AI Takes Transformers Beyond Robots In Disguise

By **Sean Li** (January 29, 2024)

Sean Li

Upon hearing the word "transformer," thoughts of change and adaptability often come to mind, sometimes evoking images of those iconic shape-shifting robots.

However, when it comes to artificial intelligence, the word transformer assumes a different, yet equally dynamic, role. Introduced in Ashish Vaswani's 2017 paper "Attention Is All You Need,"[1] a transformer in AI refers to an innovative type of AI model that has revolutionized the field.

This term gained significant attention in the AI industry following the introduction of OpenAI's ChatGPT, or generative pre-trained transformer, in November 2022, marking a pivotal moment in the commercial use of transformer-based AI models.

These models have sparked widespread interest and led to several high-profile lawsuits over the past several months, including notable cases such as Authors Guild et al. v. OpenAI Inc., filed on Sept. 19, 2023, The New York Times Co. v. Microsoft Corp., filed on Dec. 27, 2023, and Basbanes v. Microsoft Corp., filed Jan 5.

All three are pending before the U.S. District Court for the Southern District of New York. This surge in legal scrutiny, particularly regarding the use of copyrighted materials in training these transformative AI models, has brought to the forefront a discussion about the intersection of AI technology and copyright law.

The serendipitous overlap of terms is hard to ignore: Just as the term transformer in AI signifies a leap in how machines understand language, the legal world's transformative-use term represents an important concept in copyright law, where innovations in the use of copyrighted material could potentially make such use of a so-called fair use.

The critical question arises: Does using copyrighted materials to train a transformer constitute a transformative use?

Transformative Use In Google v. Oracle

The concept of transformative use was most recently discussed before the U.S. Supreme Court in the Google v. Oracle case.

The crux of the case revolved around Google's use of Java API code in developing its Android operating system. Oracle, which acquired Java, argued that Google's use of this code violated its copyrights. Google, however, contended that its use constituted fair use.

In a landmark 2021 decision, the U.S. Supreme Court held that Google's use of the Java API was indeed a fair use. Central to this determination was the concept of transformative use.

This doctrine posits that if the use "'adds something new, with a further purpose or different character, altering' the copyrighted work 'with new expression, meaning or message,'" it is likely to be considered transformative.[2]

The Supreme Court found that Google's use of the Java API was transformative because it "seeks to create new products" — Android — where Google repurposed the Java API for a new smartphone environment that expanded the Java platform's use and reach.[3]

In Google v. Oracle, a new product was clearly created, that is, the Android operating environment for smartphones. But when we focus on AI models, we observe a process that appears to be merely text input and output, seemingly not generating anything new besides the text itself.

This situation prompts a question: Does the analysis from the Google v. Oracle case align with the domain of AI?

The answer to this question hinges on a deeper understanding of how transformers utilize these literary materials. Do they simply replicate and store the original works, or do they transform them in a way that is akin to how Google repurposed Java's API for Android?

To unravel this, we first dive into the mechanics of the transformer model, exploring how it processes and reinterprets literary inputs to create something new and potentially transformative.

Transformers in Large Language Models and the Training Process

In the development of AI machine learning, the term transformer has emerged as a cornerstone concept. This term, introduced in Vaswani's paper, describes an advanced architecture that has revolutionized the way machines process language.

To grasp the concept of a "transformer," imagine every word[4] we use in our daily conversations can be mapped to a corresponding star in the sky.

Take, for instance, the sentence "Time flies like an arrow; fruit flies like a banana." Each word in this sentence can be visualized as a star with a particular position in the sky.

The positions of the stars — or words — in the sky are critical for AI models to perform their calculations. However, there is a problem. Without training, these positions are somewhat arbitrary. They do not represent the true positions of the stars in their constellations. In other words, they do not convey nuanced meanings of the words in the context.

In our "Time flies like an arrow; fruit flies like a banana." sentence, the word "flies" appears twice, but with entirely different meanings — once as a verb and once as a noun. In an untrained model, both instances of "flies" might be mapped to the same position in the sky, leading to confusion.

This is similar to how atmospheric refraction can bend starlight, distorting a star's apparent position in our naked eyes, which does not accurately represent its true location for astronomers' use.

This is where the concept of the transformer comes in. To accurately represent the meaning of words, their spatial positions need to be adjusted or transformed[5] based on their context.

Proper transformation ensures that the positions of the words more accurately reflect their true meaning in the context. Essentially, the transformer's role is to reposition every word corresponding to the input context, ensuring they align accurately with their contextual

meanings.

In our sky analogy, the transformer functions as an array of lenses through which we observe the stars, and it adjusts their positions to provide a more accurate basis for subsequent calculations.

