Vergelijking open source modellen

**Introductie**

Er zijn ambtenaren die gebruikmaken van taalmodellen zoals Chat-GPT om hun werkzaamheden te verbeteren, maar er zijn ook zorgen over het feit dat de overheid zelf geen controle heeft over die modellen en dus afhankelijk wordt van grote bedrijven. Daarom zijn er mensen die zich afvragen of we niet zelf binnen de overheid een model kunnen maken, dat de data intern houdt en specifiek is getraind voor de taken waarvoor wij het willen gebruiken. Die mogelijkheid willen we in dit project onderzoeken.

Dit vanaf 0 doen is onrealistisch, het kost gigantisch veel tijd, geld en rekenkracht om een model te bouwen en bovendien hebben we er de expertise niet voor. Wat we echter wel kunnen proberen is een **open source** model downloaden en dat voor lokale applicaties gebruiken. In ons geval focussen we het meest op **RAG** (retrieval augmented generation), meer daarover kan je [hier](https://docs.google.com/document/u/0/d/1J6k0HdlY5NMrOtSI0UhMYBZSoIlaqGwMx-FXzhio3Vk/edit) lezen. Open source-modellen zijn vrij voor iedereen beschikbaar om te downloaden en de code en werking ervan is lokaal aan te passen. Deze veranderingen en de data die je daarvoor gebruikt blijven dan ook on premise.

Het is hierbij belangrijk om te beseffen dat de beschikbare modellen en technieken snel veranderen en ook het debat erover volop in ontwikkeling is. Dit document beschrijft de recente ontwikkelingen op het gebied van open source-modellen in januari 2024. Een vorige versie kwam uit september 2023 en sindsdien is het veld fors veranderd. De snelle ontwikkelingen betekenen dat we er zo veel mogelijk voor willen zorgen dat we snel kunnen wisselen in het model waarvan we gebruik maken.

**Foundational models: LLama en Mistral.**

Zoals gezegd is het kostbaar om een model vanaf 0 te trainen, er zijn een handjevol bedrijven en organisaties die de basistraining van een model gedaan hebben en hun model open source beschikbaar hebben gemaakt. Een dergelijk model noemen we een foundational model. Niet ieder open source model is een foundational model. Er zijn ook mensen die een foundational model downloaden, dit zelf extra trainen en het resultaat weer openbaar beschikbaar maken. Ook als je van een dergelijk model gebruik wilt maken is het goed om te weten op basis van welk foundational model het getraind is, de data en methodiek waarmee het foundational model getraind blijft altijd een effect hebben op de uiteindelijke prestaties.

MPT (getraind door het bedrijf Mosaic) en Falcon modellen (getraind door technology innovation institute (TII) een onderzoekscentrum gefinancierd vanuit de overheid van Abu Dahbi) zijn voorbeelden van open source foundational models. De Falcon modellen zijn in het Engels, Duits, Spaans en Frans getraind, de MPT modellen alleen in het Engels en zijn relatief goed in programmeren.

Sinds 19 juli (2023) zijn de nieuwe modellen van Meta (het bedrijf waar Facebook een onderdeel van is) open source beschikbaar. Deze modellen, genaamd [LLama 2,](https://arxiv.org/abs/2307.09288#:~:text=In%20this%20work%2C%20we%20develop%20and%20release%20Llama,Llama%202-Chat%2C%20are%20optimized%20for%20dialogue%20use%20cases.) lijken doorgaans beter te presteren dan de MPT- en Falcon modellen. Bovendien heeft Meta, in tegenstelling tot Mosaic en TII, ook een paper uitgebracht waarin precies staat hoe het getraind is. Hierin is te lezen dat de Llama 2 modellen niet alleen getraind zijn op basis van veel online beschikbare data, maar ze ook uitgebreid aan ‘reinforcement learning with human feedback’ (RLHF) en ‘red-teaming’ gedaan hebben. RLHF houdt in dat mensen de antwoorden die het model geeft beoordelen, in dit geval op basis van de behulpzaamheid van de antwoorden en de ‘veiligheid’. Veiligheid in deze context betekent dat het model geen schadelijke antwoorden geeft, zoals antwoorden die kunnen helpen bij het uitvoeren van illegale activiteiten en antwoorden die onjuiste informatie of schadelijk advies bevatten. Op basis van de vele beoordelingen van deze mensen wordt het model verder getraind om behulpzame en veilige antwoorden te geven. Bij red teaming proberen mensen actief het model onveilige dingen te laten doen, en slimme technieken toe te passen om het zover te krijgen. Hiermee wordt de robuustheid van het model getest en iedere keer dat het model de fout ingaat wordt het aangepast om die kans in de toekomst te verkleinen. Een groot nadeel is echter dat LLama 2 alleen in het Engels getraind is, wat zowel de prestaties als het effect van de veiligheidstrainingen in het Nederlands onzeker maakt.

