EUROPEAN
COMMISSION

Brussels, 24.7.2025
C(2025) 5235 final

ANNEX

ANNEX

to the

Communication to the Commission

Approval of the content of the draft Communication from the Commission – Explanatory Notice and 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)

1. Background

(1) Regulation (EU) 2024/1689 of the European Parliament and the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain regulations[1] ('the AI Act') entered into force on 1 August 2024. Section 2 of Chapter V AI Act lays down harmonised rules for providers of general-purpose AI models, including obligations applicable to all providers of general-purpose AI models[2] and additional risk assessment and mitigation requirements for those of the most advanced general-purpose AI models posing systemic risks[3]. Those rules will apply as of 2 August 2025.

(2) Article 53(1)(d) AI Act requires all providers of general-purpose AI models to draw up and make publicly available a sufficiently detailed public summary of the content used for the training of the model (the 'Summary'), according to a template provided by the AI Office (the 'Template'). Recital 107 AI Act contains additional clarifications on the objectives of the Summary and the Template which include transparency on the data that is used for the training of general-purpose AI models, including text and data protected by copyright law.

(3) Providers of all general-purpose AI models placed on the Union market must fulfil the above obligation, including providers of general-purpose AI models released under free and open-source licenses[4], in so far as the models fall within the scope of the AI Act[5]. Recital 107 AI Act contains additional clarifications on the Summary and the Template.

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), PE/24/2024/REV/1, *OJ L, 2024/1689, 12.7.2024*.

[2] Article 53 AI Act.

[3] Article 55 AI Act.

[4] The exception for free and open-source general-purpose AI model under Article 53(2) AI Act does not apply to the obligation to make publicly available the Summary.

[5] See Article 2 AI Act and Guidelines for providers of general-purpose AI models | Shaping Europe's digital future

RELEVANT LEGAL TEXT

Article 53(1)(d) AI Act.

Providers of general-purpose AI models shall […] draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

Recital 107 AI Act.

In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the AI Office to provide a template for the summary, which should be simple, effective, and allow the provider to provide the required summary in narrative form.

(4) Given that providers are obliged to draw up a Summary according to a Template provided by the Commission, the latter holds important legal value for the proper implementation of the AI Act. This Explanatory Notice and the Template annexed to it aim to help providers of general-purpose AI models comply with their obligation under Article 53(1)(d) of the AI Act in a simple, consistent and effective manner.

(5) The Template is based on the outcome of a multi-stakeholder consultation on general-purpose AI models, organised by the AI Office from 30 July to 18 September 2024[6]. Over 430 responses were submitted from a wide range of stakeholders. Based on this input, the AI Office prepared and presented its preliminary approach to the Template and allowed participants involved in the drawing up of the Code of Practice on General-Purpose AI[7] to provide additional written feedback. The current version of the Template annexed to this Explanatory Notice takes into account comments received from 111 stakeholders, including providers of general-purpose AI models, business associations, rightsholders organisations, academia, civil society and public authorities. The draft Template was also presented and discussed with the AI Board Steering subgroup on General-Purpose AI and with the European Parliament (IMCO-LIBE Committees) working group on AI.

2. Objective of the Summary

(6) General-purpose AI models are trained with large quantities of data for which there is typically limited information available. Recital 107 AI Act explains that the objective of the Summary is to increase transparency on the content used for the training of general-purpose AI models, including text and data protected by law and to facilitate parties with legitimate interests, including rightsholders, to exercise and enforce their rights under Union law.

[6] AI Act: Have Your Say on Trustworthy General-Purpose AI | Shaping Europe's digital future
[7] General-Purpose AI Code of Practice | Shaping Europe's digital future

(7) Such legitimate interests relate to copyright and related rights and other intellectual property rights, but also to other rights protected by Union law that should benefit from increased transparency.

(8) First, in relation to intellectual property rights, including copyright and related rights, transparency of the data used for the model training should help rightsholders obtain relevant information on the content used in the training of general-purpose AI models. This information is needed to facilitate the exercise of their fundamental right to intellectual property[8] and the fundamental right to an effective remedy in the enforcement of their rights, as provided for in Union law in the area of intellectual property rights. In the case of copyright and related rights, transparency of the training data will contribute to ensuring that general-purpose AI models providers comply with Union law on copyright and related rights[9].

