

Building AI while Compensating Creators

Details of the Australian AI Public Benefit Fund

31 August 2026

Context

The Consultation Paper on the [Australian AI Training Scheme](#) proposes a model with an [AI Training Permit](#) and an [Australian AI Public Benefit Fund](#). The paper says the Government could establish the fund to support Australian creators. Its design would draw on the *Public Lending Right Act 1985* (PLR)—compensating for income loss and supporting Australian culture.

Eligibility would be limited to Australians. The Minister would appoint a committee of relevant stakeholders—supported by staff within the responsible department—to administer the fund and develop operational rules about how the fund is distributed.

The consultation paper proposed developing the fund's detailed design in consultation with creators. Because the fund is decoupled from negotiations with AI companies, the Government has substantial latitude to design it to meet creator expectations. This paper explores the next level of detail about how the fund could operate, including more details on its aims, objectives, and how it could make distributions.

Aims and objectives

The [Australian AI Public Benefit Fund](#) aims to:

- benefit Australian creators
- be fair and equitable, including supporting the social licence of AI in Australia
- reduce the harms and downsides of AI on Australians, including acknowledging uneven impacts across sectors
- value, protect, and foster Australian culture, society, and lived experience, and
- be practical to administer.

The fund's objectives could be modelled on those of the PLR, which includes recognising creators and enriching national culture. **Attachment A** provides an overview of the PLR. Based on that model, the objectives of the fund would be to:

- 1) make payments to Australian creators in recognition of contributions to, and the impacts of, AI in Australia, and
- 2) support the enrichment of Australian culture by encouraging Australians to create.

What the fund is *not* doing

The fund is not directly compensating rights holders for the value their works contribute to AI training. One reason the AI Training Scheme is necessary is that it's not currently possible to accurately track the vast number of works involved in AI training, the ownership rights associated with each work, and the value each represents to resulting AI models. Further, compensating rights holders when their works are used to train AI would see the vast majority of the money go overseas. Instead of providing compensation, the fund recognises Australian creators' contributions.

Because of this, the fund size would not increase with the number of works. Rather, the fund fixes the annual sum. This fixed sum is then distributed across a growing pool of creators. By contrast, the PLR pays a fixed rate per copy, so the total payout grows with the number of works in libraries.

Details of the governance

The fund's governance would closely mirror the PLR.

An Act and a scheme. Like the PLR, an Act would legislate the fund's objectives, establish the AI Public Benefit Fund Committee (the Committee), and authorise the Minister to approve an Australian AI Public Benefit Fund Scheme (the Scheme) on the Committee's recommendation. The Scheme would set the operational rules of the fund, including the factors the Committee must weigh when recommending distributions. This structure provides payments to creators while allowing the Minister, on the Committee's advice, to adapt the rules as technology and creative sectors change.

AI Public Benefit Fund Committee. Like the PLR Committee, the Minister would appoint a committee of relevant stakeholders, supported by staff within the responsible department. Members would include a chairperson, representatives of key collecting societies, representatives of professional bodies for creative categories without a collecting society, representatives of permit holders, at least one independent AI expert, and officers of the departments administering the *Copyright Act 1968* and the Scheme. Neither collective societies nor professional bodies should represent a majority of Committee members.

The Committee would recommend the fund's allocation, weighing the factors set out in the Scheme, and the Minister would approve it. The Committee would recommend the split between the two mechanisms: Mechanism A, payments to creators, and Mechanism B, support for the enrichment of Australian culture.

Within Mechanism A, the Committee would recommend the share for each creative category. The Committee would work with stakeholders to deliver both mechanisms and recommend Scheme modifications to the Minister.

Transparency. Like the PLR Committee, the AI Public Benefit Fund Committee would report annually to the Minister, with reports tabled in Parliament. The reports would set out the Committee's methodology for allocating shares to each mechanism, creative category, and distribution partner, and the distribution of payments. Where others make distributions (e.g., collecting societies), the report would include the methodology they used with a rationale for why it meets the Scheme's objectives and relevant statutory factors.

Sunsetting. The Act should sunset if no payments are made to creators within five years of its entry into force. A concern raised during consultation was that the overall Australian AI Training Scheme may fail to attract AI training to Australia, leaving the Australian AI Public Benefit Fund unable to achieve its objective of recognising Australian creators' contributions to AI in Australia.

