Training Content Protocol

Version 1.0

Table of Contents

1. Introduction

Within the scope of the GPT-NL project, a Dutch Large Language Model ("**GPT-NL Model**") will be created by TNO. The intention is to train the GPT-NL Model based on (i) high-quality content (free of harmful and irrelevant content), (ii) for which a valid license has been obtained (explicitly via a contract or by using content that is published under a permissive open-source license) and (iii) by ensuring that the content is prepared in a privacy-preserving manner. To achieve this, TNO has contacted Content Contributors (as defined below) that license content to TNO for the training and future updating of the GPT-NL Model.

Content Contributors each select the training content that they want to contribute to the GPT-NL project. Before this content can be used for training of the GPT-NL Model, it must be cleaned in accordance with this Training Content Protocol. TNO will follow the same requirements when preparing open-source content collected by TNO.

This Training Content Protocol provides a holistic overview for the data collection and data curation process and links to all relevant documents that further detail these processes. This document describes the steps from first contact between the GPT-NL team and the Content Contributor, to the data curation itself, and the final step: the moment in time when the training content is securely in the data storage cluster of the GPT-NL project to be used by TNO for the training and future updating of the GPT-NL Model.

This Training Content Protocol is subject to change management and version numbers and date of amendments shall be documented. The Training Content Protocol can be amended in accordance with the GPT-NL Governance Charter of the GPT-NL project only.

If you have any questions about this document, please contact the GPT-NL Lead Data consultant, Frank Brinkkemper (frank.brinkkemper@surf.nl).

2. Definitions

For the purpose of understanding this document, we use the following definitions and law references.

Content Contributors: the parties that provide content for the training and future updating of the GPT-NL Model;

Content Preparation Process: the process applied to prepare Raw Content in accordance with this Training Content Protocol, turning it into Contributor Training Content, evaluating such Contributor Training Content and combining it with TNO Training Content into the Prepared Dataset;

Contributor Training Content: Raw Content that has been prepared in accordance with the Training Content Protocol;

Data Curation Process: the process applied by Content Contributor to prepare Raw Content and by TNO to curate such Raw Content and turn it into Contributor Training Content, in accordance with the Data Curation Specifications;

Data Curation Specifications: the specifications of the Data Curation Process included in **Annex 1**;

Data Evaluation Process: the evaluation process applied by TNO to evaluate Contributor Training Content as described in **Annex 2**;

GPT-NL Governance Charter: the charter that sets out the functions, roles, and responsibilities assigned to the various participants in the GPT-NL project and the process for amending the GPT-NL project documentation;

Harmful Information: directly hurtful or offensive language, such as:

1. Violent, criminal, or unlawful content;
2. Biased or discriminatory content, hate speech, or other content hostile to individuals or groups; or
3. Fake, manipulated, or inaccurate content;

Non-Public Persons: all individuals who are not Public Persons;

Non-Suitable Data: Sensitive Personal Data, Personal Data of Non-Public Persons, Unsuitable Source Data, Protected Information, and Harmful Information;

Personal Data: any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

Prepared Dataset: means any data resulting from modifying, combining, adapting, merging or aggregating (wholly or in part) the Contributor Training Content and TNO Training Content or portions thereof for purposes of training and future updating of the GPT-NL Model;

Protected Information: any information that a Content Contributor cannot freely share, such as, know-how and other confidential business information;

Public Persons: individuals who have a public presence and have lower expectation of privacy in the capacity of the role they fulfil. For instance, a minister, a scholar, a judge in a legal proceeding, professional athletes, famous artists, or high-ranking civil servants, as further specified in **Annex 1**;

Raw Content: means the content that Content Contributor provides to TNO for preparation in accordance with the Training Content Protocol for the purpose of training and future updating of the GPT-NL Model, which has been scrubbed by Content Contributor of any content that cannot be used for training purposes, including those datasets as identified by the TNO during the Viability Assessment in accordance with the Training Content Protocol;

Revenue Sharing Mechanism: the mechanism for compensating Content Providers for providing Training Content for the purposes of training and future updating of the GPT-NL Model, including by sharing in future net revenues of professional licenses to the GPT-NL Model;

Sensitive Personal Data: the categories of Personal Data which have a format and are of a sensitive nature, such as government-issued IDs (e.g., BSN or passport number), credit card data as listed in **Annex 1**;

TNO Training Content: publicly available content that is not subject to copyright (including where the copyright has expired) or is subject to an Open-Source License that has been collected or received by TNO and has been curated in accordance with the Data Curation Process; and

Unsuitable Source Data: data collected from sources that are unsuitable for training or future updating of the GPT-NL Model, such as, gossip magazines, or social media data.

