Deelname
GPT-NL Project

De Nederlandse overheid heeft aan TNO (*de Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek*) een onderzoeks- en innovatie subsidie toegekend om in het nationaal belang een Nederlandstalig groot taalmodel te trainen dat voldoet aan Europese en Nederlandse regels en publieke waarden (*lees meer over het GPT-NL Project*). Voor het trainen van het GPT-NL Model zijn veel Nederlandstalige teksten nodig. Vele overheidsinstanties, bibliotheken, archieven en uitgevers hebben hiertoe hun content aan TNO ter beschikking gesteld. Ook wij stellen content beschikbaar aan TNO om het GPT-NL Model mogelijk te maken.

In de content die we aanleveren aan TNO zitten mogelijk persoonsgegevens. Om uw privacy te waarborgen wordt de content door ons op goed beveiligde wijze aangeleverd aan TNO, die de content vervolgens bewerkt om deze geschikt te maken voor training van het GPT-NL Model. Onderdeel daarvan is dat irrelevante persoonsgegevens worden verwijderd en het restant vergaand wordt gepseudonimiseerd. Indien u een publiek persoon bent met een eigen Wikipedia-pagina, dan gelden bijzondere regels.

Lees de Privacyverklaring GPT-NL Project

Privacyverklaring
GPT-NL Project

TNO vindt het belangrijk dat uw privacy goed wordt gewaarborgd. In deze privacyverklaring leggen we uit hoe we persoonsgegevens verzamelen en verwerken in het kader van het GPT-NL Project.

Het GPT-NL Project

TNO (*de Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek*) heeft van de Nederlandse overheid een onderzoeks- en innovatie subsidie ontvangen om in het nationaal belang een soeverein Nederlands taalmodel te ontwikkelen dat voldoet aan Europese en Nederlandse regels en publieke waarden (het "**GPT-NL Model**").

Het GPT-NL Model is bedoeld om LLM-toepassingen in Nederland mogelijk te maken die veilig, transparant en betrouwbaar zijn. Denk bijvoorbeeld aan het automatisch samenvatten van Nederlandse teksten, het beantwoorden van vragen of het herschrijven van teksten.

Digitale soevereiniteit. Met de ontwikkeling van een eigen GPT-NL Model wil de overheid bevorderen dat Nederland minder afhankelijk is van buitenlandse LLM-modellen. Het Nederlands is verder een relatief klein taalgebied en daarmee geen prioriteit voor buitenlandse aanbieders. Doordat het GPT-NL Model wordt getraind op originele Nederlandstalige teksten van hoge kwaliteit, kunnen de finesses van de Nederlandse taal beter worden geborgd.

Wetgeving en publieke waarden. Het GPT-NL Model wordt ontwikkeld in lijn met Europese en Nederlandse regels en publieke waarden, in het bijzonder de privacyregels en de rechten en belangen van auteursrechthebbenden. Bijvoorbeeld wordt het GPT-NL Model uitsluitend getraind op teksten die vrij zijn van auteursrecht of waarvoor een geldige licentie is verkregen. Waar auteursrechtelijk beschermde content wordt gebruikt wordt geborgd dat de rechthebbenden ook een eerlijke vergoeding voor hun bijdragen verkrijgen.

Samenwerking. Voor het trainen van het GPT-NL Model zijn veel Nederlandstalige teksten nodig. Om voldoende originele Nederlandstalige teksten van hoge kwaliteit te verzamelen werkt TNO samen met partijen die over zulke teksten beschikken, zoals overheidsinstanties, bibliotheken, archieven en uitgevers (de **Deelnemers**).

Privacybescherming. Alvorens het GPT-NL Model te trainen, bewerkt TNO de trainingsteksten om deze daarvoor geschikt te maken. Onderdeel daarvan is dat schadelijke en ongepaste inhoud en alle irrelevante persoonsgegevens worden verwijderd. Het restant van de persoonsgegevens wordt vergaand gepseudonimiseerd. Voor publieke personen met een eigen Wikipedia-pagina, gelden bijzondere regels zodat het GPT-NL Model bijvoorbeeld antwoord kan geven op vragen zoals: "*Wie is de koning van Nederland?*"

Op deze wijze wordt het GPT-NL Model op een verantwoorde manier getraind en kan het model veilig worden ingezet bij de overheid, het onderwijs en de wetenschap.

