Annex 1

Training Content Deep Dive Survey (To be) Completed by Content Contributor

Training Content Deep Dive Survey (To be) Completed by Content Contributor

Description of the Training Content

Dataset #01

Metadata Field	Data Contributor Response
1.1 Description Organization Dataset	
2.1 Origin Content	
2.2 Topics	
2.3 Languages	
2.4 Quality-assurance	
2.5 Date Created	
2.6 Collection Methods	
2.7 Date Collected	
2.8 Author Dataset	
2.9 Available Metadata	
2.10 Modifications made	
3.1 Personal Data in Set	
3.2 Reason for Personal Data	
3.3 Accuracy Assurance Personal Data	
3.4 Request for Personal Data Removal Implementation	
3.5 Assurance of no confidential/protected information	
3.6 Harmful Data in Set	
4.1 Ethical Review	
4.2 Author Background	
4.3 Diversity Metadata	
4.4 Potential Biases	
4.5 Bias Mitigations	
5.1 Additional Information	

Data Set #02

…

Annex 2

Training Content Protocol

GPT-NL

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 Acquisition, Jesse van Oort (jesse.vanoort@tno.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 staff of Team GPT-NL tasked with training of the GPT-NL Model. 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. After completion of the pre-training phase (i.e. completion of the pre-trained GPT-NL Model v. 1.0 or any successor pre-training version of the GPT-NL Model) TNO will delete the Raw Content and provide a copy of the Contributor Training Content to Content Contributor to use for any purpose.

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.

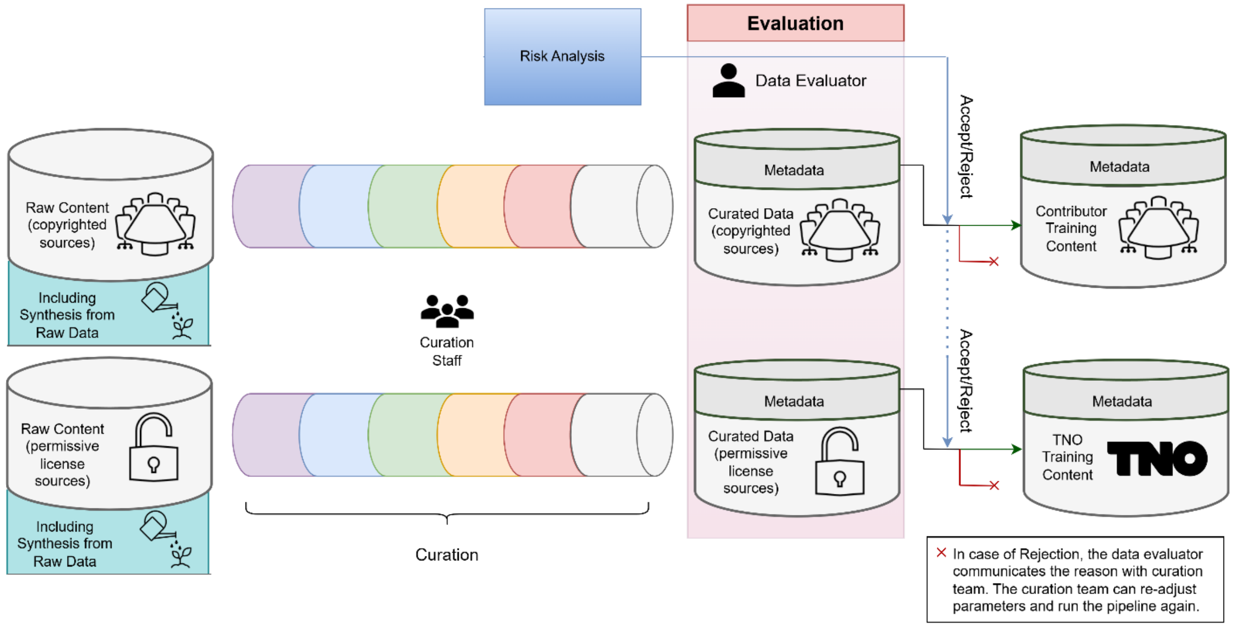

Risk Analysis
Evaluation
Data Evaluator
Raw Content (copyrighted sources)
Including Synthesis from Raw Data
Raw Content (permissive license sources)
Including Synthesis from Raw Data
Curation Staff
Curation
Metadata
Curated Data (copyrighted sources)
Metadata
Curated Data (permissive license sources)
Accept/Reject
Accept/Reject
Metadata
Contributor Training Content
Metadata
TNO Training Content
TNO
In case of Rejection, the data evaluator communicates the reason with curation team. The curation team can re-adjust parameters and run the pipeline again.

Figure 1: Data Curation and Evaluation Pipeline

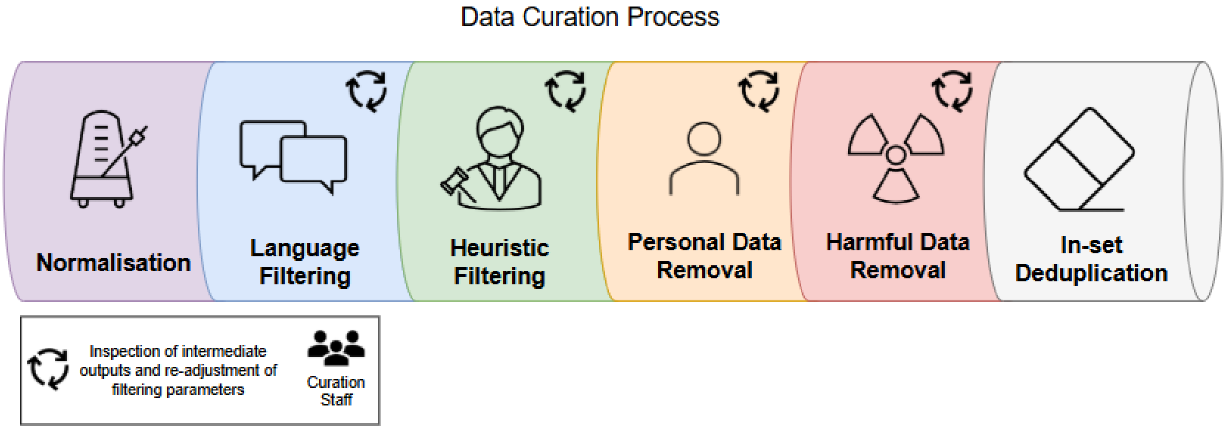


Figure 2: Data Curation Process

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.

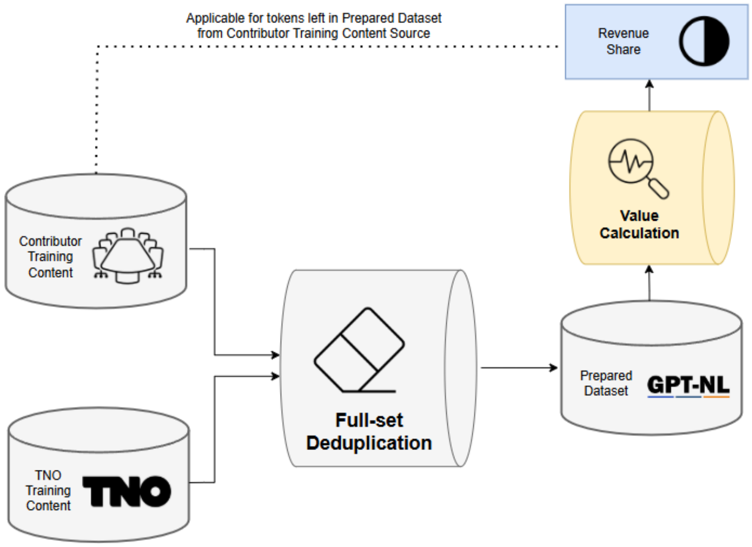

Applicable for tokens left in Prepared Dataset
from Contributor Training Content Source
Revenue Share
Value Calculation
Contributor Training Content
Full-set Deduplication
Prepared Dataset
GPT-NL
TNO Training Content
TNO

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.

Appendix 1: Data Curation Specifications

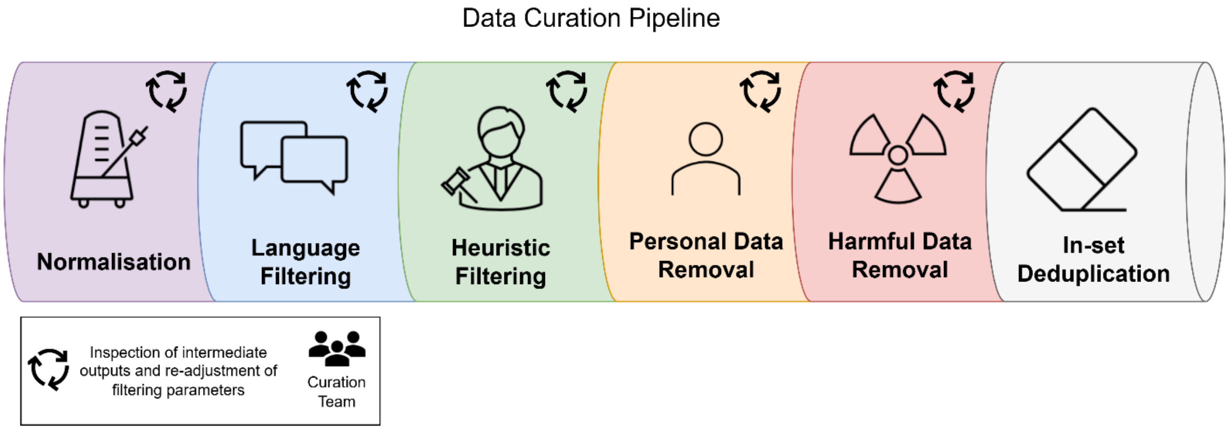


Figure 3: Overview of the Data Curation Process.

