Public Document

GenAI for Agriculture Advisory (GAIA)

AI Licensing - Pricing Guidance for Publishers and Non-profits

CABI & Zachary Haynes (Consultant)

November 2025

KNOWLEDGE FOR LIFE

AI Licensing: Pricing Guidance for Publishers and Non-profits

This report was developed by Zachary Haynes, Principal Consultant at IndieLex, an AI licensing consultancy agency, working on behalf of CABI. Zachary is a former Director of Licensing for Oxford University Press, the largest non-profit, mission led university press in the world.

The consultation and writing of this report was carried out during the second phase of the Generative AI for Agriculture Advisory (GAIA) project, led by IFPRI (CGIAR). CABI gratefully acknowledges the funding provided to this project by the Gates Foundation and the Foreign, Commonwealth and Development Office, UK (FCDO).

Pricing models included in this report are intended as a guide only. Publishers, including CABI, should adopt commercial terms or pricing models based on their own strategic considerations, the evolving market landscape, and the specific customer use case, which may differ from those described here.

Table of Contents

Goals 1
Market context 1
Legal landscape 3
Mission-based business perspective 3
Strategic approach to pricing 3
AI rights to books and journals 4
Open Access books and journals 4
Pricing for Subscription products 5
Practical Guides/Factsheets/Posters/Infographs 5
Data Portals 6
Courseware: Video content with supplemental materials (i.e. Q&A Banks) 6
General Fixed Term/Fixed Budget Model 7
Freemium Model 8
Sample AI Corpora Licensing Recommendation 8
Conclusions 8
Appendix of Embedded Links 9

Goals

Academic and non-profit publishers, societies and research institutions' portfolios of intellectual property and copyrights include books, journals, major reference works, factsheets, databases, and online resources covering vast subject areas and fields of expertise. They are trusted sources of subject-specific material, with a wealth of experience, expertise and value to offer the emerging AI (models, products, services and applications) training ecosystem.

There are any number of heretofore unknown AI applications and AI-enabled tools, at commercial, non-commercial, institutional, organizational and individual levels that may benefit from academic and research publications, and the digital data that is cultivated and curated in pursuit of that scholarship.

The copyright holders need to find a sustainable and strategic approach for pricing their knowledge bases and repositories, to meet the market expectations of these opportunities that are simultaneously business prudent, ethical and mission-based. Therefore, we want to consider and determine a pricing strategy that is multipurpose and multifocused, considerate of various factors and axes, while also aligning the appropriate market and value expectations across a variety of use cases, customers and users.

This document offers an investigation into:

- The potential to create a license pricing framework in order for publishers to contribute their curated digital and analog assets into GenAI tools.
- The variety of commercial AI opportunities that may be presented to copyright, IP holders and creators.

Market context

Many AI companies that have built large foundational models have trained them using content already available on the web, through web-crawling and scraping methods. This is a challenge and an opportunity for publishers and organizations that hold resources that are both publicly available online through open access publishing policies, as well as closed access proprietary resources available only through subscription, or within archives.

While in many cases AI companies trained on the publicly available data under the pretense that it was ‘fair-use’, because AI model outputs are ‘transformative’ (both grey areas in copyright law), publishers are clear that crawling and scraping the web for ingestion and training is a machine learning commercial exploit that requires either explicit permission or a license.

Open Access licenses of CC-BY and CC-BY-NC have either attribution or non-commercial requirements and limitations, that should require AI companies interested in leveraging the works to seek formal license agreements with the copyright holders. However, they are ultimately dependent on jurisdictional copyright laws, which vary greatly, and have been at odds with powerful and influential AI companies that are pushing the boundaries of copyright protections. More explicit updates to these licenses, to address the realities of the circumventions being employed for web crawling and AI training, may be overdue and necessary.

Publishers have traditionally taken steps to protect their content from this unauthorized usage by utilizing the robot.txt protocol, that provides instructions to web crawlers (bots), about which parts of a site they are permitted to crawl or index.

As a strategy to minimize licensing costs, content acquisitions teams at commercial AI developers typically aim to commodify the content, minimized to its lowest base unit such as a per word metric. This devalues source material by treating all content the same. It discounts

the reality that some words cost more to create than others. A reality even more prescient now that AI generated 'synthetic' content proliferates online.