3. Data Collection Actors

The various steps in the Content Preparation Process require different expertise. During the Content Preparation Process, questions need to be answered about the value of a specific dataset and the technical know-how of processing and filtering the dataset. Table 1 describes the actors involved during the Content Preparation Process.

Role	**Function**
GPT-NL Team	The team that works on the overall realization of the GPT-NL Model with members from organizations TNO, SURF, and NFI.
Data Consultant	First point of contact between the GPT-NL Team and the Content Contributor. The Content Contributor can contact their Data Consultant, or they can contact the team of Data Consultants via info@gpt-nl.nl.
Data Viability Officer	Evaluator of the viability of a specific dataset based on the quality and quantity of the dataset.
Data Agreement Officer	Responsible for making an agreement with the Content Contributor based on the general Term Sheet, the Governance Charter, and the evaluation of the Data Viability Officer. For communication, we use the same team as for initial contact with the Content Contributors (where possible, the Data Consultant will also take on the role of the Data Agreement Officer).
Data Curation Staff	Responsible for filtering Non-Suitable Data from the Raw Content.
Data Evaluation Officer	Evaluator of the (curated) dataset considering aspects such as Non-Suitable Data.
Data Transfer and Security Officer	First point-of-contact for questions about transmission of data to TNO/SURF and security of the data.
Content Board Moderator	Informs Content Contributors about the specifics of the Content Board and moderates live sessions between members of the Content Board.

Table 1: Actors in the Data Collection Process.

4. Overview of the Content Preparation Process

The Content Preparation Process involves the following steps:

I. **Viability Survey & Assessment:** The Content Preparation Process starts with a Content Contributor indicating interest in participating in the GPT-NL project and identifying Raw Content that it would like to contribute. The Content Contributor is first requested to complete a Viability Survey, which provides TNO with the information necessary to determine whether the Raw Content is viable content for training and future updating of the GPT-NL Model.

II. **Term Sheet:** When TNO concludes that the Content Contributor's Raw Content is viable for training and future updating of the GPT-NL Model, a Term Sheet is concluded between the Content Contributor and TNO that sets out the principles of the Content Contributor's participation in the GPT-NL project and outlines the rights and obligations.

III. **Deep Dive:** After the Term Sheet is concluded, the Content Contributor is requested to complete a Deep Dive Questionnaire that includes detailed information about the Raw Content identified by the Content Contributor.

IV. **Data Curation Process:**
    1. **Preparation Raw Content by Content Contributor**. Before delivering the Raw Content to TNO, Content Contributor will scrub any content from its datasets that cannot be used for training purposes (such as Protected Information and Unsuitable Sources), including those as identified by the TNO during the Viability Assessment.
    2. **Preparation of Contributor Training Content by TNO**. TNO prepares the Raw Content in accordance with the Data Curation Specifications included in **Annex 1** to this Training Content Protocol.

V. **Data Evaluation Process:** TNO evaluates the Contributor Training Content through a risk analysis and determines the risk of using specific data (sub)sets for training GPT-NL.

VI. **Combining of all Training Content**: TNO combines the Contributor Training Content with the TNO Training Content into a single Prepared Dataset.

Each step of the Content Preparation Process is described in more detail below.

I. Viability Survey & Assessment

The purpose of the Viability Survey & Assessment is for TNO to get an idea about the content included in the Raw Content and to evaluate the type and amount of Non-Suitable Data in such Raw Content compared to the value of the total Raw Content. This is important to do as soon as possible before acquiring new Raw Content because:

1. The GPT-NL Team needs to know whether the current Content Preparation Process is fit for removing the Non-Suitable Data, and if not, whether it is possible to amend the Content Preparation Process.
2. The GPT-NL Team needs to assess whether the value of the Raw Content for training of the GPT-NL Model warrants the time and resources required to conduct the Content Preparation Process.

Viability Survey

Standard way of working

A Data Consultant will be assigned to every Content Contributor as their first point of contact during the Content Preparation Process. The Viability Survey is completed by the Content Contributor, possibly with assistance of the Data Consultant. The link to this survey can also be found on the GPT-NL website (https://gpt-nl.nl/samenwerken).