Details of the scheme

We propose two mechanisms to achieve the objectives of the fund:

- 1) Mechanism A: Payments to creators
- 2) Mechanism B: Programs supporting the enrichment of Australian culture

Allocation between mechanisms

The Committee would recommend allocations between the two mechanisms to the Minister. In the case of the PLR, the PLR Committee makes a similar assessment and recommends 80% to creators and 20% to publishers. In this case, a 50:50 split might reasonably reflect the fund's objectives. The fund may reasonably bias towards Mechanism A if distribution partners can reach most or all individual creators within most or all

creative categories. The fund may reasonably bias toward Mechanism B if distribution partners can reach only some individual creators in some creative categories.

Both mechanisms could result in undistributed funds. These should be reserved and paid out via Mechanism B in the following distribution. This is necessary to prevent perverse incentives for distribution partners.¹

Mechanism A: Payments to creators

Mechanism A pays Australian creators in the categories AI draws on and affects.² This could include writers, musicians, coders, journalists, academics, and others whose work helps build AI models or is disrupted by them.

The Committee determines how Mechanism A is distributed between these creative categories by weighing a set of factors laid out in the Scheme. These could include:

- the presence and proportion of the category across the stages of AI training³
- the effect of AI on the category
- the size of the category's creative workforce in Australia
- the category's cultural importance to Australia
- the capacity of distribution partners to reach Australian creators
- the extent to which creators in the category have other avenues to be paid for AI use, such as market mechanisms and TPM-protected material
- consultation with AI experts and experts from the creative sector, and
- other factors the Committee considers relevant

The Committee weighs these factors and recommends an allocation across creative categories for the Minister's approval. This recognises creators' contributions and the impact of AI on them.

Once the Committee has set the allocation across creative categories, it uses two methods to reach creators.

Reaching creators through their collecting societies

The Committee allocates each category's share of funds among established collecting societies, in proportion to how much of that category the society represents. For instance, if the Committee allocates 20% of Mechanism A to published books, and a collecting society represents 80% of Australian authors, the society would receive 16% for distribution to those authors. The remaining 4% would be distributed via the second method below.⁴ The Committee would establish these shares for each disbursement from the fund.

Reaching creators not represented by collecting societies

The Committee must also distribute funds to Australian creators in other ways. This is for two main reasons:

1. Some creative categories, like computer code, journalism, and academic publications, have no collecting society. In these cases, the Committee would work with relevant professional peak bodies and associations, such as the Australian Computer Society, the Media, Entertainment & Arts Alliance, or the National Tertiary Education Union, to distribute the funds to creators.

¹ Distribution partners should be incentivised to have broad reach within the creative category they distribute to.

² A creative category is a kind of creative work the fund treats as a unit for distribution – for instance, published books, music, journalism, or code.

³ As part of the AI Training Permit, AI companies could be asked to indicate the broad presence and proportion of creative categories their training data draws on, including at different stages of training. The detail of such a requirement should be developed in consultation with permit holders. The Committee could draw on this as one information source to help understand which creative sectors AI draws on and affects.

⁴ This is because it would be inappropriate for creators to receive more the smaller the reach of their representing body.

2. In other cases, a collecting society exists but does not represent all creators within the category. For instance, not all Australian musicians with works in AI training datasets are represented by a collecting society. In these cases, the Committee would need to explore ways to distribute funds to unsigned or self-releasing artists.

Undistributed Mechanism A funds would be distributed via Mechanism B the following year. Undistributed funds would not count towards the distribution ratio agreed by the Minister. This is necessary to prevent perverse incentives for distribution partners.

Mechanism B: programs supporting the enrichment of Australian culture

Mechanism B supports the enrichment of Australian culture, and reaches Australians whose contribution to AI is too diffuse to trace to a category or cannot yet be reached by a distribution partner. Where Mechanism A recognises creators in categories that can be identified and reached, Mechanism B enriches Australian culture and encourages Australians to create, rather than recognising individual creators.

