WAT ZIJN FOUNDATION MODELS?

Met de komst van ChatGPT zijn generatieve AI-modellen het onderwerp van dit jaar. De basis van OpenAI's ChatGPT is een foundation model, namelijk GPT-3.5 en GPT-4, dat ook door andere bedrijven wordt gebruikt, zoals bijvoorbeeld in de toepassing Microsoft Copilot. Een foundation model is een AI-model dat de 'basis' vormt voor andere bedrijven om verder op te werken en er dus andere AI-toepassingen mee te bouwen of te trainen.

In deze brAInfood kom je meer te weten over wat foundation models zijn en waarvoor ze worden gebruikt. We lichten dit toe met enkele voorbeelden van toepassingen van foundation models. Ten slotte, gaan we ook dieper in op mogelijke voordelen ervan en maken we enkele bedenkingen bij het ontwerp en het gebruik van foundation models.

WAT IS EEN FOUNDATION MODEL?

Een foundation model is een AI-model dat wordt gevoed met een **enorme hoeveelheid en verscheidenheid aan gegevens**. Dit model kan slechts één keer worden getraind en vervolgens worden verfijnd om **allerlei verschillende soorten taken uit te voeren**, zoals het genereren van tekst, afbeeldingen of audio.

Sommige foundation models **nemen enkel input aan in een bepaalde modaliteit**, zoals tekst, terwijl andere foundation models **meerdere modaliteiten**, zoals bijvoorbeeld tekst, beeld, video, etc., kunnen aannemen en zelfs ook **meerdere soorten output kunnen genereren**. Deze output kan bijvoorbeeld het genereren van beelden, het samenvatten van tekst of het beantwoorden van vragen zijn.

Wat foundation models uniek maakt is dat zij **op zichzelf staande systemen** kunnen zijn, maar zij kunnen ook gebruikt worden **als 'basis' voor andere AI-toepassingen**. De organisatie die het foundation model maakt, stelt het product dan open voor andere organisaties, vaak mits een vergoeding, om hierop verder te bouwen en een nieuwe AI-toepassing te ontwikkelen.

TOEPASSINGEN VAN FOUNDATION MODELS

Twee bekende foundation models zijn grote taalmodellen (large language models of LLMs) en generatieve AI:

- **LLMs**, onderdeel van Natural Language Processing (NLP), worden getraind met miljarden woorden tekst en kunnen natuurlijke taal genereren op basis van tekstvoorspelling. Deze modellen berekenen de waarschijnlijkheid van het gebruik van een karakter, woord of tekenreeks op basis van de voorafgaande of omringende taalcontext. LLM's vallen onder de bredere noemer van 'generatieve AI'.
- **Generatieve AI** kan verschillende soorten inhoud (tekst, afbeeldingen, video of audio) genereren op basis van gebruikersinvoer zoals tekstaanwijzingen. Let op: niet alle generatieve AI modellen zijn gebaseerd op foundation models, aangezien sommige toepassingen ontworpen zijn voor een specifiek doel en niet hergebruikt kunnen worden in nieuwe contexten.

Foundation models worden vaak gebruikt voor interne doeleinden, zoals een portaal waar een chatbot op verzoek van een werknemer informatie over personeelszaken deelt, of een assistent die lange juridische teksten samenvat voor een advocaat.

VOORDELEN

Dankzij foundation models kunnen **nieuwe applicaties versneld worden uitgerold** doordat de basis van de applicatie al voorhanden is. Op die manier wordt dubbel werk vermeden en kan er sneller naar de markt gegaan worden met nieuwe toepassingen. Het is niet meer nodig om je als ontwikkelaar eerst volledig te verdiepen in de wereld van machine learning en artificiële intelligentie toepassingen, want dankzij foundation models kunnen dergelijke toepassingen **op grotere schaal inzetbaar worden gemaakt zonder dat een diepgaande kennis over de bouw van deze toepassingen nodig is**. Ontwikkelaars hebben zo meer tijd om zich te focussen op de training van het model op specifieke gegevenssets met gelabelde voorbeelden. Dit wil natuurlijk niet zeggen dat er bij het trainen van deze modellen geen technische- en/of programmeer-kennis nodig is.

REGULERING

AI-modellen met een aanzienlijk algemeen karakter, die een breed scala aan verschillende taken kunnen uitvoeren en downstream in verscheidene systemen kunnen worden geïntegreerd, worden in de Europese AI verordening gereguleerd onder de noemer "AI-modellen voor algemene doeleinden" of "GPAI-modellen".

Aanbieders van **alle GPAI-modellen** zijn verplicht om technische documentatie op te stellen, informatie te geven aan aanbieders die het AI-model in hun AI-systeem zullen integreren, een beleid op te stellen rond hun naleving van het auteursrecht en een samenvatting openbaar te maken over de inhoud voor het trainen van het AI-model.