Wie zijn verantwoordelijk voor het verwerken van mijn persoonsgegevens?

TNO is verantwoordelijk voor het GPT-NL Project en de verantwoordelijke voor de verwerking van persoonsgegevens die TNO verzamelt en verwerkt in het kader van het trainen van het GPT-NL Model.

Deelnemers zijn verantwoordelijk voor hun selectie van de trainingsteksten en het veilig aanleveren van deze teksten aan TNO.

Als tegenprestatie voor het bijdragen van hun trainingsteksten aan het GPT-NL Project ontvangen Deelnemers een kopie van hun door TNO opgeschoonde trainingsteksten (zie *Hoe worden mijn persoonsgegevens bewerkt?*) en een korting op hun licentievergoeding voor gebruik van het GPT-NL Model. De Deelnemers zijn zelf verantwoordelijk voor het verdere gebruik van deze opgeschoonde trainingsteksten en het gebruik van het GPT-NL Model als licentienemer.

Wat is het doel van de verwerking?

TNO. TNO verwerkt persoonsgegevens om een Nederlandstalig taalmodel te trainen dat voldoet aan Europese en Nederlandse wetgeving en publieke waarden. Het doel van het GPT-NL is om teksten automatisch samen te vatten, vragen te beantwoorden en teksten te verkorten of herschrijven. Hiervoor is het niet nodig om beeld, video, of geluid te verwerken en het GPT-NL Model wordt alleen met teksten getraind.

Voor Deelnemers is het doel van de verwerking een bijdrage leveren aan het trainen van het GPT-NL Model, en vervolgens zelf ook het GPT-NL Model te kunnen gebruiken en mogelijk verder te trainen voor hun eigen doeleinden.

Uit welke bronnen worden trainingsteksten verzameld?

TNO ontvangt trainingsteksten van Deelnemers. Afhankelijk van de Deelnemer betreft dit teksten uit kranten, tijdschriften, in openbare archieven opgeslagen documenten, en informatie van (semi-)overheidsinstellingen.

TNO verzamelt daarnaast zogenaamde *open source* teksten van het internet. Deze *open source* teksten zijn teksten waar geen intellectuele eigendomsrechten op rusten, zoals teksten van wetten en rechterlijke uitspraken, of teksten waarvan de rechthebbende heeft aangegeven dat ze door anderen vrijelijk mogen worden gebruikt, zoals de teksten op de websites van de overheid en de Europese Unie. TNO hanteert een streng protocol voor het selecteren van open source teksten, en zal ook een overzicht van alle gebruikte bronnen publiceren op de website: https://gpt-nl.nl/.

De door Deelnemers aangeleverde teksten noemen we samen met de *open source* teksten hierna de “ruwe content”.

Hoe bewerkt TNO de ruwe content voordat het GPT NL Model ermee wordt getraind?

Voordat de ruwe content kan worden gebruikt voor het trainen van het GPT-NL Model, past TNO een zorgvuldig opschoningsproces toe. Alle schadelijke, ongepaste, of irrelevante content wordt verwijderd, zoals informatie uit een roddelblad. Het doel van het GPT NL Model is het

automatisch samenvatten, verbeteren of verkorten van teksten en beantwoorden van eenvoudige vragen. Daarvoor is niet nodig dat het GPT-NL Model wordt getraind met informatie over mensen die geen publieke rol hebben ("niet-publieke personen"). Van deze niet-publieke personen worden alle persoonsgegevens uit de ruwe content weg gefilterd of gepseudonimiseerd.

Om vragen te kunnen beantwoorden zoals "*Wie is de koning van Nederland*?", is het wél nodig dat het model informatie bevat over mensen met een publieke rol ("publieke personen"). Denk hierbij aan politici, artiesten of andere personen die een eigen Wikipedia-pagina hebben. Alleen informatie die direct verband houdt met hun publieke rol wordt gebruikt om het GPT-NL Model te trainen. In bepaalde gevallen zal deze informatie ook speciale categorieën persoonsgegevens betreffen, zoals geloof, politieke overtuiging, of gezondheid. Dit betreft dan informatie die in publiek domein is (zoals het feit dat de Paus katholiek is of Mark Rutte lid is van de VVD. Andere gevoelige gegevens zoals contactgegevens, locatiegegevens, bankgegevens worden altijd verwijderd of vervangen door willekeurige voorbeelden.

