Generative AI

Tjorn Brederoo en Raoul Nedermeijer

*Klas 1 Damian Gans*

# Inhoudsopgave

[**Inhoudsopgave 2**](#_a6if92aiqs04)

[**1. Oriënteren 4**](#_3hyp3vxsnkpt)

[1.1 Bronnen 4](#_2tthbzv7g9iv)

[1.1.1 Zoektermen 4](#_5kdfk9zaqtwo)

[1.1.2 Informatie vinden 4](#_hyl1l9eyl4jw)

[1.2 Verwachtingen 5](#_wv69hkmg58q2)

[**2. Generatieve AI 6**](#_wfovz1tccxqq)

[2.1 Wat is generatieve AI 6](#_46p4emlr265i)

[2.2 Soorten generatieve AI 7](#_qu42ual27ual)

[2.2.1 Autoregressive Models 7](#_i8matp3w532j)

[2.2.2 Diffusion 7](#_bc7y7e134tyu)

[2.2.3 Generative Adversarial Networks (GAN) 7](#_99uftwevsjca)

[2.2.4 Variational Auto encoders 7](#_d38wjpcujolp)

[2.2.5 Flow-based Modellen 8](#_cn47jhte003w)

[2.2.6 Transformer 8](#_j8ukkpa2wt3x)

[2.2.7 CNN 9](#_jsm2bqlzhbc8)

[2.2.8 RNN 9](#_28nlrexmm08q)

[2.2.9 Overzicht verschillen 10](#_zajgyjmw45d3)

[**3. Diepgang 11**](#_inzulo2zqdpf)

[3.1 Diffusion 11](#_saeplatpz7ij)

[3.1.1 Werking 11](#_svhnou1vn90)

[3.1.2 Toepassingen 11](#_cuzg6p84ppew)

[3.1.3 Impact 12](#_2yzwc4atovko)

[3.2 Autoregressie 13](#_20zmqcudldtq)

[3.2.1 Werking 13](#_3insqz3b3ccg)

[3.2.1.1 Lineaire regressie 13](#_9cmliubihthn)

[3.2.1.2 Autoregressie 14](#_szv9rjnqqjgt)

[3.2.2 Toepassingen 14](#_z9lzkpou9o4r)

[3.2.3 Impact 15](#_5ym59v2yq2rb)

[3.3 Toepassen op de Haagse Hogeschool 16](#_8q8h2lkx1u5k)

[3.3.1 Tools voor de medewerkers 16](#_pxcju4fxy6iu)

[3.3.2 Onderwerpen voor toekomstige studenten 16](#_39vkk0lo56ms)

[**4. Het toepassen van generative AI 17**](#_i7ox5zmy7f9a)

[4.1 Hardware 17](#_xoq4yuf089vt)

[4.1.1 GPU 17](#_n5wszlipnk9q)

[4.1.2 CPU 17](#_t3p619douxau)

[4.1.3 AI-accelerator 17](#_gr8m16qh463b)

[4.1.4 RAM 18](#_gkfhr2sffgmi)

[4.2 Frameworks 18](#_ovb81n19cbn2)

[4.2.1 TensorFlow 19](#_hygijk1xk7tw)

[4.2.2 PyTorch 19](#_esh3lw2ywo3b)

[4.3 Demo 19](#_a6k2h3lryxce)

[4.3.1 Wat is RAG 19](#_bcrponil87h)

[4.3.1.1 Vector database 20](#_p7ywzud1kxog)

[4.3.1.2 Reranker 20](#_aruh0f3nbdf5)

[4.3.2 EasyRag 20](#_n7bqn86sp551)

[**Samenvatting en conclusie 22**](#_6dzzo4ppq9pe)

[Samenvatting 22](#_ovdye2twqtan)

[Conclusie 22](#_gfvrowqy1cvu)

[**Reflectie 23**](#_4x1kkym1cfe4)

[Reflectie Tjorn Brederoo 23](#_xo1ir0bzs9aw)

[Reflectie Raoul Nedermeijer 23](#_t0kx04o5qjyk)

[**Bronvermelding 24**](#_57ba9sx46iut)

# 

# Oriënteren

AI is een verzamelterm voor een systeem of stuk software dat is ontworpen om taken uit te voeren waar anders menselijke intelligentie voor nodig is. Generatieve AI is AI die media genereert als output (Harvard, 2023). Wellicht wel het bekendste voorbeeld hiervan is ChatGPT.

In dit onderzoek gaan wij in op de werking en de toepassingen van generatieve AI. Wij gaan verschillende typen generatieve AI kort toelichten en vervolgens dieper ingaan op twee soorten hiervan. Hierna zullen we ingaan op de toepassingen en risico’s in ons toekomstige werkveld, software-engineering, en de bredere samenleving. Hierbij zullen wij ook kort de hardware kant toelichten, om zo een beter beeld te kunnen geven over wat er allemaal nodig is voor generatieve AI.

Aan het eind van dit onderzoek zullen wij een toepassing van generatieve AI toelichten. Hiervoor hebben wij gekozen door [EasyRAG van Assistanslab](https://github.com/AssistantsLab/AssistantsLab-EasyRAG). Dit is een project waar een vriend van ons aan meewerkt en persoonlijk vinden wij dit een interessante toepassing. Dit onder andere omdat het lokaal toepasbaar is, ook zonder internetverbinding.

Ondanks onze uren aan ChatGPT ervaring, is generatieve AI niet onze expertise. Daarom is het van belang om een goede basis op te stellen voor dit onderzoek. Dit hoofdstuk beschrijft hoe wij te werk zullen gaan om een beter begrip van generative AI te krijgen.

## 1.1 Bronnen

Betrouwbare bronnen zijn essentieel voor het uitvoeren van een goed onderzoek. Om deze bronnen te vinden hebben wij het volgende stappenplan bedacht.

### 1.1.1 Zoektermen

* Generative AI
* Types of generative AI
* Applications of generative AI
* Popular generative ai models
* Impact of generative AI on society and businesses
* Role of software developers in generative AI
* Use of generative AI in education

### 1.1.2 Informatie vinden

Om betrouwbare bronnen te vinden, zullen we op de volgende manieren zoeken. Hierbij gaan we gebruik maken van de genoemde zoektermen.

* Zoeken naar academische artikelen en andere betrouwbare online bronnen die inzicht bieden in generative AI en de verschillende typen ervan.
* Zoeken naar toepassingen van generatieve ai op onder andere HuggingFace, GitHub en academische artikelen.
* Analyseren van casestudies en voorbeelden van organisaties die generative AI toepassen in verschillende sectoren.
* Informatie verzamelen over de technische vereisten, programmeervaardigheden en toolkits/frameworks (TensorFlow, PyTorch, etc.) die nodig zijn voor software engineers om met generatieve AI te werken.

Om meer informatie over EasyRag te vergaren zullen we de code bekijken en het programma uitvoeren. Verder zullen we een interview uitvoeren met één van de breinen achter EasyRag, Michiel Kamphuis. Michiel is een goede kennis van ons en heeft ongeveer anderhalf jaar gestudeerd aan de Haagse Hogeschool.

## 1.2 Verwachtingen

Door deze informatiebronnen te raadplegen, hopen we een goed begrip krijgen van wat generative AI is. Wij gaan uitzoeken wat de verschillende typen en hun toepassingen zijn. Ook willen wij de risico’s en impact in kaart brengen. De zoektermen zijn allemaal in het Engels, omdat dit een veel gebruikte taal is in de programmeerwereld. Wij verwachten meer Engelse resultaten te kunnen vinden dan Nederlandse. Voor de casestudies en ervaringen zullen we wellicht wel Nederlandse bronnen kunnen vinden.

Wij verwachten dat deze manieren van informatie vinden genoeg zal zijn om de vragen te beantwoorden, maar wij verwachten wel meer zoektermen te gebruiken dan hier benoemd. Deze zoektermen zijn globaal. Tijdens dit onderzoek gaan wij ook gerichter en dieper moeten zoeken, zoals op de twee soorten generatieve AI die wij gaan toelichten.

# Generatieve AI

Generatieve AI komt de laatste tijd veel in het nieuws, maar ermee werken is net magie. Het is voor veel mensen onbekend en onbekend is eng (Robson, 2022). Om licht te schijnen op dit magische stuk techniek, zullen we in dit hoofdstuk uitleggen wat generatieve AI is. Ook zullen we globaal ingaan op een aantal verschillende soorten generatieve AI en deze met elkaar vergelijken.

## 2.1 Wat is generatieve AI

Generatieve ai is te definiëren als een AI-systeem dat media genereert op basis van invoer media (Harvard, 2023). Wanneer je gebruik maakt van kunstmatige intelligentie om originele media als afbeeldingen, tekst, video of audio te genereren, spreek je van generatieve AI.

Deze technologie heeft de afgelopen jaren een snelle groei doorgemaakt in zowel geavanceerdheid als populariteit, vooral sinds de lancering van ChatGPT in november 2022. De capaciteit om op aanvraag inhoud te genereren heeft brede implicaties in diverse domeinen, zoals academisch onderzoek en de creatieve sector (Caulfield, 2023). Hoewel kunstmatige intelligentie al langer bestaat, maakt generative AI “nieuwe” data in plaats van een voorspelling op basis van de invoer (Explained: Generative AI, 2023).