Synthetic data could be AI-generated content, or 'slop' that is automatically generated for content marketing purposes. It could be inaccurate information drawn from an AI-powered search, or poorly supported research that is incorporated into a human authored piece in error. Or it could be derived from human authored text, with the purpose and intention of creating more training content (where training content may be lacking or hard to source), in a specific specialized domain or area of interest. The fact is that synthetic data can unintentionally result in poorer training outcomes for the web crawling and scraping practices traditionally employed for training large language models. The proliferation of synthetic data should buoy the prices of quality human-authored content. But as AI has evidently enabled a massive increase in academic and research paper submissions over the last year, publishers need to be more diligent in screening and reviewing submissions to prevent any unintended consequences from AI-generated content.

Licensees aim to reduce the risk of elevating licensing costs that value content on a per word or per book basis differently, such as content that is born from higher editorial or production costs, or on a longer timeline, with more contributors, and a higher market value, such as academic monographs, journals, textbooks or well-maintained databases.

These buyers also portend, as the AI models ingest more data and the models perform at ever higher levels, that the market for books for training is rapidly and inevitably amounting to a decline in licensing demand and value for publishers' books. The more books ingested, the lesser the need for more books to train on. And fine-tuning models to perform marginally better, requires smaller amounts of speciality content.

The risks in this argument should be weighed against the scarcity of resources and the value a publisher can retain by possibly withholding speciality materials from the AI licensing marketplace. Small publishers, with ever more specialized and unique content, may benefit more from model 'fine-tuning', 'grounding' and RAG licenses than more generalized publishers. Yet, at the same time, it is debatable whether the largest general purpose AI models trained by large consumer-first companies will disrupt the specialty market occupied by specialty publishers, where their brands and reputations bolster their reliability, position, and authority. Regardless, there is no doubt that there have been many widely publicized licensing deals where publishers have granted broad rights to their full catalogues, on a price per book basis, and although they include large gross volume discounts, in aggregate they have resulted in large windfalls, winning publishers revenue multiple times their typical annual licensing revenues. For how long these opportunities will present themselves and knowing the true risk/reward ratio of entering into them, is what we aim to calculate via prudent pricing and strategically considered negotiations of rights and terms.

Academic and large STM publishing platforms like Wiley, (to some mixed reviews), as well aggregators in the industry (Clarivate, Kortext), are increasingly aiming to disintermediate the channels to licensing for AI. They do this by claiming to add a layer of value on top of the existing data formats. By preparing content into data formats that integrate more seamlessly with the variety of AI models available, they claim to cut costs for both publishers and the AI model developers. This strategy may prove effective as the long tail of the AI developer ecosystem increases, and as RAG models (as they are assumed) become more prevalent.

One thing to be aware of though, is that in disintermediating the licensing relationship between the largest Big Tech foundational models and the speciality publishers, they seize both value (taking a share of the licensing revenue), and control (often they retain their own internal AI development and exploitation rights in the process).

Legal landscape

Many lawsuits have been filed by and/or on behalf of authors and publishers, with regard to the training of AI language models on known pirated books databases. One of them, the Anthropic Settlement, has recently been settled in court, awarding the authors $3,000 per book. This could serve two purposes in the current market. It could potentially positively influence the pricing in the market, bolstering copyrights holders in the value of their IP. It may also encourage more legitimate licensing deals from some AI companies, particularly a handful of Big Tech firms that have shown hesitancy to engage in licensing discussions while their legal cases are pending, given their claims that their use of the publicly accessible pirated books qualified as both 'fair-use' and 'transformative'.

Unfortunately, there has been recent evidence that instead of leading to more legitimate licensing for books, print distributor services have been receiving large full catalog print orders. The idea being that the AI companies then scan the books they purchase circumventing the requirement to license them for training. The curious fact behind this strategy is that for the cost of purchasing and scanning the books, they could likely just license the digital files for nearly the same cost, or less, with less time and effort expended.

Mission-based business perspective

Non-profit publishers have multifaceted long-term goals and responsibilities to their customers, authors and the academic and research communities at large. They must pursue ethical and sustainable means to generate commercial value from their publishing activities, while continuing to make the content available to those individuals and communities that cannot pay commercial fees, but that depend most on the materials.

Therefore, establishing an AI licensing strategy and framework aims to serve both needs; first, provide strategic guidance to the organization's business leadership on how to potentially convert as many new business opportunities in the field of AI licensing at the highest margin; and secondly, serving the non-profit and in-country end-users that also distribute and utilize the materials, and whom also aim to employ some of the same powerful AI models and tools that promise the same scale and efficiency.