Mechanism B could include:

- **Grants:**
 - **Grants to industries impacted by AI** - Each year, the Committee could identify occupations under acute pressure from AI and provide time-limited support. Jobs and Skills Australia's publications and the Government's ongoing monitoring of AI's employment effects would provide an evidence base. The Committee would consult the relevant Jobs and Skills Councils, publish an annual statement of affected occupations, and contribute funding to supporting entry-level positions, retraining, micro-credentials, transition programs, and self-employment support.
 - **Grants to culturally important industries** - grants could support ongoing human creation in culturally important industries. This could range from local music to children's television and First Nations art.
 - **Community creative grants** - the Committee could run grant rounds for local creative infrastructure and activities. Recipients could include live music venues, rehearsal and recording spaces, regional galleries and museums, public libraries, community radio, local festivals, and youth arts programs. Rounds could be weighted to reflect AI's uneven impacts across regions and sectors.
 - **Journalism and public interest news grants** - journalism is among the occupations most exposed to AI. The fund could make grants for regional and local newsrooms, cadetships, and court and council reporting. It could align with the News Media Assistance Program.
- **Scaling established programs.** Existing programs already reach people affected by AI, including:
 - Music Australia and Writers Australia, within Creative Australia. This could include funding scholarships and fellowships dedicated to First Nations creators.
 - The Regional Arts Fund, Playing Australia, Festivals Australia and Visions of Australia, which deliver arts activity and touring to regional communities
 - The Indigenous Visual Arts Industry Support program and the Indigenous Languages and Arts program, and
 - Digitisation funding for Trove.

- **Creating an Australian cultural corpus.** The Committee could pay Australian creators to co-create a curated national corpus protected by a TPM. The corpus would create a new income stream for Australian creators and opportunities to licence a TPM-protected curated data-set for AI training overseas – helping Australia and Australian creators also receive value when their works are used to train AI overseas under fair dealing or TDMs.
- **Creator income.** Ireland piloted a basic income for the arts from 2022 to February 2026, paying €325 a week to 2,000 randomly selected artists. The pilot was run as a randomised controlled trial. Participants showed greater capacity for creative work, less anxiety, and higher life satisfaction. A cost-benefit analysis found each €1 invested returned €1.39 to society. In February 2026, the Irish Government made the scheme permanent. A smaller Australian equivalent supporting 1,000 creators at \$500 a week would cost about \$26 million a year.
- **A lottery** - Many Australians will have made some contribution to AI through content they post online – reviews, code, forum discussions, blog posts, and community websites – collected via the Common Crawl. These contributions are too small and diffuse to track and administer individually. A public lottery could distribute this portion of the fund to randomly selected Australians. Assuming an annual fund of \$200m, 1% of the fund would be enough for 40 Australians to receive a \$50k windfall per year.

Frequently asked questions

Why can't the fund accurately pay people for their contribution to AI training?

AI training datasets contain a very large number of works from around the world. There is no way to measure the contribution an individual work makes to an AI model. The value of a training dataset comes from a vast number of works operating together, not from any work alone. The fund instead recognises contribution by creative category, which can be estimated.

Why can't the fund accurately pay people for their contribution to AI training where we can identify the rights holder?

The AI Training Scheme provides for individual payments where it's made practical by technological protection measures (TPM), like paywalls, or the rights holder chooses to opt out and enter into commercial negotiations. The permit does not allow a permit holder to bypass a TPM and use that material for training. The fund covers publicly available works, where individual payment is not practical. Knowing who created a work does not establish what the work contributed. Because an AI model is trained on billions of works, identifying a specific work or creator says little about what that work contributed to AI training.

What's the difference between compensating creators and recognising them?

These words aren't subject to strict definitions. Compensation can be about giving or receiving money in *direct exchange* for a good or service. Compensation can imply a detailed calculation or negotiation to best match the value of the good or service. Recognition acknowledges that a contribution was made and matters, without trying to calculate it precisely.

Why does the fund support creators, not rights holders?

The PLR pays authors, not copyright owners, via a personal and non-transferable payment. The fund follows that model. Its objectives are to recognise the people who made the contribution and to encourage Australians to keep creating. Paying living creators furthers these goals. Paying rights holders may not, because copyright often rests with employers, has been assigned, or has passed through estates. Money flowing there recognises ownership rather than creation. A creator does not stop being eligible because they assigned their copyright.

How does the Mechanism A factor “presence and proportion of the category across the stages of AI training” work?

This is one of several factors the Committee weighs. Presence is relevant because if a creative category is unrelated to AI, creators of that kind of work should not be recognised directly under Mechanism A. Approximate proportion is relevant because the extent of works used for AI training goes to the aims of fairness and equity. If one creative category makes up a very large proportion of AI training datasets and another makes up a very small proportion, that is a reasonable consideration for the Committee to weigh alongside other considerations.