De meerderheid van de generatieve kunstmatige intelligentie maakt gebruik van diepgaande leertechnologieën, zoals grote taalmodellen (Large Language Models(LLM’s)). Deze modellen worden getraind met een enorme hoeveelheid gegevens, zoals tekst, om patronen te herkennen. Op deze manier kunnen ze de juiste antwoorden geven op vragen van gebruikers (Caulfield, 2023).

Generatieve AI heeft meerdere toepassingen. De primaire toepassing is het genereren van nieuwe media, zoals tekst, afbeeldingen en videos. Deze gegenereerde media is vaak niet van menselijke media te onderscheiden. Generatieve AI kan echter ook gebruikt worden voor het verbeteren van andere kunstmatige intelligentie, zoals het genereren van nieuwe trainingsdata voor image recognition modellen. Ongeacht de toepassing heeft generatieve AI de kans om tijd en geld te besparen (NVIDIA, z.d.).

## 

## 2.2 Soorten generatieve AI

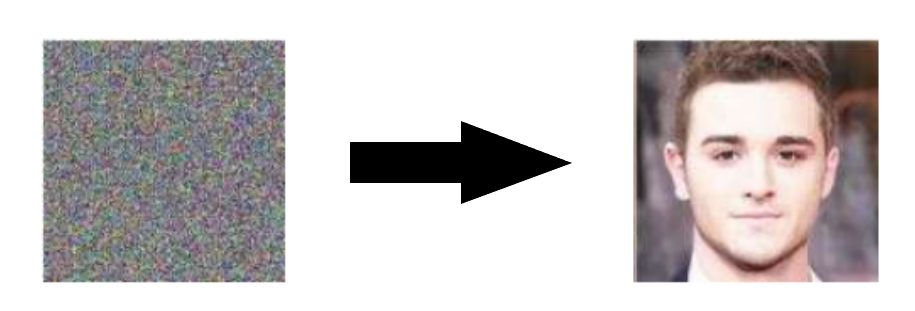
Er zijn verschillende soorten generatieve AI. We gaan een aantal soorten beknopt bespreken en deze met elkaar te vergelijken, om zo een beter begrip van de werking te krijgen.

### 2.2.1 Autoregressive Models

Autoregressieve modellen zijn statistische modellen die gebruikt worden voor het begrijpen en het voorspellen van toekomstige waarden in een tijdreeks op basis van eerdere waarden (DeepAI, 2020). Een voorbeeld van een autoregressief model is chatGPT (Paruchuri, 2023). Op autoregressive modellen zal later in dit document dieper ingegaan worden.

### 2.2.2 Diffusion

De diffusion methode wordt gebruikt om afbeeldingen en video's van hoge kwaliteit te produceren. Bekende modellen die hier gebruik van maken zijn GLIDE, DALL.E en Stable Diffusion (Ahirwar, 2023). Het model wordt getraind met bepaalde data, waarmee het vervolgens vergelijkbare afbeeldingen kan produceren (Sajid, 2023). Later in dit document zal de diffusion methode verder toegelicht worden.



*Fig 1:* [Versimpeld voorbeeld werking diffusion models](https://arxiv.org/pdf/2006.11239.pdf?ref=assemblyai.com)

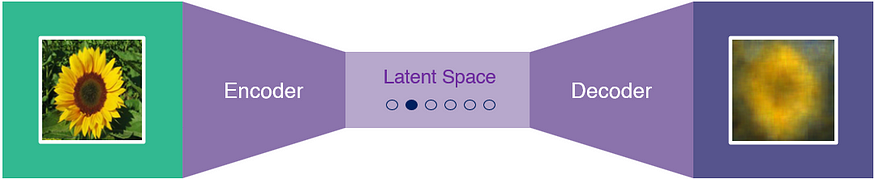
### 2.2.3 Generative Adversarial Networks (GAN)

Bij GAN worden twee modellen gebruikt, een generator en een discriminator. De generator maakt nieuwe data. De discriminator bepaalt of de data gemaakt is door de generator of dat het echte data is. Het doel van de generator is om de discriminator te laten geloven dat de neppe data echt is. De twee modellen trainen elkaar en worden zo steeds beter (Google, z.d.).

### 2.2.4 Variational Auto encoders

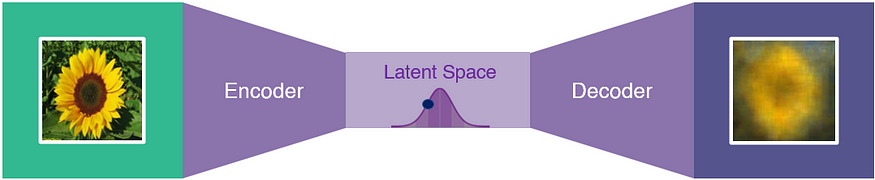
Een variationele auto encoder bestaat uit een auto encoder en een variationele inferentie. De auto encoder bestaat op zijn beurt weer uit een encoder en een decoder, met de latent space als een soort database er tussen.

De encoder zet de invoer om in een abstracte representatie met alleen de belangrijkste informatie. Deze abstracte representatie is de latent space. De decoder pakt deze informatie en zet het weer om van latent space naar een uitvoer.



*Fig 2:* [Weergave van een auto encoder](https://miro.medium.com/v2/resize:fit:1100/format:webp/1*NRkrdWes8wVWVpJJ-DLp1Q.png)

Een nadeel hiervan is dat een auto encoder alleen de invoer afbeeldingen kan reconstrueren. Dit is waar de variationele inferentie van pas komt. Hierbij wordt een beredeneerde schatting gemaakt over de uitkomst. Als je op een lichtschakelaar gedrukt wordt, verwacht je dat het licht aangaat. Hiervoor hoef je niet te weten hoe het elektriciteitsnet werkt. Door variationele inferentie kan de auto encoder wel nieuwe data genereren (Lynch, 2023).



*Fig 3:* [Een variational autoencoder met een verdeling als latent space](https://miro.medium.com/v2/resize:fit:1100/format:webp/1*tpDSDnFQNVA1GtG4p6avlA.png)

### 2.2.5 Flow-based Modellen

Flow-based modellen vormen simpele data om naar complexere data. Deze transformaties zijn omkeerbaar, waardoor je ook van complexe data naar simpele data kan gaan, wat handig is voor compressie (Kumawat, 2023). Flow-based modellen genereren data die lijkt op de invoer. Binnen generatieve AI kan dit ingezet worden om een mix van twee gezichten (de invoer) te maken (OpenAI, 2018).

### 2.2.6 Transformer

Transformers transformeren de invoerdata om zo tot een uitvoer te komen. Een transformer kan zo woorden uit de invoer gebruiken om nieuwe zinnen te genereren.

Transformers doen dit door naar verschillende stukken van de data tegelijkertijd te kijken en hier vervolgens de belangrijkste dingen uit te filteren.

Als eerst wordt de invoer opgebroken en omgevormd naar een wiskundige representatie. De opgebroken invoer wordt vervolgens gelabeld. Hierdoor kan het model verschillende stukken aan elkaar linken. Transformer modellen verwerken de input niet op volgorde.

Transformer modellen komen in verschillende soorten. Een belangrijke hiervan zijn GPT’s, zoals ChatGPT. In een GPT zorgt een transformer model voor tekst generatie die rekening houdt met stijl en toon (What Are Transformers? - Transformers in Artificial Intelligence Explained - AWS, z.d.).

### 2.2.7 CNN

Een CNN is een model voor classificatie. Het bestaat uit meerdere lagen met knooppunten, waarbij elk knooppunt een eigen gewicht en drempelwaarde heeft. Het knooppunt wordt geactiveerd als de data over die drempelwaarde heen gaat, waarna het knooppunt de data doorstuurt naar de volgende laag.

De convolutional layer is de eerste primaire laag van een CNN. Hier bewegen filters over de invoer afbeelding om kenmerken te detecteren. Dit gebeurt door middel van 2D-gewichtsmatrices die op delen van de afbeelding worden toegepast om output matrices te produceren die de aanwezige kenmerken aangeven. CNN’s kunnen meerdere convolutional layers hebben, waarbij de layers de vorige output array gebruiken als hun invoer. Hierdoor kan er steeds specifieker naar kenmerken gekeken worden. Aan het einde van de convolutional layer is de afbeelding uitgedrukt in numerieke waarden, die de rest van het neurale netwerk kan gebruiken om te kijken wat er in de afbeelding staat.

Na de convolutional layer komt een pooling layer. Het enige wat deze laag doet, is een nieuwe array maken op basis van de gemiddelden. Hiermee gaat veel informatie verloren, maar het helpt met efficiëntie en snelheid van de CNN.

De laatste laag in een CNN is de fully connected layer. Deze laag vervormt de array en pakt hier de hoogste waarde uit. Deze waarde is de classificatie die het model aan de data geeft.

CNN’s worden vooral gebruikt voor image recognition, het herkennen van afbeeldingen (Chatterjee, 2019). Binnen generative AI kunnen CNN’s gebruikt worden binnen een GAN, als discriminator.

