Summary of training data

Domyn Large

Version of the Summary: v1 - Last update: 19 March 2026

Table of contents

This page outlines the summary of the training data used for Domyn Large, in accordance with the requirements set out in Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act). The purpose of this summary is to enhance transparency and enable stakeholders to better understand the nature of the training data. This summary has been prepared following the template issued by the AI Office and reflects our commitment to complying with the AI Act's obligations while promoting responsible and transparent AI development.

1. General Information

Model Dependencies: The model is a modification of a general-purpose AI model already placed on the Union market, specifically Colosseum 355B

1.2 Date of Placement on the Market and Knowledge Cut-off Date

Date of placement on the Union market: 20 March 2026

Knowledge cut-off date: September 2024 (based on pre-training dataset cut-off)

1.3 Overall Training Data Size, Modalities and Characteristics

Data Size by Modality

Modality	Overall Size	Status
Text	Number of tokens or bytes: 11.1 Trillions	Used
Image	Number of images (or pairs with other media): 0	Not Used
Video	Number of minutes (or pairs with other media): 0	Not Used

Other	pairs with other media): 0	Used

Content Categories by Modality

Text

- Fictional texts, literature
- Scientific and educative texts
- News, journalism and opinions
- Legal and official documents
- Social communication (e.g., messages)
- Promotion, advertising, product and service reviews
- Other text

Image

- Photography
- Paintings & fine-arts
- Infographics
- Illustration & graphic design
- Social / personal images

Special

- Source code
- Structured data (e.g., calendar, maps)

performancesy	
Animated video content	User content, short videos
Documentaries	Other video content (e.g., experimental art, video effects)

Audio

Music	Radio shows and podcasts
Narrative and fiction (e.g., audiobooks)	Social communication (phone calls, voice messages)
Non-fiction educative audio content	Other (e.g., sounds and ambient)

Description of linguistic, regional, demographic and other relevant characteristics

Linguistic characteristics:

Primary language: English (majority of pre-training and CPT data)

Multilingual coverage: 56 languages extracted from FineWeb2-HQ and HPLT, with tiered weighting.

Post-training multilingual: SFT data includes dedicated multilingual splits (Spanish, French, German, Italian etc.)

Code languages: Python, C++, Java, JavaScript, SQL,

web content.

Multilingual data intentionally upweights European languages (Tier A: ES, IT, FR, DE) to align with Domyn's regulated-enterprise customer base in European markets.

Academic data (ArXiv, peS2o) reflects global English-language research output.

Wikipedia dumps cover all available language editions, with representation proportional to each Wikipedia's size.

Demographic characteristics:

No explicit demographic annotation or sampling was applied to the training data.

Web-crawled data inherits the demographic biases of internet content production — disproportionately represents populations with higher internet access, literacy, and digital participation.

Academic corpora reflect the demographics of academic publishing (predominantly English-speaking, university-affiliated researchers).

Other relevant characteristics:

Domain coverage: General web, code, academic/STEM, mathematics, Wikipedia/encyclopedic, function calling/tool use.

Temporal range: Web crawls primarily from 2023–2024; academic and code corpora reflect historical archives.

2. List of Data

publicly available datasets.

Main large publicly available datasets used for pretraining:

1. DCML (September 2024)
2. The Stack v2 (September 2024)
3. HPLT (September 2024)
4. HuggingFaceFW fineweb-2 (September 2025)

General description of other publicly available datasets

The remaining training data comprises approximately 5–6 trillion tokens in total, drawn from the following categories:

1. **Web crawls** — Text-only, multilingual web data sourced primarily from DCML.
2. **Code** — Machine-readable source code derived from The Stack v2.
3. **Academic** — Text-only scientific and mathematical content, including ProofPile 2 (arXiv).
4. **Wiki** — Text-only encyclopaedic content extracted from Wikipedia dumps.
5. **Multilingual** — Text-only web crawls in multiple languages, drawn mainly from HPLT and FineWeb-2.
6. **Synthetic** — Machine-generated text comprising more than five datasets of QA-style content, produced from high-quality web documents.