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.

Appendix 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.

Annex 3

Governance Charter

GPT-NL

GPT-NL Project

Governance Charter

Version 1.0

TABLE OF CONTENTS

1. DEFINITIONS

In this Governance Charter, the following words and expressions shall have the following meanings:

“**Advisory Board**” means the board as set up and with the roles and functions as described in Section 2.5.3;

“**Co-Chairs**” means the persons co-chairing the Advisory Board, designated as such in accordance with Section 2.5.2;

“**Content Consortium**” means the consortium of Content Contributors who collaboratively provide content necessary for the training and future updating of the GPT-NL Model;

“Content Contributor” means a party that participates in the GPT-NL Project by (i) contributing relevant content for purposes of training of the GPT-NL Model and (ii) is invited to participate in the governance of the GPT-NL Project in accordance with this Governance Charter;

“Content Board” means the board consisting of representatives of the Content Contributors, designated as such in accordance with Section 2.5.6.

“**Content Contributor Agreement**” means the agreement between TNO and a Content Contributor regarding the provision of training content for the GPT-NL Project;

“**Data Protection Protocol**” means the protocol governing obligations of TNO and the Content Contributors and potential other stakeholders with respect to processing of personal data in the context of the GPT-NL Project. The Data Protection Protocol is published on the GPT-NL website and may be updated from time to time in accordance with this Governance Charter;

“**GPT-NL Model**” means one or more Dutch-language large language model developed pursuant to the Grant and further grants or license income generated for this purpose;

"**Grant**” means the grant awarded to TNO on 26 April 2024[1] for the development of a cloud-based research infrastructure for operating, training, finetuning, testing and analyzing all aspects of large language models, which includes the creation of at least a first iteration of a Dutch large language model;

“**Governance Charter**” means this charter, which sets out the functions and responsibilities assigned to the various participants to the GPT-NL Project;

"**GPT-NL Project**” means the project to build the Research Facility, which includes the creation of the GPT-NL Model, which GPT-NL Model will be made available to third parties to serve the national public interest;

“GPT-NL website”: the website with url: www.gpt-nl.nl

“**Managing Director of TNO**” means the managing director of TNO responsible for the GPT-NL Project.

[1] See https://www.rijksoverheid.nl/documenten/kamerstukken/2024/04/10/eerste-financieringsronde-faciliteiten-voor-toegepast-onderzoek-fto.

“**NFI**” means the Netherlands Forensic Institute.

“**Project Team**” means the individuals who are designated as such in accordance with Section 2.5.4;

“**Research Facility**” means the cloud-based research infrastructure for operating, training, finetuning, testing of large language models (LLMs) and research into all aspects of LLMs to developed by or on behalf of TNO pursuant to the Grant;

“**Responsible Use Policy**” means the policy that shall set out the default responsible use restrictions that will apply under the licensing terms for use of the GPT-NL Model to both scientific non-commercial research and commercial uses. The Responsible Use Policy will contain limitations and permissions based on objective norms that are common practice in the market. The Responsible Use Policy is published on the GPT-NL website and may be updated from time to time in accordance with the Governance Charter.

“**Responsible Directors**” means the director *ICT Strategy & Policy* and the director *Defense, Safety & Security* of TNO.

“**Responsible Units**” means the TNO Unit *ICT Strategy & Policy* and TNO *Unit Defense, Safety & Security*.

“**Revenue Sharing Mechanism**” means the policy for compensating Content Contributors for providing training content for purposes of training and future updating of the GPT-NL Model. The Revenue Sharing Mechanism is published on the GPT-NL website, and may be updated from time to time in accordance with the Governance Charter;

“**SURF**” means the Dutch IT cooperation for education and research institutions;

“**Training Content Protocol**” means the protocol that sets out the requirements to be followed by the Content Contributors and TNO when preparing training content for the GPT-NL Model. The Training Content Protocol provides the objective norms, tools, and/or protocols that must be followed or applied for the compliance of (the development of) the GPT-NL Model with Applicable Laws. The Training Content Protocol is published on the GPT-NL website and may be updated from time to time in accordance with the Governance Charter;

“**TNO**” means the Netherlands Organization for Applied Scientific Research; and

“**Workgroups**” means the ad-hoc, time-limited groups as described in Section 2.5.7.

2. BACKGROUND INFORMATION

2.1. Background of the GPT-NL Project

The intention of the GPT-NL Project is to train the GPT-NL Model based on high quality content (free of harmful and irrelevant content) for which a license has been obtained (or no license is required) and by ensuring that the content is prepared in a privacy-preserving manner.in accordance with the Training Content Protocol. To achieve this, TNO is setting up a consortium of Content Contributors that provide training content for the purpose of the creation and future updating of the GPT-NL Model (the **Content Consortium**) and participate as stakeholders in the governance of the GPT-NL Project.

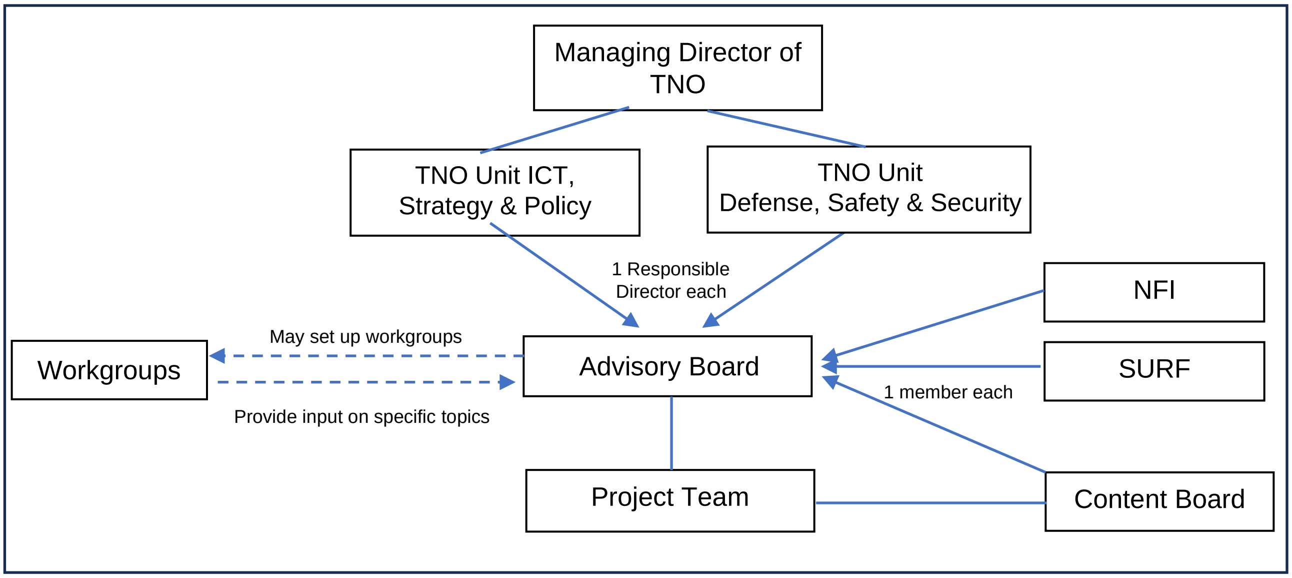
2.2. Description

This Governance Charter is an attachment to the Content Contributor Agreements (or other specific agreement) to be entered into between TNO and each of the individual Content Contributors, who collectively make up the Content Consortium. This Governance Charter sets out the functions and responsibilities assigned to each of the parties involved with the GPT-NL Project.

2.3. Amendments to this Governance Charter

This is a controlled document for which the Project Team is accountable. Changes to the Governance Charter may be proposed by TNO or any of the relevant stakeholders and will be decided in accordance with this Governance Charter.

2.4. Organizational Structure


Managing Director of TNO
TNO Unit ICT, Strategy & Policy
TNO Unit Defense, Safety & Security
1 Responsible Director each
NFI
May set up workgroups
Workgroups
Advisory Board
SURF
1 member each
Provide input on specific topics
Project Team
Content Board

2.5. Roles

2.5.1. The Managing Director of TNO

2.5.1.1. Decision-making

Within TNO, Units are tasked with executing specific scientific projects. The GPT-NL Project is a joint project of the Responsible Units. The decision-making authority in respect of the GPT-NL Project is in accordance with the existing decision-making mandates within TNO. The Responsible Directors have the power to take decisions about the GPT-NL Project as are within their respective mandates and will do so in accordance with this Governance Charter. Decisions that exceed the Responsible Directors’ mandate will be submitted to the Managing Director of TNO, accompanied by an advice of the Advisory Board (see below).