De opgeschoonde ruwe content noemen we hierna de "**Schone Dataset**". Alleen de Schone Dataset wordt gebruikt voor het trainen van het GPT NL Model.

Een gedetailleerde omschrijving van het opschoningsproces van TNO is opgenomen in het Training Content Protocol, dat beschikbaar is via deze link.

Wat is de grondslag voor verwerking van mijn persoonsgegevens?

TNO is verantwoordelijk voor het uitvoeren van het GPT-NL Project. Op grond van de TNO-wet heeft TNO een wettelijke taak om toegepast technisch-wetenschappelijk onderzoek te verrichten in het algemeen belang. TNO heeft van de overheid een onderzoeks- en innovatie subsidie toegekend gekregen om in het nationaal belang een trainingsinfrastructuur te ontwikkelen en daarmee vervolgens het GPT-NL Model te trainen. De verwerking van persoonsgegevens in het kader van het GPT-NL Project vindt derhalve plaats voor toegepast technisch-wetenschappelijk onderzoek in het algemeen belang.

De verwerking door TNO van persoonsgegevens voor het ontwikkelen van het GPT-NL Model is noodzakelijk voor de vervulling van de wettelijke taak van TNO. De grondslag hiervoor is artikel 6, lid 1, onder e, van de Algemene Verordening Gegevensbescherming (**AVG**): de verwerking is noodzakelijk voor de vervulling van een taak van algemeen belang die aan TNO is opgedragen.

Voor zover TNO *bijzondere categorieën van persoonsgegevens* (zoals politieke voorkeur, geloof of gezondheidsgegevens) verwerkt in het kader van het GPT-NL Project, gebeurt dit op basis van artikel 9, lid 2, onder g AVG, in combinatie met de TNO-wet. Deze verwerking is toegestaan vanwege het algemeen belang bij het ontwikkelen van het GPT NL Model en vindt plaats onder strikte waarborgen ter bescherming van de rechten en vrijheden van betrokkenen.

Voor zover TNO *Strafrechtelijke gegevens* verwerkt in het kader van het GPT-NL Project, gebeurt dit op basis van artikel 10 AVG in samenhang met artikelen 24 en 32 onder f van de Nederlandse Uitvoeringswet AVG ("**UAVG**"), waarin specifieke waarborgen zijn opgenomen voor wetenschappelijk onderzoek.

Een gedetailleerde omschrijving van de waarborgen die zijn opgenomen in het Training Content Protocol, dat beschikbaar is via deze link. De specifieke waarborgen ter bescherming van betrokkenen zijn verder uitgewerkt in de gegevensbeschermingseffectbeoordeling (ook wel *data protection impact assessment* of *DPIA* genoemd) voor het GPT-NL Project, die beschikbaar is via deze link.

Deelnemers zijn zelf verantwoordelijk voor het selecteren van trainingsteksten en het overdragen daarvan aan TNO ten behoeve van de training van het GPT-NL Model. Door de vergaande waarborgen die zijn getroffen door TNO ter bescherming van de privacy (zie hiervoor) zal voor Deelnemers de vertrekking een verdere verwerking van persoonsgegevens betreffen voor wetenschappelijk onderzoek. Artikel 5 lid 1 onder b AVG geeft aan dat deze verdere verwerking is toegestaan, mits voldaan wordt aan de waarborgen voor de verwerking van persoonsgegevens voor wetenschappelijke doeleinden, zoals het minimaliseren en pseudonimiseren van persoonsgegevens. Deze maatregelen zijn door TNO toegepast.

Licentienemers van het GPT NL Model zijn zelf verantwoordelijk voor het bepalen van de grondslag waarop zij persoonsgegevens verwerken. Afhankelijk van het doel waarvoor het GPT NL Model wordt gebruikt kan dit bijvoorbeeld hun gerechtvaardigd belang zijn (artikel 6, lid 1, onder f AVG). De gebruiker van het GPT NL Model zal verdere informatie geven over de grondslag en, als de grondslag gerechtvaardigd belang is, om welk belang het gaat. Door de specifieke waarborgen die TNO treft, borgt TNO dat het model in lijn met de beginselen van dataminimalisatie, privacy-by-design en transparantie wordt ontwikkeld. Hierdoor wordt het mogelijk gemaakt dat toekomstige gebruikers het model op een verantwoorde manier kunnen inzetten, met respect voor de privacy van individuen.