### 2.2.8 RNN

Een RRN is ontworpen om sequentiële (geordende) invoer te verwerken, zoals een zin woord voor woord verwerken. Wat RRN’s anders maakt dan bijvoorbeeld een CNN, is dat de lagen een geheugen hebben. Dit maakt RRN’s veel accurater in voorspellingen. Dit samen maakt een RNN een goede keuze voor sequentiële data zoals tekst en video.

De eerste laag is een input laag, die de data een voor een doorgeeft aan de volgende laag. Dan komen de zogenoemde hidden layers die een functie uitvoeren op de data. Belangrijk hierbij is dat deze layers elementen uit eerdere invoer kan onthouden. Dit komt doordat data in een soort loop door de hidden layers gestuud wordt. De data gaat dus niet recht door alle layers heen. De laatste laag is de outpul layer, die voor de uitvoer zorgt (Donges & Urwin, 2024).

Door het geheugen van RRN’s zijn ze goed voor het maken van voorspellingen. Binnen generatieve AI kan RNN toegepast worden op het genereren van media., zoals tekst, video en audio. Ze zijn echter wel lastig om te trainen en te schalen. Recent onderzoek brengt hier verandering in (De et al., 2024).

### 2.2.9 Overzicht verschillen

Elk type generatieve AI heeft zijn eigen sterke en zwakke punten, waardoor ze geschikt zijn voor verschillende taken:

* Autoregressive Models: Uitstekend in het genereren van tekst, maar kan traag zijn voor lange sequenties en gevoelig voor fouten in de beginfase.
* Diffusion: Genereert afbeeldingen en video's van hoge kwaliteit, maar vereist veel rekenkracht en data.
* Generative Adversarial Networks (GANs): Geschikt voor het creëren van realistische afbeeldingen, maar kan gevoelig zijn voor model collapse. Bij model collapse loopt het model vast en geeft het telkens dezelfde output (Common Problems, z.d.).
* Variational Autoencoders (VAEs): Goed in het leren van de onderliggende structuur van data en het genereren van variaties, maar de gegenereerde resultaten zijn mogelijk minder scherp dan GANs.
* Flow-based Modellen: Ideaal voor het transformeren van data en het genereren van variaties, maar genereert alleen uitvoer die lijkt op de invoer (OpenAI, 2018).
* Transformers: Krachtig in het verwerken van sequentiële data zoals tekst en het genereren van samenhangende en contextueel relevante resultaten.
* Convolutional Neural Networks (CNNs): Perfect voor beeldherkenning en classificatie (Chatterjee, 2019), maar minder geschikt voor het genereren van nieuwe beelden.
* Recurrent Neural Networks (RNNs): Goed in het verwerken van sequentiële data (Donges & Urwin, 2024), zoals tekst en video, maar kan lastig te trainen en te schalen zijn (De et al., 2024).

De keuze voor het beste model hangt af van de specifieke toepassing, de beschikbare data en de gewenste output. Het is ook mogelijk om verschillende modellen te combineren, zoals gebeurt bij multimodal modellen.

# 

# 3. Diepgang

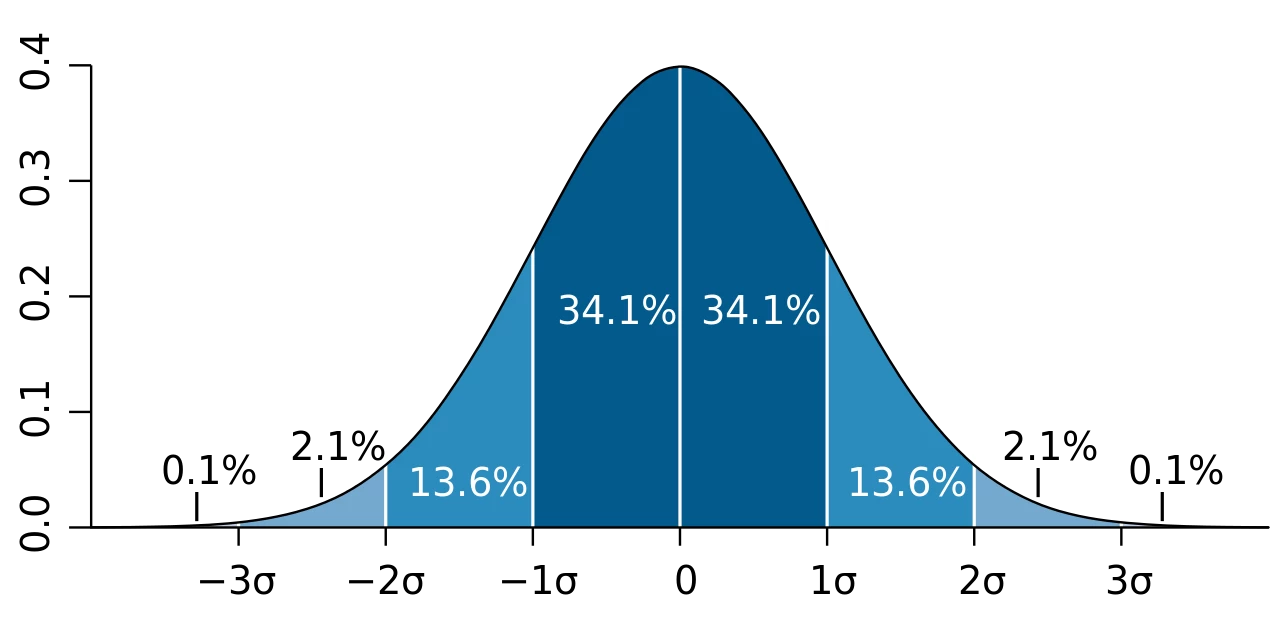
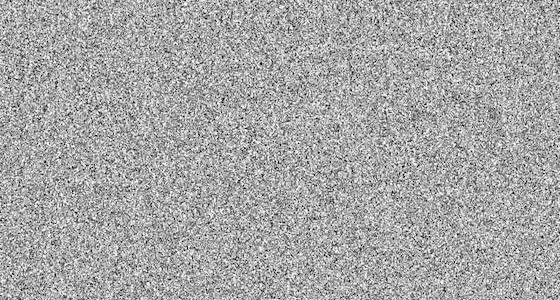
In dit hoofdstuk zullen we verder ingaan op de werking van twee soorten generatieve AI, namelijk diffusion en autoregressive models. Deze twee modellen zijn gekozen omdat wij ze interessant vinden en bekend zijn met een aantal toepassingen van beiden. We willen graag meer te weten komen over deze modellen en hun toepassingen. Dit doen we door de toepassingen van deze modellen en de impact die ze hebben op de samenleving te bespreken.

## 3.1 Diffusion

Zoals eerder verteld is de diffusion methode een manier om onder andere afbeeldingen en video’s te creëren door middel van kunstmatige intelligentie. In dit hoofdstuk zal hier verder op ingegaan worden.

### 3.1.1 Werking

|  |
| --- |

In de basis werkt de diffusion methode als volgt: tijdens het diffusieproces wordt er bij een bepaalde afbeelding (X0) geleidelijk ‘Gaussian noise’ toegepast (Adaloglou, 2022). Gaussian noise is een statistische ruis met een normale verdeling. Dit betekent dat de ruiswaarden op een normale manier worden verdeeld (*Gaussian Noise | CloudFactory Computer Vision Wiki*, z.d.). 

*Gaussian noise -* [*bron*](https://opendatascience.com/generating-data-with-random-gaussian-noise/) *Normale (Gaussian) verdeling -* [*bron*](https://wiki.cloudfactory.com/docs/mp-wiki/augmentations/gaussian-noise#:~:text=Gaussian%20Noise%20is%20a%20statistical,added%20to%20the%20original%20image.)

Het toepassen van deze ruis wordt gedaan in een serie van een aantal (T) ‘denoising’ stappen. Vervolgens wordt een neuraal netwerk getraind om de originele data te herstellen door het proces van het toevoegen van Gaussian noise terug te draaien. Door het proces van het terugdraaien is het mogelijk geworden om nieuwe data te genereren. De stappen in het denoising proces (het weghalen van de gaussian noise om zo weer tot een afbeelding te komen) werken met een Markov-keten, wat inhoudt dat elke stap in het proces alleen afhankelijk is van de vorige stap (Adaloglou, 2022).

### 3.1.2 Toepassingen

Diffusion wordt op verschillende manieren gebruikt. Zo wordt het bijvoorbeeld gebruikt voor het produceren van zowel afbeeldingen als video's. Met text-to-video is het mogelijk om door een tekstuele beschrijving en de juiste video-instellingen op te geven aan het model, zodat het vervolgens een diffusieproces kan starten. Het model bekijkt de tekst en genereert geleidelijk de video frames. Bij elke stap in het diffusieproces worden de gerenderde frames verfijnd, waardoor ze van willekeurige ruis worden getransformeerd naar visueel zinvolle inhoud die overeenkomt met de opgegeven tekst. Naast text-to-video bestaat er ook text-to-image, waarbij door middel van een tekst als invoer een afbeelding geproduceerd wordt die overeenkomt met de tekst, en image-to-image. Diffusion modellen die tegenwoordig erg populair zijn zijn Stable Diffusion, DALL-E 2, Midjourney en GLIDE (Acharya, 2024).