2.5.2. Co-Chairs of the Advisory Board

2.5.2.1. Appointment

The Responsible Directors shall act as Co-Chairs.

2.5.2.2. Roles and Responsibilities

The Co-Chairs assume overall responsibility for the proper functioning of the Advisory Board and will chair the Advisory Board meetings. They will also act as spokespersons about the GPT-NL Project to the public.

2.5.3. Advisory Board

2.5.3.1. Appointment

The representatives serving on the Advisory Board will include, but may not be limited to:

1. The Co-Chairs (set out in Section 2.4.2)
2. A member on behalf of the Content Board;
3. A member on behalf of NFI; and
4. A member on behalf of SURF.

The selection process of the members on behalf of the Content Board, NFI and SURF will require input from the other members of the Advisory Board and the Project Team. When at a certain moment in time NFI or SURF is no longer actively involved in the GPT-NL Project in a meaningful manner, their respective role as member of the Advisory Board will terminate and the Co-Chairs will invite other stakeholders in the GPT-NL Project to join the Advisory Board instead. Any selection of new members will require input from the other members of the Advisory Board and the Project Team.

2.5.3.2. Roles and Responsibilities

The Co-Chairs will determine the agenda of the Advisory Board meetings, which will include all important decisions in relation to the GPT-NL Project. The Advisory Board will discuss such material matters relating to the GPT-NL Project, which will inform the decision-making by the Responsible Directors. If a decision falls outside the decision-making mandate of the Responsible Directors, the Advisory Board will issue advice to the Managing Director of TNO for decision.

Topics that will be discussed within the Advisory Board include, but are not limited to the following topics:

1. the collective vision, the goals of the GPT-NL Project and the annual collaborative agenda;
2. selection of Workgroups, potential new Content Contributors, and service providers;
3. new licensing types and terms under which the GPT-NL Model will be made available as well as individual deviations to be agreed upon by TNO with individual licensees that would substantially affect the protection provided to Content Contributors;
4. new development ideas;
5. the continuing integrity and rigor of the GPT-NL Project;
6. potential additional funding sources and help create and maintain relationships with existing funding sources;
7. articles, presentations, press releases or other publicity;
8. revisions to the Contributor Agreement (including amendments per Section 14.8), the Governance Charter, the Training Content Protocol, the Revenue Sharing Mechanism, the Responsible Use Policy, the Data Protection Protocol and the licensing terms for the GPT-NL Model as well as additional policies in the context of the GPT-NL Project;
9. requests or proposals submitted to the Advisory Board by the Content Board in accordance with Section 2.5.6;

10. any enforcement actions to be undertaken by TNO in respect of violation by licensees of the license terms;
11. advising and assisting the Co-Chairs in their capacity as spokespersons for the GPT-NL Project to the public.

The term for representatives serving on the Advisory Board will be three years, except for the inaugural representatives, to allow for continuity. The Co-Chairs' and Advisory Board members' terms should be staggered; this is done so that there is at least one continuing representative on the Advisory Board in any year. Re-election is permitted, and there is no overarching term limit.

2.5.3.3. Meetings and advice

Meetings will be convened by teleconference, web, or face-to-face. Meetings will be regularly scheduled, monthly as needed, and no less than quarterly. Agenda items will be solicited before the meeting and circulated to attendees. Meeting minutes will be distributed following the meeting.

The Co-Chairs will ensure that the Advisory Board has a meaningful say in important decisions regarding the GPT-NL Project and that any advice of the Advisory Board will be taken into serious consideration. If at any given time a member of the Advisory Board disagrees with an important decision of the Co-Chairs, they may provide a written advice to the Managing Director of TNO requesting reconsideration of the decision of the Co-Chairs.

Decisions that exceed the Responsible Directors' mandate will be submitted to the Managing Director of TNO together with an advice of the Advisory Board. Any advice of the Advisory Board to the Managing Director of TNO requires a quorum of at least 60% of the Advisory Board. Any member of the Advisory Board that does not agree with the advice may require that the advice includes a written explanation why the relevant member does not agree with the advice. For advice requested over email, at least 60% of the Advisory Board are required to respond before the decision on the advice is considered final.

At least one Co-Chair's vote is required for decisions on advice made in meetings or remotely. The Managing Director of TNO shall take any advice of the Advisory Board into serious consideration.

The Advisory Board shall take into serious consideration any requests or proposals submitted by the Content Board to the Advisory Board in accordance with Section 2.5.6.

Advisory Board members are responsible for disclosing any conflicts of interest related to decisions on advice that affect their other roles. In these cases, representatives must recuse themselves and not participate in those discussions or attempt to influence the decision-making process. Any conflict-of-interest disclosure must be declared at the beginning of the Advisory Board meeting. Other representatives from the Board, Project Team, Workgroups, or other external advisors may be invited to Advisory Board meetings to discuss specific agenda items on an as-needed basis when the Advisory Board desires additional expertise or other input.

Decisions on advice and action items from meetings will be documented and archived by the Project Team; any representatives responsible for action items will be notified.

2.5.3.4. Decisions requiring agreement member representing Content Board

Decisions by the Co-Chairs or the Managing Director of TNO on topics 3, 8, or 9 in Section 2.5.3.2. that may have a substantial negative impact on the protection provided to Content Contributors under the Content Contributor Agreement, may only be taken with the agreement of the member on behalf of the Content Board, which will not be unreasonably withheld. In case such agreement is withheld, relevant Content Contributors may terminate the Content Contributor Agreement in accordance with Section 13.2 thereof, and from the moment notice of termination has been given, any decisions of TNO on topics 3, 8 or 9 may only be taken in respect of new versions of the GPT-NL Model which are no longer trained on the content of these Content Contributors, unless specifically agree otherwise by the member of the Content Board and the relevant Content Contributors that have provided notice of termination of the Content Contributor Agreement.

2.5.4. Project Team

2.5.4.1. Appointment

The Project Team will be designated by the Responsible Directors and comprise of individuals who provide operational support to the GPT-NL Project and the Content Board, and any other designees as appropriate. Subject matter experts may also be invited to Project Team meetings from time to time.

2.5.4.2. Roles and Responsibilities

The Project Team will be responsible for all operational activities related to the GPT-NL Project, including operational decisions that support the GPT-NL Project's priorities defined by the Responsible Directors, communications, meeting planning, budgeting/finance, and contracting with Content Contributors, service providers, and other third parties. The Project Team will report to the Responsible Directors and keep the Advisory Board informed of any issues.

Project Team meetings may be weekly or bi-weekly. In addition, the Project Team or their designee will coordinate activities (calls, meetings, communications), coordinate the development and maintain version control of all relevant documents and protocols relating to the GPT-NL Project, and develop and implement quality assurance measures that include monitoring of adherence to the Training Content Protocol, Data Protection Protocol and any other related protocols or policies as well as quality control of data.

2.5.5. Content Contributors

2.5.5.1. Appointment

Content Contributors will be invited to participate in the GPT-NL Project by the Project Team, in collaboration with the Advisory Board, based on their ability to participate in and contribute to the GPT-NL Project.

2.5.5.2. Roles and Responsibilities

Content Contributors will be responsible for adhering to the process and policies in this Governance Charter, choosing a member to the Content Board, and (if so invited) participate in Workgroups, to provide input to protocols of the GPT-NL Project and be active contributors to support the mission and goal of the GPT-NL Project.

2.5.6. Content Board

All Content Contributors will be invited to join the Content Board. The Project Team is responsible for convening the Content Board every 6 months basis, or such other interval as decided by the Advisory Board. A member of the Project Team will chair the meetings of the Content Board. The Content Board will engage in a collaborative manner with the Project Team and each other on any topics relating to the GPT-NL Project and will advise the Advisory Board on any such topics.

The Content Board will be represented on the Advisory Board with one member. The Content Board may propose a representative which is decided by voting on the basis of Relative Data Value. The selection process of the member to join the Advisory Board, will require input from the other members of the Advisory Board and the Project Team.

The Content Board may submit a request or proposal to the Advisory Board for consideration, provide the request or proposal is supported by a majority of the Content Contributors based on their collective Relative Data Value.

2.5.7. Workgroups

2.5.7.1. Appointment

Workgroups can be proposed by the Project Team, the Content Contributors, or the Advisory Board, as needed and will be discussed in the Advisory Board for advice. In any event a Content Workgroup will be set up for any topics relating to the practical and technical aspects of the preparation and provision of the Training Content.

2.5.7.2. Roles and Responsibilities

Workgroup members will engage in a collaborative manner with each other on a specific project or question/needs for which the Workgroup has been established. Any member of a Workgroup would be expected to participate in the relevant GPT-NL Project. Workgroup members must disclose any conflicts of interest that could present a bias in the design, conduct, or reporting when working on the relevant project. Workgroup meetings will be conducted on an as needed basis but no less than quarterly. Workgroups will be inactivated once projects are completed.

2.5.8. Service Providers

The Project Team will consult the Advisory Board on selecting any necessary service providers. The GPT-NL Project, as any project that involves the training of a large language model, requires a substantial amount of computational power.

