> WHITEPAPER

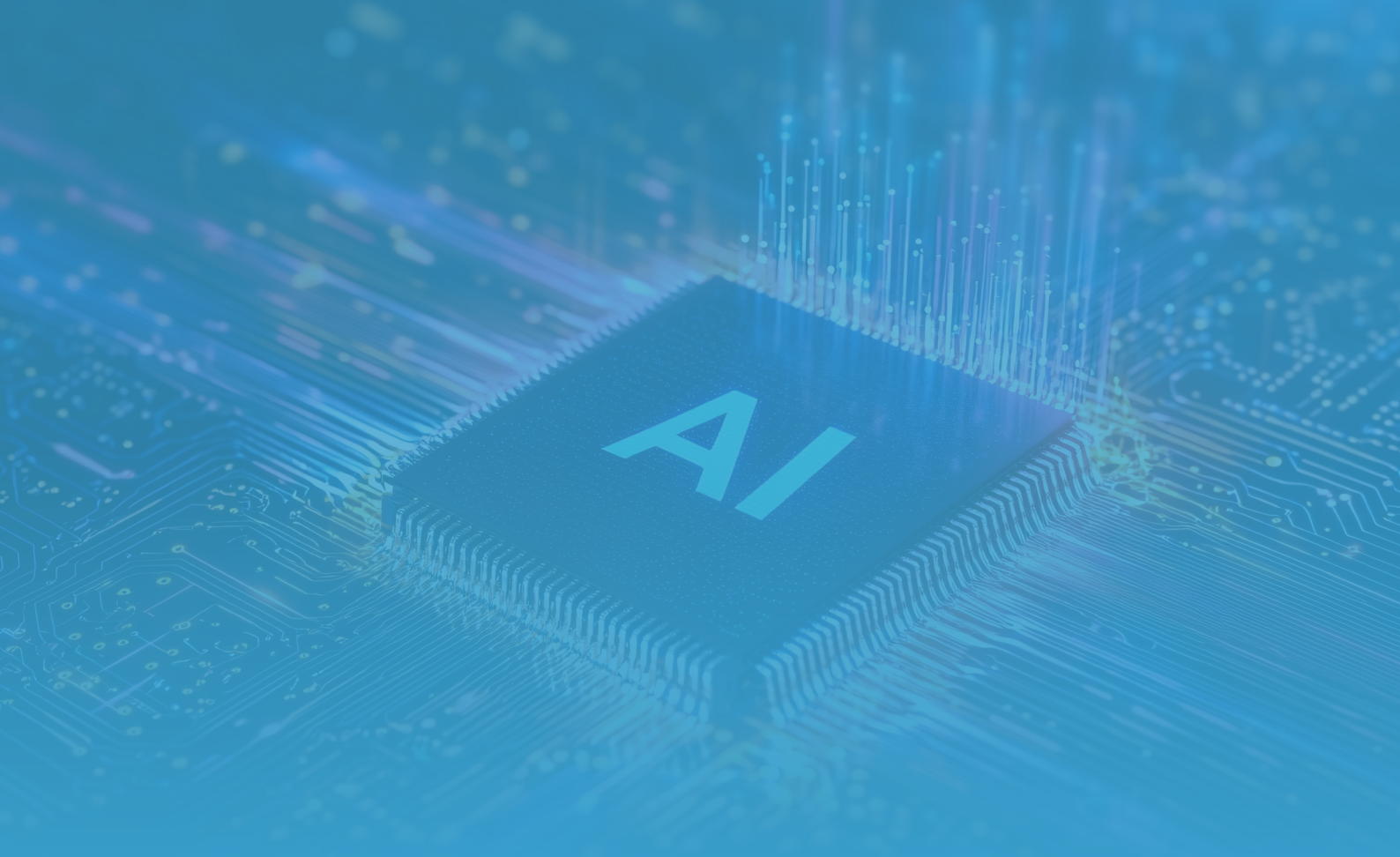
RISICO'S EN MITIGERENDE MAATREGELEN VAN GENERATIVE AI

Door Marco Lammers

AT COMPUTING

ATYPICAL OPEN SOURCE GURUS

versie: 04/25



TL;DR

Generative AI speelt een steeds belangrijkere rol in zowel ons dagelijks leven als onze professionele werkzaamheden. De technologie biedt talrijke voordelen, waaronder het efficiënter maken van processen, het ondersteunen van creativiteit en het reduceren van kosten. Toch brengt het ook aanzienlijke risico's met zich mee die niet mogen worden onderschat. Deze whitepaper bespreekt enkele van deze mogelijke risico's en onderzoekt manieren waarop deze risico's gemitigeerd kunnen worden. Eén van de meest effectieve maatregelen is het gebruik van lokale modellen in plaats van cloudgebaseerde oplossingen.