### 3.1.3 Impact

Generatieve kunstmatige intelligentie stimuleert de groei van banen in plaats van alleen automatisering, zoals blijkt uit een grote toename van AI-gerelateerde vacatures op platforms zoals Upwork. Rollen zoals deep learning engineers, Stable Diffusion en DALL-E artiesten en AI-ethici worden steeds meer gevraagd. Bedrijven huren niet alleen meer vanwege de opkomst van generatieve AI, maar zijn ook van plan om extra freelancers en voltijdse medewerkers aan te nemen om de toenemende vraag naar creatief werk dat door AI wordt gefaciliteerd, aan te kunnen (Curry, 2023). Software Engineers doen er dus goed aan om zich te verdiepen of zelfs te specialiseren in AI gerelateerde onderwerpen, zoals diffusion.

Diffusion kan ook een rol spelen bij de productie van desinformatie en inbreuken op intellectueel eigendom. Verder wordt de invloed van technologie op verschillende maatschappelijke domeinen als de zorg of het onderwijs vergroot (*Risico’s van Generatieve AI Vereisen Terughoudendheid in Gebruik | Rathenau Instituut*, z.d.). Diffusion is tegenwoordig in staat om superrealistische afbeeldingen te genereren waardoor het relatief makkelijk is om desinformatie te verspreiden. Zo is het bijvoorbeeld al gebruikt om afbeeldingen te genereren van de paus gekleed in hypermoderne kleding en een neppe arrestatie van Donald Trump (Radauskas, 2023).

## 

## 3.2 Autoregressie

In dit hoofdstuk zal verder worden ingegaan op de werking autoregressie modellen. Zoals eerder verteld zijn autoregressie modellen goed in voorspellingen.

### 3.2.1 Werking

Autoregressie modellen voorspellen automatisch het volgende onderdeel in een reeks. Dit doen ze aan de hand van eerdere invoer in de reeks. Autoregressie is een techniek die stelt dat de huidige waarde een functie van een eerdere waarde is.

Een autoregressie model maakt gebruik van lineaire regressieanalyse om het volgende onderdeel te voorspellen. Bij regressieanalyse wordt het model voorzien van verschillende onafhankelijke variabelen die gebruikt worden voor de voorspelling (What Are Autoregressive Models? - AR Models Explained - AWS, z.d.).

#### 3.2.1.1 Lineaire regressie

Lineaire regressie is een techniek die de waarde van onbekende gegevens voorspelt door gebruik te maken van gerelateerde en bekende gegevens. De bekende en onbekende gegevens worden wiskundig gemodelleerd als een lineaire vergelijking. Lineaire regressie is een veelgebruikte en gevestigde techniek voor het voorspellen van toekomstige gebeurtenissen. Binnen autoregressie modellen betekent dit bijvoorbeeld het volgende woord.

De basis van lineaire regressie is het uitzetten van een lijngrafiek tussen twee variabelen, X en Y. X wordt uitgezet langs de horizontale as en Y langs de verticale as. X is een onafhankelijke variabele en Y een afhankelijke variabele. Dit houdt in dat Y afhangt van de uitkomst van de formule, waar X in gebruikt wordt. De bijbehorende formule maakt gebruik van constante variabelen die voor alle waarden van X en Y hetzelfde zijn.

Voor het plotten van de lijn moeten de volgende stappen ondernomen worden. Hierbij gaan we uit van Y=C\*X+M, dus twee constante waarden. Verder gaan we uit van (1,5), (2,8), en (3,11) als waarden voor X en Y respectievelijk:

1. Zet een rechte lijn uit tussen 1 en 5, en meet de correlatie hier tussen. Deze lijn is verder willekeurig en dient als startpunt.
2. Verander de lijn zo dat alle andere waarden, in dit geval (2,8) en (3,11) op de lijn passen. Als dit niet lukt zorg dan dat de lijn zo dicht mogelijk langs de nieuwe punten komt.
3. De lijn die je in dit geval hebt gevonden, voldoet aan de vergelijking Y = 3X + 2. Hierbij is 3 de helling (M) van de lijn en 2 het snijpunt (C) met de y-as. Hierbij maken we gebruik van de methode van de kleinste kwadraten.
4. Met de gevonden lineaire vergelijking Y = 3X + 2 kun je voor elke x-waarde de bijbehorende y-waarde voorspellen. Als X = 4, dan is Y = 3\*4 + 2 = 14.

Y=C\*X+M is de simpelste formule voor lineaire regressie. Er bestaan verschillende formules die verschillende voor- en nadelen hebben (Amazon, z.d.).

#### 3.2.1.2 Autoregressie

Autoregressie modellen passen lineaire regressie toe met vertragingen van de uitvoer (afhankelijke variabele) uit vorige stappen. In tegenstelling tot lineaire regressie gebruikt een autoregressie model geen andere onafhankelijke variabelen, behalve de eerder voorspelde resultaten.

Wanneer we dit in probabilistische termen uitdrukken, verdeelt een autoregressie model de onafhankelijke variabelen over n mogelijke stappen, ervan uitgaande dat eerdere variabelen invloed hebben op de uitkomst van de volgende. We kunnen autoregressie modellen dan ook uitdrukken met de volgende formule:

y(t) = c + φ1y(t-1) + φ2y(t-2) + ... + φp\*y(t-p) + ε(t)

Hier is y de voorspelde uitkomst van eerdere resultaten, vermenigvuldigd met hun respectievelijke coëfficiënten φ. De coëfficiënt vertegenwoordigt de gewichten of parameters die het nieuwe resultaat beïnvloeden. De formule wordt ook beïnvloed door willekeurige ruis, wat de voorspelling kan beïnvloeden. Dit schaadt de nauwkeurigheid van een autoregressie model.

Data scientists voegen meer vertragingen toe om de nauwkeurigheid van autoregressie modellering te verbeteren. Ze doen dit door de waarde van (t) te verhogen, wat het aantal stappen in de tijdreeks van gegevens aangeeft. Een hoger aantal stappen betekent meerdere vorige voorspellingen die meegenomen worden in de huidige. Het verhogen van vertragingen resulteert echter niet altijd in verbeterde nauwkeurigheid. Meer stappen betekent ook meer hogere hardware vereisten (What Are Autoregressive Models? - AR Models Explained - AWS, z.d.). Een balans hierin is dus van belang.

### 3.2.2 Toepassingen

Autoregressie modellen worden binnen generatieve AI toegepast in verschillende gebieden, waaronder:

* Natural Language Processing (NLP) en Large Language Models (LLM): Autoregressie modellen vormen de basis voor veel grote taalmodellen die worden gebruikt voor taken zoals tekstgeneratie, machine vertaling en chat systemen. Deze modellen voorspellen de volgende woorden of tokens in een zin op basis van de voorgaande context.
* Genereren van nieuwe trainingsdata: Autoregressie modellen kunnen worden gebruikt om extra synthetische trainingsdata te genereren, wat nuttig kan zijn wanneer er beperkte hoeveelheden echte trainingsdata beschikbaar zijn. Door het model te trainen op de beschikbare data, kan het vervolgens nieuwe, realistische voorbeelden genereren die kunnen worden toegevoegd aan de trainingsset om de prestaties te verbeteren (What Are Autoregressive Models? - AR Models Explained - AWS, z.d.).
* Beeldgeneratie: Modellen zoals PixelRNN maken gebruik van autoregressieve principes voor het genereren van afbeeldingen. Deze modellen voorspellen de pixelwaarden van een afbeelding op basis van de omringende pixels die al zijn gegenereerd, een pixel tegelijk. Dit maakt het mogelijk om complexe en realistische afbeeldingen te genereren vanaf een leeg canvas (Van den Oord et al., 2016).

Autoregressie modellen hebben buiten generatieve AI ook andere toepassingen, zoals bij het voorspellen van tijdreeksen. Denk hierbij aan: voorspellen van aandeelkoersen, verwachte verkopen of weersvoorspellingen. In deze toepassingen wordt de volgende waarde in de tijdreeks voorspeld op basis van eerdere waarnemingen (What Are Autoregressive Models? - AR Models Explained - AWS, z.d.).

### 3.2.3 Impact

Door de opkomst van krachtige taalmodellen zoals Language Models (LLMs) worden software developers in staat gesteld efficiënter te coderen, aangezien deze modellen kunnen assisteren bij het genereren van code, het vinden van fouten en andere suggesties kunnen maken. LLMs bieden dus potentieel voor verhoogde productiviteit in de softwareontwikkeling. GitHub Copilot, een tool die software developers ondersteund via een LLM, claimt zelfs developers 55% sneller coderen met deze tool (Kalliamvakou, 2023).

Autoregressiemodellen kunnen AI-wetenschappers ondersteunen door het genereren van extra synthetische trainingsdata op basis van de reeds beschikbare trainingsdata. Dit vergroot de hoeveelheid data waarop de modellen kunnen leren, wat kan leiden tot betere prestaties (What Are Autoregressive Models? - AR Models Explained - AWS, z.d.).

Er zijn initiatieven zoals het Devin-project die erna streven software developers grotendeels te vervangen door AI-systemen. Deze systemen moeten zelfstandig kunnen coderen op basis van functionele specificaties en andere invoer (Introducing Devin, The First AI Software Engineer, 2024).

