POLICY BRIEF

POLLICINO & PARTNERS
AIDVISORY

Template for the Public Summary of Training Content for general-purpose AI models required by Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act)

On 24 July 2025, the European Commission published its official *Template for the Public Summary of Training Content for General-Purpose AI Models*, accompanied by an Explanatory Notice. This template serves to implement Article 53(1)(d) of the AI Act (Regulation (EU) 2024/1689), which requires all providers of general-purpose AI models (GPAIMs) placed on the Union market to make publicly available a sufficiently detailed summary of the content used to train their models.

This new obligation applies from 2 August 2025, including to models released under free and open-source licences. Providers of models placed on the market before that date have until 2 August 2027 to comply. This brief unpacks the purpose, structure, and legal significance of the template, with a particular focus on copyright, transparency, and regulatory strategy.[1]

Training summary: definition of the obligation

The training summary obligation is rooted in Article 53(1)(d) AI Act, and further substantiated by Recital 107, which outlines its dual rationale:

1. To enhance transparency regarding training content used in GPAIMs, including material protected by copyright and related rights;
2. To enable parties with legitimate interests, such as rightsholders, data subjects, downstream developers, and regulators, to better understand and, where applicable, enforce their rights under Union law.

Importantly, the template does not require granular or technical disclosure of datasets. Instead, it demands a "generally comprehensive" narrative overview of data sources and types, striking a balance between public interest transparency and the protection of trade secrets and confidential business information.

Scope and disclosure

The template divides the required information into three key sections covering the general information about the model, the sources, and the data processing aspects.

[1] See also our policy brief on the *Guidelines on General Purpose AI models.*

A. *General information*

The first section of the template should include the elements to identify the model, the type and volume of training data, the linguistic and demographic characteristics, and dependencies on upstream models.

B. *Data sources*

In the second section, providers must describe and categorise all data used across the full lifecycle of model training, from pre-training to fine-tuning. Categories include:

- Publicly available datasets (e.g. Common Crawl);
- Commercially licensed datasets;
- Private third-party datasets;
- Scraped or crawled online content;
- User-generated data from provider platforms;
- Synthetic data used for distillation or alignment;
- Other sources (e.g. digitised offline content).

A particularly sensitive area is web scraping, where providers must disclose crawler behaviour, collection periods, source types (e.g. news, social media) and, critically, a summary list of domain names scraped. For large providers, this includes the top 10% of scraped domains; for SMEs, the top 5% or the top 1000 domains, whichever is lower.

C. *Data processing aspects*

This final section relates directly to copyright compliance. In particular, given the amount of content used to train the models, providers must explain how they detect and honour reservations of rights under Article 4(3) of Directive (EU) 2019/790. Moreover, they must describe their processes for removing illegal content from training data: it seems, then, that they have an obligation to actively monitor the presence and the “moderation” of such content in their models. Optional information, instead, covers the disclosure of broader data governance practices. Not much is present on the data protection aspects of the training, if not a general reminder that “the lawful collection and processing of the data remains the responsibility of the provider under other applicable Union law (e.g. copyright and data protection)”, and that the transparency summary is clearly not replacing the obligations under the Regulation (EU) 2016/679 (e.g., Art. 15 GDPR).

Section	Content Focus
General Information	▪ Identification of the model and provider ▪ Modalities used (e.g. text, image, audio, video)
Data sources	▪ Publicly available datasets (e.g. Common Crawl) ▪ Commercially licensed datasets ▪ Private third-party datasets ▪ Scraped or crawled web content ▪ User-generated data from provider's platforms ▪ Synthetic data (e.g. from distillation) ▪ Other sources (e.g. digitised offline media)
Data processing aspects	▪ Measures to respect copyright opt-outs under Article 4(3) of Directive (EU) 2019/790 ▪ Processes to detect and remove illegal content from training data ▪ Optional disclosure of broader data governance measures (e.g. filtering techniques, handling of user data, synthetic data attribution)

Copyright

The Summary is conceptually and functionally tied to the copyright obligations under Article 53(1)(c) AI Act. While Article 53(1)(c) requires a copyright policy, Article 53(1)(d) operationalises this by compelling proactive transparency around the content used. This includes: i) public identification of datasets likely containing copyrighted material; ii) documentation of measures taken to detect opt-outs; iii) facilitation of downstream rights enforcement and accountability.
Notably, the Code of Practice for GPAIMs reinforces this by establishing structured expectations for respecting reservations of rights. Signatories are encouraged to align the summary disclosures with their commitments under the Code.

Publication

The summary of the training content should be made public and available at the time the model is placed on the Union market. As clarified by the Europan Commission, it should be published on the provider's official website, in a clearly visible and accessible manner. Clearly, providers must update the summary every six months, or whenever additional training data materially changes its content.

Enforcement

While specified that the summary should be made public at the moment of market placement, this means that for new models, the summary must be available from 2 August 2025 onward, while for older models (pre-2025), the summary must be published by 2 August 2027, with justification for any unavailable information.

Methodological note

While the term *transparency* is widely invoked across the AI Act, its practical and normative meaning remains context-dependent. In the case of Article 53(1)(d), transparency is not about technical replicability or full data traceability.

One aspect should be, therefore, clarified: being transparent about the processing and disclosing the summary does not equal legal compliance. Rather, it creates a presumptive evidentiary environment, where providers must justify their data use and rightholders are better positioned to assert claims.

The obligation to publish a training data summary, therefore, reflects a shift from technical opacity to normative accountability. It is designed to enable third parties to understand the *general nature* of what went into model training, not with the aim of full replication, but to permit scrutiny, challenge, and informed interaction.

This mechanism serves multiple overlapping policy objectives. It empowers rightsholders to assess whether their works may have been used and to assert reservations of rights where applicable. It allows data subjects and downstream developers to evaluate the linguistic, regional, and cultural representativeness of training data, especially in sensitive or high-impact domains. And it provides regulatory authorities with a benchmark for monitoring compliance with Article 53(1)(c), particularly the effectiveness and credibility of copyright compliance policies.

In this light, transparency here is intended not as exhaustive disclosure, but as a framework for accountable claims-making, and, hopefully, as an instrument to strengthen the access and safeguard of individuals' fundamental rights.