Deze verplichtingen gelden niet voor open-source GPAI-modellen zonder systeemrisico. Aanbieders van **GPAI-modellen met systeemrisico** moeten bijkomend hun modellen evalueren om systeemrisico's in kaart te brengen, te evalueren en te beperken, informatie over ernstige incidenten rapporteren aan bepaalde autoriteiten en een passend niveau van cyberbeveiligingsbescherming voorzien voor het GPAI-model.

GPAI-modellen hebben een systeemrisico als ze beschikken over capaciteiten met grote impact of als ze door de Commissie, op basis van bepaalde criteria, als GPAI-model met systeemrisico wordt aangemerkt.

BEDENKINGEN

Er zijn enkele kanttekeningen die we kunnen maken bij het ontwikkelen en gebruiken van foundation models en AI in het algemeen, zoals:

- **Machtsverhouding in onbalans**: doordat ontwikkelaars kunnen starten vanaf een bestaand foundation model en hierop verder werken, geeft dit natuurlijk wel een bepaalde macht aan de aanbieders van die foundation models omdat zij bepalen hoe ze functioneren. De gebruikers van de foundation models moeten steeds kritisch het model analyseren en beoordelen.
- **Gebrek aan contextueel begrip**: foundation models hebben geen begrip van de output die ze genereren. Ze beschikken niet over het vermogen om nuances of onderliggende emoties van een situatie te begrijpen en in rekening te brengen (zoals sarcasme).
- **Hallucinaties**: deze modellen missen "gezond verstand", het vermogen om nuances van menselijke ervaringen en de echte wereld te begrijpen. Dit kan leiden tot een fenomeen dat bekend staat als "hallucinaties", waarbij het model plausibele maar feitelijk onjuiste informatie genereert. Bijvoorbeeld: een model vat een nieuwsartikel samen, maar verzint citaten of details die niet in de originele bron staan.
- **Bias - garbage in, garbage out**: een term die stelt dat een model maar zo goed werkt als de data die gebruikt is om het te trainen. Als een foundation model wordt getraind op data die bevooroordeeld en/of negatief is, kan het discriminerende, haatzaaiende en/of andere vormen van schadelijke inhoud produceren. Een model voor gezichtsherkenning dat enkel getraind is op voorbeelden van blanke personen, zal moeite hebben om personen met een andere huidskleur (correct) te herkennen.
- **Kwetsbaarheden in de beveiliging**: voortbouwend op de derde bedenking, zijn er verschillende scenario's die zich voor kunnen doen bij het trainen van modellen. Omdat deze modellen getraind worden met ontzettend veel data bestaat de kans dat bij het importeren van nieuwe trainingsdata er per ongeluk of moedwillig schadelijke, gevoelige of vertrouwelijke informatie wordt ingevoegd.
- **Moeilijkheden bij het controleren en/of begrijpen van het systeemgedrag**: foundation models zijn zeer complex, waardoor het voor gebruikers niet altijd duidelijk is om hun besluitvormingsprocessen en uitkomsten te begrijpen. Dit gebrek aan transparantie wordt verder uitvergroot doordat de modellen vaak niet open source zijn. Het gebruik van de modellen is (relatief) vrij, maar dat betekent niet meteen dat de broncode vrij toegankelijk is voor controle of aanpassingen.
- **(Milieu)kost**: het ontwikkelen, implementeren en onderhouden van foundation models vereist een hoge kost, zowel monetair als op vlak van energieverbruik. Hoewel er geen officiële cijfers zijn vrijgegeven wordt het energieverbruik voor het trainen van ChatGPT geschat op meer dan 1 Gigawattuur (GWh), vergelijkbaar met het energieverbruik van 120 huishoudens voor een jaar.
- **Regulering**: Doordat foundation models moeilijk of niet 'leesbaar' zijn, is het moeilijk voor wetgevers om het gebruik en functioneren van deze technologieën te reguleren en vast te leggen in wetgeving. De EU heeft dit aangepakt met een tweeledig regelgevend kader. De AI act maakt onderscheid tussen generieke modellen en systeemmodellen. Voor generieke modellen gelden transparantieverplichtingen (technische documentatie en naleving van de wetgeving inzake auteursrecht). Systeemmodellen moeten, naast de verplichtingen die gelden voor generieke modellen, evaluaties uitvoeren, systeemrisico's beoordelen en beperken, incidenten melden en cyberbeveiliging waarborgen.

Kenniscentrum Data & Maatschappij (maart 2024). Wat zijn foundation models? brAInfood van het Kenniscentrum Data & Maatschappij. Brussel: Kenniscentrum Data & Maatschappij.

Kenniscentrum Data & Maatschappij

Artificiële Intelligentie Vlaanderen