De goede prestaties en veiligheidstrainingen zorgen ervoor dat sinds het model gepubliceerd is, het verreweg het meest gebruikt wordt als foundational model door anderen die hun eigen aanpassingen maken. Er is een LLama 2 model met 7, 13, 34 en 70 miljard parameters. In het algemeen geldt: hoe meer parameters een model heeft, hoe meer rekenkracht (en dus geld en energie) het kost om het model te downloaden, aan te passen en te gebruiken maar ookhoe beter het model is. Wij doen een onderzoeksopdracht, waarbij we beperkte rekenkracht ter beschikking hebben. Daarom hebben wij een sterke voorkeur voor modellen met minder parameters. Bovendien wil de overheid in het algemeen zuinig met energie en belastinggeld omgaan, dus alleen meer parameters als het ook echt veel beter resultaat oplevert.

Vandaar dat het Falcon model dat kort na Llama2 uitkwam voor ons ook niet interessant is. ITT heeft een model open source uitgebracht dat volgens sommige maatstaven net iets beter presteert dan het LLama2 model van 70 miljard parameters, maar 180 miljard parameters heeft.

Wat voor ons echter wel heel interessant is, is het MistralAI model. Het Franse bedrijf MistralAI heeft op 27 september (2023) een model uitgebracht van slechts 7 miljard parameters dat op veel [testen beter presteert](https://mistral.ai/news/announcing-mistral-7b/) dan het LLama2 model van 13 miljard parameters. Van dit model is de exacte mix van talen in de training onbekend, maar het lijkt het puur anekdotisch[[1]](#footnote-0) ook iets beter in Nederlands te zijn. Sinds 11 december 2023 heeft Mistral ook een nieuw model uitgebracht, het zogenaamde Mixtures of Expert model. Dit heeft een bijzondere architectuur waarvan vermoed wordt dat GPT4 die ook heeft, niet iedere vraag gaat door het hele model. In plaats daarvan is het model eigenlijk een combinatie van 8 aparte modellen (in het geval van Mistral allemaal met 7 miljard parameters) met hun eigen gewichten en zodra een prompt ingevuld wordt, worden voor iedere token 2 modellen gekozen om deze token te verwerken. Dit zorgt ervoor dat, hoewel het model 46.7 miljard parameters in totaal heeft, er per token slechts 12.9 miljard parameters gebruikt worden. Hierdoor krijgen we een model dat qua competentie op een groot model lijkt (het doet het zelfs beter dan het 70 miljard parameter model van Meta op veel testjes), maar waarvoor de (energie) kosten per verzoek lijken op die van een 13 miljard parameter model. Bovendien is dit model naast het Engels ook op Duits, Frans en Spaans getraind.

In tegenstelling tot Meta heeft Mistral geen gebruik gemaakt van RLHF of red teaming, in plaats daarvan hebben ze zogenaamde ‘guardrails’ ingebouwd. Ten eerste zorgen ze ervoor dat bij iedere vraag van een gebruiker er automatisch bij gezegd wordt dat er geen schadelijke, discriminerende of onethische reactie gegeven moet worden. Daarnaast classificeert het systeem zelf of een bepaalde vraag acceptabel is of bijvoorbeeld discriminerend of illegaal. Het classificeert ook het eigen antwoord. Als de vraag ofwel het antwoord als onacceptabel geclassificeerd is, wordt het verzoek geweigerd.

Sinds kort zijn er bovendien een aantal Chinese open source modellen, zoals Yi getraind door 01-AI een bedrijf opgericht door taiwanese AI expert [Kai-Fu Lee](https://www.youtube.com/watch?v=ajGgd9Ld-Wc), gefinancierd door Sinovation Ventures. Bij het maken van deze modellen is er intensief samengewerkt met Meta en veel van de Llama architectuur gebruikt, het is niet duidelijk of RLHF en Red teaming ook op vergelijkbare wijze toegepast is.