De inzet van LLMs als chatbots voor klantenondersteuning in diverse sectoren kan ertoe leiden dat menselijke medewerkers hun banen verliezen, omdat de AI-systemen veel van deze taken goedkoper en mogelijk zelfs beter kunnen vervullen. Klarna meldt dat het gebruik van een LLM in hun support chats tot $ 40 miljoen US-dollars extra winst zal leiden in 2024 (Klarna AI Assistant Handles Two-thirds Of Customer Service Chats in Its First Month | Klarna International, 2024).

Hoewel ChatGPT en soortgelijke taalmodellen een schat aan kennis kunnen overdragen en mensen kunnen helpen met leren, bestaat er een risico op hallucinaties, waarbij het model onjuiste of ongefundeerde informatie genereert alsof het waar is (Risico’s van Generatieve AI Vereisen Terughoudendheid in Gebruik | Rathenau Instituut, z.d.). Daarnaast is plagiaat een probleem, omdat een model teksten kan genereren die sterk lijken op bestaande bronnen zonder correcte attributie. Tijdens het trainen gaat het ook mis, de data wordt zomaar van het internet afgehaald, meestal zonder expliciete toestemming. Hierdoor spande onder andere The New York Times een rechtszaak tegen OpenAI en Microsoft aan (Mac & Grynbaum, 2023).

## 

## 3.3 Toepassen op de Haagse Hogeschool

Generatieve AI is op dit moment erg populair. Toepassingen zoals ChatGPT worden door veel leerlingen gebruikt. Wij verwachten dat deze trend voort blijft houden. In dit hoofdstuk gaan we in op hoe generatieve AI op de Haagse Hogeschool gebruikt kan worden.

### 3.3.1 Tools voor de medewerkers

Binnen de Haagse Hogeschool kan generatieve AI op verschillende manieren worden toegepast. Zo zijn er voor docenten bijvoorbeeld een heel aantal tools die hen kunnen helpen in het onderwijs. Gradescope is hier een van. Gradescope is een tool die gebruikt kan worden voor het beoordelen van het schoolwerk van de studenten. Met deze tool wordt het makkelijker gemaakt om studenten te voorzien van feedback (*Gradescope | Save Time Grading*, z.d.).

Een andere tool die sommige docenten nu al gebruiken zijn AI-detectors. Deze worden gebruikt om te controleren of het ingeleverde werk van studenten door AI gegenereerd is of niet. Echter, op dit moment is geen enkele AI-detector in staat om met 100% zekerheid te zeggen of een tekst door AI gegenereerd is. Hierbij ontstaat er dus het gevaar dat teksten onterecht bestempeld worden als een door AI gegenereerde tekst (Scribbr, 2023). Het is dus belangrijk dat docenten hiervan op de hoogte zijn wanneer ze documenten van studenten beoordelen.

### 3.3.2 Onderwerpen voor toekomstige studenten

Niet voor elke opleiding is generative AI even relevant. Wij vinden het echter wel belangrijk dat er op elke opleiding in zekere mate les of voorlichting gegeven wordt over het gebruik van generatieve AI. ChatGPT wordt tegenwoordig veel gebruikt door studenten voor allerlei school opdrachten. Het is belangrijk om studenten op de hoogte te stellen hoe je ChatGPT het beste kunt gebruiken en wat de gevaren zijn. Een van de gevaren die hierbij komt kijken is dat het makkelijk is om plagiaat te plegen door hele teksten over te kopiëren vanuit ChatGPT (*ChatGPT: Wat Kan (en Mag) Er? - Universiteit Leiden*, z.d.).

Naast het risico op het plegen van plagiaat zijn er meer gevaren die komen kijken bij generatieve AI. Zo kan generatieve AI mogelijk zorgen voor productie van desinformatie en inbreuken op intellectueel eigendom. Verder wordt de invloed van technologie op verschillende maatschappelijke domeinen als de zorg of het onderwijs vergroot (*Risico’s van Generatieve AI Vereisen Terughoudendheid in Gebruik | Rathenau Instituut*, z.d.). Doordat generatieve AI naast positieve toepassingen ook genoeg risico’s met zich meebrengt, vinden wij het verstandig om voorlichting te geven hierover. Elke opleiding moet dan kijken naar wat relevant is voor het werkgebied dat van toepassing is.

# 4. Het toepassen van generative AI

Het trainen van generatieve AI-modellen kost veel hardware. Nvidia, op dit moment de grootste producent van dit soort hardware (Bloomberg - Are You A Robot?, 2023), heeft hier veel profijt van. De H100, een computer chip van Nvidia, kost ongeveer evenveel als een nieuwe auto, ongeveer 40.000 dollar (Shilov, 2024). In dit hoofdstuk beschrijven we wat voor hardware gebruikt wordt voor het trainen van generatieve AI-modellen. Verder zullen we bespreken wat voor frameworks gebruikt worden voor generatieve AI en bespreken we hoe EasyRag werkt. Dit doen we om te laten zien dat je ook zelf met je eigen generatieve AI aan de slag kunt gaan.

## 4.1 Hardware

Generatieve AI-modellen hebben, net zoals alle andere kunstmatige intelligentie, hardware nodig om op uitgevoerd te worden. Kunstmatige intelligentie maakt voornamelijk gebruik van drie soorten hardware; de GPU, de CPU en een AI-accelerator. Alle drie bestaan tegenwoordig uit meerdere cores, wat je kan zien als het brein. Deze cores zijn verantwoordelijk voor het uitvoeren of het geven van opdrachten. De cores kunnen deze opdrachten los van elkaar uitvoeren, maar ze kunnen ook samenwerken als het desbetreffende programma daar support voor heeft (Computer Hope, 2022). Het is echter niet zo dat een GPU een CPU of AI-accelerator compleet vervangt, of andersom. Ze hebben allemaal hun eigen voor- en nadelen. Wel geld dat voor het uitvoeren van kunstmatige intelligentie een GPU of AI-accelerator sneller zijn dan een CPU.

### 4.1.1 GPU

GPU staat voor graphical proccesing unit. GPU’s zijn ontworpen voor parallel-processing, waardoor GPU’s meerdere berekeningen tegelijk kan uitvoeren. Dit kunnen ze, doordat GPU’s ten eerste veel meer cores hebben en ten tweede omdat de cores hier speciaal voor zijn gemaakt. GPU’s hebben vaak honderden tot duizenden cores, vele malen meer dan een CPU (Lee, 2020). GPU’s waren van origine ontwikkeld voor het versnellen van 3D grafische toepassingen, zoals videogames en video’s bewerken. GPU’s worden nog steeds veel gebruikt voor videogames en content creation, maar ook voor AI (Intel, n.d.).

### 4.1.2 CPU

CPU staat voor central processing unit. Een CPU is het brein van de computer en stuurt de andere componenten aan, waaronder dus ook de GPU (Martindale, 2023). Een CPU is beter in sequentieel processing, waardoor ze een reeks van taken snel kunnen uitvoeren. CPU’s zijn interessant bij kunstmatige intelligentie waarbij geen gebruik gemaakt wordt van parallel-processing. Verder zijn ze handig bij RNNs, omdat een RNN-gebruik maakt van sequentieel processing (ThinkML Team, 2021).

### 4.1.3 AI-accelerator

AI-accelerators zijn gespecialiseerde hardware of software die de efficiëntie van AI-applicaties en algoritmes verbeteren. Het zorgt bijvoorbeeld voor een betere accuraatheid, snellere verwerking en een verminderd energiegebruik (Lark Editorial Team, 2023). Een Neural Processing Unit (NPU) is een voorbeeld van een AI-accelerator. In tegenstelling tot traditionele CPUs en GPUs zijn NPUs specifiek ontworpen om efficient om te gaan met ingewikkelde wiskundige berekeningen die cruciaal zijn voor kunstmatige neurale netwerken (Peru, 2023).

Google Cloud Tensor Processing Units (TPUs) is een ander voorbeeld van een AI-accelerator. TPUs zijn geoptimaliseerd voor het trainen van grote AI-modellen. Ze zijn speciaal ontwikkeld om kostenefficiënt de taken, zoals training en fine-tuning, die AI uitvoert te kunnen schalen. Cloud TPUs bieden de flexibiliteit om de werklasten van verschillende AI-frameworks, zoals PyTorch, en TensorFlow, te versnellen. TPUs worden gebruikt voor bijvoorbeeld chatbots, het genereren van code, synthetische spraak en meer (*Tensor Processing Units (TPUs)*, z.d.).

### 4.1.4 RAM

De Random Access Memory (RAM) en Video Random Access Memory (VRAM) zijn van groot belang wanneer je gaat werken met AI. RAM is in principe het kortetermijngeheugen van de computer. Hier worden gegevens opgeslagen waarmee de computer op een bepaald moment mee te maken heeft (Freda, 2022). VRAM is ook een soort geheugen in de computer, maar dan direct op de GPU (Bijl, 2023). Omdat VRAM direct op de grafische kaart zit, kan de grafische kaart hier veel sneller bij dan bij de normale RAM (Bradley, 2023).

Met name bij Large Language Models (LLM’s) zijn RAM en VRAM belangrijk. LLMs, zoals GPT-3, zijn getraind met parameters en hebben enorme hoeveelheden parameters. Een parameter is een gewicht dat is toegewezen aan een aantal nummers. Hiervoor geldt: hoe groter het aantal parameters, hoe meer geheugen er nodig is. LLMs worden elke dag groter en groter. Voorbeelden van LLMs met veel parameters zijn BLOOM met 176 miljard parameters en PaLM met 540 miljard parameters (Hamza, 2023).

