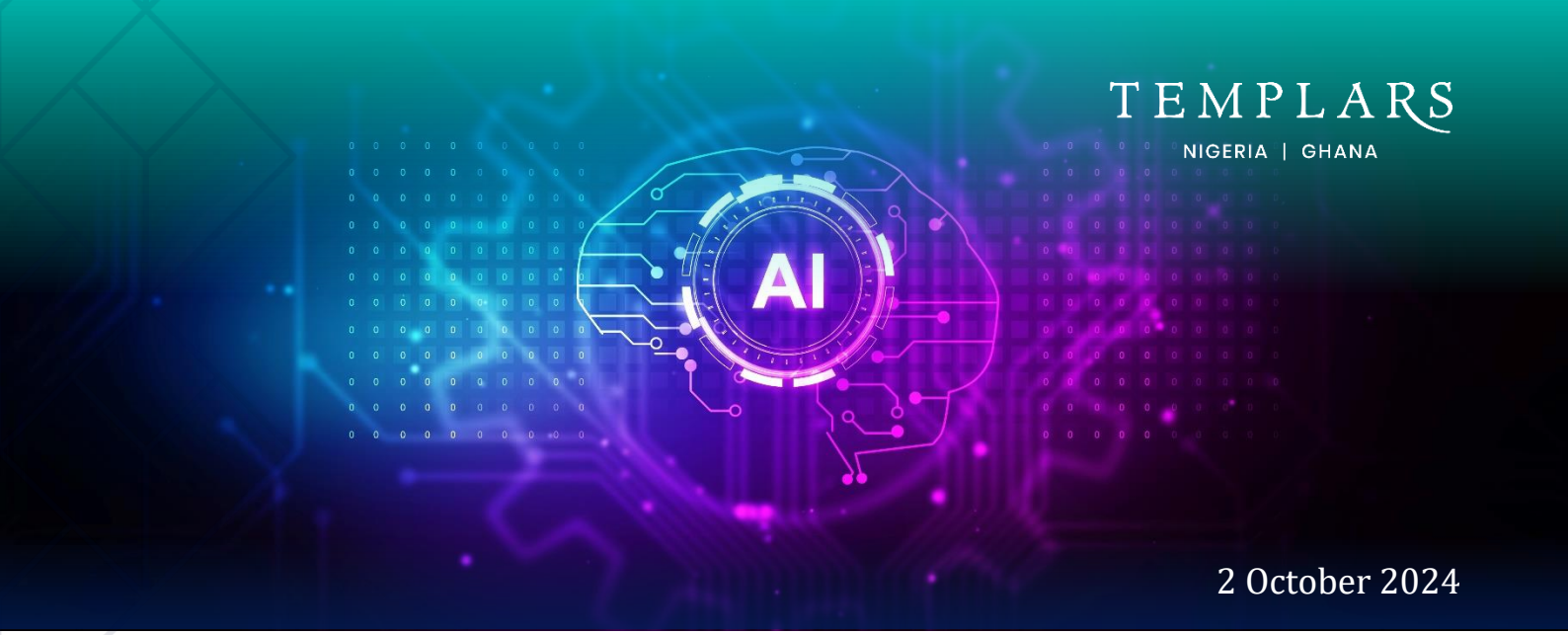

TEMPLARS
NIGERIA | GHANA

2 October 2024

Key contacts

Augustine B. Kidisil
Managing Partner and Head,
Dispute Resolution (Ghana)
augustine.kidisil@templars-law.com

Daniel Akuoko Darkwah Jnr.
Associate,
Dispute Resolution (Ghana)
daniel.akuoko@templars-law.com

TEMPLARS ThoughtLab

Navigating Copyright Infringement Risks in Training Large Language Models: A Ghanaian Perspective

Introduction

The release of ChatGPT in 2020 marked the beginning of widespread adoption of Artificial Intelligence (AI) systems. Since then, other AI systems like Gemini, Claude, and Copilot have emerged, all belonging to a category of AI models known as Large Language Models (LLMs). At the same time, the number of copyright infringement claims against AI companies has increased. Training these LLMs requires vast amounts of text data, which often includes copyrighted material. This reliance on extensive text data exposes AI companies to potential copyright infringement claims, particularly in jurisdictions such as Ghana, where copyright protection is clearly defined.

This article explores the key considerations for AI companies in Ghana when collecting and using text data for training LLMs. It highlights the importance of securing proper licenses, understanding permitted uses, and ensuring compliance with public domain requirements. We suggest that by incorporating a thorough legal risk evaluation into the dataset collection process, companies can better manage potential legal issues and ensure their AI training practices align with copyright laws in Ghana.

Large Language Models (LLM)

Large Language Models (LLMs) are advanced Artificial Intelligence (AI) systems designed to understand and generate human language. These models are developed using a technique known as deep learning, which enables them to identify and learn complex patterns in unstructured data such as texts, images, and videos. For LLMs specifically, the focus is on text data. By analyzing vast amounts of text data, the model learns to recognize patterns and relationships between words, allowing it to generate coherent text.