Daarnaast zijn er de Qwen modellen getraind door Alibaba Cloud. Ook dit model is zeer vergelijkbaar getraind als de rest, waarbij het gehele trainingsproces [gedocumenteerd](https://arxiv.org/abs/2309.16609) is. Er is net als bij Llama 2 gebruik gemaakt van RLHF.

In de onderstaande tabel zijn een aantal modellen met elkaar vergeleken.

De LLM leaderboard score en Arena leaderboard score komt van het platform [huggingface](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard), een platform waar mensen hun modellen open source beschikbaar maken.

LLM leaderboard score is een gewogen gemiddelde van hoe goed modellen scoren op veel gebruikte tests die gaan over dingen als kennis, redeneervaardigheden, programmeervaardigheden, common sense logica. Deze scores zijn geen perfecte afspiegeling van hoe goed het model in de taken zal zijn waarvoor wij ze gaan gebruiken, maar ze zeggen wel iets over het algemene competentieniveau van een model.

De arena leaderboard score gaat niet om automatische tests maar daar kunnen mensen zelf een vraag stellen, geven 2 willekeurig gekozen modellen een antwoord en kiest de gebruiker welke beter was. Dit geeft alle modellen een win percentage tegen andere modellen, hoe hoger de score hoe vaker ze ‘winnen’. Deze score zegt dus iets over welke modellen gebruikers als beter beschouwen.

Veel ontwikkelaars hebben modellen van verschillende omvang uitgebracht, in dat geval hangt prestatie vaak af van omvang en zijn de scores van het kleinste naar grootste model opgeschreven met / tussen de verschillende scores. Voorbeeld: het 13 miljard Llama2 model heeft een LLM leaderboardscore van 58.66 en 70 miljard Llama2 model van 67.35.

**Tabel: vergelijking foundational models**

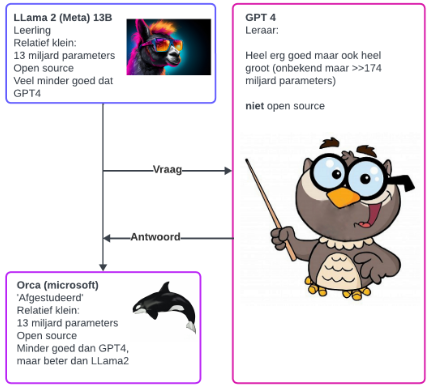
|  | Yi | Llama 2 | Falcon | Qwen | MPT | Mistral |
| --- | --- | --- | --- | --- | --- | --- |
| Getraind door | 01-AI | Meta (facebook) | TII | Alibaba Cloud | Mosaic | Mistral AI |
| Aantal parameters beschikbaar | 6 miljard / 34 miljard | 7 miljard / 13 miljard / 70 miljard | 7 miljard / 40 miljard / 180 miljard | 7 miljard /14 miljard / 72 miljard | 7 miljard / 30 miljard | 7 miljard / 8x7 miljard |
| LLM leaderboard score | 56.76 / 70.81 | 54.32 / 58.66 / 67.35 | 47.01 / 61.48 / 68.74 | 59.19/ 65.86 / 73.6 | 47.38 / 56.15 | 62.4 /  68.42 |
| Arena Leaderboard score | ?? /  1110 | ?? /  1033 / 1077 | ?? /  ?? /  1055 | ?? /  1041 /  ?? | ?? | 1023 / 1121 |
| Talen in getraind en getest | Engels, Chinees | Engels | Engels, Duits, Spaans, Frans | Engels, Chinees | Engels | Engels /  Engels, Spaans, Duits, Frans |
| Safety training | Geen voor zover bekend | RLHF + Red teaming | Geen voor zover bekend | RLHF | Geen voor zover bekend | System prompt + self-reflection |
| Informatie | [Huggingface](https://huggingface.co/01-ai/Yi-34B-200K), [website](https://01.ai/?ref=maginative.com) | [Paper](https://arxiv.org/abs/2307.09288) | [dataset](https://arxiv.org/abs/2306.01116) | [Paper](https://arxiv.org/abs/2309.16609) | [Blog](https://www.mosaicml.com/blog/mpt-30b) | [Training](https://arxiv.org/abs/2310.06825)  [Blog 7B](https://mistral.ai/news/announcing-mistral-7b/)  [Blog 8x7B](https://mistral.ai/news/mixtral-of-experts/) |

**Je laten trainen door GPT 4: Orca.**

Microsoft heeft recent een nieuwe techniek bedacht om modellen met relatief weinig parameters resultaten te laten bereiken die in de buurt komen van modellen met extreem veel parameters. Het idee is dat je een klein open source model kan laten verbeteren door de beste, maar ook grootste, modellen die beschikbaar zijn. Op dit moment is dat GPT4.