(9) Second, transparency of the training data in the Summary may facilitate data subjects' rights and more broadly support the enforcement of the Union data protection rules. In particular, this can be done by summarising all the relevant information together, such as information about the data scraped from the internet or collected by the provider through interactions with the model or other services and products. The information in the Summary is not meant to replace, nor affect the respective information the providers of general-purpose AI models should make available to data subjects under Union data protection law. In the context of the Summary, the interests of consumers and the protection of their consumer rights under Union law may also be relevant.

(10) Third, transparency of the general characteristics of the content used for training may also assist providers integrating these models into downstream applications to assess the diversity of the data. This, in turn, will allow them to implement, where appropriate, mitigating measures to ensure that the fundamental rights to non-discrimination[10] and language and cultural diversity[11] are respected.

(11) Fourth, greater transparency of the training data may also facilitate the fundamental right to receive and impart information[12] and allow researchers to exercise their freedom of science[13] to conduct scientific research. It can allow academic institutions and organisations to critically evaluate the implications and limitations of a particuler general-purpose AI model and the potential risks and harms associated with the data used.

(12) Finally, transparency of the training data may also contribute to more transparent and competitive markets. For example, information about whether publicly available general-purpose AI models have been used to train other models, in particular through model distillation, or whether a model has been trained on user data collected from provider's own products and services, may help users and companies better understand how their data and models have been used and avoid potential lock-in effects.

3. Comprehensive scope of the training data and sufficient details

(13) Information about the general-purpose AI model provided in the Summary should cover data used in all stages of the model training, from pre-training to post-training, including model alignment and fine-tuning. This covers all sources and types of data, regardless of whether the data are protected or not, including by an intellectual property right. Since Article 53(1)(d) AI Act mentions explicitly 'training', other input data used during the model's operation (e.g. through retrieval augmented generation) are not

[8] Article 17(2) of the EU Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391–407.
[9] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.), PE/51/2019/REV/1, OJ L 130, 17.5.2019, p. 92–125.
[10] Article 21 of the EU Charter of Fundamental Rights.
[11] Article 22 of the EU Charter of Fundamental Rights.
[12] Article 11(1) of the EU Charter of Fundamental Rights.
[13] Article 13 of the EU Charter of Fundamental Rights.

required in the mandatory sections of the Template[14], unless the model actively learns from this input data.

(14) Recital 107 AI Act explains that the information about the training content should be comprehensive in its scope and sufficiently detailed to achieve the objective of the Summary of providing meaningful public transparency and facilitating parties with legitimate interests to exercise and enforce their rights under Union law.

(15) The Template annexed to this Explanatory Notice aims to provide a common minimal baseline for the information to be made publicly available in the Summary. It consists of three main sections:

1. **General information**: this section requires information allowing identification of the provider and of the model, and information on modalities, the size of each modality within broad ranges, as well as general characteristics of the training data.
2. **List of data sources:** this section requires disclosure of the main datasets that were used to train the model, such as large private or public databases, and a comprehensive narrative description of the data scraped online by or on behalf of the provider (including a summary of the most relevant domain names scraped) and a narrative description of all other data sources used (e.g. user data or synthetic data) to ensure completeness of the summary regarding the content used for the model training[15].
3. **Relevant data processing aspects:** this section of the Template requires disclosure of certain data processing aspects that are relevant for the exercise of the rights of parties with legitimate interests under Union law. This is especially important for compliance with Union law on copyright and related rights and for the removal of illegal content to mitigate the risk that such illegal content may be reproduced and disseminated at scale by the general-purpose AI model.

(16) On the basis that the Summary aims to provide sufficient details and to facilitate parties with legitimate interests, including rightsholders, exercising their rights under Union law, the Template requires a disclosure of a summary of the list of top domain names crawled and scraped from online sources in a summarised narrative form[16]. At the same time, it does not require disclosure of the details for the specific data and works used to train the model as this would go beyond the requirement in Article 53(1)(d) to provide just a 'summary', which in line with Recital 107 AI Act must be "generally comprehensive" but not "technically detailed". Providers may nevertheless decide on a voluntary basis to go beyond the minimum requirements in the Template and disclose in the Summary more details than what is required by Article 53(1)(d) AI Act and the Template[17]. Furthermore, for domain names scraped or crawled from the internet that are not listed in the Summary, it is recommended that providers act in good faith and on a voluntary basis enable parties with a legitimate interest including rightholders, upon requests, to obtain information whether the provider has scraped and used for training content which includes protected works and other subject matter that rightholders have made available on specific internet domains. This recommended and voluntary 'upon request' mechanism does not affect other available remedies for rightholders under Union law on enforcement of intellectual property rights (e.g. Article 8 of the Intellectual Property Rights Enforcement Directive)[18].