The training of an LLM typically involves three stages. The first is the pre-training stage, where the model is fed a vast and diverse collection of texts, often comprising hundreds of billions of words. These texts comes from a wide range of sources – books, websites, and other publicly available content – covering topics from science and history to everyday conversations.[1] The goal is for the model to learn patterns in language, such as grammar, context, and relationships between words.

One technique to achieve this is having the model predict the next word in a sequence. For example, if the sentence is "The capital of Ghana is...", the model learns to predict that "Accra" should come next based on patterns it has observed during training. This process is repeated millions of times, allowing the model to develop a general understanding of how language works. In the development of most LLMs, nearly all models undergo this first stage of training.

At this early stage, the model is not yet very useful and may produce nonsensical or incoherent texts, as it has only learned basic language structures without any understanding of context or specific tasks. To make the model useful, it undergoes further training in a process known as instruction fine-tuning. At this stage, the model is trained using high-quality text data consisting of instruction-response pairs, which are obtained from carefully curated sources or texts generated by human annotators. The goal is to teach the model to follow instructions and respond to user queries in a way that aligns with their intent. It helps the model move from simple word prediction to performing more complex tasks, such as summarizing articles, answering questions, or offering advice. Essentially, this is when the model learns to act as a helpful assistant.

The final stage of training is known as Reinforcement Learning from Human Feedback (RLHF). At this stage, human trainers provide feedback on the model's responses to guide its behavior. This feedback helps the model learn what makes a good response versus a poor one.

Reinforcement learning works by using a reward system. When the model generates a response, human evaluators rate or critique it, providing positive feedback (rewards) for useful answers and negative feedback (penalties) for poor ones. The model adjusts based on this feedback, improving its ability to generate more accurate, helpful, and aligned responses over time. For example, if the model initially responds to a question with a lengthy, off-topic answer, human feedback might indicate that a shorter, more focused response is preferred. The model then learns to prioritize concise and relevant answers in similar situations in the future.

This iterative process of refining its performance through human guidance allows the model to handle more complex tasks and better meet user expectations.

Copyright protection in Ghana

Without rich, varied, and well-curated text data, an LLM would struggle to understand language or provide helpful assistance. Text data acts as both the foundation and the continual source of learning throughout the model's development, ensuring that it evolves to meet the complex demands of users.

Yet, AI companies cannot on their own generate the vast and diverse texts needed to train LLMs. So, they rely on texts created by others. In Ghana, such text data may be

[1] GPT-3, developed by OpenAI, was trained on several terabytes of text data from externally generated sources like Common Crawl, WebText2, Books1, Books2, and Wikipedia, with approximately 570GB coming from Common Crawl alone. Similarly, Llama 3, developed by Meta, used text data from Stack Exchanges, wikiHow articles, Pushshift Reddit, and manually authored papers.

protected under the Copyright Act, 2005 (Act 690), which grants copyright protection to authors of literary works.[2]

The definition of literary work in Act 690 covers a wide range of materials, including novels, textbooks, stories, biographies, computer programs, and other content that, through the selection or arrangement of its contents, qualifies as an intellectual creation.[3]

A literary work is eligible for copyright protection if it is original and fixed in a definite medium. For works created by individuals, copyright protects the economic rights of the author for their lifetime plus seventy years after their death. For works created by corporate entities, copyright protects the economic rights of the entity for seventy years from the date the work is made or published, whichever is later.[4] The moral rights of the author of a work exists in perpetuity.[5]

While copyright subsists, the author enjoys exclusive moral and economic rights.[6] Moral rights allow the author to claim authorship and ensure their name or pseudonym is credited when the work is used. Authors can also object to and seek remedies for any alterations to their work that could damage their reputation or discredit it.[7] Economic rights give the copyright owner control over the sale, reproduction, translation, adaptation, and distribution of the work, or its copies.[8]

Copyright protection in Ghana does not depend on registration, although registration is desirable.[9] Copyright protection extends to works created by Ghanaian citizens or residents, works first published in Ghana or published in Ghana within 30 days of their initial publication outside the country, and works for which Ghana has an obligation to provide protection under international treaties.[10] Ghana is a signatory to key international copyright treaties, including the Berne Convention for the Protection of Literary and Artistic Works (Berne Convention), the TRIPS Agreement and the WIPO Copyright Treaty. Under the Berne Convention, copyright owners in one member country receive the same level of protection and legal remedies against infringement in other member countries as those enjoyed by the nationals of those countries.[11] Thus, Ghana law protects copyright literary works from infringement.