>

> DE BELANGRIJKSTE RISICO'S BIJ HET GEBRUIK VAN GENERATIVE AI

1. PRIVACY- EN BEVEILIGINGSRISICO'S

Wanneer je generative AI een vraag stelt of een opdracht geeft, wordt jouw input vaak gebruikt om het model verder te trainen. Dit betekent dat vertrouwelijke informatie onbedoeld gedeeld kan worden met derden en mogelijk in toekomstige output aan andere gebruikers terechtkomt. Sommige aanbieders geven opties om dit te beperken, vooral bij betaalde abonnementsvormen, maar dit staat niet altijd standaard aan. Hierdoor lopen gebruikers die hier niet bewust van zijn onnodige privacy- en veiligheidsrisico's. Daarnaast maken steeds meer organisaties (soms ongemerkt) gebruik van Amerikaanse of Chinese cloudproviders die AI-diensten aanbieden, zoals ChatGPT. Hoewel veel bedrijven een beleid hebben dat medewerkers niet zomaar gevoelige bedrijfsinformatie mogen delen, gebeurt dit in de praktijk vaak toch. Hierdoor kan confidentiele informatie niet alleen bij de provider, maar ook in andere jurisdicties terechtkomen. Daarnaast kan lokale wet- en regelgeving in landen als de Verenigde staten en China ertoe leiden dat deze data verplicht gedeeld moet worden met overheidsinstanties wat in strijd kan zijn met de Europese GDPR/AVG wetgeving.

2. VOOROORDELEN (BIAS) IN AI MODELLEN

AI-modellen baseren hun antwoorden op trainingsdata die ze verwerken. Hierdoor nemen ze vaak ook (onbewuste) voorkeuren en vooroordelen uit die datasets over. Dit kan leiden tot gekleurde of bevooroordeelde antwoorden, vooral bij gevoelige onderwerpen zoals politieke of culturele kwesties. Bijvoorbeeld kan het antwoord van een Amerikaans ontwikkeld model over gevoelige geopolitieke onderwerpen sterk verschillen van dat van een Chinees ontwikkeld model, simpelweg doordat de gebruikte trainingsdata verschillen.

3. GEBREK AAN GEDEGEN TRAINING IN SPECIFIEKE VAKGEBIEDEN

Generative AI werkt goed wanneer er voldoende kwalitatieve data beschikbaar is om op te trainen. Echter, voor specialistische vakgebieden of heel specifieke niches is er vaak slechts beperkte data aanwezig. Dit betekent dat AI in deze domeinen minder effectief functioneert en vaker foutieve of onvolledige informatie zal produceren. Dit probleem geldt bijvoorbeeld voor esoterische wetenschappelijke domeinen of specifieke bedrijfstakken met weinig publieke data.

4. BEPERKTE ACTUALITEIT DOOR "CUT-OFF DATUM"

Iedere generative AI heeft een bepaalde "cut-off datum" waarop haar training stopt. Dit betekent concreet dat gebeurtenissen, innovaties of ontwikkelingen die plaatsvinden nadat het model is getraind niet bekend zijn bij het model. In het beste geval geeft het model aan hierover geen informatie te hebben, maar helaas kan het ook voorkomen dat een model informatie "verzint" of verouderde informatie aanbiedt als ware deze actueel. Het verschijnsel waarbij AI incorrecte informatie fantaseert wordt "hallucineren" genoemd.

> EFFECTIEVE MITIGERENDE MAATREGELEN

1. TOEPASSING VAN LOKALE AI-MODELLEN

Het gebruik van lokaal draaiende modellen (AI-modellen binnen je eigen infrastructuur zonder internetverbinding met derden) is één van de veiligste keuzes voor gevoelige of vertrouwelijke data. Hierbij worden jouw instructies, documenten en broncode niet gedeeld met derde partijen en belanden ze niet in publieke trainingssets. Hoewel dit het probleem van privacy en deels dataveiligheid oplost, brengt het ook eigen verantwoordelijkheden met zich mee. Zo dien je zelf zorg te dragen voor het onderhoud, de beveiliging, het monitoren van prestaties, evenals regelmatige updates van je lokale model.

