

Voor deze weekopdracht heb ik onderzoek gedaan naar Large Language Models, oftewel LLM's. Vooraf wist ik al ongeveer hoe ze werkten en door een heleboel podcasts te luisteren de afgelopen paar jaar had ik ook een goed idee van wat ze allemaal konden. Tijdens het onderzoek merkte ik dat het gebruiken van een chatbot iets anders is dan begrijpen hoe het systeem werkt. Vooral attention en de manier waarop een model van fouten leert, vond ik in het begin lastig te begrijpen.

Een bron die tijdens het onderzoek naar voren kwam, was de paper over reinforcement learning. De bron hiervan is: Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://doi.org/10.48550/arXiv.2203.02155>. Ik vond het interessant dat mensen mee bepalen wat wel of niet een goed antwoord is.

Nieuw voor mij was hoe de training precies werkt. Een model krijgt tekstvoorbeelden en voorspelt steeds het volgende token. Het werkelijke vervolg staat al in de trainingstekst. De foutscore hangt af van hoeveel kans het model aan dat vervolg geeft. Daarna worden de interne getallen, de parameters, een beetje aangepast. Door dit met veel voorbeelden te herhalen, leert het model taalpatronen. Ook werd het verschil tussen instructietraining en training met feedback duidelijker. Bij instructietraining leert het model van opdrachten met voorbeeldantwoorden. Bij feedbacktraining kunnen beoordelingen van antwoorden helpen om het gedrag te verbeteren.

Wat mij aan het denken zette, was de mogelijkheid van het uitsterven van de mensheid. Het is een serieuze mogelijkheid, waar eerlijk gezegd niet genoeg aan gedaan wordt. Iets anders wat me ook aan het denken zette, was de mogelijke misbruik van hele goede AI modellen. Op bijvoorbeeld cybersecurity en de risico's daarvan.

Bij de presentatie over Internet of Things wist ik al dat het gaat om fysieke apparaten die met een netwerk verbonden zijn. Nieuw voor mij was de uitleg over het versturen van gegevens en dat niet alle bijbehorende software goed beveiligd is. Het voorbeeld van sensoren die vocht in landbouwgrond meten, bleef mij bij. Daarbij dacht ik aan het delen van die gegevens met grote bedrijven. Een toepassing kan praktisch zijn, terwijl er tegelijk vragen ontstaan over wie de verzamelde informatie ontvangt en gebruikt.

Bij de presentatie over zoekmachines kende ik Google en DuckDuckGo al. Ook wist ik dat zoekresultaten een bepaalde volgorde hebben. De uitleg over crawlers, ook wel spiders genoemd, was nieuw voor mij. Die ontdekken pagina's, waarna zoekmachines de inhoud verwerken en indexeren. Het onderscheid tussen advertenties en de gewone zoekresultaten vond ik interessant. Vooral dat betalen voor advertenties niet betekent dat je daarmee de eerste organische positie koopt. De uitleg over personalisatie en het volgen van gebruikers via advertenties zette mij aan het denken over de invloed daarvan op wat je online tegenkomt.

Bij de presentatie over online betalen kende ik iDEAL al als betaalmethode. Nieuw voor mij was de uitleg dat iDeal een doorverwijzingssysteem is. Vooral de voorbeelden van misleiding bleven hangen. Een link kan de indruk wekken dat je geld ontvangt, terwijl je juist een betaling goedkeurt. Het verschil tussen echt en nep is ook erg klein. Dat zette mij aan het denken over hoe belangrijk het is om niet alleen naar het uiterlijk van een betaalpagina te kijken, maar ook naar welke handeling je daadwerkelijk bevestigt.