If a Content Contributor wants to contribute multiple sets of Raw Content, this can be done by completing a separate Viability Survey for each set of Raw Content. Questions can be left empty if the Content Contributor does not know the answer. The Content Contributor is invited to explain in the survey why such questions are not answered. The Data Consultant will contact the Content Contributor about the open questions. It is important that all relevant information about the Raw Content is provided.

Initial Content Contributors

For the initial Content Contributors, the Viability Survey is completed by a Data Consultant based on an interview with a Content Contributor. The interviews will be planned by the Data Consultants. When the process is optimized, the standard way of working will be applied.

Viability Assessment

The evaluation of the Raw Content will be performed by the Data Viability Officer based on the Viability Survey by completing the Viability Assessment. Via this assessment, Data Viability Officers estimate the value and effort/risk of a specific dataset in a structured manner. The Data Viability Officer will communicate the result of the evaluation to the Content Contributor as soon as possible, and the Content Contributor will have the opportunity to comment on the Data Viability Officer's conclusions.

If during the assessment it appears that the Viability Survey did not provide enough information about the Raw Content:

1. The Data Viability Officer will explain the missing information to the relevant Data Consultant. The Data Consultant will then discuss this with the Content Contributor.
2. If certain information in the Viability Survey is not clear (e.g., because of the way the questions are formulated in the survey), the Data Viability Officer will provide this feedback to the Data Consultant. Possibly, the Viability Survey will be amended to improve the questions.

II. Term Sheet

If the Raw Content is deemed viable pursuant to the Viability Assessment, the case progresses to the agreement phase. The Data Agreement Officer will now communicate with the Content Contributor (where possible, this will be the same person who fulfilled the role of Data Consultant). The agreement phase is mostly uniform: the same Term Sheet will be used for everyone.

The steps during the agreement phase will be roughly as follows. First, an email will be sent to the Content Contributor with the next steps. This email will contain:

1. A completed Viability Assessment to confirm with the Content Contributor that the Raw Content is suitable for being used as training data of the GPT-NL Model.

2. The Term Sheet that includes the high-level terms and conditions that pertain to all Raw Content used in the scope of the GPT-NL project, and which will form the basis for the Content Contributor Agreement that will be concluded between the Content Contributor and TNO. The Content Contributor will indicate in the Term Sheet when the completed Deep Dive Questionnaire will be shared with TNO.
3. Instructions on signing the Term Sheet and returning a signed copy to TNO.

III. Deep Dive

Standard way of working

The dataset Deep Dive Questionnaire needs to be completed by the Content Contributor. This questionnaire includes more detailed questions than the Viability Survey, such as, about the origin of the original data and the processing undertaken to create the Raw Content. TNO uses this information to generate meta-data for the Raw Content that is necessary for the GPT-NL Team to collect and publicize as part of the transparency commitments made.

Initial Content Contributors

For the initial Content Contributors, the Deep Dive Questionnaire is completed by a Data Consultant based on an interview with a Content Contributor. The interviews will be planned by the Data Consultants. When the process is optimized, the standard way of working will be applied.

IV. Data Curation Process

a) Contributor Training Content

1. Filtering by Content Contributor

Before Raw Content is ready to enter the Data Curation Process, the Content Contributor needs to scrub its dataset of any content that cannot be used for AI training purposes (e.g., Protected Information, Unsuitable Source Data, content to which the Content Contributor cannot grant a valid license to be used for training purposes), including those datasets as identified by the GPT-NL Team during the Viability Assessment.

2. Filtering by TNO

When Content Contributors have completed step 1), the Content Contributor will provide the Raw Content to the dedicated Data Curation Staff. The Data Curation Staff will then prepare the Raw Content in accordance with the Data Curation Process described in this chapter and in **Annex 1** of this Training Content Protocol.

By completing the Data Curation Process, the Raw Content is turned into Contributor Training Content. A copy of the Contributor Training Content is provided by the Data Curation Staff to the Content Contributor and to the staff of Team GPT-NL tasked with training of the GPT-NL Model. The Data Curation Staff will then delete all remaining copies of the Raw Content and the Contributor Training Content. No staff of TNO other than the Data Curation Staff, will have access to the Raw Content of the Content Contributor. Data

Curation Staff may be redeployed to the GPT-NL team tasked with training of the GPT-NL Model, but upon such redeployment, such Data Curation Staff will no longer have access to the Raw Content.

