Annex 4

Baseline Responsible Use Policy

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;