Blueprint of the template for the summary of content used to train general-purpose AI models (Article 53(1)d AIA) – v.2.0[1]

Preface

This blueprint for the template for the AI Act's training data transparency requirement for general-purpose AI models is based on the initial version presented in the policy brief *"Sufficiently detailed"? A proposal for implementing the AI Act's training data transparency requirement for GPAI models*, published in June 2024. The second version of the blueprint presented in this document and published in September 2024 has been further developed and refined based on feedback from various stakeholders and a workshop conducted with experts from industry, academia, and civil society.

Blueprint

1. Overall size

1.1. The total size of the training data (e.g., the total number of works and tokens, including information about the tokenization process for better comparability), the total number of data sets, and their size in GB.

2. Data sets and data sources

2.1. List of all data sets that were used in training the model, including the following information for each data set:

2.1.1. Data set identifier: the data set's name, along with, if applicable, a link to the data set, as well as the name and URL for the data collection of which the data set is a part.

2.1.2. Dataset size: The total size of the data set and information about what percentage of the total training data the data set represents, using an appropriate measurement methodology for consistency and comparability.

2.1.3. Owners or curators: groups or teams that own or compile the data set; if applicable, authors associated with the dataset.

2.1.4. Data set domain: humans (e.g., data set about people), objects (e.g., data set about places or objects), etc.

2.1.5. Type/modality: text, image, audio, video, code, etc. If multimodal, what combination (e.g., image-text pairs).

2.1.6. If language data, what language/languages and, where feasible, other relevant characteristics (such as, for example, slang, dialects, usage contexts, etc.).

[1] The blueprint was developed by Zuzanna Warso (Open Future), Maximilian Gahntz (Mozilla Foundation), and Paul Keller (Open Future). The authors would like to thank Yacine Jernite, Lucie-Aimée Kaffee, Shayne Longpre, Abeba Birhane, Stefan Baack, Kris Shrishak, Aviya Skowron, Mark Dingemanse, Frederike Kaltheuner, Andreas Liesenfeld, Claire Pershan, Michał "rysiek" Woźniak, Natali Helberger, João Pedro Quintais, Toni Lorente, and Rania Wazir for their valuable feedback on this blueprint.

2.1.7. If applicable, information about the specific purpose for which the data set was used in the training process (e.g., for instruction tuning, multilingual capabilities, etc.).

2.2. For each data set (including, if applicable, the constituent data sets of a composite data set), information on the sources from which data was obtained – i.e., whether it was:

- scraped from the internet. If so, information about the crawling methodology and names/identifiers of the crawlers used to obtain the data, including, for example, the seed and link selection criteria, as well a weighted list of the top 5 percent or 100 000 domains by data modality (e.g., text, images, video).
- collected from public repositories. If so, information on the source and, where applicable, the license should be provided.
- copyright-protected content licensed from rightholders or third-party intermediaries. If so, information about the source and the license, including whether the licensing arrangement is exclusive or not, should be provided.
- obtained from proprietary databases. If so, information about the source and, where applicable, the license, including whether the licensing arrangement is exclusive or not, should be provided.
- generated by users of products or services offered by the provider. If so, by which products or services.
- synthetically generated by the provider. If so, by what methods.
- generated by the provider by other means. If so, by what methods (e.g., crowdsourced or in a lab setting).
- or obtained through other means. If so, details of the means should be provided.

2.3. For each data source, the date range of the training data. This should include data cutoffs for online data and, if applicable, for other data sources such as archival data. For synthetic data, information about the time/date range of generation should be provided.

2.4. For each data source, where applicable, information on mechanisms and policies that have been implemented to comply with Union law on copyright and related rights, including respect for opt-outs under Article 4 of the CDSM.

2.5. If applicable, information on the legal basis for the processing of personal data and measures to protect data subjects' rights under the GDPR.

3. Data diversity

3.1. Specific information on steps taken to ensure diversity and representativeness of training data across relevant categories (e.g., demographics, languages).

3.2. Where feasible, for each data set and, where feasible, in the aggregate, estimates of proportions of relevant categories and characteristics (such as linguistic or regional diversity) included in the training data.

4. Data processing

4.1. Description of (pre-)processing steps applied to the different types of training data (text, image, speech, etc.), such as filtering, data cleaning, tokenization, feature extraction, detoxifying, etc.

This should include information on the:

- filtering processes and methods used for inclusion in a dataset (e.g., metrics or cosine similarities used to decide the cut-off point for what content makes it into a data set), with a sufficient level of detail to understand the motivations behind using these methods and how they have been employed; and

- filtering processes and methods used for removing data from a dataset (e.g., outliers, duplicates, potentially harmful content), with a sufficient level of detail to understand the motivations behind using these methods and how they have been employed.

4.2. Information on anonymization techniques that were implemented. When personal data is not anonymized, a justification for not anonymizing and information on how long the data is kept before being deleted should be provided.

4.3. Explanation of how (pre-)processing may have affected the characteristics of the data used for training, e.g., how it might have affected feature representation, class imbalance, etc.

4.4. If applicable, data sampling: Information on sampling methods used to select training data from a larger dataset (e.g., random sampling, systematic sampling, stratified sampling, under- or over-sampling, or other techniques).

4.5. Information on the methodology and processes used for annotation and labeling (e.g., crowdsourcing, supervised learning, etc.).

4.6. If crowdsourcing was used, information about the recruiting criteria of the annotators, and the specific task they were given.

4.7. Explanation of measures and evaluation methods to ensure reliability and accuracy of annotations and labeling.