Colossyan
NEO 2

*Context: This combined template is designed to satisfy both the **EU AI Act (Article 53)** and **California AB 2013**. I have merged them so you only have to maintain one public disclosure page.*
*The instructions in italics are based on the final **EU AI Office Template** (released July 2025) and the statutory language of **AB 2013**.*

AI Training Data Transparency Report

Model Name: NEO
Version: 2
Date of Publication: 18-12-2025

Section 1: General Model Overview
• **Model Provider:** Colossyan
• **Intended Purpose:** The purpose of the model is to create realistic AI avatars that can dictate scripts so that customers can incorporate them into videos they create with the Colossyan platform. This model also allows for hand movements, body movements and facial expressions. The datasets described below were curated specifically to improve the model's performance in temporal consistency and high-fidelity lip-syncing across multiple languages.
• **Date of First Use (California Mandate):** 18-12-2025
• **Knowledge Cut-off Date:** 01-11-2025

Section 2: Dataset Composition (The "High-Level Summary")

Data Category	Description & Modality	Data Point Count (Range)	Source / Ownership
Licensed Data Sets	Videos of individuals speaking to the	~ 800 hours	Licensed by a third party who

	camera, generally front facing		gave Colossyan with full ownership rights.
Colossyan-procured data sets	Videos of individuals speaking to the camera, generally front facing	~ 50 hours	Captured by Colossyan staff during data collection exercises. Consent forms were collected.

Section 3: Legal & Privacy Status

- **Intellectual Property Status:** Data sets were licensed to Colossyan through third parties or through the data subjects themselves. IP ownership in the recordings including derivative works was transferred to Colossyan.
- **Personal Information (PI) Disclosure:** Yes. All personal information was obtained in accordance with data protection laws.
- **Aggregate Consumer Information:** No
- **Collection Timeframe:** 01-01-205 to 01-11-2025

Section 4: Data Processing & Ethics

- **Use of Synthetic Data:** No

    The dataset is filtered using a data pipeline that is developed in-house. The pipeline employs automated filters that discard the section of the videos that do not meet the required criteria:

    - Age Filter.
    - NSFW filter.
    - Noise/Blur filter.
    - Face to image density filter.
    - Face oclusion filter.
    - Motion blur filter
    - Minimum duration of single ID filter

Using the above filters, the pipeline cleans the data and discards the samples that do not meet the requirement. And only the filtered output data is then used for the model training.

- **Copyright Opt-Out Compliance (EU Article 4 CDSM):** The training datasets for this model were acquired exclusively through direct licensing and proprietary capture. As this model was not trained on data crawled or scraped from the public web, the automated 'opt-out' mechanisms (such as robots.txt or machine-readable TDM reservations) under Article 4(3) of Directive (EU) 2019/790 are not applicable to the acquisition of this dataset.