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).

Definitions

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. TNO 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.