Hoe worden mijn persoonsgegevens beveiligd?

TNO verwerkt zowel de ruwe content als de opgeschoonde versie daarvan (de Schone Dataset) binnen een streng beveiligde infrastructuur. Deze infrastructuur is opgezet conform de geldende normen voor informatiebeveiliging, waaronder ISO 27001. Er worden zowel technische als organisatorische maatregelen genomen om de persoonsgegevens te beschermen tegen verlies, onbevoegde toegang, ongeoorloofde wijziging of andere vormen van onrechtmatige verwerking.

Zo heeft alleen een selecte groep geautoriseerde medewerkers van TNO die deel uitmaken van het *Data Curation Team*, toegang tot de ruwe content. Deze medewerkers hebben een geheimhoudingsplicht en krijgen alleen toegang tot de gegevens voor zover en zolang dat strikt noodzakelijk is voor hun werkzaamheden.

Hoe lang worden persoonsgegevens bewaard?

De ruwe content wordt verwijderd zodra de Schone Dataset is samengesteld en de pre-training fase van het GPT-NL Model is afgerond. De Schone Dataset wordt zeven jaar bewaard. Dit is de wettelijke bewaartermijn die voor TNO geldt met betrekking tot door de overheid gesubsidieerd onderzoek.

Rol van SURF

SURF is de IT-partner van TNO in het GPT-NL Project. SURF stelt de rekenkracht en IT-capaciteit ter beschikking die nodig is om het GPT NL Model te trainen (zie: Snellius: de

Nationale Supercomputer | SURF.nl). SURF verwerkt gegevens uitsluiten ten behoeve van TNO als een 'verwerker'. TNO heeft hiervoor een contract afgesloten met SURF, waarin waarborgen zijn opgenomen om ervoor te zorgen dat de gegevens goed worden beschermd.

Rechten van betrokkenen

Onder de AVG hebben betrokkenen het recht op inzage in de verwerking van hun persoonsgegevens, en verder in bepaalde gevallen het recht op rectificatie, verwijdering, beperking, data portabiliteit, en bezwaar. Er gelden uitzonderingen op deze rechten wanneer persoonsgegevens worden verwerkt door TNO als instelling voor wetenschappelijk onderzoek. Hieronder leggen we uit welke rechten u kunt uitoefenen als u een *niet-publiek persoon* bent (een persoon zonder een eigen Wikipedia-pagina) en welke rechten u kunt uitoefenen als u een *publiek persoon* bent (een persoon met een eigen Wikipedia-pagina).

U bent een niet-publiek persoon
TNO doet onderzoek naar de ontwikkeling van taalmodellen zoals het GPT-NL Model en verzamelt hiervoor de ruwe content. Om uw privacy te beschermen treft TNO vergaande privacybeschermende maatregelen. Dit houdt onder andere in dat TNO uw persoonsgegevens zo snel mogelijk verwijdert of pseudonimiseert en deze niet gebruikt voor het trainen van het GPT-NL Model (zie voor meer informatie hierboven onder "*Hoe bewerkt TNO de ruwe content voordat het GPT NL Model ermee wordt getraind?*"). In het kader van het trainen van het GPT-NL Model is niet mogelijk om specifieke informatie over u in de ruwe content op te zoeken, aan te passen, of te verwijderen, omdat deze gegevens juist nodig zijn om te onderzoeken of de opschoning van de ruwe content op de juiste manier plaatsvindt. Voor TNO geldt hiervoor als instelling voor wetenschappelijk onderzoek een wettelijke uitzondering om aan verzoeken van betrokkenen te voldoen onder artikel 44 UAVG en artikel 17 lid 3 sub d AVG

Als uw verzoek ziet op de vertrekking van ruwe content door Deelnemers aan TNO, dan kunt u uw verzoek bij de betreffende Deelnemer indienen.