Ze hebben een dataset, genaamd Orca, gemaakt dat bestaat uit inputs gecombineerd met de outputs die GPT4 zou geven, en dit kan gebruikt worden om een model te leren hoe GPT4 op een prompt zou reageren. Deze dataset is online beschikbaar voor iedereen. In [het paper](https://arxiv.org/abs/2306.02707) dat Microsoft hierover geschreven heeft, hebben LLama2 getraind met de Orca dataset.. Het blijkt dat je op deze manier een model kan krijgen dat veel beter functioneert dan je zou verwachten op basis van het aantal parameters. Met een fractie van de rekenkracht die het GPT4 gekost heeft om diens competentie te bereiken kan een groot deel van de kennis en kunde overgedragen worden. De Orca dataset laat ook zien dat er vaak minder data nodig is dan voorheen gedacht werd, als de data heel slim gemaakt is.

**Figuur: Overzicht van het trainen van Orca**



Inmiddels zijn er ook mensen die het 7 miljard model van Mistral hebben getraind met de dataset van Orca, dus in feite Mistral door GPT4 hebben laten trainen. Het resultaat is [Mistral-7B-OpenOrca](https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca), een model dat erg goed lijkt voor iets van 7 miljard parameters en het model is wat wij daarom als basis voor onze training zullen gebruiken. [Dit](https://huggingface.co/spaces/Open-Orca/Mistral-7B-OpenOrca) is een demo waar je het model zelf kan testen. Op het eerste gezicht lijkt die ook beter te werken dan de alternatieven, onze ervaring is dus in lijn met de officiële testresultaten.

**Fine-tunen voor RAG**

Wij willen het model gebruiken voor retrieval augmented generation, dus het is logisch om te kijken of er ook modellen zijn die hiervoor gefinetuned zijn.

Dit model: [Photolens/OpenOrcaxOpenChat-2-13b-langchain-chat · Hugging Face](https://huggingface.co/Photolens/OpenOrcaxOpenChat-2-13b-langchain-chat) is een gefinetunede versie van een Llama 2 - 13 miljard model gefinetuned op de Orca dataset. Dit laat zien dat finetunen iteratief is, zolang het resulterende model open source gemaakt wordt kan ook dat model weer verder gefinetuned worden. Vervolgens hebben ze het getraind op een dataset die bedoeld is om het beter te maken voor Langchain applicaties (onze RAG applicatie maakt gebruik van Langchain). het is in totaal op meer dan 100.000 voorbeelden getraind waarvan 628 in het Nederlands.

Ook dit model: [SciPhi/SciPhi-Self-RAG-Mistral-7B-32k · Hugging Face](https://huggingface.co/SciPhi/SciPhi-Self-RAG-Mistral-7B-32k) zou voor betere RAG prestaties kunnen zorgen. Dit is het 7 miljard parameter model van Mistral dat gefine-tuned is op de [self-RAG dataset](https://arxiv.org/abs//2310.11511). Hierin leert een systeem itteratief beoordelen wat relevante en minder relevante context is en wanneer kennis nog ontbreekt. Hierbij wordt ook gebruik gemaakt van GPT4 om deze beoordeling te toetsen

**Trainen in het Nederlands: in gesprek met Bram Vanroy**

Bram Vanroy is een postdoctoraal onderzoeker met expertise op het gebied van natuurlijke taalverwerking (NLP) en machinale vertaling aan de katholieke universiteit Leuven. Hij is bovendien een actieve bijdrager aan open-source projecten en heeft datasets, modellen en demonstraties ontwikkeld die verband houden met NLP, met name in het Nederlands.

