GPT-NL Data Pipeline

Important: This schematic is not final yet. The steps outlined here will be the general steps to take, but not all details are set in stone
~ 10-06-2024

This schematic aims to explain the process of data collection for the GPT-NL project. This schematic is also useful to check documents that are relevant to you. The link for all materials will redirect to the location where the document is published. Most documents that are not accessible yet are in draft or review phase.

If you have any questions. Don't hesitate to ask on info@gpt-nl.nl

First Contact

You get to know the GPT-NL project

Through presentations, social media, email or our website you get to know the GPT-NL project and our ambitions. Our data consultants can tell you about our vision and the benefits of joining the effort.

gpt-nl.nl | LinkedIn | Commitments | Data Pipeline

Viability Check

Viability Survey

To check whether the data you own can be used for the project, please fill in the the viability survey as much as you can. We need to know how much personal or harmful information is in your dataset.

We have Data Collection Viability Officers that are responsible for evaluating the potential of your data for the GPT-NL project based on your surveys response. They will contact you if there are any open questions about the filled in survey.

Agreement

Deep Dive Questionnaire | Term Sheet | Licensing

Together we create a plan on how and when we are going to filter out harmful data and we make an agreement on when the data will be transferred. The term sheet outlines the terms of use for the data. You will be informed about licensing terms of the GPT-NL model for you and for third parties. You will also be asked to fill in a questionnaire that dives into the origin and prior processing steps the data went through to realize full data transparency.

The agreement is made in consultation with the Data Agreement Officers from the GPT-NL team. One of 5 data cleaning scenarios is decided on based on input from the viability survey and a data is picked for when the questionnaire and data will be delivered.

Data Clean

Definitions of harmful data | Data Cleaning modules with documentation

If there is a need for filtering out certain data, we will do that according to the data processing agreement made.

Depending on the scenario, a member of the GPT-NL data curation staff will assist with data cleaning activities.

Data Cleaning scenarios (in order of preference)

1. No Data Cleaning Activities necessary
2. Data Cleaning is done by Data Contributor
3. Data Cleaning is done by Data Contributor with the help of tools from team GPT-NL.
4. Data Cleaning is done by a GPT-NL staff member at the premises of the Data Contributor
5. Raw data is transferred immediately and cleaned on GPT-NLs premises

Data Transfer

Data Transfer Guide | Data Management Plan

The data gets transferred to the storage unit of GPT-NL following the agreed upon principles

The Data collection Team will ensure the data will be stored on the SURF cluster along with the filled in checklists, surveys, metadata and classification report. The Data Transfer officer can answer all your questions about this process.

Evaluation

Dataset Classification Report Code | Data Evaluation Checklist

After the data is transferred, we do one final check. The GPT-NL team determines whether we can use the set based on the output of the deep dive questionnaire and a generated classification report with details about the amount of harmful data in the set.

Our Data Evaluation Officers are responsible for checking whether the data can be used in the current state and will (a) accept the data. (b) ask for modifications or (c) reject the data completely

Modifications to the Agreement

Description for currently unpublished documents

- Data Pipeline: a document outlining this schematic pipeline in more detail.
- Deep Dive Questionnaire: a survey that asks questions about the origin and processing steps that the original data went through. We ask data providers to fill this in to create transparency about the data and to comply to the European AI act.
- Term Sheet: a document outlining the general terms & conditions for data usage and management. Includes information about licensing of the GPT-NL model. Can be adjusted via the content board.
- Governance Charter: Document outlining how the content board will function.
- Data Cleaning modules: Open-source code which we use to filter unwanted data, including guides on how to run them.
- Definitions of Harmful data: Explanations for the decisions we made with respect to what type of data we consider harmful to use with respect to e.g. privacy, offensive language, etc.
- Data Management Plan: GPT-NLs protocol with respect to managing/updating/deleting/etc. data of data providers.
- Data Transfer Guide: Practical guide about transferring data to GPT-NLs servers.
- Data Classification Code: Open-source code which we use to semi-automatically generate a report about how "clean" the dataset is.
- Data Evaluation Checklist: Checklist to determine whether the data is ready to use as training data for GPT-NL.

Content Board

Governance Charter

The content board is there to provide a forum for data contributors to make broadly supported changes to the general terms & conditions. The rules are outlined in the Governance Charter.