The staff of Team GPT-NL tasked with training of the GPT-NL Model will evaluate the Contributor Training Content, before adding it to the dataset that will be used to train the GPT-NL Model.

The costs associated with the Content Preparation Process will be borne by the GPT-NL project (i.e., TNO).

b) <u>TNO Training Content</u>

The raw content that team GPT-NL acquires from public sources that is not subject to copyright or is subject to a permissive license, goes through the same Data Curation Process, with a few key differences:

- There is no strict separation between Data Curation Staff and GPT-NL training staff for these data sources.
- The curated output of these sources is called "TNO Training Content" and it will be open-sourced.

Data Curation and Evaluation Pipeline

The Data Curation and Evaluation Pipeline is visually represented in **Figure 1**. This process has two main components:

(1) The **Data Curation Process (Figure 2)** is used to make Raw Content ready for training GPT-NL
(2) The **Risk Analysis (Annex 2)** is done by the GPT-NL Data Evaluator to check whether the Pipeline acted as expected and to judge the risk of including a dataset in the Prepared dataset.

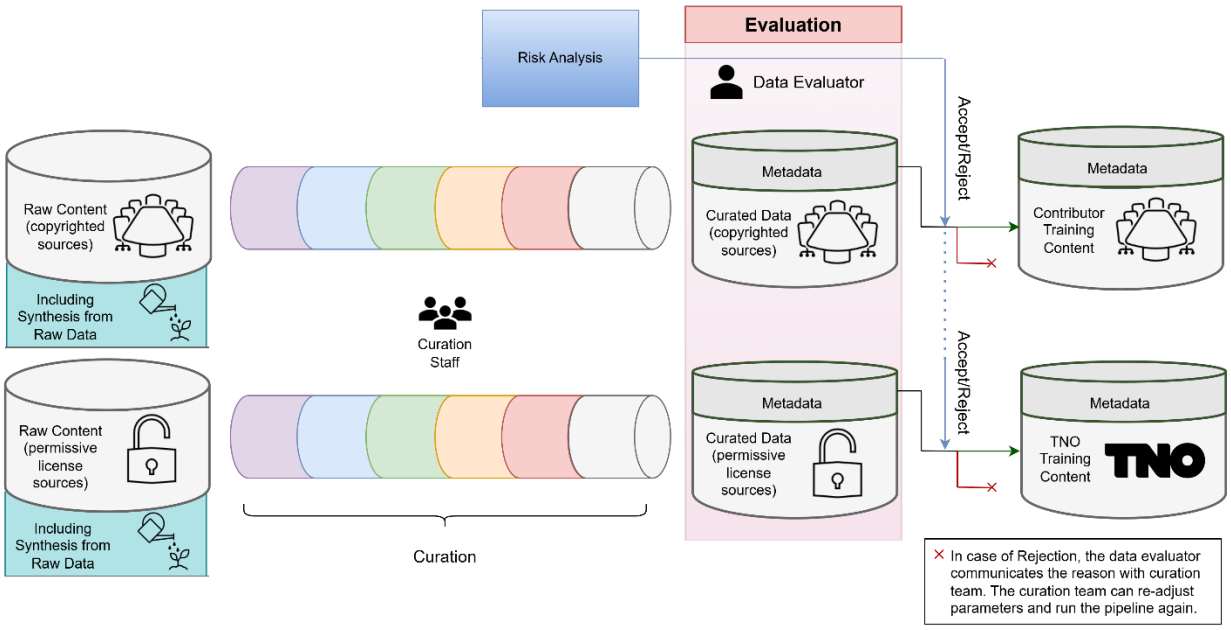

Figure 1: Data Curation and Evaluation Pipeline

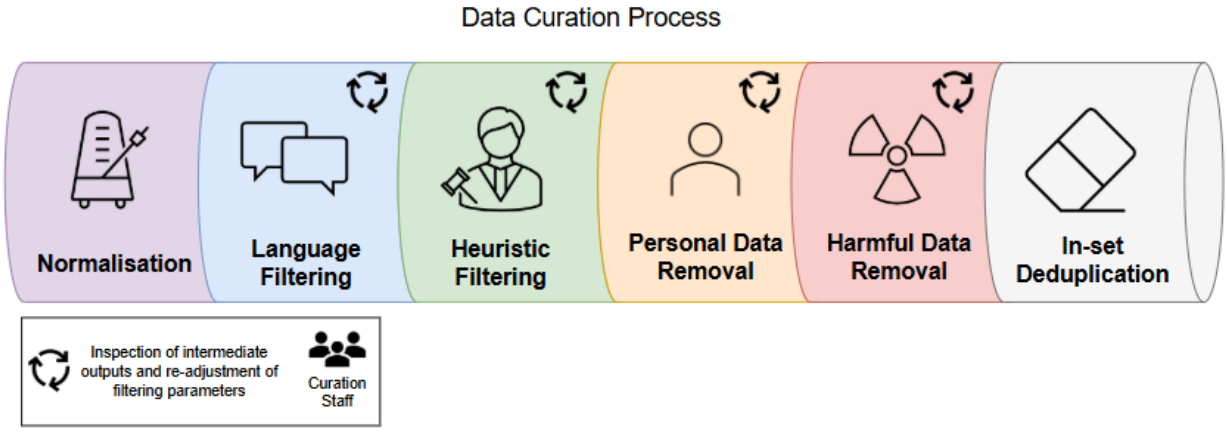

Figure 2: Data Curation Process

The Data Curation Process is shown in **Figure 2**. The goal of the Data Curation Process is to turn sets of Raw Content into Contributor Training Content or TNO Training Content. During this process, the output of every curation stage is inspected by the Data Curation Staff to check if the execution of the stage was handled as expected. Raw Content is used as a starting point for the Data Curation Process. This process is described in detail in the Data Curation Specifications included in **Annex 1**. In short, the Data Curation Process involves the following stages:

