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 and a proportionate discount on their professional license fee, or 3) a one-time upfront compensation and 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.

For example:

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 $I_d = 1.25$.

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

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