[14] Since such data is used as input for the generation process, its influence on the outputs of the model may be significant, and relevant for the exercise of the rights of parties with legitimate interests. For this reason it may be disclosed by the provider on a voluntary basis in the optional Sections in the Template which allow the provision of additional information.
[15] See in this context Recital 107 of the AI Act.
[16] See Section 2.3. of the Template that requires a list of the internet domain names (top and second-level domain, e.g. "example.com") in the top 10 % of all domain names determined by the size of the content scraped (in a representative manner across modalities, where applicable). For small and medium-sized enterprises (SMEs), including start-ups, the Template requires the internet domain names in the top 5%, or the top 1000 domains to ensure proportionality of the burden on SMEs in line with recital 109 AI Act.
[17] See in the end of each sub-section of the Template, optional part with possibilities to provide additional information.
[18] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (OJ L 157, 30.4.2004), OJ L 195, 2.6.2004, p. 16–25.

4. Balance with trade secrets and confidential business information

(17) As explained in Recital 107 AI Act, the Template should seek to strike a balance between serving the interests of parties with legitimate interests and promoting increased transparency of the training content in a meaningful way, while respecting the rights of all parties concerned, in particular taking due account of the need to protect trade secrets and confidential business information.

(18) Since the Commission is bound by the Charter on fundamental rights, this careful balancing exercise has been implemented in relation to the information that the Template requires to be disclosed in order for providers to fulfil their obligation under Article 53(1)(d) AI Act and provide a 'sufficiently detailed' public summary of the training content. The provision of information regarding more specific details about the content used for the training of the general-purpose AI models is required in the Template only where it is necessary to enable the exercise of rights protected under Union law in a meaningful manner as required by Article 53(1)(d) and Recital 107 AI Act. Determining which details should be disclosed has been the result of a careful balancing exercise carried out by the Commission in drawing up the Template to ensure that relevant information on the training data is provided to meet the Template's objectives, while confidential commercially sensitive information about the data sources and the precise manner in which providers curate the data and train their models is preserved.

(19) To protect providers' trade secrets, different levels of detail are required in the Template depending on the source of data considered. In particular, limited disclosure of information is required for licensed data given that the rightsholders concerned are parties to the licensing agreements (see Section 2.2.1 of the Template). Furthermore, private datasets not commercially licensed by rightsholders and obtained from other third parties have to be listed only if publicly known (or the provider wants to make them publicly known), and otherwise described in a general manner (see Section 2.2.2 of the Template). Considering the public nature of the information contained in publicly available datasets, more detail is required about those datasets, including the disclosure of 'large' datasets (defined in the Template), in line with Recital 107 AI Act (see Section 2.1 of the Template).

(20) For data scraped from online sources, the Template requires disclosure of relevant information such as the crawlers used, their purpose and behaviour, the period of collection and a comprehensive description of the type of content and online sources scraped (see Section 2.3 of the Template). In addition, the Template requires disclosure of a summary list of most relevant domain names crawled and scraped from online sources by or on behalf of the provider in a summarized narrative form[19], in so far as their content has been used for the training of the general-purpose AI model. Such a summary of the domain names scraped from the internet aims to provide a meaningful information about the most relevant top domain names scraped, while striking a balance with the trade secrets ensuring the Summary remains non-technical and in a summarised narrative form, as required by Recital 107 AI Act.

(21) Furthermore, the Template (see Section 2.4) requires minimal information about user data collected through user interactions with all services and products of the provider, including interactions with the providers' AI models. This category excludes data licensed by users based on commercial transactional agreements already covered under Section 2.2.1, or customer data used for fine-tuning a model for specific purposes. For synthetic data generated by AI model(s) used for training purposes and in particular for model distillation (see Section 2.5 of the Template), the information is also limited to names of the general-purpose AI model(s) used if those models have been placed on the market or, if other AI models have been used, including models owned by the providers, information about the model (including a general description of the model training data if known and in so far as this may be needed for the exercise of the rights of parties with legitimate inetersts and to avoid circumvention of the disclosure obligations in the other Sections of the Template).