Strategic approach to pricing

Here we lay out some of the more common AI use cases and terms that commercial companies, and non-commercial organizations, are licensing for presently, and that may be presented to publisher-owned, licensed-in, or partner content. Note that foundational model training is assumed to be perpetual as the nature of the process effectively takes advantage of the content 'forever'. A crude, albeit common industry annual metric for 'perpetual,' when considering the life or usefulness of a piece of research may be 15 years, though some STEM content may have a shorter lifespan.

- Commercial Foundational Model training: Perpetual (15 years)
- Commercial RAG/Fine-Tuning
    - 1 Year/RAG. Fine Tuning
    - 3 Year/RAG. Fine Tuning
    - Individual AI Application (per year-basis)
- Open Access Repository
- Closed Research-Only (Internal Only)
- Closed Enterprise Commercial (Internal Only)

Customer size is a common approach to defining prices, although assumes the buyer's budget corresponds to their size, and thus applies a Gold/Silver/Bronze multiplier or discount to the book price benchmark for a given book type:

- Large/Gold: Generally reserved for multinational, Big Tech and Fortune 500 commercial businesses.
- Medium/Silver: A commercial enterprise or an established philanthropy or global mission-aligned NGO; while
- Small/Bronze: Start-up, small business, pre-seed or underfunded, (whether commercial, mission or charitably focused or not).

AI rights to books and journals

Buyers are negotiating licenses using a variety of pricing metrics including per book, per journal, per journal article, and per word calculations. Furthermore, one may then consider or expect additional volume discounting principles for larger bundles.

Here are broad High/Medium/Low tiers of some deals that we have come across in the academic licensing industry. These are 'Foundational Model' licenses with the following limitations: training-only, non-display, non-attribution rights entitling the licensee to perpetual rights to the content and outputs created from it, without the right to display or redistribute the content, in either verbatim, or original format.*

Type	High	Medium	Low
Best Sellers	$3,000	$1500	$100
Academic Textbooks (PDF/EPUB)	$2,500	$1250	$350
Academic Monographs (PDF/EPUB)	$1,500	$550	$100
Academic Journals (PDF/XML) *Per article	$15.00	$7.00	$3.00
Journals OCR'd Scanned PDFs *Per article	$5.00	$3.00	1.00

Price per book deals have ranged anywhere from $2000 on the high-end, for PDF and ePUB, to $100 per book on the low-end for non-native, OCR'd or scanned PDFs, to as low as $3.25 per book for public domain works out of copyright but not previously available in a digital format online.

As for market realities for Journals, deals are being struck with pricing anywhere from $15 per article, on the high-end for PDF and XML, to $1.00 per article.

[*LLM developers and publishers are reluctant for models to produce verbatim outputs of the content for a variety of reasons. Primarily because the models are unreliable and may hallucinate or misrepresent the content. And so while academics and publishers of scholarly content would generally want citation and attribution to their work, in this case, it continues to be an unquantifiable risk for both parties. (With the exception of some platforms like Perplexity.com, where this is their strategic USP, and they have a publisher partner strategy where they direct web traffic to partner content in exchange for a license fee and the right to display and attribute licensed content). It could potentially undermine traditional distribution outlets of publishers, so for this reason, if verbatim display is contemplated in any contracts, it is usually an acknowledgment from both sides that there may be outputs that appear verbatim, but that will not be more than 500 characters (100 words), and will be unattributable.]

Open Access books and journals

Open access journals, generally do not carry as high a licensing value as the non-OA books and journal copyrights. It may be that the 'attribution' and 'non-commercial' requirements are limiting or restrictive. But this does not mean they do not have value when included in licensable packages of books and journals. For publishers with a large percentage of content that is open, OA books and journals could be considered as part of the mix and included in any combination of licensable packages delivered. Especially considering RAG

implementations where attribution is viable (RAG/Grounding more so than in foundational LLM model training).

Also, aspects of the data format (XML), meta-data, annotation, linking or delivery mechanism (such as API) will have inherent costs and value that could be charged for.

Here is a model to consider for Open Access Content, with attribution, for both foundational model training or RAG/Grounding. These figures span ranges that could be agreeable, depending on size or number of books and articles, based on known industry costs to bring these to market, and subsidiary rights licensing rates that occur in aggregator databases or that have been used in some AI training marketplaces.