2.2 Private non-publicly available datasets obtained from third parties

We have not crawled, scraped, or otherwise directly compiled data from online sources ourselves or through third parties on our behalf (excluding publicly available third-party datasets covered under 2.1. above).

2.4 User data

No user data collected by our services and products, including through mail services, social media platforms, content platforms or interaction with our' AI models and/or systems was used to train the Model.

2.5 Synthetic Data

A portion of the training data consists of synthetic AI-generated data created by us or on our behalf.

Modality of the synthetic data: Text only

Name of AI model **used to generate the synthetic data:**

1. Colosseum 355B
2. Phi 3 Medium
3. Llama-3_1-Nemotron-Ultra-253B-v1

2.6 Other sources of data

No data falling outside the categories described in the previous sections was used for training.

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

We are a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the text and data mining TDM exception or limitation

Measures implemented to respect reservations of rights from the text and data-mining exception under Art.4(3) DSM Directive during data collection:

1. **Specification of opt-out protocols:** All open dataset have used web crawlers that honor machine-readable opt-out signals—such as robots.txt and standard metadata
2. **Solutions honoured by the provider:** A public feedback channel is maintained for rights-holders to request removal. Updates are applied dynamically to domain-/URL-level blocks based on requests.
3. **Measures implemented after data collection is completed to identify and remove content for which rights have been reserved by the rightsholders:** URL/domain-based blocking of flagged sources, ML classifiers trained to detect protected content, Manual review and removal triggered by rights-holder notifications.

3.2 Removal of Unwanted Content

Description of content deemed unwanted by the provider as part of the training data:

such content:

Blacklists: Dynamic domain/URL blacklists maintained based on rights-holder input or legal considerations. Also, We leverage the publicly maintained Université Toulouse 1 Capitole URL blacklists—the UT1 blacklists—for filtering undesirable domains.

Keywords: We maintain a curated list of sensitive keywords and phrases used to identify and filter out unwanted content, including but not limited to:

1. Copyright-related markers (e.g. "©", "All rights reserved", "Unauthorized reproduction")
2. Adult content indicators ("porn", "hard pornography", "adult site", "XXX")
3. Hate speech slurs or explicit offensive terms
4. calls for violence or extremist content
5. gambling-related terms ("betting", "casino", "gambling site")
6. malware/phishing language ("download here", "free crack", "keygen")

These keywords are matched both in URLs and in page content; any document exceeding a threshold score is excluded.

List of measures taken to avoid and/or remove such content:

Model-based classifiers: We curate both positive and negative examples (e.g., unwanted vs. acceptable content) as well as general classifiers for high-quality data, often using authoritative sources such as Wikipedia to define positive examples. We have found that combining multiple classifiers yields strong performance with a low error rate. For training, we primarily use transformer-based models such as BERT

reliably estimate real-world error rates and ensure robust generalization.

Other measures: Manual review upon flagging.

Measures applied by the curators of listed datasets:

1. DCML: Curated source domains with known copyright compliance. They employ a combination of model-based filtering and heuristic methods to curate content, aiming to exclude material that may be subject to copyright reservations.
2. The Stack v2: Mainly includes code under open-source licenses. To facilitate compliance with licensing requirements, The Stack v2 dataset provides provenance information for each data point, allowing users to trace the origin and licensing terms of the included code.
3. ProofPile 2 (arXiv): Includes academic content with explicit author/reviewer permissions.
4. HPLT / multilingual web crawls: Apply open-license filtering and language-based exclusion.
5. Synthetic QA datasets: Generated from high-quality web documents cleared for reuse; measures to respect rights reservations respected.

Agents Platform Models Compute

Contact Us

Governance Company Resources

Legal address
Piazza Gae Aulenti, 8, 20154
Milano MI

North America HQ
115 Broadway, New York, NY
10006

VAT number: 08523180969

AI Sentinel

Technology

Models

Supercomputer

Board

Careers

Gender Equality Plan

Resources

Blog

News

Research

Media Kit

Cookies

Software Licer

Disclaimer

Compliance

Whistleblowin