3. CONFLICTS OF INTEREST AND SERIOUS MISCONDUCT

3.1. Financial Disclosure and Conflict of Interest

The Advisory Board, the Project Team, the Board, and other key personnel will be required to disclose all financial interests and working relationships with any entity whose financial interests potentially could be affected by the conduct or outcome of the GPT-NL Project.

3.2. Serious Breach of Policies or Misconduct

Major violations or misconduct under an agreement with a Content Contributor Agreement, the Training Content Protocol, the Data Protection Protocol or the Acceptable Use Policy that may jeopardize privacy or safety, and/or the integrity of the GPT-NL Project should be reported to TNO via abuse@gpt-nl.nl, and may result in the infringing party no longer being able to participate in the GPT-NL Project. This is especially a concern when a certain individual or entity is making more mistakes than expected. Assessment of any suspected or alleged serious misconduct will be discussed by the Project Team first and then escalated to the Advisory Board if there is evidence of serious misconduct.

4. COLLABORATION AND TRANSPARENCY

4.1. Requests to Use General Information

Organizations wishing to publish or present general information about the GPT-NL Project, which do not include confidential or proprietary information, may do so without formal approval. Examples of general information include the number and identity of participating Content Contributors, information (e.g., protocols and milestones) about the GPT-NL Project, and summaries of publications and presentations. Because information about the GPT-NL Project changes frequently, such organizations are encouraged to use frequently updated slides from the Project Team and send a courtesy advance notification to the Project Team via info@gpt-nl.nl about the intended publication or presentation.

***

Annex 4

Baseline Responsible Use Policy

GPT-NL

Responsible Use Policy

Introduction

This Responsible Use Policy (R**UP**) sets out the default use restrictions that apply to the use of the GPT-NL large language model (**GPT-NL Model**) subject to various license terms. TNO can divert from these default use restrictions for specific use cases, such as, for potential use by Dutch law enforcement agencies or by Dutch/NATO armed forces. This RUP may be changed from time to time in accordance with the license terms that apply to your use of the GPT-NL Model (**Applicable License Terms**).

Please take note that this RUP focuses primarily on the ethical aspects of the use of the GPT-NL Model. Other aspects, such as specific intellectual property or other legal considerations, are set out in the Applicable License Terms

The GPT-NL Model is created as part of a project aimed to serve the public interest. We have therefore considered and incorporated use restrictions in this RUP that are broadly carried by relevant societal stakeholders. Documents that inspired this RUP include:

- The RAIL-contract models for responsible AI Licensing;
- The “living guidelines on the responsible use of GenAI in Research” (ERA Forum March 2024);
- The ethics guidelines on trustworthy AI (EU High level expert group on AI)

Use restrictions

You agree not to use the GPT-NL Model for any of the following:

1. **General**

   (a) To defame, disparage, or otherwise harass others.

   (b) To intentionally deceive or mislead others, including failing to appropriately disclose to end users any known dangers of your system.

   (c) To automatically or programmatically extract outputs whether to generate training content for other large language models or AI algorithms, or otherwise.

   (d) To reverse engineer the GPT-NL Model to extract content the GPT-NL Model was trained upon.

   (e) To circumvent safeguards or use restrictions included in the GPT-NL Model and/or prescribed by TNO, unless supported by TNO (e.g., for red teaming or testing purposes).

2. **Discrimination**

   (a) To discriminate or exploit individuals or groups based on legally protected characteristics and/or vulnerabilities.

(b) For purposes of administration of justice, law enforcement, immigration, or asylum processes, such as predicting that a natural person will commit a crime or the likelihood thereof.

(c) To engage in, promote, incite, or facilitate discrimination or other unlawful or harmful conduct in the provision of employment, employment benefits, credit, housing, or other essential goods and services.

2. Military and national intelligence services

(a) For weaponry or warfare

(b) For purposes of building or optimizing military weapons or in the service of nuclear proliferation or nuclear weapons technology.

(c) For purposes of military or national intelligence surveillance, including any research or development relating to such surveillance.

(a) – (c) unless a specific license has been obtained for such use.

3. **Legal**

(a) To engage or enable fully automated decision-making that adversely impacts a natural person's legal rights without expressly and intelligibly disclosing the impact to such natural person and providing an appeal process.

(b) To engage or enable fully automated decision-making that creates, modifies or terminates a binding, enforceable obligation between entities; whether these include natural persons or not.

(c) In any way that violates any applicable law or regulation.

4. **Disinformation**

(a) To create, present or disseminate verifiably false or misleading information for economic gain or to intentionally deceive the public, including creating false impersonations of natural persons.

(b) To synthesize or modify a natural person's appearance, voice, or other individual characteristics, unless prior informed consent of said natural person is obtained.

(c) To autonomously interact with a natural person, in text or audio format, unless disclosure and consent is given prior to interaction that the system engaging in the interaction is not a natural person.

(d) To defame or harm a natural person's reputation, such as by generating, creating, promoting, or spreading defamatory content (statements, images, or other content).

(e) To generate or disseminate information (including - but not limited to - images, code, posts, articles), and place the information in any public context without expressly and intelligibly disclaiming that the information and/or content is machine generated.

5. **Privacy**

   (a) To utilize personal information to infer additional personal information about a natural person, including but not limited to legally protected characteristics, vulnerabilities or categories; unless informed consent from the data subject to collect said inferred personal information for a stated purpose and defined duration is received, except where this is explicitly allowed under relevant personal privacy legislation.

   (b) To generate or disseminate personal identifiable information that can be used to harm an individual or to invade the personal privacy of an individual.

   (c) To engage in, promote, incite, or facilitate the harassment, abuse, threatening, or bullying of individuals or groups of individuals.

6. **Health**

   (a) To provide medical advice or make clinical decisions without necessary (external) accreditation of the system; unless the use is (i) in an internal research context with independent and accountable oversight and/or (ii) with medical professional oversight that is accompanied by any related compulsory certification and/or safety/quality standard for the implementation of the technology.

   (b) To provide medical advice and medical results interpretation without external, human validation of such advice or interpretation.

   (c) In connection with any activities that present a risk of death or bodily harm to individuals, including self-harm or harm to others, or in connection with regulated or controlled substances.

   (d) In connection with activities that present a risk of death or bodily harm to individuals, including inciting or promoting violence, abuse, or any infliction of bodily harm to an individual or group of individuals

8. **Research**

   (a) In connection with any academic dishonesty, including submitting any informational content or output of the GPT-NL Model as your own work in any academic setting.

9. **Malware**

   (a) To generate and/or disseminate malware (including - but not limited to - ransomware) or any other content to be used for the purpose of Harming electronic systems;

Annex 5

Data Protection Protocol

GPT-NL project – Data Protection Protocol

Arrangement between TNO and Content Contributors in respect of the processing of personal data in the context of training the GPT-NL large language model, including an arrangement determining their respective data protection responsibilities

Background

A. TNO has been awarded a research grant from the Dutch government to develop a *state-of-the-art* research infrastructure for training large language models ("**LLMs**") that comply with European and Dutch legislation and public values. This research facility will be kept operational for three years and the design documentation and operating software will be publicly made available to enable third parties to set up their own LLM training environments to serve the national public interest. Deliverable of the research grant (and further funding) will further be the creation of a competitive Dutch LLM that complies with European legislation and public values ("**GPT-NL Model**"). These activities are hereafter referred to as the "**GPT-NL Project**".

B. TNO is setting up a consortium of organizations that are willing to participate in the GPT-NL Project by contributing relevant content for purposes of initial training (and potential later updating) of the GPT-NL Model (the "**Content Contributors**"). TNO, and the Content Contributors are hereafter referred to as each a "**Party**" and collectively the "**Parties**".

C. Content Contributors each select the training content that they want to contribute to the GPT-NL Project. Before delivering the Raw Content (as defined below) to TNO Content Contributor will scrub any content from its datasets in accordance with the Training Content Protocol (as defined below). This involves scrubbing for example Unsuitable Sources (as defined in the Training Content Protocol) including those identified by TNO during the Viability Assessment in accordance with the Training Content Protocol.

D. TNO will prepare the Raw Content provided by Content Contributor in accordance with the Training Content Protocol (the so cleaned content: "**Contributor Training Content**"). This involves scrubbing sensitive categories of personal data which have a structured format (e.g., BSN, passport phone numbers, email addresses and geolocation data), the contextual anonymization of information about non-public persons, the removal of harmful language, and the removal of other information that is not relevant to LLM training. TNO will implement *privacy-by-design* measures to reduce risk of unnecessary access of TNO staff to the Raw Content provided by a Content Contributor. The cleaning activity will be performed by dedicated TNO data curation staff ("**Data Curation Staff**"). After cleaning the Raw Content, the Data Curation Staff will provide a copy of the Contributor Training Content to the staff of the GPT-NL team tasked with training the GPT-NL Model. No staff of TNO other than the Data Curation Staff will have access to the Raw Content of a 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. After completion of the pre-training phase (i.e. completion of the pre-trained GPT-NL Model v. 1.0 or any successor pre-training version of the GPT-NL Model) TNO will delete the Raw Content and provide a copy of the Contributor Training Content to Content Contributor to use for any purpose.