Term	**1 year (RAG)**	**3 year (RAG)**	**Perpetual**
Open books (per book)			
Gold	$50	$100	$150
Silver	$25	$50	$75
Bronze	$5	$10	$22.50
Open journal articles (per article)			
Gold	$5	$10	$20
Silver	$2.50	$5	$10
Bronze	$0.50	$1	$5

Pricing for Subscription products

In order to limit risks to cannibalisation of a publishers' institutional subscription business, we should start with the current institutional benchmark pricing, and then apply a discount or multiplier to that price, depending on the term, and customer size and type. Here is a suggestion for where to start the pricing, and then gauge whether or not it is acceptable, or appropriate, given the particular scenario and current market/customer fit.

Discounting /Multiplying Factors

- Bronze Member and Developing Countries (-50% Discounting)
- Silver: Academic (-33% discount)
- Gold Corporate (+50% Premium)
- 1 Year (RAG): rate x1
- 3 Year (RAG): rate x3
- Perpetual (Foundational Model/ Pre-Training): rate x 15

Practical Guides/Factsheets/Posters/Infographs

Practical Guides and reference resources that fall under Open Access licenses, where the average size may be relatively small, just a page or two, but the variety of specialty subject areas may be broad, can be particularly attractive for licensing, as they may often present information in a clearly formatted layout. Examples of these could be fact sheets, slides, posters or infographics. While these assets are significantly smaller in word count than other Open Access content types like books and journals, their specialized usefulness and presentation, as well as the costs to create, maintain and distribute them may be significant, and thus increase their value. Furthermore, the large quantity of them makes the viability and flexibility of volume discount bundles potentially attractive; therefore, the following pricing guidelines (per sheet/guide) may be a good starting point:

Term	1 year (RAG)	3 year (RAG)	Perpetual
Gold	$10	$20	$150
Silver	$5	$10	$75
Bronze	$2.50	$5	$37.50

Data Portals

A proprietary data portal is often a unique resource, especially if there are limited competitors or similar databases in existence. Whether compiled from a combination of open publicly available, or closed proprietary data sources, it's the expert curation of this rare and specialized data that makes it such a valuable resource to experts in any specialty field.

If it's a free service, we must find a way to quantify the value of the product in order to price it for general AI opportunities, since we believe this asset is of particular interest and usefulness. This value can be estimated based on the current annual operating costs, let's use $500,000 for illustrative purposes. Other factors include the number of monthly visitors, and the competitive position of the data portal. It is recommended to use the operating costs as a measure for its simplicity, and because it represents the other components (it takes investment to create and maintain high visitor numbers and competitive position).

This premise gives us the following perpetual pricing by data record and by geographic/country, depending on how we choose to offer the database to an AI developer. Additional content to accompany the data, like evergreen blogs/SEO marketing content can be used to add further value, but likely not additional price.

	Units	Price per unit
Data records	7000	$71
Countries	50	$10,000

Under this premise, we are assuming all data records hold equal value, but not all data records or country data records are relevant to every licensee. And in the interest of creating subsets of the full database that both meet the needs of the licensee, or at least work with their defined budget, we can create subsets of the database. Here would be an example based on the above pricing calculus:

A) 5% of the products delivered in the database would be a fee quote of $24,850

or

B) 5% of countries delivered would be a fee quote of $25,000

Courseware: Video content with supplemental materials (i.e. Q&A Banks)

Educational video and Q&As have been of particular interest for AI applications and model training as they are useful for benchmarking AI models. Here are pricing suggestions for both foundational and RAG terms. The pricing is arrived at starting with existing institutional license pricing, and then applying our AI multipliers, depending on AI term and use-case.

*RAG Pricing Models with Attribution, based in part on factoring of the established institutional pricing guidelines. The table indicates two equivalent courses (A and B) and a bundling discount (A + B).

	Gold (2x)	Silver	Bronze (-25%)
Course A			
1 year (RAG)	$20,000	$10,000	$7,500
3 year (RAG)	$35,000	$17,500	$13,125
Course B			
1 year (RAG)	$20,000	$10,000	$7,500
3 year (RAG)	$35,000	$17,500	$13,125
Course A+B (bundle discount)			
1 year (RAG)	$34,000	$17,000	$12,750
3 year (RAG)	$59,500	$29,750	$22,312

AI Foundational Model Training Question Bank (Perpetual Buyout):

This type of Q&A (video recall) content can often accompany video courseware. Here are broad price ranges for similar real-world content licensed in today's market.