It's not currently possible to detail the vast number of works involved in AI training, but achieving the aims does not require a precise measure. Tokens are likely the best basis for the Committee's consideration. Training tokens measure the units of text, audio, or image data the model processes during training.

The best measure may change as AI technology changes. The Scheme can be updated without amending the Act, and the Committee can adjust how it weighs the factors.

How does the fund interact with creators being able to opt-out of AI training?

The opt-out mechanism and the fund are separate. The opt-out is a permit obligation. Permit holders must let Australian rights holders opt out of training, and the permit does not authorise training on opted-out works. The fund is a separate support mechanism, resourced from consolidated revenue.

Opting out does not affect a creator's eligibility for the fund. The fund recognises Australian creators as a group and distributes payments by creative category. It does not pay for the use of individual works.

On what basis will distribution partners distribute funds under Mechanism A?

The Committee allocates each organisation its share. The organisation designs the distribution method, applying its sector knowledge to meet the fund's aims and objectives. This means recognising the creators it represents for their contribution to AI, not compensating them for individual works. Each organisation provides its methodology to the Committee and explains how the methodology it uses meets the aims and objectives of the fund. The methodology is published in the annual report. Organisations may retain a portion for administration costs, which is also published.

How will the Committee be appointed? Who will be on it?

The Minister would appoint all members, as with the PLR. Members would include a chairperson, representatives of key collecting societies, representatives of professional bodies for creative categories without a collecting society, representatives of permit holders, at least one independent AI expert, and officers of the departments administering the Copyright Act 1968 and the Scheme. Permit holders and AI experts are necessary so the Committee has the expertise to assess Mechanism A factors, like presence and proportion of the category across the stages of AI training and the effect of AI on the category.

How will the process be transparent?

Like the PLR, the Committee would report annually to the Minister, with reports tabled in Parliament. The reports would set out the Committee's methodology for applying the Scheme, the shares allocated to each category and distribution partner, and the distribution of payments. Where collecting societies or peak bodies distribute funds, the report would include their methodology and how it meets the Scheme's objectives. Committee decisions would be subject to judicial review.

How will complaints be handled?

The Committee may receive two kinds of complaints.

1. Complaints about the distribution methods in Mechanism A.
2. Complaints from creators who believe they were entitled to a payment under an agreed method that they did not receive.

For the first kind, creators have two avenues.

- Committee decisions about distribution would be subject to judicial review, allowing aggrieved parties to challenge a decision that failed to properly consider the Scheme's criteria.
- The Committee would maintain a channel for feedback on the distribution, which it would consider in the next cycle.

The Committee would hold back a portion of the fund to deal with the second kind of complaint.

Once partners have made their payments, a complaints window would open for a set period. A creator who believes they were entitled to a payment and did not receive one contacts the Committee. The Committee decides whether the person is eligible and pays from the reserved portion. Complaints made after the window closes would not be accepted.

Routing these complaints to the Committee rather than the partners avoids burdening collecting societies and professional associations, which may not be equipped to run a review process.

Does the fund, in combination with the permit, weaken Australian copyright laws?

No. Australian creators who do not want their works used in AI training, or who would prefer to negotiate within current frameworks, can opt out. The permit and fund are additive to existing copyright laws. They create new, optional mechanisms. They don't weaken copyright laws.

Attachment A: Public Lending Rights Scheme Explainer

The Public Lending Right Scheme (PLR) compensates Australian authors and publishers for income lost because of public lending libraries. Creators and publishers register through the [Lending Rights](#) online portal and lodge claims for their works. The relevant department then surveys a sample of public libraries. From this, they estimate how many copies of each claimed book are in libraries across Australia. From that, they calculate a payment per creator.

In 2024–25, each copy earned a creator between \$1.75 and \$2.35 per book, depending on how many total copies of the book are held. Most earned around \$2.30 per book. Publishers earned 25% of the creator rate. In total, the PLR paid \$14,785,462 to 7,044 creators and publishers in 2024–25. The median creator received around \$550.

How is the Public Lending Rights Scheme Administered?

The overarching legislation is the [Public Lending Right Act 1985](#) (the **Act**). Under the Act, the Minister for Arts can approve a Scheme to operationalise the purpose of the Act.

The [Public Lending Right Scheme 2016](#) (the **Scheme**) details who counts as an eligible creator or publisher, what makes a book eligible, how the annual library survey is designed and conducted, the rates per book, and the survey thresholds below which no payment is calculated.

