	Given the way AI models are trained, how could each copyrighted work used in the process be identified and remunerated? What opportunities and challenges are involved in this task?
9	In cases where copyrighted content is present, are there techniques to ensure that AI models do not memorize or reproduce it? Could you explain this issue in more detail?
10	If a mandatory remuneration system for copyrighted works were implemented, what would be the impacts on companies and startups developing AI?
11	If restrictions on data use for AI were enacted in Brazil, could model training simply be relocated to other countries? Does this already occur in other contexts?
12	In your view, what would an ideal system to balance data use for AI and copyright protection look like? Are there viable technical solutions to this problem?
13	Is there anything we have not asked but that you consider important regarding the use of data in AI training?