	Foundation Assessment (Multiple Choice Questions)	Practitioner assessment/ Longer form questions	Video (per hour)
Gold	$4−8	$7−14	$50−150
Silver	$2−4	$3.50−7	$25−80
Bronze	$1−2	$1.75−3.50	$5−25

General Fixed Term/Fixed Budget Model

For scenarios when a potential customer or licensing partner has a fixed budget committed to licensing a broad array of content, here is a suggested approach:

- One aim might be to make as much content available to the public, for as long as possible, therefore this request might be met by including mostly Open Access content, for a longer period.
- Alternatively, the opposite may be the goal: More bespoke and specialized content, for limited distribution or term, or a compromise somewhere in the middle.

Modeling out the two goals would be helpful for then finding compromises, for example:

Consider if the client is a Gold/Silver/Bronze and apply (_X#_Books or _X# _ Journals or __ data Products) @ (Corresponding market per word/unit rate) x (Term period) x (Discount Rate) = Total License fee

Then repeat until the correct balance or product mix is achieved within the allocated budget.

Or, alternatively we may consider the standalone knowledge bases are of most applicable value, then determine whether subsets of the databases, that meet the specific desired end-users needs, are achievable within the allocated budget. A 'subset' approach could be a Freemium Model.

Freemium Model

Another suggested AI licensing market approach, to both meet the broad swathe of AI licensees needs, both commercial and mission led, while also generating new opportunities and seeding the market, could be the commonly used Freemium Product Marketing model for web services and apps in general.

Here a publisher makes subsets of their broad array of licensable offerings available to developers through either their website or known developer portals like Hugging Face. The goal is that once discovering and experimenting with the content, the AI developers build proof of concepts or pilot programs that are then a business case for licensing more content/data.

Sample AI Corpora Licensing Recommendation

If the goals are to make content available through an open AI development platform such as Hugging Face, we need to assume the broadest access, to a potentially unlimited audience of AI developers, therefore we would suggest pricing the content at the GOLD tier, at the Perpetual Foundational Model/Pre-Training rate.

This approach gives the AI developers the broadest rights and terms to incorporate and utilize the data, in order to serve the largest number of downstream beneficiaries.

Sample AI license terms:

- The term of the license will be a 3-year contract term.
- During the term licensee may distribute the content via the open AI development platform via a controlled method where developers that access and choose to utilize the content must register for the purposes of knowing where the content is being utilized.
- During the Term the community of developers may incorporate the licensed content into both RAG and foundational model use-cases.
- During the term, the developers may/must attribute the licensed copyright (with date) content where utilized.
- At license expiry, the content must be removed from the open AI development platform.
- After license expiry, developers will not be required to remove the content from their apps or services.

Conclusions

The AI licensing marketplace that has emerged in just the last few years is rapidly trying to keep pace with the technology and expanding specialty use cases. While the large foundational model license opportunities steeply increased in number, the per book, article (unit) prices came down. They have now plateaued, and the frequency and number of licenses has leveled off, as has the pricing. Publishers are now expecting a second wave of grounding, fine-tuning and RAG license opportunities that will be higher in value, per unit, with fewer units licensed. This will resemble more closely the types of traditional limited term, cited and attributed digital licenses that publishers are more familiar with. However, it's fair to assume that the rights granted on the back-end integration of the content will be expected to be as permissive as for the foundational model licenses. So how publishers choose to value, and how successful they are in leveraging, the front-end display and attribution of their works remains to be seen. Until that happens, the recommendations here-in may prove to be a helpful guide for strategically navigating the current opportunities, while also preparing for where the market may go.

Appendix of Embedded Links

The following links are embedded in the text, and full URLs included here for clarity.

- https://www.nature.com/articles/d41586-024-04018-5
- https://creativecommons.org/using-cc-licensed-works-for-ai-training-2/#:~:text=If%20AI%20models%20or%20outputs,license%20as%20the%20original%20works.&text=CC%20BY%2DNC%20and%20CC,permission%20for%20NonCommercial%20uses%20only.
- https://www.businesswire.com/news/home/20251013329806/en/Wiley-Launches-Interoperable-Platform-to-Power-Scientific-Discovery-in-Worlds-Leading-AI-Technologies
- https://www.ce-strategy.com/the-brief/gateways/
- https://kortext.com/partnerships/microsoft/
- https://www.anthropiccopyrightsettlement.com/
- https://www.publishersweekly.com/pw/by-topic/digital/copyright/article/96590-an-ai-licensing-primer-for-book-publishers.html