Section 7 of the Act establishes the 'Public Lending Right Committee' (the **Committee**). The Committee's legislative function is to determine eligibility and approve payments under the scheme. In practice, day-to-day operation is delegated to staff in the Creative Industries Branch of the Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts.

The Committee is consulted before the Minister changes the payment rates, and reports annually to the Minister on the operation of the scheme, including any recommendations it wishes to make. The Committee is comprised of:

- A chairperson.
- Two members representing Australian authors.
- One member representing publishers of books written by Australian authors.
- One member representing libraries.
- One officer of the department administering the Copyright Act 1968, nominated by that Minister.
- One officer of the National Library of Australia, nominated by the Director-General.

The Minister for the Arts appoints all members. The chairperson and the author, publisher and library representatives are appointed for a period specified by the Minister not exceeding four years. The two nominated officers hold office until the Minister terminates the appointment.

What are the transparency and reporting obligations?

Section 19 of the Act requires the Committee to submit an annual report to the Minister. These reports are tabled in Parliament and published. They are the principal public source of operational detail on the Scheme.

The [annual reports](#) contain the payment rate table for the year, the format composition of the survey, the distribution of payments by amount, and the names of creators and publishers receiving the largest payments, among other information.

Claimants can review decisions of the Committee. A person affected by a decision may request a review in writing within 28 days. If the Committee does not respond within 90 days, the original decision stands, but the applicant may then apply to the tribunal for external review.

How are the rates set?

The Minister is ultimately responsible for approving the payment rates for each program year. Neither the Act nor the Scheme prescribe a method for arriving at a rate. The annual report lists the rates but lacks an explanation of how they are calculated.

However, changing a rate requires the Minister to amend the Scheme under section 5(1)(b) of the Act. These changes come with an explanatory statement. [Recent explanatory statements](#) provide that:

- The department reviews the rates annually, “taking into consideration the results of the Committee's library surveys and the available Commonwealth budget”.
- The tiered structure – whereby the amount per book depends on the total number of books – is to “provide responsibly targeted funding towards emerging and mid-career authors”, a move “widely welcomed by key stakeholders, including industry representatives on the Public Lending Right Committee”.

Beyond this, there is no public detailed arithmetic about how the rate is determined.

Section 16 of the Scheme also permits the Minister to determine minimum and maximum payment amounts, and to set different amounts for creators and publishers. Currently, a minimum of \$100 applies. No maximum is evident in the published payment data.

Attachment B: Analysis of Presence and Proportion of Creative Categories in AI Datasets

The Committee needs some information about the presence and proportion of creative categories in AI training data to inform its decisions about the division of funds under Mechanism A. Frontier AI companies do not publish their dataset composition. The Committee, once established, has different mechanisms to inform itself on presence and proportion:

- Article 53 of the EU AI Act requires some transparency about dataset composition for AI models available in Europe.⁵
- Some open-source information is available.
- The Committee would include an independent AI expert and representatives from permit holders.

This attachment explores ‘presence and proportion’ from two kinds of public disclosure:

- 1) published breakdowns of the data used to train models, and
- 2) the composition of open-source training datasets.

The disclosures **cover text models**. They provide little information about the images, music, sound recordings, or audiovisual material used to train multimodal systems.

Available Information on Dataset Composition⁶

Table 1. Disclosed training mixes of selected models, as a share of training tokens

Creative category	GPT-3 ⁷ (2020)	GLaM (2021)	Gopher (2021)	Chinchilla (2022)	PaLM (2022)	LLaMA 1 (2023)	Llama 3 ⁸ (2024)
General web material ⁹	82%	42%	58%	55%	27%	82%	50%
Conversations and forums ¹⁰	—	30%	—	—	50%	2%	—
News ¹¹	—	2%	10%	10%	1%	—	—
Wikipedia and reference	3%	6%	2%	1%	4%	4.5%	—
Books	16%	20%	27%	30%	13%	4.5%	—
Code (e.g. Github)	—	—	3%	4%	5%	4.5%	17%
Scholarly and mathematical (e.g. ArXiv)	—	—	—	—	—	2.5%	25%

⁵ European Union. (2024). [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#), Article 53.