E. Besides training content contributed by the Content Contributors, the GPT-NL Model will be trained based on publicly available that is not subject to copyright (including where the copyright has expired) or that is subject to a valid open-source license. TNO will clean such content in accordance with the Training Content Protocol ("**TNO Training Content**") on the same basis as TNO cleans training content for Content Contributors. TNO will subsequently compile the TNO Training Content and the Contributor Training Content into a single dataset and further prepare this dataset for use for the training of the GPT-NL Model ("**Prepared Dataset**").

F. TNO will use the Prepared Dataset to train a first version of the GPT-NL Model. The deliverables will consist of the technical source code of the GPT-NL Model ("**Source Code**") and the model weights necessary to use the GPT-NL Model generate responses to inputs ("**Model Weights**"). The Source Code will be made publicly available under an open-source license. The Model Weights will be licensed by TNO to researchers for non-commercial research purposes ("**Research License**") and further to Content Contributors and other parties for all other purposes, including commercial purposes ("**Professional License**"). The parties that acquire a license to both the Source Code and Model Weights are referred to as "**Licensees**". Licensees can use the Source Code and Model Weights to run and host their own version of the GPT-NL Model and use it in accordance with the Responsible Use Policy (defined below). TNO will not be hosting a web-version of the GPT-NL Model for use by Licensees. The TNO Training Content will be made publicly available under an open-source license, but not the Contributor Training Data and the Prepared Dataset.

G. TNO being a public research organisation having obtained a government grant in relation to the GPT-NL Project, is subject to mandatory retention requirements in respect of any project materials, which provide that TNO will retain the Contributor Training Content, the TNO Training Content and the Prepared Dataset for a period of 7 years after the GPT-NL Project is completed.

H. The Parties wish to record in this protocol (the "**Data Protection Protocol**") for which parts of the potential processing personal data in the context of the GPT-NL Project they are responsible, as well as a division of their respective controller responsibilities for the fulfilment of their obligations under the Data Protection Laws (as defined below in Section 1).

<u>Definitions</u>

1. In addition to the terms defined above, the terms set out in this Section have the following meaning in this Data Protection Protocol:

   "**Content Contributor Agreement**" means the commercial agreement concluded between TNO and the Content Contributor that incorporates the terms applicable to the Content Contributor's provision of content for the training and future updating of the GPT-NL Model;

   "**Data Protection Laws**" means all laws applicable to the Parties' processing of personal data under the Term Sheet and this Data Protection Protocol, including the GDPR;

   "**DPIA**" means a data protection impact assessment in accordance with the GDPR;

   "**GDPR**" means Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation);

   "**Governance Charter**" means the charter that sets out the functions, roles, and responsibilities assigned to the various participants to the GPT-NL Project.

   "**TNO**" means the *Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek* (Netherlands Organization for Applied Scientific Research), a legal entity established by public law with its principal place of business at The Hague, Netherlands;

   "**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 the GPT-NL Model in accordance with the Training Content Protocol;

   "**Responsible Use Policy**" means the policy that shall set out the permitted uses of the GPT-NL Model and the restrictions and limitations that will apply to both business and academic use;

"**Training Content Protocol**" means the protocol that sets out the requirements to be followed by the Content Contributor when preparing Training Content for the GPT-NL Model. The Training Content Protocol shall provide the objective norms, tools, and/or protocols that must be followed or applied for the compliance of (the development of) the GPT-NL Model with applicable laws; and

"**controller**", "**data subject**", "**personal data**", "**personal data breach**", "**processing**", "**processor**", and "**supervisory authority**" have the meaning given to them under the GDPR.

Scope; Role of the Parties

2. **Scope**. The obligations under this Data Protection Protocol only apply to the Parties to the extent that their processing of Raw Content, Contributor Training Content, TNO Training Content or the Prepared Dataset contains personal data subject to Data Protection Laws.

3. **Role of Content Contributors:**

   a. Each Content Contributor qualifies as an independent controller for its processing of personal data as part of the (i) collection and preparation of Raw Content and transfer of such Raw Content to TNO; and (ii) any subsequent processing of the Contributor Training Content for Content Contributor's own purposes outside the GPT-NL Project.

4. **Role of TNO:**

   a. TNO qualifies as an independent controller for its processing of personal data (i) when cleaning Raw Content to generate Contributor Training Content; (ii) to generate the TNO Training Content; (iii) to generate the Prepared Dataset using TNO Training Content and Contributor Training Content; (iv) to train the GPT-NL Model using the Prepared Dataset; and (v) to operate and use any of TNO's own instance(s) of the GPT-NL Model.

5. **Role of Licensees**

   a. Each Licensee qualifies as an independent controller for its processing of personal data in the context of its operation and use of any instance(s) of the GPT-NL Model for which it has obtained a Research License or Professional License.

Allocation of the Parties' responsibilities

6. Each Party shall comply with its respective controller obligations under applicable law (including but not limited to the Data Protection Laws), on the understanding that in respect of the Parties' processing of Raw Content, Contributor Training Content, TNO Training Content, and the Prepared Dataset in the context of the GPT-NL Project, the controller obligations are allocated as set out in Sections 8 through 10 of this Data Protection Protocol.

7. Each Content Contributor is responsible for:

   a. **Informing data subjects.** Informing data subjects about the processing of their personal data in the Raw Content and the Contributor Training Content and the division of controller obligations in respect of the GPT-NL Project either by using the GPT-NL Information Statement or by including this information in the Content Contributor's own privacy statement.

   b. **Cleaning training content.** Ensuring that the Raw Content is cleaned in accordance with the Training Content Protocol before transferring such Raw Content to TNO.

   c. **Compatible use or legal bases.** Establishing and documenting the compatible use assessment and/or the legal bases for processing Raw Content and the transfer of such Raw Content to TNO for the purposes of training the GPT-NL Model, taking into account the mitigating measures that are implemented by TNO as documented by TNO in accordance with its DPIA for the GPT-NL Project.

d. **Record-keeping and DPIAs.** Maintaining a record of the processing activities of the Raw Content and the transfer of such Raw Content to TNO and performing and documenting a DPIA (where required under Data Protection Laws).

e. **Security.** Implementing appropriate technical, physical and organization security measures to protect the preparation and provision of Raw Content to TNO.

f. **Responding to data subjects' requests.**

   i. Responding to complaints or requests of data subjects ("**DSRs**") in relation to the Raw Content and the Contributor Training Content;

   ii. As soon as possible, but no later than two (2) weeks after receipt of such DSR inform TNO of the DSR and the personal data to which it pertains, enabling TNO to take proportionate and appropriate measures to ensure any legitimate DSRs are complied with, for example by removing the relevant personal data from the Contributor Training Content and the Prepared Dataset (if these data sets will be re-used for further training of the GPT-NL Model), ensuring such personal data will not be present in future Contributor Training Content and the Prepared Dataset, or requesting Licensees to implement safeguards to prevent such personal data being included in output of the GPT-NL Model.

g. **Requests of public authorities**. Responding to requests of supervisory authorities, in accordance with Section 11 of this Data Protection Protocol, in relation to the processing of Raw Content and Contributor Training Content for the Content Contributor's own purposes.

h. **Assisting TNO**. Assisting TNO where necessary to (i) carry out a DPIA for the processing of the Contributor Training Content within the Prepared Dataset in relation to the GPT-NL Project, and (ii) otherwise achieve compliance with Data Protection Laws.

i. **Personal data breaches**. Responding to personal data breaches affecting the Raw Content and the Contributor Training Content processed for its own purposes in accordance with Section 11-13 of this Data Protection Protocol.

8. <u>TNO</u> is responsible for:

a. **Legal bases.** Establishing and documenting the legal bases for its processing of Raw Data, Contributor Training Content, TNO Training Content, and the Prepared Dataset for purposes of training the GPT-NL Model.

b. **Security.** Implementing appropriate technical, physical and organization security measures to protect the processing of the Raw Data, Contributor Training Content, TNO Training Content, and the Prepared Dataset (including adequate access controls) against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure, or access, and against all other forms of unlawful processing.

c. **Privacy-by-design.** TNO will implement *privacy-by-design* measures to reduce the risk of unnecessary access of TNO staff to the Raw Content provided by a Content Contributor, which measures include: (i) no staff other than the Data Curation Staff has access to Raw Content; (ii) after cleaning the Raw Content, the Data Curation Staff will provide a copy of the Contributor Training Content to both the relevant Content Contributor and the staff of the GPT-NL team tasked with training the GPT-NL Model; (iii) 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 any Raw Content; (iv) after completion of the pre-training phase (i.e. completion of the pre-trained GPT-NL Model v. 1.0 or any successor pre-training version of the GPT-NL Model) the Data Curation Staff will delete the Raw Content; and (v) any Contributor Training Content, TNO Training

Content and Prepared Datasets are retained by TNO for a period of 7 years after completion of the GPT-NL Project..

d. **Processors**. Engaging SURF or other processors for purposes of hosting and compute or other processing activities in the context of the GPT-NL Project, using appropriate contractual terms, including – in the case of cross-border transfers of personal data – ensuring adequate safeguards and transfer mechanisms in accordance with Data Protection Laws.