Copyright infringement under Ghana law

Infringement of copyright for literary works occurs when someone performs an act that violates the economic or moral rights of the author.[12] However, the law allows certain exceptions where such acts do not constitute infringement. These exceptions are specific and limited.[13]

Additionally, works that have entered the public domain can be used freely, although a prescribed fee may be required.[14] Public domain works include those whose copyright

[2] Section 1(1) of Act 690
[3] Section 76 of Act 690
[4] Section 12 and 13 of Act 690
[5] Section 18 of Act 690
[6] Moral rights always belong to the author, but economic rights can be held by others. For example, under section 7 of Act 690, copyright for work created by an employee belongs to the employer unless agreed otherwise.
[7] Section 6 of Act 690
[8] Section 5 of Act 690
[9] Section 39(4) of Act 690
[10] Section 1(2)(c) of Act 690
[11] As of the date of this article, there are 181 member countries of the Berne Convention.
[12] Section 41 of Act 690
[13] Section 19 of Act 690
[14] Section 38 of Act 690

protection has expired, works where the authors have given up their rights, and foreign works not protected by copyright in Ghana.[15]

Copyright infringement has severe consequences. A person convicted of infringement may face substantial fines or imprisonment.[16] Additionally, the copyright owner can seek civil remedies, such as damages or an injunction.[17] The law permits both criminal and civil actions for copyright infringement to be pursued at the same time.[18]

The purpose of awarding damages for copyright infringement is to fairly compensate the author for their loss. What constitutes fair compensation is determined case by case, and the Ghanaian courts may consider factors such as the license fee the copyright owner would have charged.[19]

For AI companies developing LLMs, facing multiple copyright infringement claims can be financially devastating. Therefore, it is crucial to be aware of potential pitfalls when collecting text data for training LLMs to avoid such risks.

Key considerations for collecting LLM training datasets

From a copyright perspective, LLM training data can be obtained from three main sources: copyrighted literary works for which licenses are required, copyrighted works for which licenses are not required, and literary works that are in the public domain.

Under the law, only copyright owners and those they authorize have the right to reproduce their works in any form.[20] Therefore, it is crucial to obtain a license or authorization from the copyright owner before including their work in the training data. Given that Ghana is a signatory to the Berne Convention, a significant number of literary works are likely to be eligible for copyright protection in Ghana. Therefore, as a preliminary step in sourcing data, it is important to determine whether a work is protected under Ghanaian law.

While obtaining proper licenses is the ideal approach, the vast amount of text data required to train an LLM often makes this nearly impossible. As a result, companies may rely on permitted uses of copyrighted works under the law and works that have entered the public domain. In both cases, authorization from the copyright owner is not required to use these works.

In the context of training LLMs, the only relevant permitted use of copyrighted material is the reproduction of a work that has been made public for the exclusive personal use of an individual.[21] However, this use is highly restricted: the individual cannot reproduce the entire work or a substantial part of it.[22] What constitutes a "substantial part" can only be determined on a case-by-case basis, leaving AI companies with little clear guidance on the issue. This permitted use is also limited in another key respect – it applies only to personal use by an individual, and the definition of personal use has not yet been clarified by the Ghanaian courts. It can be argued that training LLMs for commercial purposes

[15] Section 38 of Act 690
[16] The Copyright Act, 2005 provides penal sanctions for the infringement of copyright. Additionally, the Electronic Transactions Act, 2008 (Act 772) also provide sanctions for unauthorized access of electronic records.
[17] Section 47 of Act 690
[18] Section 47(4) of Act 690
[19] Rex Owusu Marfo v. Joy Industries Ltd (Civil Appeal No. H1/34/2020) (delivered by the Court of Appeal on 17th June 2021)
[20] Section 5(a) of Act 690
[21] Section 19(1)(a) of Act 690
[22] Section 19(2) of Act 690

may not qualify as personal use under the law. Therefore, these restrictions significantly narrow the scope of permitted use for training LLMs.

For works in the public domain, a prescribed fee must be paid to use them without any restrictions.[23] However, it is crucial to verify whether a particular work has entered the public domain. Such verification can be challenging when the literary works are not registered.

Recommendations

AI companies that train LLMs face significant risks of copyright infringement claims. For instance, there are currently around 20 pending AI copyright lawsuits in the US. In the absence of specific legislation on AI and copyright that clearly defines the rights of all stakeholders, copyright owners are likely to use the courts to protect their rights. In Ghana, the courts have the authority to order the production of evidence in intellectual property infringement cases, which means that an AI company could be required to disclose its training dataset.

Given these risks, it is crucial to incorporate a legal risk evaluation into the dataset collection process. This evaluation should focus on determining whether licenses or authorizations are needed for copyrighted works to be included in the dataset and ensuring that works which do not require the consent of the copyright owner meet all required conditions for unrestricted use. By including this evaluation step in the training process, companies can reduce the likelihood of legal disputes and make informed decisions about the sources of text data used for training LLMs.

[23] Section 38(3) of Act 690