

Digitale AI-assistent



Inhoud

INLEIDING	3
CONTEXT: VERANDEREND RETAILLANDSCHAP, DIGITALISERING EN PERSONEELSTEKORT	3
DOEL VAN DE WHITEPAPER	4
VOOR WIE IS DEZE WHITEPAPER BEDOELD	5
HANDLEIDING BIJ HET LEZEN VAN DEZE WHITEPAPER	6
DE DIGITALE ASSISTENT IN EEN NOTENDOP	8
PERSOONLIJKHEIDSKENMERKEN VAN DE AI-ASSISTENT	10
AVATARS: WIE ZIJN ZE, WAT DOEN ZE?	13
AI-ASSISTENT IN DE FYSIEKE WINKEL	15
MEERWAARDE VOOR FYSIEKE WINKELS	16
PRAKTISCHE TOEPASSINGEN IN DE WINKEL	19
VALKUILEN EN MISVERSTANDEN	20
VAN DATA TOT DIALOOG: DE TECHNISCHE FUNDERING	22
HOE MAAK JE INTERNE INFORMATIE BESCHIKBAAR?	23
DE PRODUCTCATALOGUS VAN DE RETAILERS INTERACTIEF MAKEN	24
VOLDOEN AAN DE NODEN VAN DE RETAILER	30
ALL YOU NEED IS RAG?	36
AANDACHTSPUNTEN	38
AANDACHTSPUNTEN BIJ IMPLEMENTATIE	39
CONCLUSIE EN AANBEVELINGEN	41
BRONNEN	42
OVER ONS	45

Inleiding

Context: veranderend retaillandschap, digitalisering en personeelstekort

De nasleep van de pandemie, recente oorlogen, stijgende energie- en grondstoffenprijzen, buitenlandse concurrentie, e-commerce en de krappe arbeidsmarkt zetten **de retailsector onder druk**. Marges worden daardoor alsmaar kleiner, **kosten stijgen** en het tekort aan personeel op de werkvloer wordt alsmaar groter. Winkels moeten steeds vaker tijdelijk sluiten door tekort aan personeel: zo bijvoorbeeld ook supermarkt Spar Genk die in een week tijd twee keer de deuren vroeger moest sluiten door **personeelstekort**. Handelsfederatie Comeos spreekt van een **structureel probleem**. Veel werknemers gaan met pensioen en er zijn onvoldoende jongeren om hen op te volgen, aldus Klaas Soens van Handelsfederatie Comeos.

Verkopers en verkoopondersteuners prijken al enkele jaren op de lijst van **knelpuntberoepen** in Vlaanderen. De invulling van deze vacatures verliep ook in 2023 nog zeer moeizaam. Zo woedt er een ware war for talent. De reden ligt niet enkel bij de specifieke arbeidsomstandigheden (zoals weekend- en avondwerk, lage verloning) maar ook kwalitatief: de vraag naar **specifieke competenties** neemt toe. Eén van de belangrijkste competenties daarbij zijn **digitale vaardigheden** die verkopers moeten aanleren. Zo moeten verkopers niet enkel goed met klanten kunnen omgaan, maar moeten ze ook een uitstekende kennis hebben van alle producten in de winkel en met de steeds toenemende digitalisering kunnen omgaan. En die combinatie is vaak heel moeilijk te vinden.

Het is dan ook niet verwonderlijk dat het **aantal fysieke winkels daalt**. In 2023 waren er in Vlaanderen 42.177 winkels. Dat zijn 3.000 handelspanden minder dan in 2022, zo berekende Locatus. Toch was er een daling van de leegstand van 10.6 naar 9.6%. Dit komt vooral door de herbestemming van winkel naar woning, kantoor of een andere functie. Als we dit in rekening brengen, vertoont het aantal winkelvevestigingen nog steeds een afname, met een daling van 1.500 winkels in Vlaanderen in het jaar 2023. Ook worden winkels doorgaans kleiner. Fysieke winkels blijven nochtans belangrijk. Comeos berekende dat Belgische consumenten **88 miljard in fysieke winkels besteden** tegenover 12 miljard online.

Doel van de whitepaper

Hoewel vaak wordt gesproken over de vrees dat Artificiële Intelligentie (AI) jobs zal innemen, toont de bovenstaande problematiek van fysieke winkels een ander perspectief. Er is een overschot aan taken die moeten uitgevoerd worden, maar er zijn te weinig mensen om die taken correct uit te voeren. Dit leidt niet alleen tot een **verhoogde werkdruk**, maar heeft ook een **negatieve invloed op de klantbeleving**. Klanten verwachten immers gepersonaliseerd advies en service als ze een winkel bezoeken. Moderne ontwikkelingen zoals AI die kunnen ook in een fysieke retailomgeving zinvol ingezet worden om de druk op ondernemers te verlichten. Een **digitale AI-assistent kan het keuzeproces van de klant faciliteren** en zo de druk op winkelmedewerkers verlichten. Op deze manier kunnen verkopers enerzijds routinematige taken efficiënter uitvoeren en zo focussen op taken die menselijke vaardigheden vereisen, zoals het leveren van persoonlijke service en het oplossen van complexe klantvragen. Anderzijds kunnen **nieuwe medewerkers** die nog niet alle producten in de winkel al kennen, **bijgestaan worden door de AI-assistent wanneer ze de klant helpen**.