e. **Record-keeping and DPIAs.** Maintaining a record of the processing activities of Raw Content, Contributor Training Content, TNO Training Content and the Prepared Dataset for purposes of training the GPT-NL Model and performing and documenting a DPIA in respect thereof.

f. **Providing notice to data subjects and responding to DSRs.**

   i. Informing data subjects of TNO's processing of personal data included in Raw Content, Contributor Training Content, TNO Training Content and the Prepared Dataset for purposes of training the GPT-NL Model and the division of controller responsibilities in respect thereof by publishing the GPT-NL Information Statement on the website of the GPT-NL Project and further notices and policies issued by TNO and updated from time to time.

   ii. Notifying the relevant Content Contributor or Licensee as soon as possible but no later than two (2) weeks following receipt of any DSR relating to Contributor Training Content, or the use of a GPT-NL Model by a Licensee and not responding to such DSR, except to redirect the relevant data subject to the relevant Content Contributor or Licensee.

   iii. Taking proportionate and appropriate measures to ensure any DSRs that are considered legitimate by the relevant Content Contributor or Licensee are complied with, for example by removing the relevant personal data from the Contributor Training Content and the Prepared Dataset (if these data sets will be re-used for further training of the GPT-NL Model), ensuring such personal data will not be present in future Contributor Training Content and the Prepared Dataset, requesting Licensees to implement safeguards to prevent such personal data being included in output of the GPT-NL Model; or ensuring compliance with such DSRs when training a new version of the GPT-NL Model.

g. **Personal data breaches**. Responding to personal data breaches affecting Raw Content (when in possession of TNO), Contributor Training Content, TNO Training Content and the Prepared Dataset in accordance with Sections 11 -13 of this Data Protection Protocol.

h. **Requests of public authorities**. Responding to requests of supervisory authorities, in accordance with Section 11 of this Data Protection Protocol, in relation to the processing of Raw Content, Contributor Training Content, TNO Training Content and the Prepared Dataset.

9. In addition to the obligations set out in Section 7 above, each Licensee is responsible for:

   a. **Transparency.** Informing data subjects about the processing of their personal data in the context of its operation and use of any instance(s) of the GPT-NL Model.

   b. **Legal basis.** Establishing a legal basis for the processing of personal data in the context of its operation and use of any instance(s) of the GPT-NL Model.

   c. **Record-keeping and DPIAs.** Maintaining a record of the processing activities data in relation to the Licensee's operation and use of any instance(s) of the GPT-NL Model and performing and documenting a DPIA (where required under Data Protection Laws).

d. **Security.** Implementing appropriate technical, physical and organizational security measures to protect any personal data processed in the context of the Licensee's operation and use of any instance(s) of the GPT-NL Model; and

e. **Responding to DSRs.**

   i. Responding to DSRs in relation to the Licensee's operation and use of any instance(s) of the GPT-NL Model, including - where proportionate and appropriate - implementing safeguards to prevent personal data from being included in outputs produced by the Licensee's instance of the GPT-NL Model.

   ii. Notifying TNO as soon as possible but no later than two (2) weeks following receipt of any such DSR;

   iii. Where so requested by TNO, implement safeguards to prevent personal data from being included in outputs produced by the Licensee's instance of the GPT-NL Model.

f. **Requests of public authorities**. Responding to requests of supervisory authorities, in accordance with Section 11 of this Data Protection Protocol, in relation to the processing of personal data in the context of the Licensee's operation and use of any instance(s) of the GPT-NL Model.

g. **Assisting TNO**. Assisting TNO where necessary to achieve compliance with Data Protection Laws.

h. **Personal data breaches**. Responding to personal data breaches in relation to the Licensee's use of the GPT-NL Model in accordance with Section 11-13 of this Data Protection Protocol.

Requests from Authorities

10. With regard to requests from supervisory authorities, including judicial authorities and law enforcement (each, an "**Authority**"), the Parties agree in addition to the following:

a. If TNO receives a request from an Authority with regard to Contributor Training Content obtained by TNO, TNO will inform the relevant Content Contributor of the request as soon as possible but no later than three (3) business days, unless it is prohibited to do so under applicable law.

b. If a Content Contributor or Licensee receives a request from an Authority with regard to Raw Content or Contributor Training Content of its use of the GPT-NL Model under a Professional License, the Content Contributor/Licensee will inform TNO of the request as soon as possible but no later than three (3) business days, unless it is prohibited to do so under applicable law.

c. If necessary, the Parties will assist each other in collecting the information required to handle the request from the Authority.

d. If the execution of the Authority's request has consequences for, or impacts, more than one Party, the request will be handled jointly and by mutual agreement between the affected Parties (each acting in good faith). In that case, the affected Parties will jointly prepare a response, without prejudice to each other's individual responsibility under this Data Protection Protocol and applicable law.

Personal data breaches

11. **Internal notification.**

a. Content Contributors. Upon becoming aware of a personal data breach affecting the Contributor Training Content, the Content Contributor shall promptly (and in any event within 24 hours) inform TNO.

b. Licensees. Upon becoming aware of a personal data breach in relation to the GPT-NL Model, the Licensee shall promptly (and in any event within 24 hours) inform TNO.

c. TNO. Upon becoming aware of a personal data breach affecting Contributor Training Content obtained by TNO or personal data included in Prepared Dataset that can be linked back to Contributor Training Content, TNO shall promptly (and in any event within 24 hours) inform the Content Contributor of whom TNO received the relevant Contributor Training Content.

12. **Notification to supervisory authorities/data subjects.**

    a. If required under applicable law, the affected Content Contributors are responsible for notifying competent supervisory authorities and impacted data subjects of a personal data breach affecting their Raw Content (before being transferred to TNO) and the Contributor Training Content (as used for their own purposes).

    b. If required under applicable law, the affected Licensees are responsible for notifying competent supervisory authorities and impacted data subjects of a personal data breach affecting their instance of the GPT-NL Model.

    c. If a personal data breach affects the Raw Content (as in the possession of TNO), the Contributor Training Content, the TNO Training Content or the Prepared Dataset, TNO is responsible for notifying the competent supervisory authorities and data subjects, to the extent such data subjects can be identified.

13. **Remedial measures.** If a personal data breach pertains to personal data or an instance of the GPT-NL Model maintained or operated by a Party, such Party shall take appropriate remedial measures in response to the personal data breach.

Information and assistance

14. The Parties will inform each other and provide each other with assistance reasonably required to ensure compliance with the relevant obligations under Articles 32 to 36 of the GDPR, as well as other requirements applicable to the Parties under Data Protection Laws.

Notices

15. All notices under this Data Protection Protocol must be sent to the following persons:

    TNO: privacy@gpt-nl.nl

    Licensee and/or Content Contributor: as set out in the underlying Content Contributor Agreement.

Amendment of the Data Protection Protocol; Duration of the Data Protection Protocol

16. This Data Protection Protocol can be amended in accordance with the Governance Charter. The version number and date of amendments shall be documented.

17. This Data Protection Protocol is in effect for as long as the Parties process Raw Content, Contributor Training Content, TNO Training Content, and/or the Prepared Dataset in the context of the GPT-NL Project.

Annex 6

Revenue Sharing Mechanism

GPT-NL

Revenue Sharing Mechanism

Version 1.0

Introduction

TNO, the Netherlands Organization for Applied Scientific Research, has been awarded a grant from the Dutch government to develop a *state-of-the art* research infrastructure for training large language models that comply with European and Dutch legislation and public values (**LLMs**). This research facility will be kept operational for three years. The design documentation and operating software will be publicly made available to enable third parties to set up their own LLM training environments to serve the national public interest (**Research Facility**). Deliverable of the GPT-NL project will further be the creation of at least one competitive Dutch LLM that complies with European legislation and public values (**GPT-NL Model**). These activities are hereafter referred to as the "**GPT-NL Project**".

TNO is responsible for delivery of the GPT-NL Project. TNO cooperates for the project with SURF, the IT cooperation for education and research, and the Netherlands Forensic Institute (**NFI**). TNO has set up a consortium of organizations that are willing to participate in the GPT-NL Project by contributing relevant content for purposes of initial training (and potential later updating) of the GPT-NL Model (**Content Contributors**). As a participant in the GPT-NL consortium, Content Contributors will be involved in important decisions relating to the licensing and future of the GPT-NL Model and compensated for their contribution of training content. For this purpose, TNO has set up a project governance as described in the Governance Charter (as defined below).

The GPT-NL Project is a research project. The initial part of the GPT-NL Project is funded from the grant and will deliver the Research Facility, and a first training run of the GPT-NL Model. To deliver a compatible GPT-NL Model, however, multiple development iterations will be required. To fund this further development of the GPT-NL Model, income must be generated via licenses (or further grants).

TNO, being an independent administrative body and having obtained a government grant in relation to the GPT-NL Project, is subject to state aid limitations under EU and Dutch law. As such, any licensing of the GPT-NL Model will have to be in accordance with market-standard terms, conditions, and prices.