U bent een publiek persoon
Anders dan voor niet-publieke personen worden niet al uw persoonsgegevens gepseudonimiseerd of verwijderd uit de ruwe dataset. Irrelevante persoonsgegevens, zoals uw contactgegevens, worden wel verwijderd zoals uitgelegd onder *Hoe bewerkt TNO de ruwe content voordat het GPT NL Model ermee wordt getraind?*"

Doordat bepaalde persoonsgegevens van u in de Schone Dataset worden opgenomen, kan het GPT-NL Model vragen over u beantwoorden. Als u bijvoorbeeld een bekende politicus bent, dan zal het GPT-NL Model in staat zijn vragen te beantwoorden over de politieke partij waaraan u gelieerd bent.

Ook als u een publiek persoon bent, geldt de wettelijke uitzondering voor TNO die hierboven is beschreven bij niet-publieke personen wat betreft de ruwe data en de Schone Data Set. Wel kunt u rechten uitoefenen indien antwoorden van het GPT-NL Model uw het GPT-NL Model onjuiste of niet langer relevante antwoorden geeft die

U kunt wel bezwaar maken onder artikel 21 AVG. Dit kan indien bij gebruik van het GPT-NL Model de antwoorden van het GPT-NL Model uw persoonsgegevens bevatten en er vanwege uw specifieke situatie redenen zijn om daartegen bezwaar te maken, bijvoorbeeld omdat de gegevens onjuist zijn of niet langer relevant. Als u bezwaar instelt, zal TNO uw verzoek

honoreren, tenzij er zwaarwegende andere belangen zijn, zoals het recht op vrijheid van informatie voor de gebruikers van het GPT-NL Model. Dit is dezelfde afweging die wordt gemaakt wanneer u een zoekmachine verzoekt om informatie over u niet langer in de zoekresultaten te tonen.[1] Als het resultaat van deze afweging is dat uw verzoek zwaarder weegt, zal TNO alle licentienemers instrueren tijdelijke maatregelen nemen om de desbetreffende antwoorden te voorkomen (zoals opgenomen in het Data Protection Protocol) en zal TNO uw verzoek definitief verwerken in de volgende versie van het GPT-NL Model.

Gebruik van het GPT NL Model door licentienemers
Als uw verzoek ziet op het gebruik van het GPT NL Model *door licentienemers*, dient u uw verzoek te richten tot deze licentienemer. TNO is namelijk niet betrokken bij het gebruik van het GPT NL Model door gebruikers, en kan uw verzoek daarom niet voor u doorsturen naar de juiste gebruiker.

Verdere vragen of klachten
Als u verdere vragen heeft over de verwerking van uw persoonsgegevens in het kader van het GPT-NL Project, dan kunt u contact opnemen met TNO via de onderstaande contactgegevens. Daarnaast heeft u altijd het recht om een klacht in te dienen bij de Autoriteit Persoonsgegevens.

Contact

Heeft u vragen over deze privacyverklaring of over de verwerking van persoonsgegevens in het kader van het GPT-NL project?

Neem dan contact op met: privacy@gpt-nl.nl, of schriftelijk via Anna van Buerenplein 1, 2595 DA Den Haag, onder vermelding van het GPT-NL project.

Verdere informatie

Het GPT-NL Project is in detail uitgewerkt in meerdere documenten die publiekelijk beschikbaar zijn of worden gemaakt via de GPT-NL website: https://gpt-nl.nl/. Zie specifiek:

- Het Data Protection Protocol, waarin wordt uitgelegd welke taken en verantwoordelijkheden TNO, de Deelnemers, en licentienemers van het GPT NL Model hebben met betrekking tot de verwerking van persoonsgegevens;
- Het Training Content Protocol, waarin wordt uitgelegd hoe (persoons)gegevens worden opgeschoond en gebruikt om het GPT NL Model te trainen;
- Het Data Protection Impact Assessment (DPIA), waarin de verwerking van persoonsgegevens in het kader van het GPT-NL Project in detail wordt omschreven, en waarbij wordt uitgelegd welke maatregelen er zijn getroffen om de impact op de privacy van betrokkenen te beperken; en
- Het Bronnenoverzicht, waarin wordt uitgelegd van welke bronnen TNO trainingsteksten heeft verzameld en gebruikt voor het GPT-NL Project.

[1] https://curia.europa.eu/juris/documents.jsf?language=NL&critereEcli=ECLI:EU:C:2019:773