## 4.2 Frameworks

Python is een van de populairste programmeertalen (Stack Overflow, 2023). Ook binnen kunstmatige intelligentie wordt Python veel gebruikt.

Vanwege de populariteit zijn er verschillende frameworks ontworpen voor kunstmatige intelligentie. Deze frameworks bieden bouwstenen voor het ontwerpen, trainen en valideren van modellen (Nvidia, z.d.). Ze maken het sneller en gemakkelijker om kunstmatige intelligentie modellen te maken, zonder dat je je hoeft te verdiepen in de onderliggende algoritmes. Dit doen ze door het bieden van vooraf gebouwde en geoptimaliseerde componenten (Sharma, 2024).

In dit hoofdstuk gaan we twee van deze frameworks bespreken: Tensorflow en Pytorch. Wij hebben deze frameworks gekozen omdat ze populair zijn en allebei gebruikt worden in het project wat wij willen laten zien.

### 4.2.1 TensorFlow

TensorFlow is een machine learning framework ontworpen door Google. De eerste versie werd gepubliceerd in 2015. Hoewel de hoofdtaal Python is, is TensorFlow ook in andere programmeertalen te gebruiken, zoals JavaScript (Yegulalp, 2024).

Modellen gemaakt met TensorFlow kunnen op verscheidene hardware uitgevoerd worden, van Smartphones en Raspberry Pi tot volledige datacenters, afhankelijk van de complexiteit van het model natuurlijk. Zo kan TensorFlow ook uitgevoerd worden in de browser met JavaScript (Introduction To TensorFlow, z.d.).

Officieel heeft TensorFlow ondersteuning voor Python 3.8 tot en met 3.11. Python staat niet bekend om zijn snelheid. Gelukkig wordt de wiskunde niet door Python gedaan, maar met C++ (Yegulalp, 2024).

### 4.2.2 PyTorch

PyTorch, ontworpen door Meta (PyTorch | Meta Open Source, z.d.), stamt uit 2016. PyTorch is ontworpen om een imperatief framework te zijn, dat betekent dat het doel van het framework is om stap voor stap te kunnen programmeren en elke stap individueel te kunnen testen. Als je op deze manier een framework ontwerpt kan de gebruiker veel sneller een project beginnen en afronden. Het is dus niet nodig om het hele programma uit te voeren om te testen of een nieuw deel werkt (Shetty, 2018).

Net zoals TensorFlow heeft PyTorch ondersteuning voor verschillende programmeertalen. Officieel ondersteunt PyTorch Python 3.8-3.11, maar het heeft ook een versie voor C++ en Java. Ondersteuning voor smartphones is op dit moment experimenteel (PyTorch, z.d.).

## 4.3 Demo

Als demo hebben wij gekozen voor EasyRag van Assistanslab. Hiervoor hebben wij gekozen, omdat dit een lokale toepassing van generatieve AI is. Je bent hierdoor niet afhankelijk van een service van een derde partij. De privacy die hieruit voortkomt kan voor personen, bedrijven en instanties van belang zijn.

In dit hoofdstuk gaan we in op de werking van EasyRag en de achterliggende systemen. We bespreken wat EasyRag anders maakt dan de competitie en stellen een aantal vragen aan Michiel Kamphuis.

### 4.3.1 Wat is RAG

Retrieval-Augmented Generation (RAG) is een methode waarbij de output van een Large Language Model (LLM) wordt geoptimaliseerd door te verwijzen naar een betrouwbare kennisbron buiten de bronnen waarop het model is getraind, voordat het een antwoord genereert. LLM’s worden getraind met enorme hoeveelheden gegevens en gebruiken complexe parameters om nieuwe inhoud te produceren, zoals het beantwoorden van vragen, het vertalen van talen en het voltooien van zinnen. RAG breidt deze capaciteiten verder uit naar specifieke gebieden of interne kennisbanken van organisaties, zonder dat het model opnieuw getraind hoeft te worden. Dit is een efficiënte aanpak om de output van taalmodellen te verbeteren, zodat deze toepasbaar, nauwkeurig en relevant blijft in diverse situaties (*What Is RAG? - Retrieval-Augmented Generation Explained - AWS*, z.d.).

#### 4.3.1.1 Vector database

RAG wordt meestal gekoppeld aan een vector database. Een vector database is een collectie van data opgeslagen als wiskundige representaties. Data wordt gelinkt aan hoe veel het op elkaar lijkt en niet op exacte overeenkomsten. Hierdoor wordt er context aan de data gegeven en kan het gekoppelde model vergelijkbare data voorstellen.

Een vector is een array van nummers, meestal kommagetallen (floats). Deze array is multidimensionaal en representeert een locatie. Elke vector wijst naar een object, de data. Vanwege de multidimensionaliteit van een vector kan de data op meerdere manieren met elkaar gelinkt worden.

Vectors die voor kunstmatige intelligentie gebruikt worden, worden embeddings genoemd (Cloudflare, z.d.). RAG kan een vector database gebruiken om de context in op te slaan als vectoren en de juiste vectoren mee te geven aan het LLM.

#### 4.3.1.2 Reranker

Een reranker gaat bij elke vraag alle data door en ze op een rij zetten van hoe belangrijk ze zijn. Dit doet het door een score te geven aan de data, gebaseerd op hoeveel het overeenkomt met de vraag. Dit maakt rerankers slomer dan het gebruiken van embeddings, maar wel accurater, omdat er per vraag opnieuw wordt bekeken welke documenten het relevantst zijn. Ook gaat bij het omzetten van de data naar vector informatie verloren, dit is niet het geval bij een reranker (Pinecone, z.d.).

### 4.3.2 EasyRag

EasyRag is een implementatie van het Retrieval Augmented Generation (RAG) model door AssistantsLab. Dit project maakt het mogelijk om lokaal een groot taalmodel (LLM) toegang te geven tot informatie in een door de gebruiker gespecificeerde map. EasyRag bevindt zich momenteel nog in de beginfase van ontwikkeling.

De interactie met het systeem verloopt via de terminal. De gebruiker geeft een bestandspad, waar EasyRag doorheen kijkt voor bruikbare bestanden. Zo werkt EasyRag op dit moment alleen met een aantal soort tekst bestanden, zoals txt, word en json bestanden. Na het verwerken van de bestanden kan de LLM bevraagd worden. Bij een vraag worden, als dat nodig is, de relevante bestanden automatisch als context meegegeven worden.

De voordelen van EasyRag zijn dat het privé en lokaal is. EasyRag verschilt van andere RAG-implementaties. In plaats van een vector database die vectoren intensief met elkaar vergelijkt, vergelijkt EasyRag integers.

RAG-systemen komen in twee hoofdvarianten voor: rerankers en vector databases. EasyRag maakt echter gebruik van een andere methode, namelijk "anchor zinnen". Deze worden omgezet naar vectoren. Het verschil in afstand van deze vectoren ten opzichten van andere tekst wordt omgezet in integers. Het voordeel van integers ten opzichte van vectoren is dat het minimale RAM-vereisten heeft en de mogelijkheid biedt om op disk op te slaan. Het embedden van teksten is een eenmalig proces, in tegenstelling tot rerankers. Het eenmalig embedden duurt wel een stuk langer, maar hier staat tegenover dat het stellen van vragen dan weer vlugger gaat.

Om meer informatie over de werking van EasyRag te vergaren, hebben wij Michiel Kamphuis vragen gesteld. Michiel is onderdeel van AssistantLabs en heeft op dit moment de meeste EasyRag code geschreven.

Michiel beschrijft de werking werkt als volgt:

1. Stel er zijn 4 vectoren en 3 anchor zinnen.
2. Voor een nieuwe zin die ge-embed wordt, berekenen we de afstanden tot de 3 anchor zinnen, resulterend in 3 integers.
3. Dit proces wordt herhaald voor alle andere embeddings, telkens met dezelfde 3 anchors.
4. Wanneer een vraag gesteld wordt, embedden we de vraag en berekenen de 3 integers ten opzichte van de anchors.
5. Vervolgens zoeken we de embedding die het dichtst bij deze 3 integers ligt, wat de relevante tekst oplevert.
6. Entity extraction wordt toegepast om de precisie te verhogen voor teksten die dicht bij elkaar liggen.
7. Zolang de domeinen van de anchor teksten verschillen, is deze methode redelijk accuraat.

Op dit moment maakt EasyRag gebruik van [flan-t5-large](https://huggingface.co/google/flan-t5-large), een model van Google. Dit model is relatief klein en heeft hierdoor goede performance, maar de gegeven antwoorden zijn hierdoor niet altijd het “mooist”. In de toekomst kunnen verschillende LLM's aan het systeem gekoppeld worden.

Michiel vertelde dat ze gaan kijken naar de mogelijkheid om twee modellen tegelijkertijd te gebruiken, één voor rag en één voor het beantwoorden van de vragen. Dit zou voor een beter balans tussen snelheid en een goed gesprek. Hij benadrukte wel dat het doel van dit rag systeem is om informatie uit bestanden te halen, niet om de beste en meest creatieve gesprekken te voeren. Grotere modellen kosten meer hardware en het is nog maar de vraag of dat het waard is (M. Kamphuis, persoonlijke communicatie, 27 maart 2024).