2. BEWUSTE SELECTIE EN VOORTDURENDE MONITORING VAN AI-MODELLEN

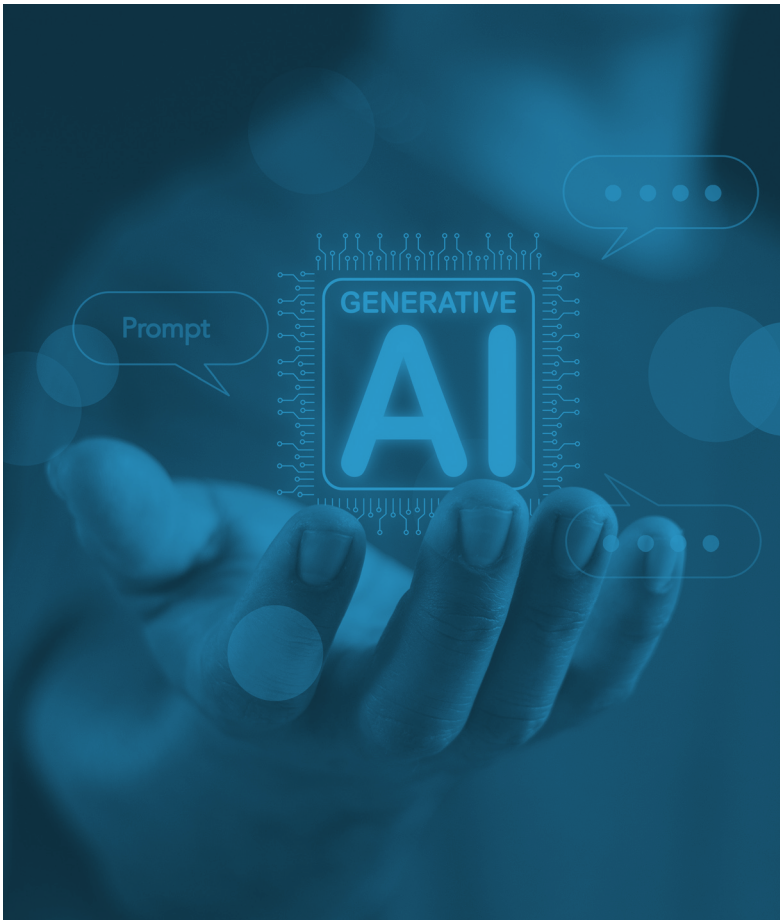
Wanneer je kiest voor een lokaal model, is het belangrijk om vooraf te onderzoeken welk model het meest geschikt is voor jouw doeleinden en ethische standaarden. Controleer regelmatig de gegenereerde output op mogelijke biases of verdachte code (“backdoors”). Er zijn diverse voorbeelden bekend waarin kwaad-willenden AI-output manipuleren met verborgen prompts (zie bijvoorbeeld het artikel van Dennis Kruyt: Phishing with Large Language Models: Backdoor Injections). Door output kritisch te analyseren en periodieke audits uit te voeren, voorkom je dat dit soort risico's onopgemerkt blijven.

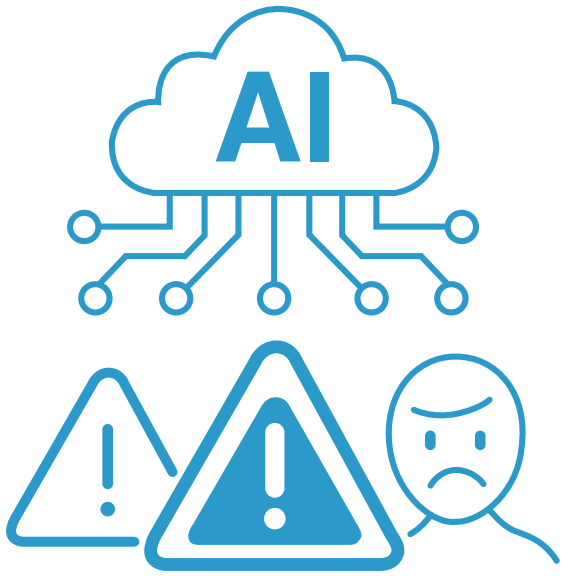
3. VERRIJKING DOOR AANVULLENDE CONTENT (RAG)

Het probleem dat AI niet adequaat getraind is voor heel specifieke niches kan deels gemitigeerd worden door gebruik te maken van Retrieval-Augmented Generation (RAG). Bij deze techniek krijgt het model extra gerichte context om betere, specifiekere antwoorden te genereren. Met behulp van eigen documentatie, interne kennisbanken of gespecialiseerde datasets binnen je organisatie, verbetert de relevantie en de betrouwbaarheid van AI-gegenereerde informatie substantieel. Hierbij is het wel noodzakelijk dat er een goed Retrieval Model wordt gekozen (zie artikel door Nayab Hassan: Mastering Retrieval Engine in RAG).

4. INTEGRATIE ACTUELE DATABRONNEN

Het probleem van beperkte actualiteit door de cut-off datum is deels op te lossen door actuele informatie via API's of RAG oplossingen aan het model te koppelen. Denk hierbij aan koppelingen met databases, nieuwsfeeds of real-time bronnen. Hoewel een model hierdoor beter in staat zal zijn actuele informatie te presenteren, blijft het model beperkt tot de informatie die momenteel beschikbaar is in aangesloten databronnen of kennisbanken. Daarom blijft handmatige controle cruciaal.



RISK	POSSIBILITIES TO MITIGATE THEM
	

Conclusie

Generative AI kan, zoals in andere publicaties beschreven (zie bijvoorbeeld: Generative AI is poised to unleash the next wave of productivity. We take a first look at where business value could accrue and the potential impacts on the workforce van Mckinsey), enorme voordelen bieden aan bedrijven, overheden en particulieren mits zij zich bewust zijn van de beperkingen en mogelijke risico's. Vooral voor organisaties die werken met gevoelige of vertrouwelijke data zoals overheidsinstanties en bedrijven in sterk gereguleerde sectoren, is het essentieel om zorgvuldig met databeveiliging, privacy en actuele informatievoorziening om te gaan.

Arnhemsestraatweg 33
6881 ND Velp (GLD)
Nederland

www.atcomputing.nl