a. Normalization: Data gets normalized to a uniform format. This is mostly a practical step, and no metadata about this process will be saved during curation.
b. Language Filtering: Data will be filtered based on whether the data is in the desired language. Aggregated statistics of data removed will be saved.
c. Heuristic Filtering: Data will be filtered based on specific rules. Aggregated statistics of data removed will be saved.
d. Personal Data Detection and Removal: Sensitive Personal Data of Public Persons and all Personal Data of Non-Public Persons is contextually anonymized. Aggregated statistics of data anonymized will be saved.
e. Harmful Language Detection: Harmful Language is detected and removed in the dataset. Aggregated statistics of data removed will be saved.
f. Deduplication: Data gets deduplicated. This is mostly a practical step, and no metadata about this process will be saved during curation.

V. Data Evaluation Process

Although some evaluation is already done at every stage of the Data Curation Process by the Data Curation Staff, this is not enough to get a good idea of the risk of using specific curated datasets as input for the Prepared Dataset. For this reason, a risk analysis is done by a data evaluator outside the Data Curation Staff. Details of this Risk Analysis are given in **Annex 2**.

VI. Combination of Training Content

TNO conducts a second de-duplication of the combined Contributor Training Content and TNO Training Content to filter out content that is the same but originates from different sets of Contributor Training Content or TNO Training Content (see Figure 3). After this step, we have a unified, cleaned dataset (the Prepared Dataset) that can be used to train the GPT-NL Model.

