

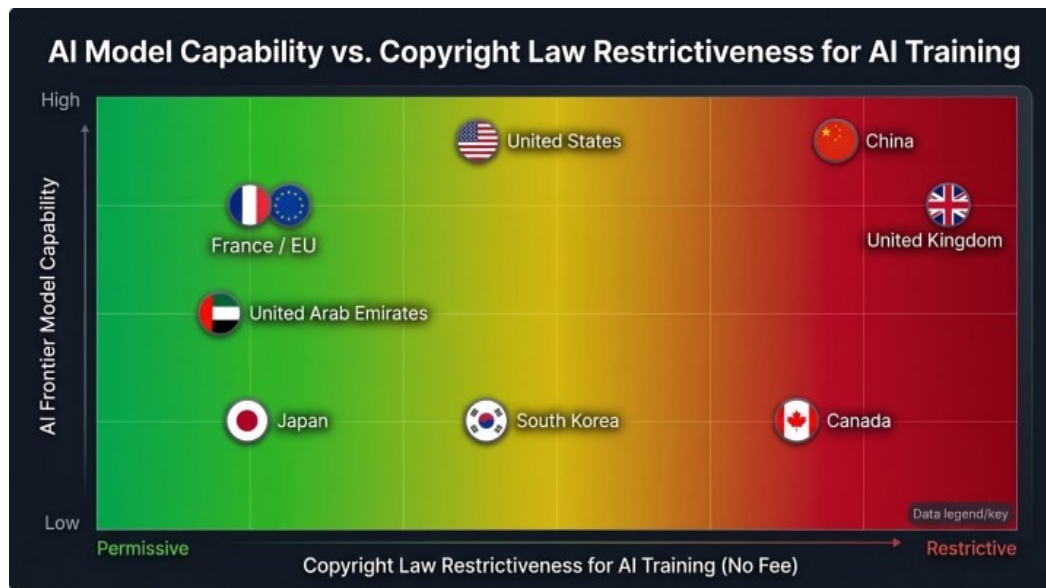
# Training in the Safe Zone: Building AI Models in Copyright Havens

October 5, 2026 – The TMCA

by Lena Petrovic

The global artificial intelligence race is no longer defined solely by compute power and algorithmic breakthroughs. It is increasingly shaped by national copyright policies governing the ingestion of training data. As model builders race toward artificial general intelligence, they can run into a messy geopolitical question: how free are you to scrape the open web without paying the creators of the content behind it? Across the globe, the legal permissiveness surrounding copyright and AI training varies dramatically, leaving foundation model companies walking a legal tightrope.

The below visualization shows an overview of permissiveness of various laws among the countries with the leading AI models:



A natural strategic question arises: why not set up compute clusters in a legally lenient haven to scrape and train, and then deploy the finished model into major commercial markets? To evaluate, one has to examine the patchwork of laws governing the world's

## Related People

- Lena Petrovic

## Related Capabilities

- Trademark, Copyright + Advertising
- Artificial Intelligence

leading AI jurisdictions and the availability of training data.

## **The Superpowers: High Capability, Fragile Legal Ground**

Leading the pack in raw capability is the United States, yet, despite its unmatched private capital and hyperscaler infrastructure, operating in the U.S. rests on precarious legal footing. Lacking a clear statutory safe harbor written specifically for AI training, American labs must rely entirely on the judge-interpreted copyright fair use doctrine. The fair use analysis consists of consideration of four factors set forth in the U.S. Copyright Act. As major publishers, news organizations, and artists drag developers into federal court over market substitution and synthetic outputs, uncompensated training remains a high-stakes litigation gamble until appellate judges establish binding precedent.

Close behind in model performance is China, whose near-parity challengers—including DeepSeek, Alibaba Cloud, and Baidu—excel in compute efficiency and post-training architectures. However, China’s regulatory landscape is far more rigid on paper. Under state generative AI measures and copyright regulations, training data must come from legitimate sources, and developers face strict liability if commercial outputs infringe on protected works. Because China maintains an exhaustive, narrow list of fair use exemptions rather than an open-ended standard, authorities are steering the domestic generative AI sector toward state-backed data registries and compulsory collective licensing.

