Bijlage Overeenkomst Data Donatie

Datum: 22 november 2025

Uitleg onderstreepte termen

Trainen van AI-modellen
Het proces waarbij een AI-model leert door grote hoeveelheden voorbeelden te bestuderen. Zo leert een AI-model door patronen in data te herkennen. Concreet betekent dit het iteratief aanpassen van een AI-model door het herhaaldelijk bloot te stellen aan voorbeelden, waarbij het model steeds beter leert welke antwoorden of voorspellingen correct zijn door zijn fouten te corrigeren. Op technisch niveau gaat het om het optimaliseren van parameters in een neuraal netwerk door middel van algoritmes die de voorspellingen van het model aanpassen op basis van trainingsdata, zodat het model patronen kan generaliseren naar nieuwe situaties.

Open source
Software of AI-modellen waarvan de onderliggende code openbaar beschikbaar is. Iedereen kan deze bekijken, gebruiken en aanpassen, net zoals een recept dat je vrij mag delen en variëren. Dit betekent dat de broncode vrij toegankelijk is voor iedereen, waardoor ontwikkelaars en onderzoekers kunnen zien hoe het werkt, verbeteringen kunnen voorstellen en het kunnen aanpassen voor eigen doeleinden binnen de voorwaarden van de licentie. Formeel gezien is het software waarvan de broncode publiekelijk toegankelijk is onder een licentie die modificatie, distributie en gebruik toestaat, waardoor transparantie en community-gedreven ontwikkeling mogelijk worden.

Veiligheid
Maatregelen om ervoor te zorgen dat AI-modellen geen schadelijke of gevaarlijke dingen doen. Dit omvat bescherming tegen misbruik en het voorkomen van onbedoelde negatieve gevolgen. In de praktijk betekent dit het bouwen van beschermingslagen in AI-systemen om te voorkomen dat ze worden misbruikt, foutieve of gevaarlijke output produceren, of gevoelige informatie prijsgeven, door middel van filters, testen en beperkingen. Op technisch niveau gaat het om het implementeren van technische en procedurele controles om AI-systemen te beschermen tegen adversarial attacks, prompt injection, data poisoning en andere kwetsbaarheden die kunnen leiden tot ongewenst of schadelijk gedrag.

Alignment
Het afstemmen van AI-gedrag op menselijke waarden en bedoelingen. Het zorgen dat de AI doet wat we werkelijk willen, in plaats van alleen maar letterlijk te doen wat we vragen. Dit houdt in dat AI-systemen zo worden getraind dat ze niet alleen technisch correcte antwoorden geven, maar ook begrijpen wat gebruikers echt bedoelen en zich gedragen volgens menselijke normen en waarden, zelfs in situaties die niet expliciet zijn voorgeprogrammeerd. Technisch gezien betekent dit het ontwikkelen van methoden zoals reinforcement learning from human feedback (RLHF) om te zorgen dat de doelfunctie van een AI-systeem overeenkomt met menselijke intenties en ethische normen, ook in onvoorziene situaties.

Bias
Vooroordelen of scheefheid in AI-systemen die bepaalde groepen mensen kunnen benadelen. Bijvoorbeeld wanneer een AI beter werkt voor mannen dan voor vrouwen, of voor bepaalde etnische groepen. Dit ontstaat door systematische afwijkingen in AI-prestaties of output doordat de trainingsdata of het ontwerp van het model bepaalde groepen ondervertegenwoordigt of verkeerd representeert, wat leidt tot ongelijke of oneerlijke behandeling. Preciezer geformuleerd gaat het om

systematische afwijkingen in AI-output die ontstaan door onevenwichtigheden in trainingsdata, proxy-variabelen of algoritmische beslissingsregels, wat kan leiden tot discriminerende of oneerlijke behandeling van bepaalde demografische groepen.

Best practices voor verantwoord gebruik
Richtlijnen en aanbevelingen over hoe je AI op een goede, ethische en veilige manier kunt gebruiken. Denk aan een gebruiksaanwijzing voor verantwoord AI-gebruik. Concreet is dit een verzameling van bewezen methoden en richtlijnen die organisaties helpen om AI-systemen in te zetten op een manier die rekening houdt met privacy, eerlijkheid, transparantie en veiligheid, en die risico's voor gebruikers minimaliseert. Op professioneel niveau betreft het gestandaardiseerde protocollen en frameworks voor ethische AI-implementatie, inclusief impact assessments, privacy-waarborgen, transparantievereisten en procedures voor monitoring en auditing van AI-systemen in productieomgevingen.

Modelkaart
Een document dat uitlegt hoe een AI-model werkt, waarvoor het bedoeld is, en wat de beperkingen zijn. Vergelijkbaar met een bijsluiter bij medicijnen. Het is een gestandaardiseerd informatiedocument dat belangrijke details over een AI-model beschrijft, zoals het beoogde gebruik, de prestaties in verschillende situaties, bekende zwakke punten en aanbevelingen voor verantwoorde toepassing. Formeel omvat dit een gestructureerde documentatie die de architectuur, trainingsmethode, prestatiemetrics, beoogde use cases, bekende beperkingen, evaluatieresultaten en potentiële risico's van een AI-model beschrijft volgens gestandaardiseerde rapportageformaten.

Datakaart
Een beschrijving van de dataset die gebruikt is om een AI-model te trainen. Dit bevat informatie over waar de data vandaan komt en welke kenmerken deze heeft. Uitgebreider is het een systematische beschrijving van een dataset die informatie geeft over de bron, samenstelling, verzamelmethode en mogelijke beperkingen van de data, zodat gebruikers kunnen beoordelen of de data geschikt is voor hun toepassing en welke vooroordelen erin kunnen zitten. Op professioneel niveau betreft het documentatie over de samenstelling, herkomst, verzamelmethoden, preprocessing-stappen, demografische verdeling, potentiële bias-bronnen en licentievoorwaarden van een dataset, conform data governance-principes en reproduceerbaarheidseisen.