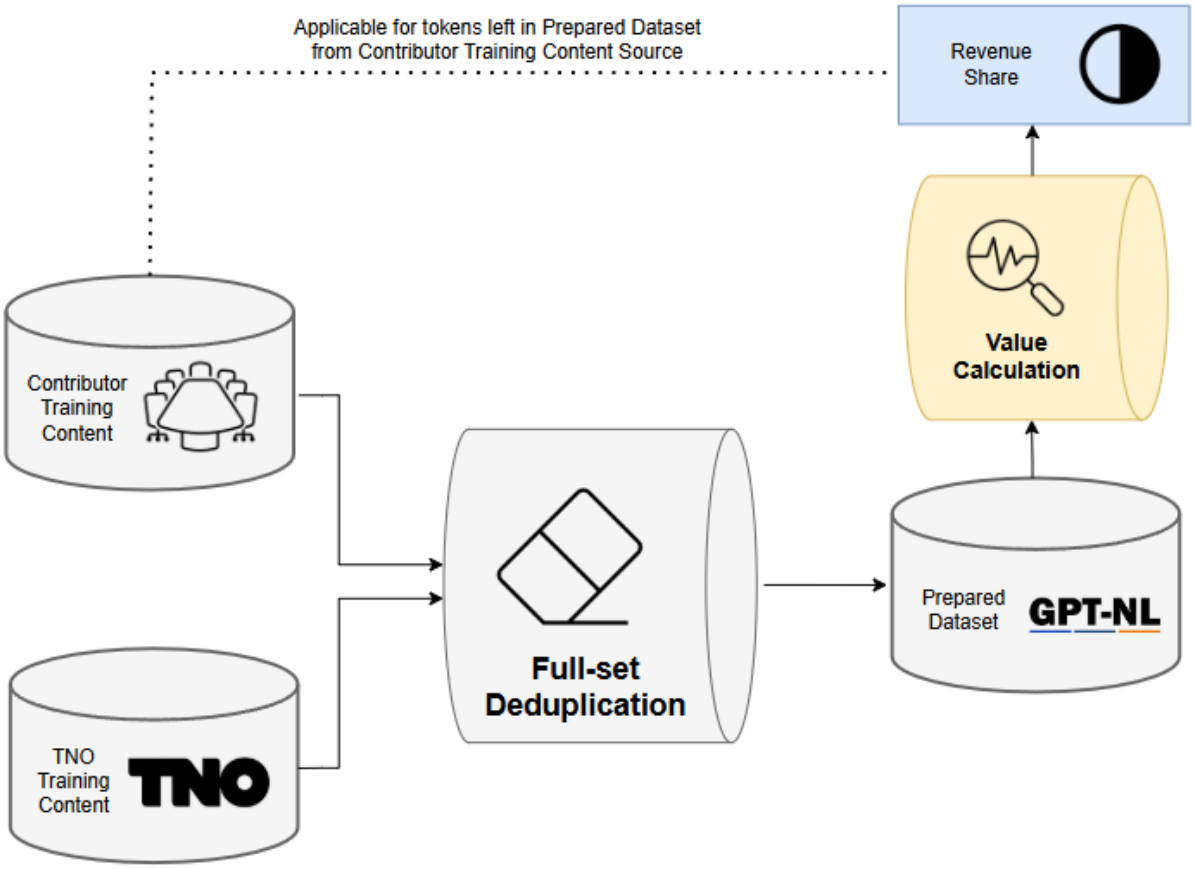

Figure 3: Process for global deduplication.

Value Calculation

The complete Prepared Dataset is collected now, and thus it is possible to calculate the relative value of every individual contributing dataset as part of the Prepared Dataset in accordance with the Revenue Sharing Mechanism. If content is the same between two different Content Contributors, the Content Contributor that has signed its Content Contributor Agreement first will get their claim on the revenue sharing of that content. If the same content is included in the TNO Training Content and one or more sets of Contributor Training Content, no value is attributed to the content because the content was already available in the public domain under an open-source license.

Read everything about the revenue sharing and data value calculation in the Revenue Sharing Mechanism.

Annex 1: Data Curation Specifications

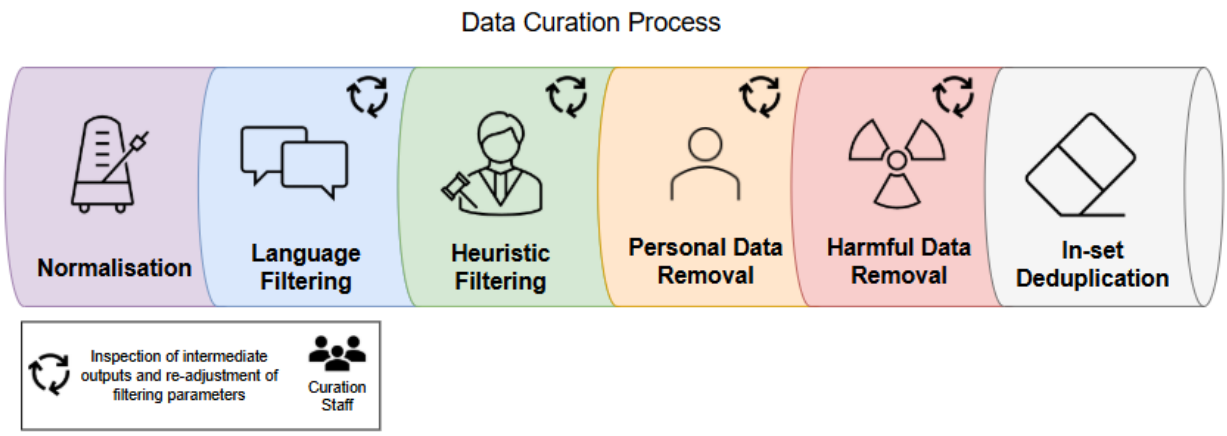

Figure 3: Overview of the Data Curation Process.