As compensation for the contribution of their copyrighted content and related efforts and costs, Content Contributors can opt for one of three options: 1) no compensation is claimed, 2) a proportionate share of 50% of the Net Revenues (as defined below) generated with the GPT-NL Model <u>and</u> a proportionate discount on their professional license fee, or 3) a one-time upfront compensation <u>and</u> a proportionate discount on their professional license fee. All compensation options are calculated using the same data value computation scheme, which is based on both the quantity and quality of the contributed training content. This ensures fairness and transparency in the distribution of funds. t

Note that TNO will retain 50% of Net Revenues generated with the GPT-NL Model. Under relevant state aid limitations, TNO may exploit the GPT-NL model in a non-profit fashion only. TNO's share of the Net Revenue will be re-invested to cover the cost of the further development and maintenance of the GPT NL Model and the Research Facility (**Not-for-Profit Purposes**).

This document discusses the types of licenses of the GPT-NL model, explains the different options for compensation of Content Contributors for contributing their training content, and outlines how the data value of such training content is to be calculated.

Definitions

In addition to the terms defined above, capitalized terms used in this document have the following meanings:

Contributor Training Content: Raw Content that has been prepared by TNO in accordance with the Training Content Protocol.

Data Curation Process: the process applied to curate Raw Content and turn it into Contributor Training Content, in accordance with the Data Curation Specifications (as defined in Section 2 of the Training Content Protocol).

Data Evaluation Process: the evaluation process applied by TNO to evaluate Contributor Training Content as described in the Content Preparation Process (as defined in Section 2 of the Training Content Protocol).

Data Value: the value of the datasets provided by a Content Contributor as determined by TNO in accordance with section 2 in this document.

Net Revenues: the total license fees generated from the Professional Licenses (or other future paid license types) in respect of a specific version (or finetuned version or other derivatives of such specific version) of the GPT-NL Model, after deduction of (i) any discounts on license fees or one-time compensations granted or paid to Content Contributors under options 2 and 3 of Section 2, and (ii) taxes and costs directly attributable to the collection of the license fees, including transaction costs, administrative costs, compliance costs, recovery costs or the fee charged by an organization that performs this activity on behalf of GPT-NL.

Prepared Dataset: 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.

Raw Content: the content that Content Contributor intends to provide to TNO for the purposes of training and future updating of the GPT-NL Model, which has not yet been prepared in accordance with the Training Content Protocol.

Tokenizer: a tool or algorithm that breaks text into smaller units called "**Tokens"**, which are the basic elements that a Large Language Model (LLM) processes and understands. The Tokenizer converts raw text (words, sentences, paragraphs) into a sequence of tokens that the model can interpret and use for training or generating text.

Training Content Protocol: the protocol that describes how Raw Content is curated in order to turn it into Contributor Training Content.

1. Licensing of the GPT-NL model

The GPT-NL model will initially have two separate sets of licensing terms.

1. Research license: A license for research on and with the GPT-NL Model, from here on the '**Research License'**. The Research License is free to use by individual researchers and research institutes for scientific non-commercial research purposes. Researchers are requested to provide their feedback to the GPT-NL Project. The GPT-NL Model will be made available on request.
2. Professional license: for all purposes other than scientific non-commercial research purposes, including commercial purposes ("**Professional Licenses").** The costs of the Professional License will be in line with market prices. Organizations that want to use the Professional License must enter into a Professional License with TNO.

TNO can create new license types to benefit usage of the GPT-NL Model in accordance with Section 6.2.3 Content Contributor Agreement. Any such new paid license types will be considered Professional Licenses for purposes of this Revenue Sharing Mechanism.

2. Compensation options

As compensation for the contribution of their copyrighted content and related efforts and costs, Content Contributors can opt for one of four options:

1. **No compensation claimed**

2. **Net Revenues Sharing:** a proportionate share of 50% of the Net Revenues generated with the Professional Licenses calculated in accordance with the principles set out in section 2.1; **and**

   Discount Professional License Fee: Content Contributor's proportionate share of 50% of the Net Revenues (generated with the Professional Licenses, calculated in accordance with the principles set out in Section 2.1 of the Revenue Sharing Mechanism will be set-off against up to 100% of the license fee due and payable by Content Contributor for Content Contributor's Professional License under the applicable license agreement with TNO. For the purpose of calculating the Net Revenues, the full license fee due for such Professional License before discount will be taken into account. If a Content Contributor's proportionate share of 50% of the Net Revenues is 40% or more, such Content Contributor will be entitled to a free Professional License.

   If at any moment in time other paid license types become available, the discount will apply against the license type relevant for Content Contributor. If a Content Contributor has more paid license types, the discount will apply against one of the licenses only, at the choice of Content Contributor. The discount will not be applicable to any reseller agreements in respect of the GPT-NL Model.

3. **One Time Upfront Compensation:** Instead of receiving a proportionate share of 50% of the Net Revenues, Content Contributor will receive an upfront one-time payment that is calculated and due and payable as set forth in section 2.1.5 of the Revenue Sharing Mechanism **and** a discount on the Professional License Fee on the terms set out in Option 2.

If Content Contributors opt for compensation option 1 (no compensation), their Relative Data Value claim in respect of the Net Revenues will be awarded to TNO, meaning TNO receives more than 50% of the Net Revenues. This additional share of the Net Revenues will also be re-invested by TNO to cover the cost of the further development and maintenance of the GPT NL Model and the Research Facility (**Not-for-Profit Purposes**).

2.1. Net Revenues sharing with Content Contributors

To reward Content Contributors that have helped making the GPT-NL Model possible, they can opt for a compensation by means of a proportioned share of 50% of the Net Revenues in accordance with the following principles:

- Any Content Contributor that contributes a dataset and that is used as part of the Prepared Dataset for the GPT-NL Model can claim its share of the Net Revenues of the GPT-NL Model.
- The Content Contributors that choose to claim their share will be compensated with a proportioned part of 50% the Net Revenues calculated in accordance with this section 2.
- The Data Value of an individual dataset has multiple components, which are explained in section **Error! Reference source not found.**.
- The share that a Content Contributor can expect is proportionate to the Data Value of the individual dataset as a relative percentage of the Data Value of the Prepared Dataset.
- Every new GPT-NL Model with a Professional License will have its own revenue sharing calculation.

2.1.1. Individual Dataset Value Determination

For every individual dataset contributed by a Content Contributor, the GPT-NL Team will determine a 'Data Value' based on Quantity and Quality of the dataset. The size of a dataset will be measured in Tokens because Tokens represent the actual workload the model processes, in other words, how much data the model learns from. This metric is more precise than counting words or characters. However, not all Tokens are equal: A dataset's worth also depends on factors such as the relevance, uniqueness and cleanliness of the texts.

When the full GPT-NL training-set is complete (**Prepared Dataset**), every dataset received from the Content Contributor will receive a 'Relative Data Value' (RDV) between 0.0 and 0.5.

$$d\ = individual\ dataset \qquad [1]$$

$$DV_d = Quantity\ Dataset(Qn_d) \cdot Quality\ Dataset\ (Ql_d)\ \ \cdot Information\ about\ Dataset\ (I_d) \qquad [2]$$

$$RDV_d\ = 0.5 \cdot \frac{DV_d}{\sum DV} \qquad [3]$$

The Relative Data Value determines the amount the Content Contributor is entitled to for the relevant dataset in respect of the 50 % of the Net Revenues and is used when a Content Contributor has opted for either the Net Revenues Sharing option or the Discount on the Professional license of the GPT-NL model.

$$Claim_d(€) = RDV_d \cdot\ Total\ Revenue\ (€) \qquad [4]$$

2.1.2. Dataset Quantity

The quantity of an individual dataset is determined by the number of Tokens that ends up in the Prepared Dataset. This means that it is only possible to receive monetary value for the part of the dataset that is useful for training the model. Furthermore:

- All the data that is received via the Data Curation Process will be marked with metadata describing the source of the data. From the resulting set of Tokens at the end of the Data Curation Process the data quantity of every individual source is determined.
- The Data Curation Process involves removing low quality data, incorrect or old data and deduplication (for more information see the Training Content Protocol).

- At the end of the Data Curation Process, Training Content data from all sources gets pooled and the GPT-NL Team will do deduplication on the full set. If there is data in a set of a Content Contributor that looks very similar to data in the set of another Content Contributor, the data from one of the sets gets deleted. In this event, the data for which a Content Contributor Agreement was signed latest, will get deleted. The Tokens of the deleted set will not count for Token quantity. If content of a Content Contributor is also included in the TNO Training Content, no value is attributed to such content of the Content Contributor because the content was already available in the public domain under an open-source license.
- At the end of the Data Curation Process, all data will go through a custom-made GPT-NL Tokenizer. Therefore, only after the curation process has run and the Prepared Dataset is complete, the absolute number of Tokens for an individual set is known.
- To give Token estimates for an individual dataset before that has happened, we estimate that every word roughly equates to 1.5 Tokens.
- Currently, we estimate to get 40B Tokens from Content Contributors. This can be used as a reference to estimate the proportional amount that the Content Contributor is entitled from the 50% of Net Revenues.