[19] See footnote 16 above and Section 2.3. of the Template.

(22) The Template does not require disclosure of the exact mix and composition of data sources, but only high-level information about the training data size per modality (selection amongst three very broad ranges) and aggregated across all sources[20] (see Section 1.2 of the Template).

5. Simple, uniform and effective reporting

(23) The information requested by the Template is to be provided in a narrative, simple and effective form. The Template aims to ensure the reported information is useful and understandable to the public and to the parties concerned, while avoiding unnecessary burden on providers of general-purpose AI models, including SMEs.

(24) Each Section of the Template includes clear and short instructions to allow providers to report the required information in an easy and uniform manner. The Commission aims to provide the Template as an online form and to publish it on its website.

(25) Providers should ensure that the information included in the Summary is reported in good faith and in an accurate and comprehensive manner. Flexibility is provided under specific sections, as indicated in the Template, to disclose only information that is relevant, necessary for the purpose of the Summary, and practicable to obtain (e.g. regarding the categorisation of some of the content or the characteristics of the training data, or the period of data collection).

(26) The AI Office may verify whether the Template has been filled in correctly in order to assess if the provider has complied with Article 53(1)(d) AI Act. In this context, the AI Office has all enforcement powers under the AI Act and and may request corrective measures. Non-compliance may be sanctioned with fines of up to 3% of the provider's annual total worldwide turnover in the preceding financial year or EUR 15 000 000, whichever is higher. 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). The AI Office will supervise the implementation of the obligation to provide a compliant summary under Article 53(1)(d) AI Act without performing a work-by-work assessment or checks whether specific content has been used or not for the training of the general-purpose AI model (Recital 108 AI Act).

(27) In case of disputes, providers and parties with a legitimate interest, including rightsholders, are encouraged to use alternative dispute resolution mechanisms available at national level (such as mediation) and other available remedies provided for by Union and national law (e.g. under Article 8 of the Intellectual Property Rights Enforcement Directive)[21].

6. Modifications of existing general-purpose AI models and updates

(28) An existing general-purpose AI model already placed on the Union market may be modified by a downstream entity in such a way that the downstream entity becomes the provider of the resulting general-purpose AI model, as specified in the Commission guidelines on General-Purpose AI models[22]. In such cases, the information reported by the modifying entity in the Template should be limited to the training content used for the model modification only[23], and the name of the model(s) that was modified should be clearly indicated in the Summary (see Section 1.2 of the Template).

(29) The Summary should also be updated whenever the provider further trains its own general-purpose AI model placed on the market on additional data that requires an update of the content of the Summary. In those cases, the Summary should be updated at six-month intervanls or if in the meantime the additional

[20] Ref to ECJ case-law aligned with this approach.
[21] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (OJ L 157, 30.4.2004), OJ L 195, 2.6.2004, p. 16–25.
[22] Guidelines for providers of general-purpose AI models | Shaping Europe's digital future
[23] See also Recital 107 AI Act.

data used to further train the model requires a materially significant update of the content of the Summary, whichever is sooner. In such cases, the Summary should be updated to reflect this additional data, as well as the date of the update. The updated Summary should be made publicly available alongside the modified model.

(30) The same Summary may be used for different models or different model versions if the content of their respective Summaries is identical. In this case, the Summary should clearly specify the different models and model version to which it applies. In addition, if different models or model versions are based on the same general-purpose AI model that has already been placed on the Union market, and the Summaries for each model and model version are different so that they cannot be covered by a single Summary, the Summaries for each of those models or model versions only need to cover the training data specifically used to further modify (including fine-tune) them out of the original model. In this case, a clear reference should be made to the original model in the Summary for each relevant model or model version, and a link to the Summary of the original model included (see Section 1.2).

(31) Where the same Summary is used for multiple models or model versions, in accordance with point (30), reference in the template to the 'model' should be understood as a reference to each model or model version covered by the Summary. Reference to 'training data' should be understood as a reference to the training data for each of these models or model versions.