All components of the Data Curation Process are embedded in a DataTrove pipeline. DataTrove normalizes, filters, and deduplicates text. The Data Curation Staff has used this DataTrove pipeline as a basis. The Data Curation Staff then 1) modified the DataTrove components to suit GPT-NL's particular needs and 2) added new modules not present in DataTrove: third party software to detect and remove Personal Data and a module to detect and remove Harmful Information. This **Annex 1** contains a summary of the functionality of each of the components of the Data Curation Process.

1. Normalization

The purpose of text normalization is to ensure that all text has the same format. Our text normalization module consists of the following elements:

- **DataTrove – FTFY:** This filter, which stands for "Fixes Text For You," fixes Unicode-related issues.
- **DataJuice – Punctuation Normalizer:** This filter (taken from DataJuice, an alternative to DataTrove) normalizes punctuation. For instance, it turns this bracket: 【 into this bracket: [to ensure that all brackets are uniform.
- **DataJuicer – Whitespace Normalizer:** This filter (also taken from DataJuice) normalizes whitespace in text so that each whitespace has the same format.

This is the first step of the Data Curation Process, because having a uniform text format is essential for all other modules to function optimally.

2. Language filtering

The GPT-NL Model will be trained on a combination of Dutch and English text. It is therefore necessary to detect the language of each document that passes through the Data Curation Process. After having conducted a small-scale experiment to compare the performance of several language detectors, the GPT-NL Team decided to use the FastText language detector to determine whether a text is written in Dutch or in English. This language information is added to the metadata so that subsequent modules can use the information to perform the right actions. For instance, if the text is Dutch, the Harmful Language module uses a Dutch model to detect harmful language, and if the text is English, it uses the English equivalent. Texts that are neither in English nor in Dutch are removed from the dataset.

3. Heuristic filtering

The purpose of heuristic filtering is to remove low-quality data, such as data with many symbols and very few words. A total of 15 of these filters have been implemented in the Data Curation Process:

- **Symbol-to-word ratio**: drops texts that contain too many symbols in comparison to the number of words
- **Filter bullets**: drops documents with more than 90% of the lines starting with bullet points
- **Filter ellipsis**: drops documents with more than 30% of the lines ending in an ellipsis (i.e. "…")
- **Non alpha words**: drops documents with less than 80% of words containing at least one alphabetic character
- **Stop words**: drops document with fewer than 2 stop words (such as "the" and "and")
- **Filter duplicate lines**: drops documents where 35% or more of the lines are duplicates of other lines
- **Filter duplicate paragraphs**: drops documents where 35% or more of the paragraphs are duplicates of other paragraphs
- **Filter duplicate character lines**: drops documents where 20% or more of the characters in a line are duplicates of one another
- **Filter duplicate character paragraphs**: drops documents where 20% or more of the characters in a paragraph are duplicates of one another
- **Filter top *n*-grams**: drops documents with a high proportion of duplicated sets of words containing 2-4 words
- **Filter duplicate *n*-grams**: drops documents with a high proportion of duplicated sets of words containing 5-10 words
- **Maximum digit fraction**: drops documents where 20% or more of the characters are digits
- **Minimum character**: drops documents with fewer than 50 characters
- **Median characters per line**: drops documents with a mean median of fewer than nine characters per line
- **Median words per line**: drops documents with a mean median of fewer than 2.1 words per non-empty line

To test this configuration, a test set was filtered using this set of filters. The perplexity scores of the original text and the filtered text were then compared. Perplexity measures how well a probabilistic model predicts a sample of text, with lower values indicating better predictive performance. Perplexity can measure data quality by identifying how predictable and coherent the text is, with lower perplexity suggesting high-quality data and higher perplexity indicating low-quality text that is harder for language models to process effectively. The team found that the text that had been filtered

by the set of filters mentioned above had a much lower perplexity than the set of text that had not undergone any filtering. This indicates that the filters improve the quality of the data.

