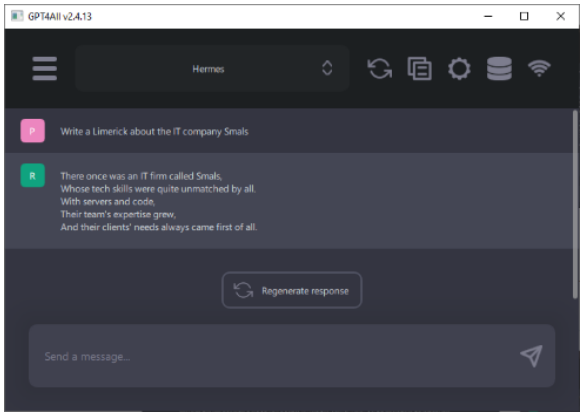
	Systeemvereisten:	Recente CPU, 8GB+ RAM
	Ontwikkeld door:	Nomic AI , https://gpt4all.io/
MIT License	Contactpersoon:	joachim.ganseman@smals.be



GPT4All is een ChatGPT-achtige user interface bovenop verschillende populaire taalmodellen, waaronder LLAMA varianten. De beschikbare taalmodellen hebben met elkaar gemeen dat ze open source zijn (i.e. men kan ze downloaden), en dat ze zodanig zijn bewerkt dat er geen GPU meer nodig is om ze te gebruiken: een CPU is voldoende. GPT4All omvat een bibliotheek met taalmodellen die als bestand gedownload kunnen worden, waarna ze op de eigen computer gebruikt kunnen worden zonder dat een internetverbinding of API-key nodig is.

De interface laat toe allerlei parameters vrij aan te passen, waaronder de verschillende prompts, de “temperatuur” (mate van creativiteit), de batch size, een “penalty” voor teveel herhalingen, of de maximale antwoordlengte. Optioneel kan men ervoor kiezen om de input en output die men verkrijgt, te delen met Nomic.ai, de makers van GPT4All, voor toekomstige verbeteringen aan hun eigen modellen.

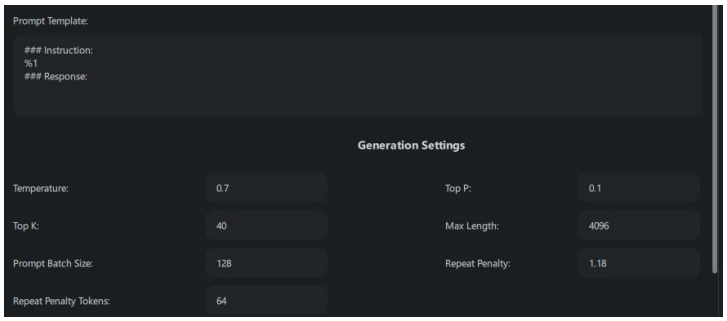
Daarnaast bevat GPT4All ook een ingebouwde server die ingeschakeld kan worden, waardoor het lokale model via API aan te spreken is. Ook de populaire toolkit LangChain bevat een integratiemodule om lokaal geïnstalleerde GPT4All taalmodellen te gebruiken in eigen pipelines.

GPT4All laat je toe om zelf met open taalmodellen te experimenteren zonder teveel afhankelijkheid van grote AI-bedrijven. Experimenteren met de parameters verschaft inzicht in de werking, de benodigde rekenkracht en de outputkwaliteit. De ingebouwde server en de ondersteuning in de LangChain library, maakt het ook tot handig vervangmiddel voor (betalende) API-toegang, voor wie applicaties met recente taalmodellen wil ontwikkelen. De tool focust wel erg op het dialoog-formaat, en mist nog ondersteuning voor andere manieren om taalmodellen aan te wenden (zoals *fill-in-the-middle* taken).

GPT4All is een normale desktop-applicatie voor Windows, MacOS en Linux. Na installatie kan men direct enkele taalmodellen downloaden. Deze download-sectie geeft ook een beknopt overzicht van o.a. de herkomst, de grootte, de hardware-vereisten en de licentievoorwaarden. Gedownloade modellen worden lokaal opgeslagen.

Na het downloaden kan men onmiddellijk aan de slag. Bovenaan de interface selecteert men uit de gedownloade modellen, daarna kan men in dialoogvorm de conversatie aangaan. De interface zelf is vrij minimalistisch en niet alle opties staan even logisch geplaatst – het is bijvoorbeeld even zoeken welke knop toelaat om de prompt (= de “voorafgaande algemene instructies”) te wijzigen.

Tijdens het genereren van een antwoord krijgt men rechtsonder te zien hoe snel het model draait. Men merkt al snel dat het op een doorsnee PC heel wat trager gaat dan met betalende cloud-toegang tot een commerciële AI-provider, maar dit schept wel een realistisch beeld van de enorme hoeveelheid rekenkracht (en energie) die verstookt wordt bij generatieve AI-taken.



De beschikbare taalmodellen zijn relatief klein in vergelijking met het beroemde GPT-3 of GPT-4, en de outputkwaliteit is dan ook wat minder. De website van GPT4All vermeldt metrieken over de performantie van de beschikbare taalmodellen op allerlei benchmarks, maar zelfs het beste model heeft nog moeite met iets virtuozere vormen van taalgebruik zoals het volgen van rijmschema’s. We merken ook op dat de grotere modellen soms enigszins in het Nederlands of Frans kunnen antwoorden, maar de kleinere vaak beperkt zijn tot het Engels.

De evolutie gaat echter zeer snel en regelmatig verschijnen er updates en nieuwe modellen. Ongetwijfeld gaan er nog modellen bijkomen met aparte eigenschappen en/of specialiteiten. Eens beschikbaar voor download, is GPT4All een gedroomde tool om ze direct aan een eerste evaluatie te onderwerpen, zonder dat men zelf moet programmeren of een expert moet zijn in deze *Large Language Models*.

GPT4All is een open source project en is gratis en vrij beschikbaar onder MIT licensie. Op de beschikbare taalmodellen kunnen andere licenties van toepassing zijn, de *model repository* licht dit verder toe. Om met enige performantie een taalmodel lokaal te kunnen draaien, zijn een krachtige CPU en een ruime hoeveelheid RAM-geheugen noodzakelijk.