7. Publication of the Summary

(32) The Summary should be made publicly available at the latest when the model is placed on the Union market. It should be published on the provider's official website in a clearly visible and accessible manner, clearly indicating which model(s) (and possibly model version(s)) the Summary covers subject to the conditions specified in point (30) above. The Summary should also be made publicly available together with the model across all its public distribution channels (e.g. online platforms).

8. Entry into application of the obligation and special rules for models placed on the market before 2 August 2025

(33) The obligation for making the Summaries publicly available becomes applicable as of 2 August 2025. For models placed on the market before 2 August 2025, providers should take the necessary steps to make the corresponding Summary publicly available **no later than 2 August 2027.** Where a provider of a model placed on the market before 2 August 2025 cannot, despite their best efforts, provide parts of the information required to prepare the Summary because the information is not available or its retrieval would impose a disproportionate burden on the provider, the provider should clearly state and justify the corresponding information gaps in its Summary[24]. The supervision and enforcement by the AI Office for compliance with the rules for general-purpose AI models will start as of 2 August 2026.

9. Review of the Explanatory Notice and the Template

(34) The Commission will monitor the implementation of the Template annexed to this Explanatory Notice and where necessary review the Notice and the Template, in view of practical experience gained and of the pace of technological, societal and market developments in this area. If the Commission deems it necessary, such a review may take place before the entry into application of the enforcement powers of the AI Office on 2 August 2026.

[24] See under each Section of the Template a box for possible additional information (optional).

Annex

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)

Version of the Summary: *Version of the summary, with link(s) to previous versions where applicable*

Last update: *DD/MM/YY*

1. General information

1.1. Provider identification

Provider name and contact details:

Authorised representative name and contact details: *Only applicable if the provider is established outside the Union (see Article 54 AI Act).*

1.2. Model identification

Versioned model name(s):	*Provide the unique identifier(s) for the model(s) or model version(s) covered by this Summary (e.g. Llama 3.1-405B). In accordance with point 30 of the Commission Explanatory Notice to the Template, the same Summary may be used for different model(s) or model version(s) provided the content of their respective Summaries is identical. Where available, provide link(s) to additional publicly available documentation, such as the model card, for the model(s) or model version(s).*
Model dependencies:	*If the model is the result of a modification, including fine-tuning, of one or more general-purpose AI models already placed on the Union market, specify the model (version) name(s) of that/those models and provide a link to their Summary(ies) where available.*
Date of placement of the model on the Union market:	*Indicate the date on which the model was placed on the Union market (including the dates each model (version(s)) was placed on the market, if the Summary applies to more than one model or version (see point 30 of the Commission Explanatory Notice to the Template).*

1.3. Modalities, overall training data size and other characteristics

This section requires general information about the overall training data after pre-processing and before the training of the model.

Modality *Select the modalities present in the training data, to the extent that they are identifiable*	**Training data size** *For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.*	**Types of content** *For each selected modality, provide a general description of the type of content that has been included in the training data.*
☐ Text	☐ Less than 1 billion tokens ☐ 1billion to10 trillions tokens ☐ More than 10 trillions tokens Alternatively, specify the approximate size in a different measurement unit: _____	*Examples of possible types of content include fiction and non fiction text, scientific text, press publications, legal and official documents, social media comments, source code.*
☐ Image	☐ Less than 1 million images ☐ 1Million to1 billion images	*Examples of possible types of content include photography, visual art works, infographics, social media images, logos, brands.*

	☐ More than 1 billion images	
☐ Audio[25]	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*Examples of possible types of content include musical compositions and recordings, audiobooks, radio shows and podcasts, private audio communication.*
☐ Video	☐ Less than 10 000 hours ☐ 10 000 to1 million hours ☐ More than 1 million hours	*Examples of possible types of content include music videos, films, TV programmes, performances, video games, video clips, journalistic videos, social media videos.*
☐ Other	*Specify the modality and for each one indicate approximate size and unit of measurement*	

Latest date of data: acquisition/collection for model training:

Indicate the latest date when data was collected/obtained for the model training: MM/YYYY
Additionally, indicate if the model is continuously trained on new or dynamic data after this date.

Description of the linguistic characteristics of the overall training data:

Where applicable, describe the languages covered by the training data (e.g., text, videos or speech), focusing in particular on EU official languages.

Other relevant characteristics of the overall training data:

Where such information is readily available and in so far as it is relevant and practicable, describe other relevant characteristics of the overall training data, such as national/regional or demographic specificities of the training data.