4. Personal Data detection and removal

Pseudonymization

Pseudonymization is done by the software that is used locally via a docker, obtained from an external company PrivateAI Installation - Grabbing the image | Private AI Docs , which specializes in this subject. PrivateAI has trained its own statistical model for the specific purpose of removing privacy-sensitive information.

The list of types of Personal Data that are removed by the software of PrivateAI are:

- Name (non-public persons only)
- Address
- Birth date
- URL
- Phone number
- IP address
- File path
- Email address
- IBAN
- Dutch identity document number
- Dutch identity number (BSN)
- Dutch “vreemdelingendocumentnummer”
- Belgian identity number
- Dutch/Belgian driver’s license number
- Belgian passport number
- Dutch license plate number
- Belgian license plate number
- Belgian phone number
- MAC address
- European VAT number
- Dutch phone number
- Credit card number
- Credit card CVV code
- Credit card expiry date
- UK national insurance number
- U.S. social security number
- UK driver’s license number
- Australian bank account number
- UK unique taxpayer reference number
- Australian driver’s license number
- Australian passport number
- Australian tax file number
- Canadian passport number
- UK/U.S. passport number

- Canadian bank account number
- U.S. bank account number
- U.S. individual taxpayer identification number
- Geolocation

The software of PrivateAI replaces these Personal Data elements by synthetically generated alternatives. This ensures the retention of a well-running text as well as the removal of Personal Data. When the same name occurs multiple times, that name is replaced by the same synthetic alternative ('grouping'). Grouping is done on a document-level (e.g., a newspaper article, a report, etc), to prevent that too much information is linked to the same person, which may lead to easier identification of that person. If the document is longer than 512 tokens, the grouping is done in 512 token chunks.

Public and non-public persons

Considering the different expectation of privacy of public and non-public persons, TNO applies a different approach to Personal Data detection and removal of public persons and of non-public persons. Because the model needs to have world knowledge concerning well-known individuals (e.g. the model needs to know about the war in Ukraine and therefore about individuals like Zelenskyy and Putin). To determine whether an individual qualifies as a public person, TNO has taken a conservative approach and relies on the Wikidata database. For people who have a Wikipedia page, it may be assumed that they qualify as public persons. The GPT-NL Team extracted a list of all such persons from Wikidata and whitelisted them using the PrivateAI software, meaning these names are kept in the data as they are. Other than the name of public persons, all other elements of Sensitive Personal Data listed above will be removed for public persons by the software of Private AI.

2. Harmful language detection and removal

The aim of this module is to remove as much harmful language from the Contributor Training Content and TNO Training Content as possible to prevent the GPT-NL Model from producing such language. Examples of types of text that are removed by this module are racist, sexist, or homophobic utterances and death threats.

For the Dutch texts, the IMSyPP Hate Speech model was used. For English, the team used the ToxiGen model. Both are statistical models that have been trained on both harmful and non-harmful annotated data. Using information learned from these data, the models determine for each sentence they come across whether or not the sentence is likely to contain harmful language. The module is set up so that sentences that are marked as harmful are removed, while the surrounding text is kept.

3. Deduplication

When training content is collected from multiple sets of Contributor Training Content, it is likely that the resulting combination will contain duplicates or near-duplicates, which has been shown to negatively affect model performance. For this reason, a deduplication module has been implemented. The team used the MinHash algorithm that is implemented in DataTrove. This algorithm groups documents in buckets and then checks the similarity of the documents within those buckets in order to reduce the number of documents that need to be compared with each other.

Annex 2: Risk Analysis Specifics

After the Raw Content has passed through the Data Curation Process, it constitutes Contributor Training Content. A risk analysis is done by a Data Evaluation Officer to estimate the severity of potential Non-Suitable Data remaining in the dataset.

The analysist has a couple of components:

- The report has aggregated dataset statistics such as the overall percentage of data flagged as personal, sensitive, harmful, or low quality.
- The deep dive survey is used as input for the data evaluation report.
- The results of the report get added to the datasheet of the data set.

Based on the Evaluation Report, the Data Evaluation Officers can either accept the dataset, request re-adjustments of curation parameters to the Data Curation Staff, or fully reject the dataset. If a dataset is accepted, the dataset is judged with a “risk-score”. Low “risk-score” datasets will be used more during training.