The parameters of this lens-like transformer, i.e., the extent and direction word positions are adjusted in the "sky," are stored in extensive matrices with billions of parameters.[6]

These parameters are established in training. Throughout the training, an AI model receives a vast number of texts and articles. Each article provides the AI model with a unique context. In each context, the AI model would let each word sense its true position by observing every other word in such context.[7]

The process of sensing the true position of a word by observing other words is known as "attention." To draw a parallel with our earlier analogy of stars, this attention can be seen as a gravitational force exerted by each star upon the others.

This gravitational interaction helps determine the star's true position in the sky, analogous to how attention helps words find their true positioning in the linguistic universe.

The "lens" utilizes this gravitational information to set its parameters for adjusting the positions of the stars. Each article in the training slightly alters the transformer's lens parameters, reflecting the specific context provided by such article.

Over time, after analyzing millions or billions of articles, the transformer's parameters would evolve to accurately represent the transformation across a wide array of contexts.

When a user interacts with a trained AI model, such as asking a question, the lens-like transformer stored in matrices helps the computer interpret each word by repositioning them. Consider the earlier sentence: "Time flies like an arrow; fruit flies like a banana."

In this context, a trained AI model would distinguish between the two instances of "flies." Although the word is the same, the transformer, or the lens, adjusts each occurrence to distinct positions in the conceptual space. In this sense, training the transformer resembles shaping a piece of quartz into an array of mathematical lenses stored in matrices.

AI Training and Transformative Use

Having delved into the technologies of transformer AI models, we now turn our attention to the intersection of AI with legal considerations.

At first glance, one might contend that employing copyrighted material in AI training does not constitute transformative use because AI training does not yield any new, tangible creation but rather facilitates the AI's comprehension of language.

In this view, the AI merely uses these texts to decipher language patterns without contributing any new expression. The essence of the text, as an informational or expressive entity, appears to remain unaltered.

Exploring technology in greater detail, however, reveals a more complex picture. Unlike human reading, which involves understanding themes, narratives and factual content, the training phase of an AI model is fundamentally different.

It does not appear to read text in the human sense; instead, it processes and understands text at a much more granular level, focusing on the contextual relationships, or so-called attentions, between words.

This method appears very different from human cognition, at least at a conscious level. AI delves into the subtle dynamics of how words relate to each other within specific contexts, capturing this information in extensive matrices.

In the training process, one might see that the AI is indeed creating something new. The creation of matrices, which store the learned transformation parameters, could be seen as a form of new expression.

These are not mere reproductions of the original texts; they represent a transformative layer of information, distinct from the original texts.

Much like the extensive use of sanding blocks to refine a quartz lens, the training articles serve as the sanding blocks and the matrices are the final lenses, a new and different form of creative expression than the training articles.[8]

Indeed, if the use, as the Supreme Court in Google writes, "'adds something new, with a further purpose or different character, altering' the copyrighted work 'with new expression, meaning or message,'" then such use could be considered transformative.[9]

Sean Li is a partner at Benesch Friedlander Coplan & Aronoff LLP.

[1] Vaswani, Ashish, et al. "Attention Is All You Need." Advances in Neural Information Processing Systems 30 (2017): 5998-6008.

[2] Google v. Oracle, 141 S.Ct. at 1202.

[3] See id. at 1203 ("It copied the API (which Sun created for use in desktop and laptop computers) only insofar as needed to include tasks that would be useful in smartphone programs... . To repeat, Google, through Android, provided a new collection of tasks operating in a distinct and different computing environment.").

[4] To be more precise, it should be referred to as "token," but we are using the term "word" in this article for the sake of simplicity and ease of understanding.

[5] The "transformer" includes, among other features, the linear transformations performed within the Attention function, denoted as Attention(QW_i^Q, KW_i^K, VW_i^V), where the matrices W_i^Q, W_i^K and W_i^V are utilized to transform the three identical input matrices Q, K, and V following the embedding step. See "Attention Is All You Need" at 4-5. Additionally, the embedding matrices also serve a transformative role by converting input tokens into vector representations. These matrices, integral to the transformer's operation, are part of the trainable parameters, unless they are specifically frozen during training. See id. at 5.

[6] In "Attention Is All You Need", the matrices include W_i^Q, W_i^K and W_i^V, as well as the embedding matrices. See "Attention Is All You Need" at 4-5. These matrices can contain billions of parameters. For example, GPT-3 contains about 175 billion parameters. See https://en.wikipedia.org/wiki/GPT-3.

[7] The "observing" action is described in the equation, where Q "observes" itself through a dot multiplication with its transpose K^T. See "Attention Is All You Need" at 4.

[8] Of course, under certain prompts, an AI model may output the identical text used in training. However, the output is a separate process from the training process, and because it depends on the users' prompts, it involves distinct legal considerations.

[9] Google v. Oracle, 141 S.Ct. at 1202.