Q&A - Using Intellect's Content to Train Large Language Models (LLMs)

AI companies are approaching publishers seeking to use their content for training LLMs, in exchange for payment. While we haven't entered into any such agreements, we recognise that Intellect content along with other publisher content is most certainly already being used in this way without permission.

We do see potential value in partnering with AI companies for both authors and Intellect including increased visibility, recognition, impact and accessibility of your work, as well as increased revenue and royalty potential. However, to ensure these benefits are realised, we would need to make sure mechanisms for transparency and fair attribution are established. We also believe authors should have a say in whether their content is used, and revenue-sharing models should compensate them for their contributions.

If you are not familiar with Large Language Models (LLMs) you can find an introduction to the subject in our document *What's all the fuss about Large Language Models*. We recommend you read this first.

Below is a Q&A which we hope will address most of the questions and concerns you may have about any contract we might make with an AI company in the future.

Has Intellect signed an agreement with an AI company yet or is it in discussions with one?

No, we haven't signed an agreement with an AI company, nor are we in discussions with any. However, AI companies are approaching publishers to licence their content to train their LLMs so we believe it's only right to include you in the conversation.

Will there be a mechanism to opt in/out?

Yes. If Intellect signs an agreement with an AI company, we will ensure that authors would need to **opt in** to have their work included. By default, author content would not be included.

Will I retain ownership of my work, or would the agreement transfer any rights to the AI company?

You would retain full ownership of your work if that is what you have currently. The agreement will not transfer any rights to the AI company; any contract will only allow for limited use of the content for training purposes under specific conditions.

How will the agreement align with copyright laws?

The agreement must comply with copyright laws. It must be structured to ensure that your content is not replicated verbatim or distributed outside the terms allowed by the contract.

Will I receive any compensation or royalties?

The training use of your work is not monetised in a way that generates direct royalties for authors. However, we believe that authors should share in any revenue derived from a contract

so we will look at that at the time. Even so, authors will indirectly benefit by increasing the visibility and accessibility of their work, which could lead to sales of your work.

What specific parts of my work would be used?

Any agreement would include access to full texts for training purposes, but no raw content would be shared publicly. Only aggregated and anonymized data are used in LLM model training.

What purposes would my work serve?

Your content would be used to train language models to improve their understanding of arts and humanities scholarship. This includes creating tools for summarizing, analysing, or generating insights based on broad academic trends.

Could the LLM tool generate outputs that replicate or paraphrase my original work?

LLMs do not replicate or paraphrase content verbatim. Outputs are designed to be original and transformative. However, we would take into consideration what mechanisms would be in place to prevent plagiarism or misuse.

Could the LLM generate derivative works?

While LLMs may produce outputs inspired by trained content, such outputs are not normally considered derivatives of specific work, although this remains a point of discussion. Ownership of these outputs typically reside with the tool's users, but we would look at how attribution mechanisms could work.

My work is published Open Access under a CC licence. Can this be used for training an LLM?

Creative Commons licenses require that the original creator is credited when their work is used, distributed, or modified. However, LLMs are typically trained on massive datasets that may include millions or billions of individual works. Tracking, storing, and reproducing detailed attribution information for every piece of training data would be computationally and logistically difficult. Even so AI companies are looking at solutions to this and we would discuss attribution with any AI company we worked with.

How would this align with ethical standards in academia and publishing?

We believe this would enhance accessibility and innovation in scholarship, aligning with our mission to disseminate knowledge. We would need to assure ourselves that any AI company we worked with was committed to upholding academic integrity and transparency.

Could this lead to devaluation of original research or misrepresentation?

We would need to assure ourselves that any AI company we worked with was committed to supporting rather than undermining academic scholarship. For example, outputs would not replace original texts but complement them by providing summaries or exploratory insights.

What safeguards would be in place?

We would need to assure ourselves that any AI company we worked with was committed to rigorous testing to prevent biased or inaccurate outputs. We also expect some oversight to address issues if they arise.

Who would be accountable if the AI misuses my work?

We would expect the AI company and our organization to share accountability. We would expect mechanisms for reporting and addressing misuse would be put in place.

Will this technology compete with my work?

LLMs are designed to support researchers, not replace scholarly works. Summaries or insights generated by the AI will more likely drive more readers to the original texts.

Could it discourage new readership?

We expect the opposite. By making research more discoverable, these tools can broaden the audience for arts and humanities scholarship.

How might this affect the value of arts and humanities research?

We believe this initiative underscores the relevance of arts and humanities by embedding their knowledge into cutting-edge technology. However, we are mindful of potential risks and will continue to monitor the technology to help ensure the sector's value is preserved.

What measures are you taking to educate authors?

We are organising informational videos, Q&A documents, and already provide resources about LLMs and their implications such as *What's all the fuss about Large Language Models*.

Does Intellect have a long-term strategy?

Yes, our strategy focuses on ensuring that AI is used responsibly and transparently to enhance scholarship. We are committed to adapting our policies as the technology evolves.

Final Note to Authors

We understand the complexity of these issues and value your feedback. Any agreement will be part of our effort to stay at the forefront of innovation while upholding our shared commitment to academic excellence. Your concerns are important to us, and we are committed to addressing them collaboratively.

Please share your thoughts and questions with us, we value your input.

Mark Lewis | Managing Director
mark@intellectbookls.com