Additional comments (optional):

Providers may also disclose other relevant information on a voluntary basis (e.g. the compression or tokenization methodologies applied for the data size calculation, the sampling frequency/rate plays for audio or video content).

2. List of data sources

This section requires information about specific sources of data used to train the general-purpose AI model. In this section "dataset" should be understood as a single, pre-packaged collection of data. The filtering and pre-processing of data collected from the same pre-packaged collection should not be considered a new dataset to be disclosed separately in the sections below. If a particular dataset can be assigned to more than one of the categories below, providers should select the most relevant category and only report the dataset in that category, except in the case of synthetic data (see Section 2.5).

2.1. Publicly available datasets

This section requires information about datasets that were used to train the model and which have been compiled by a third party, are made available publicly for free, and are readily downloadable as a whole or in predefined chunks, such as datasets and collections available on public repositories and online platforms, specialised websites, or snapshots of common crawl. The public availability of the datasets for free does not mean that the content at issue is necessarily free of rights since it may be subject to licensing arrangements or conditions of use (e.g., certain free and/or open licenses may determine the scope of the uses, including prohibiting uses relating to model training).

A dataset is considered to be "large" if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. The size of the dataset should be based on its size after pre-processing (for example filtering), and without splitting the dataset to prevent reporting circumvention.

Have you used publicly available datasets totrain the model? ☐ Yes ☐ No

[25] Excluding audio that is part of video, as this should be reported under the "video" modality instead. Furthermore, the Commission understands the modality of 'audio' to include 'speech'.

If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
List of large publicly available datasets:	*For each large dataset, provide the identifier/name of the dataset and a link through which the dataset can be accessed. If a link is not available, provide a general description of the dataset, including the approximate start and end dates of the data collection if known (otherwise indicate "not known"). If only part of the datasets has been used for the training, indicate the general approach to selecting those parts.*
General description of other publicly available datasets not listed above:	*For other publicly available datasets that are not listed above, provide a general description of their content. The description could include indication of: (i) the types of modality (e.g. text, images), (ii) nature of the content (e.g. personal data, copyright protected content, machine generated data such as Internet of Things or synthetic data), (iii) its linguistic characteristics, where applciablem (iv) the approximate start and end dates of the data collection if known (otherwise indicate "not known").*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis, e.g. size of the datasets and other relevant details.*

2.2. Private non-publicly available datasets obtained from third parties

This section requires information about private non-publicly available datasets of third parties that are not publicly available and not disclosed under Section 2.1. These include:

1) *datasets for which transactional commercial licensing agreements were concluded between the provider and the rightsholders or their representatives, including by collective management organisations and legitimate content aggregators who have the right to collectively license works on behalf of rightsholders (Section 2.2.1);*
2) *other private datasets obtained through data intermediaries, non-publicly available databases and datasets of third parties for which transactional commercial licenses have not been concluded with rightsholders or their representatives (Section 2.2.2).*

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?	☐ Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)

2.2.2. Private datasets obtained from other third parties

Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?	☐ Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
If publicly known, list private datasets obtained from other third parties:	*If publicly known, list the identifiers/names of the main private datasets from third parties that are not licensed as described in Section 2.2.1 and that are used to train the model, and provide links to relevant information, where available.*

General description of non-publicly known private datasets obtained from third parties	*For those private datasets used to train the model that are not publicly known and whose identifiers are not listed above, provide a general description of their content. The description should indicate (i) the modalities (e.g,. text, images), (ii) nature of the content (e.g., personal data, copyright protected content, machine generated data such as Internet of Things or synthetic data) and (iii) its linguistic characteristics, where applicable.*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis, e.g. the period of data collection, size of the datasets and further details.*

2.3. Data crawled and scraped from online sources

This section requires information about crawled, scraped data, orotherwise compiled from online sources directly by the provider of the model or on their behalf (i.e. excluding publicly available datasets already compiled by third parties and made available on platforms such as common crawl that are covered under Section 2.1).