Stel je het volgende **scenario** voor: een klant is op zoek naar een paar schoenen maat 39 met een open neus dat past bij een outfit voor een lentefeest. De medewerkers zijn druk bezig. Een Nederlandstalige digitale AI-assistent waarmee je kan communiceren via een fysiek scherm in de winkel - en die binnen het beschikbare aanbod producten voorstelt - biedt hier een oplossing. De klant omschrijft waar die naar op zoek is, net zoals de klant dat bij een medewerker zou doen. De digitale AI-assistent gaat op zoek naar beschikbare oplossingen binnen het winkelaanbod. Terwijl de klant zo in eerste instantie geholpen wordt, kunnen medewerkers inspringen wanneer ze weer tijd hebben of wanneer de klant nog extra ondersteuning nodig heeft. De klant, medewerker en AI-assistent werken op die manier samen om tot een betere fysieke winkelervaring te komen.

Luc Ardies, algemeen directeur Winkelraad Unizo en gedelegeerd bestuurder Buurtsuper.be, beaamt de noodzaak aan dergelijke digitalisering binnen de retail. Hij stelt dat **digitalisering en automatisering de fysieke winkel** niet alleen kunnen **ondersteunen**, maar de klant ook vooral **goesting kan geven om zich nog te begeven naar de fysieke winkel**. Het Vlaams beleidsplan voor Artificiële Intelligentie bevestigt dit. Bedrijven die AI succesvol weten in te zetten, zullen een enorm voordeel krijgen t.o.v. bedrijven die achterblijven. Wanneer fricties en drempels de omarming van AI vertragen of bemoeilijken zullen deze bedrijven en werknemers de boot missen en nog meer achterophinken.

Helaas tonen de bevindingen in de imec AI-barometer (2022) dat ondanks de duidelijke meerwaarde van AI, de adoptie ervan nog grotendeels uitblijft: 75% van de Vlaamse bedrijven heeft op dit moment namelijk nog geen enkele AI-technologie geïmplementeerd in hun werking. Het merendeel van het beperkt aantal bedrijven die wel een AI-technologie geïmplementeerd hebben, blijken daarbovenop beperkt tot grote ondernemingen en middelgrote bedrijven. Kleinere spelers vallen dus uit de boot. Waar grote retailketens vaak zelf investeren in innovatieonderzoek, hebben kleine en middelgrote retailondernemers vaak noch de kennis noch de middelen om uitgebreid onderzoek te doen naar de mogelijkheden om AI succesvol in te zetten.

Dit bleek ook uit onze eigen bevraging van de Vlaamse retailmarkt eind 2020 waaraan 369 retailondernemers deelnamen. 60% zag het belang van digitale transformatie van hun winkel in maar mist de handvaten om dit doordacht aan te pakken en hun businessplan hierop aan te passen. Hun oriëntatie naar digitalisering beperkt zich echter nog vooral tot de informatiefase. Slechts 17% bleek recent een technologie geïmplementeerd te hebben, hiervoor concrete doelstellingen bepaald te hebben en de inzichten toe te passen om hun bedrijfsprocessen te optimaliseren. Kortom, de digitale transformatiematuriteit van Vlaamse retailondernemers staat nog in de kinderschoenen.

Voor wie is deze whitepaper bedoeld

Deze publicatie richt zich voornamelijk tot kmo's uit **2 doelgroepen**:

1. Vlaamse retailondernemingen met minstens één fysieke winkel

In 2023 waren er in Vlaanderen 42.177 fysieke winkels. De branche waarin ze actief zijn is minder relevant aangezien een digitale AI-assistent zowel in een winkel met levensmiddelen, kleding en mode, bruin- en witgoed als vrije tijd ingezet kan worden. Belangrijk is dat ze aan volgende voorwaarden voldoen:

- Kmo's met maximaal 50 werknemers en minstens 1 fysieke winkel met een sterk competitief fundament die het belang van digitalisering inzien, over een innovatiemindset beschikken maar niet weten hoe de juiste oplossing te integreren
- Voldoende omvangrijk en divers productgamma dat keuzestress bij klant oplevert (schoenen, kleding, multimedia, ...) en ondersteuning van de winkelmedewerker vergt
- Beschikken over een online productgamma/digitaal CRM-systeem waarmee gekoppeld kan worden

Startpunt zijn de noden van de klant die in zijn zoektocht naar een juiste oplossing gepersonaliseerd advies nodig heeft.