Vanwege het feit dat dit project nog in de beginfase zit, is het installatieproces nog niet zo gemakkelijk. Verder de [codebase](https://github.com/AssistantsLab/AssistantsLab-EasyRAG) op dit moment nog privé en gaat deze nog flink veranderen. Hierdoor is het zelf uitvoeren van de demo op dit moment nog niet mogelijk.

# Samenvatting en conclusie

## Samenvatting

Dit onderzoek heeft zich gericht op generatieve AI. Generatieve AI maakt media zoals tekst, afbeeldingen, video en audio en heeft toepassingen in diverse sectoren, van kunst tot softwareontwikkeling. Er bestaan verschillende soorten generatieve AI:

* Autoregressive Models: Gebruikt lineaire regressie om toekomstige waarden te voorspellen, met toepassingen in NLP, LLM's en beeldgeneratie.
* Diffusion: Werkt door ruis toe te voegen en te verwijderen, met toepassingen in beeld- en videogeneratie.
* GANs: Creëert realistische output door een generator en discriminator te laten concurreren.
* VAEs: Leert de onderliggende structuur van data en genereert variaties.
* Flow-based modellen: Transformeert data en genereert variaties, maar genereert alleen output die lijkt op de invoer.
* Transformers: Krachtig in het verwerken van sequentiële data zoals tekst en genereert samenhangende en contextueel relevante resultaten.
* CNNs: Perfect voor beeldherkenning en classificatie, maar minder geschikt voor het genereren van nieuwe beelden.
* RNNs: Goed in het verwerken van sequentiële data zoals tekst en video, maar kan lastig te trainen en te schalen zijn.

Generatieve AI heeft impact op werkgelegenheid, ethiek, onderwijs en wetenschap. Er zijn kansen voor productiviteitsverbetering, maar ook risico's zoals desinformatie, plagiaat en banenverlies. Tools voor docenten (vb. Gradescope) en voorlichting over ethisch en effectief gebruik voor studenten zijn essentieel.

GPU's, CPU's, AI-accelerators kunnen allemaal gebruikt worden voor het trainen van kunstmatige intelligentie. Hierbij is voldoende RAM zijn cruciaal. Frameworks zoals TensorFlow en PyTorch vereenvoudigen de ontwikkeling. EasyRag is een lokale RAG-implementatie die LLM's toegang geeft tot informatie in een specifieke map, met focus op privacy. RAG is een methode om een LLM de juiste extra informatie te geven.

## Conclusie

Generatieve AI is een snel evoluerend veld met veel potentieel en uitdagingen. Het is belangrijk om de verschillende types, toepassingen en impact te begrijpen om deze technologie effectief en verantwoord te gebruiken. Er zijn veel vragen etnische vragen, met name om hoe de datasets tot stand komen en in hoe verre generatieve AI banen gaat vervangen. Wij vinden dat de Haagse Hogeschool en andere scholen les moeten geven over generatieve AI om ervoor te zorgen dat deze technologie bewust gebruikt wordt.

De ontwikkeling van lokale oplossingen zoals EasyRag draagt bij aan privacy-bewuste implementaties. De toekomst van generatieve AI zal afhangen van verdere technologische vooruitgang, ethische overwegingen en de manier waarop we deze tools integreren in ons leven en werk.

# Reflectie

## Reflectie Tjorn Brederoo

Generatieve AI is niet helemaal nieuw voor mij. Ik vind kunstmatige intelligentie interessant en houd het globaal bij, maar ben verre van een expert. Wel heb ik mijn PWS over image recognition geschreven, waardoor ik een aantal onderliggende principes al kende.

Tijdens dit onderzoek heb ik veel geleerd over verschillende soorten generatieve AI. Dit is toch anders dan image recognition en alle soorten generatieve AI buiten diffusion modellen waren nieuw voor mij. Ik vond het interessant om te leren over deze voor mij nieuwe soorten AI. Ik ben tijdens dit onderzoek ook op nieuwe toepassingen gekomen die mij interessant lijken.

Kunstmatige intelligentie draait achter de schermen op heel veel wiskunde. Ik merkte dat dit voor mij best lastig te begrijpen is. Ik heb mijn best gedaan om de wiskundige werkingen te beschrijven waar nodig.

## Reflectie Raoul Nedermeijer

Na onderzoek gedaan te hebben naar generatieve ai heb ik meer geleerd over de verschillende vormen hiervan en hoe het in grote lijnen werkt. Eerder heb ik mij nog niet verdiept in de werking van generatieve ai, dus ik heb hier redelijk wat nieuwe kennis van opgedaan. Doordat ik af en toe afbeeldingen genereer met diverse diffusion models als Midjourney en leonardo.ai, leek het mij leuk om me te verdiepen in diffusion.

Na redelijk wat artikels en video's benaderd te hebben kwam ik erachter dat de werking van diffusion en generatieve ai in het algemeen erg ingewikkeld in elkaar zit. Ik merkte dan ook dat ik erg moeite had om het goed te begrijpen en om het te documenteren. Uiteindelijk is het mij toch redelijk gelukt om een goede bijdrage te leveren aan het onderzoek.

# Bronvermelding

Harvard. (2023, 19 april). The Benefits and Limitations of Generative AI: Harvard Experts Answer Your Questions. Harvard Online. Geraadpleegd op 25 maart 2024, van <https://www.harvardonline.harvard.edu/blog/benefits-limitations-generative-ai>

Ahirwar, K. (2023, 19 oktober). A Very Short Introduction to Diffusion Models - Kailash Ahirwar - Medium. Medium. Geraadpleegd op 25 maart 2024, van <https://kailashahirwar.medium.com/a-very-short-introduction-to-diffusion-models-a84235e4e9ae>

Caulfield, J. (2023, 30 augustus). Wat is generative AI? | Betekenis & Voorbeelden. Scribbr. Geraadpleegd op 25 maart 2024, van <https://www.scribbr.nl/ai-tools-gebruiken/generative-ai-uitgelegd/>

Sajid, H. (2023, 31 maart). Diffusion Models in AI – Everything You Need to Know. Unite.AI. Geraadpleegd op 25 maart 2024, van <https://www.unite.ai/diffusion-models-in-ai-everything-you-need-to-know/>

Explained: generative AI. (2023, 9 november). MIT News | Massachusetts Institute Of Technology. Geraadpleegd op 25 maart 2024, van https://news.mit.edu/2023/explained-generative-ai-1109

DeepAI. (2020, 25 juni). Autoregessive model. DeepAI. Geraadpleegd op 25 maart 2024, van <https://deepai.org/machine-learning-glossary-and-terms/autoregressive-model>

Google. (z.d.). Overview of GAN Structure. Google For Developers. Geraadpleegd op 25 maart 2024, van <https://developers.google.com/machine-learning/gan/gan_structure>

Lynch, M. (2023, 1 juni). Generative AI with Variational Autoencoders - Morgan Lynch - Medium. Medium. Geraadpleegd op 25 maart 2024, van <https://medium.com/@morgan_lynch/generative-ai-with-variational-autoencoders-86d1926df6e8>

Kumawat, T. (2023, 16 mei). Deep Learning Part 6: Generative Modelling through Normalizing Flows. Medium. Geraadpleegd op 26 maart 2024, van <https://medium.com/@tejpal.abhyuday/deep-learning-part-6-generative-modelling-through-normalizing-flows-c79fffc90091>

OpenAI. (2018, 9 juli). Glow: Better reversible generative models. Geraadpleegd op 26 maart 2024, van <https://openai.com/research/glow>

Chatterjee, H. S. (2019, 16 juli). A Basic Introduction to Convolutional Neural Network. Medium. Geraadpleegd op 26 maart 2024, van <https://medium.com/@himadrisankarchatterjee/a-basic-introduction-to-convolutional-neural-network-8e39019b27c4>

NVIDIA. (z.d.). Generative AI – What is it and How Does it Work? | NVIDIA. Geraadpleegd op 26 maart 2024, van <https://www.nvidia.com/en-us/glossary/generative-ai/>

Intel. (sd). What Is a GPU? Geraadpleegd op 26 maart 2024, van https://www.intel.com/: <https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html>

Computer Hope. (2022, 18 oktober). Core. Geraadpleegd op 26 maart 2024, van https://www.computerhope.com/jargon/c/core.htm

Lee, K. C. (2020, 27 augustus). Parallel Computing — Upgrade Your Data Science with GPU Computing. Medium. Geraadpleegd op 26 maart 2024, van <https://towardsdatascience.com/parallel-computing-upgrade-your-data-science-with-a-gpu-bba1cc007c24>

Martindale, J. (2023, 2 september). What is a CPU? Here’s everything you need to know. Digital Trends. Geraadpleegd op 26 maart 2024, van <https://www.digitaltrends.com/computing/what-is-a-cpu/>

Team, T. (2021, 18 juli). CPU vs GPU in Machine Learning Algorithms: Which is Better? ThinkML. Geraadpleegd op 26 maart 2024, van <https://thinkml.ai/cpu-vs-gpu-in-machine-learning-algorithms-which-is-better/>

NVIDIA. (z.d.). Deep learning frameworks. NVIDIA Developer. Geraadpleegd op 27 maart 2024, van <https://developer.nvidia.com/deep-learning-frameworks#:~:text=Deep%20learning%20(DL)%20frameworks%20offer,a%20high%2Dlevel%20programming%20interface.&text=This%20eliminates%20the%20need%20to,build%20DL%20frameworks%20from%20source>.