Were crawlers used by the provider or on behalf of?	☐ Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	
Purposes of the crawler(s):	
General description of crawler behaviour:	*For example, this includes respect of captchas, password protected websites and paywalls, respect of robot.txt and other protocols, while crawling.*
Period of data collection:	*From MM/YYYY to MM/YYYY*
Comprehensive description of the type of content and online sources crawled:	*Provide a comprehensive description of the type of content crawled, including its geographical, linguistic or demographic characteristics, as well as an indication of the type of websites scraped (e.g. news, blogs, social media, forums, community websites, other user-generated content platforms, websites of cultural heritage institutions, educational sites, government portals, personal blogs, streaming, gaming platforms, online TV platforms, synthetic data libraries, etc).*
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
Summary of the most relevant domain names crawled:	*In so far as any content from internet domains has been crawled or scraped and used for the training of the model, provide a list of those most relevant internet domains names (top and second-level domain, e.g. "example.com") by listing the top 10 % of all domain names determined by the size of the content scraped (in a representative manner across all modalities where applicable). Small and medium-sized enterprises (SMEs), including start-ups, should disclose top 5% of all domain names or 1 000 internet domain names, whichever is lower. You can provide this list for instance as a downloadable file or provide here information on how to access it.*
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis, for instance more domain names than those required in the list above and/or URLs and the sources of individual works.*

2.4. User data

This section requires information about user data collected by all services and products of the provider, including through mail services, social media platforms, content platforms or interaction with the providers' AI models and/or systems. This does not cover data licensed by users based on commercial transactional agreements described inSection 2.2.1., or customer data to fine-tune models for specific purposes.

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	☐ Yes ☐ No
Was data collected from user interactions with the orovider's other services or products used to train the model?	☐ Yes ☐ No
If yes, provide a general description of the provider's services or products that were used to collect the user data:	
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
Additional comments (optional):	*Providers may also disclose other relevant information on a voluntary basis.*

2.5. Synthetic data

This Section requires information about synthetic data created by or on behalf of the provider for training the model directly on the outputs of another AI model, in particular through model distillation or model alignment (e.g. AI feedback through reinforcement learning). This does not include the use of AI models to clean or enrich data (e.g. AI-generated metadata to enrich or modify a dataset, such as creating depth maps or text descriptions of images). In case this concerns publicly available datasets as described in Section 2.1, these should be reported in that Section of the Template. In case this concerns synthetic datasets created by third parties on behalf of the provider, these should be reported in this Section of the Template instead of in Section 2.2.2.

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	☐ Yes ☐ No
If yes, modality of the synthetic data:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other (please specify)
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	*Specify the name of the general-purpose AI model(s) and provide a link to their Summary(ies) where available.*
Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic data to train the model to which this Summary applies:	*Provide information about other AI models used to generate synthetic data, including provider's own models if not available on the market. This includes a general description of the model training data if known and in so far as this may be needed for the exercise of the rights of parties with legitimate inetersts and to avoid circumvention of the disclosure obligations in the other Sections of the Template.*
Additional comments (optional):	*Providers may also disclose on a voluntary basis other relevant information.*

2.6. Other sources of data

This Section requires information about data that does not fall under any of the categories in the previous Sections, for example data collected from offline sources, self-digitised media (e.g., digitised analog text context, images), datasets labelled by humans commissioned by the provider, or human generated data through reinforcement learning.

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☐ Yes ☐ No

If yes, provide a narrative description of these data sources and the data:

Additional comments (optional):

Providers may also disclose other relevant information on a voluntary basis.

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

This Section concerns measures implemented by the provider to identify and comply with the reservation of rights from the text and data mining (TDM) exception or limitation expressed pursuant to Article 4(3) of Directive (EU) 2019/790, as outlined in the copyright policy put in place by the provider in accordance with Article 53(1)(c) AI Act.

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?

☐ Yes ☐ No

Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:

Additional comments (optional):

Providers may also disclose other relevant information on a voluntary basis. Providers are also encouraged to disclose a summary of their copyright policy under Article 53(1)(c) AI Act, if made publicly available (e.g., by providing a link to the relevant web-page).

3.2. Removal of illegal content

This Section concerns measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers), without requiring disclosure of specific details about the provider's internal business practices or trade secrets. Such measures are advisable if the training data is likely to include illegal or unlawful content under Union law, in particular child sexual abuse material and terrorist content and the non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices, for example to increase the capability of the model.

General description of measures taken:

3.3. Other information (optional)

Other relevant information about data processing (optional):

Providers are also encouraged to disclose on a voluntary basis other relevant information about relevant data processing aspects and measures taken before or after the training of the model that is relevant for the respect and exercise of rights protected under Union law.