## **The European Divide: French Opt-Outs vs. British Fortress**

Across the Atlantic, European nations present contrasting philosophies. France stands as continental Europe’s primary AI powerhouse through ventures like Mistral AI and Kyutai, benefiting from a balanced European Union framework. Under EU digital copyright directives—further reinforced by the EU AI Act—commercial AI teams can scrape and train on lawfully accessible web data free of charge, as long as rightsholders have not embedded a machine-readable opt-out tag. Across the English Channel, the United Kingdom hosts foundational research giants like Google DeepMind and Stability AI, yet enforces one of the most restrictive commercial environments. Following fierce pushback from its domestic creative sector, the UK government refused to create broad commercial exceptions.<sup>[1]</sup> Uncompensated data mining is limited strictly to non-commercial scientific research, leaving unlicensed commercial model training potentially exposed to infringement claims.

## **The Outliers: Sovereign Havens and Cautious Challengers**

Outside the Western and Chinese ecosystems, strategic legal approaches diverge even further. The United Arab Emirates has surged forward as an open-source pioneer with its Falcon model family, supported by sovereign wealth and a permissive domestic posture where enforcement against state-backed AI data ingestion is practically non-existent. Meanwhile, Japan has deliberately established one of the world’s most developer-friendly environments: its copyright laws broadly exempt machine learning and statistical processing from licensing fees under a sweeping “non-enjoyment” rule, even for commercial endeavors.<sup>[2]</sup> Rounding out the landscape are South Korea and Canada.

While South Korea combines a flexible fair use standard with active legislative plans to introduce explicit statutory data-mining carve-outs, Canada's strict, closed-category fair dealing rules leave commercial web scraping vulnerable to significant legal risk.

## The 200-Million-Book Wall

So, can developers simply avoid restrictive jurisdictions by training exclusively on data from permissive havens like Japan, France, and the UAE and then deploy the model in the major markets like the U.S.? When it comes to frontier general-purpose models, the math makes that nearly impossible. Training a cutting-edge frontier model requires massive corpora equivalent to roughly 200 million standard 300-page books.<sup>[3]</sup> Across the major permissive jurisdictions, the entire readily digitized, machine-readable literary supply amounts to only about 1 million books.<sup>[4]</sup> Attempting to train exclusively on data from permissive jurisdictions would starve a model of the scale, coding depth, and global English fluency required to compete at the frontier.

## The True Safe Harbor: Specialized & Sovereign Models

Where this jurisdictional arbitrage **does** work is in the targeted world of specialized, domain-specific AI with approximately 3 billion parameters. Training such a model requires approximately 60 billion tokens<sup>[5]</sup> which at quarter of a word per token<sup>[6]</sup> and 250 words per page, translates into approximately 600,000 standard book equivalents—within the digital inventories of permissive countries. For specialized enterprise tools, regional LLMs, and sovereign architectures, leaning into permissive legal frameworks offers a viable, defensible path forward. For frontier builders chasing raw general capability, however, navigating the legal minefield of global copyright and the ever-evolving laws and judicial interpretations of laws and regulations remains an unavoidable cost of doing business.

[1] [The Times – No 10 Is Giving a Green Light to Music Laundering and Parliament UK – Abandon Artificial Intelligence Copyright Exemption to Protect UK Creative Industries, MPs Say.](#)

[2] See [Article 30-4 at Japanese Law Translation – Copyright Act \(Act No. 48 of 1970\).](#)

[3] “GPT-4 was trained on at least 10 trillion tokens, or around a hundred million books’ worth of data.” See [The Information Difference – Under the Hood: Inside a Large Language Model.](#)

[4] In the EU there are “24 million pages of full-text content and more than 7 million digital objects.” See [Wikipedia – European Library.](#) If you divide 24 million pages by a standard 300-page benchmark, it roughly equates to 80,000 normalized 300-page machine-readable books. In Japan, 620,000 publications are available to everybody on the Internet, albeit of unspecified length. See [The Japan News – National Diet Library Rapidly Digitizes Publications; Plans to Digitize 450,000 This Year.](#) In the United Arab Emirates “a vast database of more than 20 million words,” is available. See [Gulf Business – 20 million words and counting: UAE’s grand plan to power Arabic with AI.](#) At

250 words per page this translates to roughly 270 books. Even if we assume that in Japan all the publications are 300 page books, the total number of 300 page books across these three jurisdictions amounts to less than one million.

[5] “Rule of thumb is that you need ~20 tokens per parameter.” See [Hacker News](#).

[6] [OpenAI – Understanding and Counting Tokens](#).