2.1.3. Dataset Quality

Not every Token of data is equally valuable for training an LLM. The quality of the data will be determined in two ways:

1. Via the answers given in the Training Content Deep Dive Survey.
2. Via the Data Evaluation Process conducted done at the end of the Data Curation Process.

Technical Context - Oversampling

The quality of a dataset depends on the oversampling multiplier used. In LLM training, oversampling involves repeating the same Tokens multiple times. In the case of GPT-NL, high-quality Tokens can be repeated up to roughly 10 times for optimal performance. However, repeating low-quality Tokens may amplify biases and is less desirable.

Quality Calculation

Team GPT-NL will assess dataset quality on the aspect given in **Error! Reference source not found.**. The metrics in this table will determine how much we will oversample (part of) a given dataset. For every dataset we will calculate a 'Quality Score', which is the total number of Tokens originated from a specific dataset (including Tokens sampled multiple times) divided by the number of Tokens from unique text originated from a specific dataset (5).

$$Ql_d = \frac{Qn_d(incl.\,oversampling)}{Qn_d} \qquad [5]$$

This 'Quality Score' will always be higher than one and will not likely be higher than 7 for very high-quality sources.

In Table 1 we describe how certain aspects of datasets influences how much the data is going to be up-sampled for the GPT-NL training set. This table will give an idea of the Quality Score of your data.

Aspect	Motivation	How do we measure
Topics (T)	Some topics are more relevant than others. The first iteration of GPT-NL will mostly be applied in work settings. Important sectors for the GPT-NL model are e.g. healthcare, governmental, education, law etc. We are also mostly looking for fact-driven sources.	Deep Dive survey input, verified by Text Analysis in Data Evaluation Stage
Risk Profile (RP)	We will do a risk analysis of the dataset in the evaluation stage. Parts of a dataset can be determined to be more risky than other. (Parts of) the dataset will be classified as low or high risk.	Risk Analysis in Data Evaluation Stage
Recency (R)	Recent Data has a much higher chance of still being relevant.	Preferably based on granular metadata provided in individual data articles. Second option is by description in Deep Dive survey.
Perplexity (F)	Data with a higher perplexity score is an indication for a better written professional text.	Text Analysis in Data Evaluation Stage.

Table 1: Aspects of GPT-NL Data Quality

2.1.4. Information about Dataset

Regardless of the quality of the dataset, following the commitments of GPT-NL and the EU AI Act, it is important for us to give accurate information about data in the set, also if the contents of that set are not openly available. The questions included in the Deep Dive Survey, which accompany the Term Sheet, are essential for ensuring transparency regarding individual datasets. GPT-NL team members will interview Content Contributors based on the answers provided in deep dive survey.

Measured with: Deep Dive Survey			
Value Multiplier	1.0	1.0 – 1.25	1.25

Requirements	10 or more (sub)questions in the deep dive survey are not sufficiently answered.	For every (sub)question in the deep dive survey not sufficiently filled in (there is one of more questions from team GPT-NL left unanswered), the multiplier gets deducted by 0.025.	Dataset metadata is sufficiently filled in in the deep dive and all follow-up questions are answered.

Table 2: Deep Dive Survey multiplier

2.1.5. One-time upfront compensation Calculation

You can decide to opt for a one-time upfront compensation instead of opting for the revenue sharing model. The exact compensation that you can expect will roughly be in accordance with formula 8, stating that for every 1 billion tokens, you can receive upfront compensation of €333,33 times the Relative Quality Score (RQl_d) of your dataset. The Relative Quality Score will always be a number between 1.0 and 10.0 (equation 7).

$$Ql_{max} = upsample\ rate\ for\ the\ highest\ quality\ data\ in\ the\ GPTNL\ set \qquad [6]$$

$$RQl_d = \max\left(1.0\,, \frac{8 \cdot Ql_d \cdot I_d}{Ql_{max}}\right) \qquad [7]$$

$$Claim_d(€) = (333.33\ \cdot RQl_d)\ for\ every\ 1B\ tokens \qquad [8]$$

It follows the expected upfront compensation will be between 333€ and 3333€ per 1 billion tokens.

<u>For example:</u>

A very high-quality dataset gets up-sampled maximally, so $Ql_d = Ql_{max}$.

Information for all questions in the Deep Dive Survey is sufficiently answered, so $I_d = 1.25$.

$$RQl_d = \max\left(1.0, \frac{8 \cdot Ql_{max} \cdot 1.25}{Ql_{max}}\right) = 10$$

$$Claim_d = (333.33\ \cdot 10) = €3333.33\ for\ every\ 1B\ tokens$$

Annex 7

License Restrictions & Other Terms

License Restrictions & Other Terms

A. License Restrictions

The license terms of the Professional Licenses are subject to the following restrictions:

1. The GPT-Model may not be reverse engineered, decompiled, or disassembled, whether to extract the training content of the GPT-NL Model or otherwise.
2. The GPT-NL Model may not be used to generate synthetic training content for other LLMs.
3. The GPT-NL Model may not be rented, leased, sublicensed or made available for use in any other way to any third party.
4. The GPT-NL Model may not be used for any unlawful purpose or in violation of any applicable laws (including but not limited to uses that qualify as prohibited AI practices under the EU AI Act).
5. The GPT-NL Model may not be used as the basis for a high-risk AI system under the AI Act, without the prior consent of TNO.
6. The GPT-NL Model may only be used in combination with any other data to generate output of the GPT-NL Model to the extent that such use would not infringe any third-party rights in and to such other data, including but not limited to Intellectual Property Rights and privacy rights (e.g., by including such other data in the respective user prompts or by using technical methods such as retrieval augmented generation in connection with use of the GPT-NL Model). In case of uncertainty due to pending litigation about the lawfulness of any use of Intellectual Property Rights, this clause shall be interpreted in a way that efficiently protects Intellectual Property Rights.
7. Without limiting the general provision in Section 6,

   A. licensee shall not, directly or indirectly, use the GPT-NL Model, or any agent, tool, plugin, interface, or technical pipeline that integrates or combines the GPT-Model with any content that:

      (i) infringes Intellectual Property Rights applicable to that content, for example because it has been obtained through unauthorized or unlawful means, e.g. is not lawfully accessible for the purposes of text and data mining under Article 4 of Directive (EU) 2019/790.

      (ii) was obtained through automated scraping, crawling, or harvesting from websites or digital services where the rightsholder has implemented opt-out mechanisms, including but not limited to robots.txt, HTTP headers, metadata, or other technical protocols, and/or where the website's terms of service or legal notices explicitly prohibit automated access or reuse of content for AI training or processing (including plain-language visible rights declarations).

B. licensee shall not deploy or integrate the GPT-NL Model into any system, agent, or pipeline that uses or relies on third-party tools, APIs, or services that ingest or process non-compliant or opt-out-protected content; circumvents or ignores opt-out signals or technical restrictions imposed by content providers; or enables downstream users to indirectly benefit from such prohibited content. This prohibition extends to any future technological method, whether known or unknown at the time of this Agreement, that enables the ingestion, transformation, or synthesis of content derived from non-compliant sources.

8. GPT-NL Model may not be used to provide Media Integrator Services to any third party. "**Media Integrator Services**" mean any services that retrieve, analyze, process, summarize and/or display in the GPT-NL Model's output to any user (whether upon a user's request or otherwise) any content of any press publication (in the sense of the DSM Directive) of a third party (whether as stand-alone content or in combination with any other content retrieved from other sources) that has been published online (regardless whether publicly available on a website or only accessible through a paywall) within six (6) weeks preceding such use in connection with the GPT-NL Model. Examples of Media Integrator Services include:

(i) providing media updates;

(ii) providing input relevant for compliance purposes such as know your customer services;

(iii) to enrich services such as enabling house for sale websites to enrich their offering with a safety score for the relevant neighborhood;

(iv) synthetic news generation, news aggregation, media monitoring of any type (e.g. adverse media monitoring); or

(v) any other service or activity that would normally require a license for up-to-date content from copyright, neighbouring right and/or database right holders.

A separate license type may ultimately be made available for the use of the GPT-NL Model for purposes of Media Integrator Services.

B. Other license terms

TNO will ensure that the license terms will in any event require licensees:

1. To comply with the Technical Requirements to securely host their own version of the GPT-NL Model;
2. To implement logging software to log any use of the GPT-NL Model;
3. To report to TNO any such use in accordance with the instructions of TNO;
4. To implement any safeguards as required by TNO to prevent any Contributor Training Content being included in output of the GPT-NL Model in order to address any claims of copyright infringement in a similar manner as is provided for in respect of complaints of data subjects in accordance with the Data Protection Protocol;
5. To implement new versions of the GPT-NL Model issued by TNO ultimately within one year of the date such new version was issued, whereby the license will expire if this is not done; and
6. To accept audits by TNO (whether at the request of Content Contributor or otherwise) to review compliance with any of the terms in this Annex 7 including the Technical Requirements.

The license terms will further include notice that by accepting the license, the licensee confirms its understanding that content from third-party providers on which the GPT-NL Model is trained constitutes very valuable data for TNO and such third-party providers, that any breach of the license terms may cause considerable damage, and will provide that the licensee will immediately notify TNO of any such breach.