Sharma, P. (2024, 25 januari). Exploring the Best Deep Learning Frameworks: A Comprehensive Guide for 2024. Analytics Vidhya. Geraadpleegd op 27 maart 2024, van <https://www.analyticsvidhya.com/blog/2019/03/deep-learning-frameworks-comparison/>

Stack Overflow. (2023, mei). 2023 Developer Survey. Geraadpleegd op 27 maart 2024, van <https://survey.stackoverflow.co/2023/>

Yegulalp, S. (2024, 5 januari). What is TensorFlow? The machine learning library explained. InfoWorld. Geraadpleegd op 27 maart 2024, van https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

Gradescope | Save time grading. (z.d.). Geraadpleegd op 27 maart 2024, van

<https://www.gradescope.com/>

Introduction to TensorFlow. (z.d.). Geraadpleegd op 27 maart 2024, van

TensorFlow. <https://www.tensorflow.org/learn>

PyTorch | Meta Open Source. (z.d.). Geraadpleegd op 27 maart 2024, van

<https://opensource.fb.com/projects/pytorch/>

Shetty, S. (2018, 20 september). What is PyTorch and how does it work? Packt Hub. Geraadpleegd op 27 maart 2024, van <https://hub.packtpub.com/what-is-pytorch-and-how-does-it-work/>

PyTorch. (z.d.). Get Started. Geraadpleegd op 27 maart 2024, van <https://pytorch.org/get-started/>

ChatGPT: Wat kan (en mag) er? - Universiteit Leiden. (z.d.). Search. <https://www.bibliotheek.universiteitleiden.nl/studenten/chatgpt#ik-ben-student-do-s-and-dont-s,ik-ben-docent-hoe-ga-ik-om-met-het-gebruik-van-chatgpt>

Risico’s van generatieve AI vereisen terughoudendheid in gebruik | Rathenau Instituut. (z.d.). <https://www.rathenau.nl/nl/digitalisering/risicos-van-generatieve-ai-vereisen-terughoudendheid-gebruik#:~:text=Risico's%20van%20generatieve%20AI%20vereisen%20terughoudendheid%20in%20gebruik,-Artikel&text=terughoudendheid%20in%20gebruik-,De%20opkomst%20van%20generatieve%20AI%2Dsystemen%20versterkt%20bekende%20risico's%20van,intellectueel%20eigendom%20en%20menselijke%20ontwikkeling>

Scribbr. (2023, 2 juni). How accurate are AI detectors? <https://www.scribbr.com/frequently-asked-questions/how-accurate-are-ai-detectors/#:~:text=AI%20detectors%20work%20by%20looking,'t%20guarantee%20100%25%20accuracy>

Donges, N., & Urwin, U. (2024, 28 februari). A Complete Guide to Recurrent Neural Networks (RNNs). Built In. <https://builtin.com/data-science/recurrent-neural-networks-and-lstm>

What is RAG? - Retrieval-Augmented Generation Explained - AWS. (z.d.). Amazon Web Services, Inc. <https://aws.amazon.com/what-is/retrieval-augmented-generation/>

Paruchuri, V. (2023, 20 maart). How does ChatGPT work? Dataquest. <https://www.dataquest.io/blog/how-does-chatgpt-work/>

Hamza, S. (2023, 2 juli). Why LLMs are consuming more memory than high end PC games? | Medium. Medium. <https://medium.com/@syedhamzatahir1001/why-llms-are-consuming-more-memory-than-call-of-duty-warzone-42e56d570498>

Lark Editorial Team. (2023, 25 december). Ai Accelerator. www.larksuite.com. Geraadpleegd op 27 maart 2024, van <https://www.larksuite.com/en_us/topics/ai-glossary/ai-accelerator>

Peru, G. (2023, 27 december). What is an NPU? Here’s why everyone’s suddenly talking about them. Digital Trends. <https://www.digitaltrends.com/computing/what-is-npu/>

Cloudflare. (z.d.). What is a vector database? Geraadpleegd op 28 maart 2024, van <https://www.cloudflare.com/learning/ai/what-is-vector-database/>

Freda, A. (2022, 20 juni). Wat is RAM-geheugen en hoe werkt het? Wat Is RAM-geheugen en Hoe Werkt Het? <https://www.avg.com/nl/signal/what-is-ram#:~:text=RAM%20staat%20voor%20Random%20Access,scherm%20naar%20de%20andere%20kant>

Bijl, M. (2023, 13 september). Hoe werkt VRAM en hoeveel heb je er van nodig? ikbouwjepc.nl. <https://www.ikbouwjepc.nl/post/hoe-werkt-vram-en-hoeveel-heb-je-er-van-nodig>

Pinecone. (z.d.). Rerankers and Two-Stage Retrieval. Geraadpleegd op 28 maart 2024, van <https://www.pinecone.io/learn/series/rag/rerankers/>

Tensor Processing Units (TPUs). (z.d.). Google Cloud. Geraadpleegd op 28 maart 2024, van <https://cloud.google.com/tpu>

What are Autoregressive Models? - AR Models Explained - AWS. (z.d.). Amazon Web Services, Inc. Geraadpleegd op 28 maart 2024, van <https://aws.amazon.com/what-is/autoregressive-models/>

Adaloglou, S. K. . N. (2022, 29 september). How diffusion models work: the math from scratch | AI Summer. AI Summer. <https://theaisummer.com/diffusion-models/>

Acharya, A. (2024, 20 maart). An Introduction to Diffusion Models for Machine Learning. <https://encord.com/blog/diffusion-models/#:~:text=Diffusion%20models%20are%20a%20class%20of%20generative%20models%20that%20simulate,a%20sequence%20of%20invertible%20operations>.

Curry, R. (2023, 10 augustus). How A.I. can help create jobs for humans, not just automate them. CNBC. <https://www.cnbc.com/2023/08/10/how-ai-can-help-create-jobs-for-humans-not-just-automate-them.html>

Amazon. (z.d.). What is Linear Regression? - Linear Regression Explained - AWS. Amazon Web Services, Inc. Geraadpleegd op 28 maart 2024, van <https://aws.amazon.com/what-is/linear-regression/>

Radauskas, G. (2023, 29 maart). Swagged-out pope, arrested Trump, and other AI fakes. www.cybernews.com. Geraadpleegd op 28 maart 2024, van <https://cybernews.com/news/ai-fake-photos-earthquake-oregon/>

Van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. Arxiv, 1601.06759. https://doi.org/10.48550/arXiv.1601.06759

Kalliamvakou, E. (2023, 29 september). Research: quantifying GitHub Copilot’s impact on developer productivity and happiness - The GitHub Blog. The GitHub Blog. <https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/>

Introducing Devin, the first AI software engineer. (2024). Geraadpleegd op 28 maart 2024, van <https://www.cognition-labs.com/introducing-devin>

Mac, R., & Grynbaum, M. M. (2023, 27 december). The Times sues OpenAI and Microsoft over A.I. use of copyrighted work. New York Times. Geraadpleegd op 28 maart 2024, van <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

Klarna AI assistant handles two-thirds of customer service chats in its first month | Klarna International. (2024, 27 februari). Geraadpleegd op 28 maart 2024, van <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>

Robson, D. (2022, 25 februari). Why we’re so terrified of the unknown. BBC. Geraadpleegd op 1 april 2024, van <https://www.bbc.com/worklife/article/20211022-why-were-so-terrified-of-the-unknown>

De, S. D., Smith, S. L. S., Botev, A. B., Christian-Muraru, G. C., Gu, A. G., Haroun, R. H., Berrada, L. B., Chen, Y. C., Srinivasan, S. S., Desjardins, G. D., Doucet, A. D., Budden, D. B., Whye Teh, Y. W. T., Pascanu, R. P., De Freitas, N. D. F., & Gulcehre, C. G. (2024). Griffin: Mixing gated linear recurrences with local attention for efficient language models. arXiv. Geraadpleegd op 11 april 2024, van <https://doi.org/10.48550/arXiv.2402.19427>

Common problems. (z.d.). Google For Developers. Geraadpleegd op 11 april 2024, van <https://developers.google.com/machine-learning/gan/problems>

Bloomberg - Are you a robot? (2023, 21 juni). Geraadpleegd op 11 april 2024, van <https://www.bloomberg.com/news/articles/2023-06-21/what-s-the-h100-the-chip-driving-generative-ai-quicktake>

Shilov, A. (2024, 2 februari). Nvidia’s H100 AI GPUs cost up to four times more than AMD’s competing MI300X — AMD’s chips cost $10 to $15K apiece;… Tom’s Hardware. Geraadpleegd op 11 april 2024, van <https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000>

Bradley, A. (2023, 14 augustus). What Is VRAM? Why Games Are Using it More, and How to Optimize Your PC. PCMAG. https://www.pcmag.com/how-to/what-is-vram-why-games-are-using-it-more-and-how-to-optimize-your-pc