⁶ Brown, T., et al. (2020). [Language Models are Few-Shot Learners](#); Du, N., et al. (2021). [GLaM: Efficient Scaling of Language Models with Mixture-of-Experts](#); Rae, J., et al. (2021). [Scaling Language Models: Methods, Analysis & Insights from Training Gopher](#); Hoffmann, J., et al. (2022). [Training Compute-Optimal Large Language Models](#); Chowdhery, A., et al. (2022). [PaLM: Scaling Language Modeling with Pathways](#); Touvron, H., et al. (2023). [LLaMA: Open and Efficient Foundation Language Models](#); Grattafiori, A., et al. (2024). [The Llama 3 Herd of Models](#)

⁷ GPT-3’s published figures add to 101%.

⁸ Llama 3’s displayed categories add to 92%. The remaining 8% is multilingual, omitted here.

⁹ Combines general web crawls and their refinements (Common Crawl, C4, MassiveWeb, WebText2).

¹⁰ When separately identified.

¹¹ When separately identified. These figures cover only material expressly labelled as news. News articles inside Common Crawl and other general web collections remain classified as web material, so the true news share is higher but cannot be calculated from published information.

Table 2. Composition of open training datasets, as a share of the corpus¹²

Creative category	SlimPajama	RedPajama-1T	The Pile	Dolma	Common Pile
General web material	78.9%	87.0%	28.1%	81.0%	8.7%
Conversations and forums	3.3%	1.7%	6.8%	2.9%	14.2%
News ¹³	—	—	—	—	0.04%
Wikipedia and reference	3.8%	2.0%	1.5%	0.1%	9.9%
Books	4.2%	2.1%	15.0%	0.2%	4.5%
Code	5.2%	4.9%	7.6%	13.4%	19.0%
Scholarly and scientific	4.6%	2.3%	27.1%	2.3%	34.2%
Legal and government	—	—	10.5%	—	8.5%
Other ¹⁴	—	—	3.4%	—	1.0%

The two tables use different measurement bases.

- Model disclosures report the share of tokens the model processed during training, counting material sampled more than once.
- Dataset disclosures report the share of the corpus as distributed.

The same source can hold a larger share of a training run than of the corpus it came from, because labs oversample material they judge to be high quality.

There are a number of observations to make from these tables:

- Web material dominates every mix, ranging from 27 to 87%.
 - The Committee may be able to work with AI experts and permit holders to break down web material with more granularity. For instance, the Committee may be able to estimate what share is news, and factor that into its Mechanism A distribution.
- Books range from 2 to 30%.
- Code ranges from 3 to 19% and has grown over time.
- Scholarly material is small in web-scale mixes and large in curated ones, reaching 34% of the Common Pile.
- Film and television appear in text corpora only as subtitles. Music does not appear at all.

How these Datasets Might Have Changed

A frontier training run in 2026 uses around 10-30x more training datasets than the 2023 LLaMA 1 run.

Fixed-size sources shrink as a share when training scales. Wikipedia was 4.5% of LLaMA 1 and can only be a fraction of one per cent of a frontier run.

Recent recipes weight code and mathematics more heavily. Llama 3 used 17% code and 25% mathematical and reasoning data. The fully open-source model Olmo 3 (2025) discloses a mid-training mix of 20% code and 19% mathematics.¹⁵

Implications for the Fund

¹² Gao, L., et al. (2020). [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#); Shen, Z., et al. (2023). [SlimPajama-DC: Understanding Data Combinations for LLM Training](#); Soldaini, L., et al. (2024). [Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research](#); Kandpal, N., et al. (2025). [The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text](#)

¹³ When separately identified. These figures cover only material expressly labelled as news. News articles inside Common Crawl and other general web collections remain classified as web material, so the true news share is higher but cannot be calculated from published information.

¹⁴ The Pile's "Other" consists mainly of film and television subtitles (OpenSubtitles and YouTube subtitles) and mathematics. Subtitles are the only trace of audiovisual works in any text corpus that was clearly identified.

¹⁵ Team Olmo. (2025). [Olmo 3](#). Ai2.

Overall, this data suggests that some creative categories have natural distribution partners, while others do not. In the disclosed mixes, the creative categories with a natural Australian delivery partner sum to roughly 15% to 60%, depending on how curated the mix is.

The limited data above supports a split favouring Mechanism B as the main distribution mechanism. But better data, and consideration alongside the other factors, may change this balance. For instance, assistance from permit holders and AI experts may give the Committee more insight and confidence to identify distribution partners and allocate funding. If so, 50:50 may be a reasonable, evidence-based starting point for allocation between the two mechanisms.

DRAFT