Am I in AI?

Description

Iâ??m a member of the Authorâ??s Licensing and Collecting Society (ALCS). They collect money for â??secondary usesâ?• of publications â?? such as photocopies, digital reproduction and educational recordings. These rights bring in only small amounts of money, so authors donâ??t usually take the effort to collect them. Nor do they know how to do this.

The ALCS gathers such funds collectively and distributes them to member authors. The amount collected and distributed is based on data such as stats on university photocopying as monitored by the Copyright Licensing Agency (CLA).

I have been a member of the ALCS since 2006. I joined after a friend told me it was a source of â??free moneyâ?•! The average payout has been about Â£200 a year since I joined and registered my publications on the ALCS database.

The ALCS has also been active in protecting author rights in relation to the use of uncredited author material in training large language models. I recall responding positively in their questionnaire about the use of published outputs in AI training, whether or not as an author I am acknowledged.

See the ALCS survey report. Authors canvassed include both writers who depend on income from their writing and academics for whom the relationship between publishing and income is less direct. Attitudes to AI seem to reflect this difference, though the report does not highlight this.

Class action

Anthropic is the AI company that developed the Claude chatbot. It was founded by former members of OpenAI, the creators of ChatGPT. I have used both LLMs. Did Anthropic use my published works as part of the extensive cache of data they used to train its large language model (LLM), Claude? It seems Anthropic mined books from pirate servers LibGen and PiLiMi that infringe copyright and breach publishing paywalls.

The ALCS mounted a class legal action seeking compensation from Anthropic for using author content without permission. The result was a $1.5 billion settlement between Anthropic and authors. The website **www.anthropiccopyrightsettlement.com** hosting a searchable list of all works eligible for compensation went live this month. It appears that six of my own authored books are in the training cache.

How AI training works

Knowing that training data gets chunked and processed in various ways as part of the process I am interested in the extent to which the â??piratedâ?• texts retain their integrity as singular authored works during the process. I concede that it is illegal to process, distribute and make money from an author or copyright holderâ??s text without permission, irrespective of how it is processed.

But I would also like to know whether the texts are used in **pre-training** or **fine tuning**. I donâ??t think the answers make any difference to the legitimacy of the claim and the settlement, but a wider understanding of the process by authors may temper some of the alarm about AI training and synthetically generated text evident amongst ALCS survey respondents.

ChatGPT on training

I put these issues to ChatGPT that assisted me in a response. For accounts of how LLMs are trained see post Training your AI and my recent book *AI and Language in the Urban Context* (open access).

Itâ??s unlikely that my books or articles survive inside the model as intact texts. Training generally involves breaking works into small fragments of text â?? â??tokensâ?• â?? which are stripped of formatting and context, then fed into the system in vast numbers.

ChatGPT reminded me that these fragments lose the boundaries of authored works and become part of statistical patterns that the model uses to predict sequences of words and hence create coherent sentences. So my sentences, paragraphs, or arguments may well have been included in Claudeâ??s (Anthropicâ??s) training, but as dispersed shards rather than recognisable wholes.

The model cannot recall my text as a coherent authored work, but it can generate new passages that echo my phrasing, style, or arguments â?? in consort with the influence of countless other texts processed during training.

This is one reason why questions of attribution and compensation are difficult. Authors donâ??t just contribute to an identifiable repository; they contribute to an ocean of fragments from which new outputs are synthesised.

As to whether such texts are used in pre-training or fine-tuning, the general pattern is that large, indiscriminate text collections are employed at the pre-training stage, while fine-tuning involves more selective, curated material to shape the modelâ??s behaviour. See post: The human in AI.

ChatGPT suggested that if my writing appeared in an online database or repository scraped for training, it is more likely to have entered the model at the pre-training stage. Fine-tuning tends to rely on

licensed or purpose-built data sets, but ChatGPT volunteered that there is no transparency about the extent to which academic works or published books are included.

AI influencers

The LLM cannot reproduce my books directly, but it can generate something that looks like them, based on patterns absorbed during training. One of the respondents reported in the ALCS survey expressed misgivings about just this aspect of LLMS: â??A writing style is nearly like a fingerprintâ?¦ it helps identify authenticity of a piece of writing. If AI was to copy this, you reduce the uniquenessâ?• (p.16).

Irrespective of the legality of training on pirated content, thereâ??s the curious pleasure for an author: the giddying thought that through LLM training, my own writing might exert some positive influence on interactions people have with AI tools and hence with the ever-churning oceans of the intellect.

Reference

- ALCS. *A Brave New World? A survey of writers on AI, remuneration, transparency and choice*. London: Authorâ??s Licensing and Collecting Society, 2025. https://d16dqzv7ay57st.cloudfront.net/uploads/2024/12/A-Brave-New-World-_ALCS_AI_Report.pdf

Note

- Featured image is by ChatGPT: Please provide a postapocalyptic image of a fingerprint reader.

Category

1. Artificial Intelligence
2. Podcast

Tags

1. ALCS
2. author
3. copyright
4. training

Date Created
October 4, 2025
Author
rcoyne99