Bram heeft meerdere keren foundational models verder getraind op basis van Nederlandstalige gegevens en de resulterende modellen, die in theorie beter zijn in Nederlands, weer open source gemaakt en op huggingface voor de wereld beschikbaar gemaakt. Bovendien ishij zo vriendelijk geweest met ons te praten en ons van advies te voorzien.

Bram heeft onder andere het [LLama 2 model](https://huggingface.co/TheBloke/Llama-2-13B-Chat-Dutch-GGUF) van 13 miljard parameters en het [Falcon model](https://huggingface.co/BramVanroy/falcon-40b-ft-alpaca-dolly-dutch) van 40 miljard parameters op het Nederlands getraind. Hij gaf hierbij zelf aan dat de resultaten bij Llama 2 niet optimaal waren. In de praktijk blijkt het lastig om een model dat heel weinig in het Nederlands pre-trained is, goed Nederlands te leren met de datasets die op dit moment beschikbaar zijn. Falcon 40 miljard ging iets beter, maar dat heeft ook meer parameters. Hij adviseerde ons dat we beter gebruik konden maken van de Mistral-modellen dan van de zijne.

Hiernaast gaf hij aan dat op dit moment de beschikbare datasets in het Nederlands van lage kwaliteit zijn, hij had zelf altijd de insteek om gewoon zoveel mogelijk data te gebruiken. Echter, het Orca project heeft laten zien dat je ook met relatief weinig data van hoge kwaliteit hele goede resultaten kan bereiken. Vanroy zei dat het voor de gehele ontwikkeling van open source taalmodellen in het Nederlands waarschijnlijk van grote meerwaarde zou zijn om de gehele Orca dataset naar het Nederlands te vertalen. Hij had hier echter zelf het budget (hij gokte ongeveer 5000 euro aan rekenkracht en daarbuiten heb je expertise nodig om het te vertalen en vervolgens ook de vertaling ook te gebruiken voor training) niet voor. Mocht er overheidsbreed een wens zijn om taalmodellen te hebben die beter Nederlands kunnen dan kan het de moeite waard zijn om de Orca dataset te vertalen en dit publiekelijk beschikbaar te maken. Het is een open vraag wie dit dan het beste uit kan voeren.

Inmiddels zijn we een aantal maanden verder en is door Rijsenberg het Mistral model op het Nederlands gefine-tuned met een grote dataset die grote delen van het Nederlandstalige internet bevat, het is relatief veel data maar niet per se van hoge kwaliteit. Dit is het model: [Rijgersberg/Mistral-7B-v0.1-chat-nl · Hugging Face](https://huggingface.co/Rijgersberg/Mistral-7B-v0.1-chat-nl). Dit lijkt beter te presteren dan de modellen van Bram Vanroy.

**Conclusie**

Het veld is razendsnel in ontwikkeling, dus dit document zal snel verouderd zijn. Sterker nog dit document is al meerdere keren geüpdatet omdat vorige versies in een paar maanden tijd hopeloos verouderd waren. Op dit moment lijkt Mistral de beste foundational modellen te hebben, waarbij de nieuwe Chinese modellen niet ver achterlopen. Ik gok dat bij de keuze tussen een Europees en Chinees model van gelijke prestatie het Europees model binnen de overheid geprefereerd wordt, al zou het bij open source modellen weinig uit moeten maken omdat je ze toch download. Het 7 miljard model van Mistral is echter ook al op wat meer dingen gefine-tuned, in mijn optiek zijn de meest kansrijke modellen:

* Mistral 7x8B mixture of expert
* Geitje, het op nederlands gefinetunede mistral 7B model van Rijserberg
* SciPhi SelfRAG, het voor RAG gefinetunede mistral 7B model met de self-rag dataset
* Mistral-7B-open orca, de met Orca gefinetunede versie van Mistral
* Qwen 14 miljard, om ook een niet Mistral model te bekijken

1. Wij heb zelf de verschillende modellen ook een beetje getest, een model op basis van het mistral model was bijvoorbeeld goed in staat onze vragen (in het Nederlands) te beantwoorden over een stukje (Nederlandse) tekst die ik hem meegaf of te redeneren over wat I&W (alleen afkorting gebruikt) mogelijk van de stikstof crisis zou vinden. Llama 2, zelfs extra getraind op het Nederlands, had hier meer moeite mee. [↑](#footnote-ref-0)