Door in te zetten op sensibiliseren, faciliteren en accelereren van beschikbare AI-kennis rond digitale AI-assistenten bij Vlaamse retailondernemers wordt de digitale kloof tussen de hedendaagse gedigitaliseerde klant en de vaak traditionele retail gedicht. Vlaanderen bestaat uit een versnipperd landschap van kleine kmo-ondernemers met weinig tot geen toegang tot R&D, expertise, tijd en middelen om de innovatiespurt te nemen.

2. Vlaamse technologiebedrijven

Agoria telt ruim 1.950 Belgische technologiebedrijven die oplossingen voor de retailsector (kunnen) aanbieden, zoals digital signage of software. Waarbij een digitale AI-assistent een uitbreiding op hun aanbod kan zijn of die momenteel retail nog niet tot hun doelgroep rekenen.

Door hierop in te zetten, kunnen technologieaanbieders hun **producten en diensten** nog **beter afstemmen** op de behoeften van zowel consument, maar zeker ook de retailondernemers die vaak een landschap van versnipperde, kleine kmo-ondernemers vertegenwoordigen. Dit leidt tot besparing op R&D-kosten, een snellere go-to-market, grotere investering en/of hogere tewerkstelling.

Door de inzichten uit deze whitepaper kunnen ze sneller en via diverse marktintroductie het marktpotentieel voelen en hun ontwikkelproces en/of vermarkting hierop afstemmen.

Handleiding bij het lezen van deze whitepaper

Dit naslagwerk is ontworpen om zowel technologieaanbieders als retailers waardevolle inzichten te bieden. We raden aan dat je als retailer zowel de secties over retail leest als over de technologieaanbieder. Hetzelfde geldt voor de technologieaanbieder. Om deze twee industrieën samen te brengen, is het zeer belangrijk om elkaar te begrijpen. Als technologieaanbieder dien je dus niet alleen te weten welke technische aspecten je nodig hebt bij het bouwen van de AI-assistent, maar ook wat de noden zijn bij een retailer. Als retailer heb je dan weer naast de implementatiemogelijkheden in retail meer zicht nodig op wat een technologieaanbieder allemaal nodig heeft om voor jouw winkel(s) een goed werkende AI-assistent te bouwen.

Inleiding



De digitale assistent in een notendop



De digitale AI-assistent in een notendop

Een digitale AI-assistent is een **interactieve toepassing** waarmee klanten in een fysieke winkel op een natuurlijke manier kunnen communiceren. Het systeem gebruikt hiervoor een **Large Language Model** (LLM) dat menselijke taal begrijpt en duidelijke antwoorden formuleert. De interactie vindt vaak plaats via een **scherm**, waar de AI niet alleen tekst en beelden kan tonen, maar ook een visuele representatie heeft om de communicatie vlotter te laten verlopen. Typisch is dit in de vorm van een **avatar**, een digitale representatie die klanten begeleidt en informatie op een toegankelijke manier presenteert. Afhankelijk van de implementatie kan deze avatar een **menselijk uiterlijk** hebben, zoals een Bekende Vlaming of een winkelmedewerker, maar net zo goed kan het een eenvoudige grafische weergave zijn of zelfs enkel via spraak communiceren. Klanten kunnen hun **vragen invoeren** of **inspreken**, waarna de AI-assistent **informatie** ophaalt uit **productcatalogi**, **voorraadsystemen** of **documentatie** om een passend antwoord te geven.

De precieze, fysieke **opstelling** varieert afhankelijk van waar en hoe de assistent wordt ingezet. Sommige winkels kiezen voor een **vaste informatiezuil** of kiosk waar klanten terecht kunnen voor ondersteuning, terwijl andere winkels kiezen voor **tablets** of **interactieve schermen** die verspreid door de winkel staan. In meer geavanceerde installaties kiest men misschien voor spraakherkenning omdat dit het gebruiksgemak verhoogt: klanten kunnen zo eenvoudig met de assistent spreken in plaats van te typen. De assistent kan meer dan alleen geschreven of gesproken antwoorden geven. Het kan bijvoorbeeld foto's of korte video's van producten tonen, een plattegrond van de winkel weergeven of een vergelijkende tabel van verschillende producten geven, zodat klanten snel de informatie kunnen vinden die ze nodig hebben.

De AI-assistent kan bijvoorbeeld het **eerste aanspreekpunt** voor veelgestelde vragen zijn en op deze manier snelle, juiste informatie aan klanten bezorgen. Het beantwoordt vragen over productkenmerken, prijzen en beschikbaarheid en kan, indien nodig, gepersonaliseerd advies geven, zoals het voorstellen van producten op basis van klantvoorkeuren. Heeft de klant extra hulp nodig, dan verwijst de assistent door naar een medewerker en geeft hij meteen de relevante context mee, zodat het gesprek efficiënter verloopt. Kortom, een digitale AI-assistent biedt een **extra laag ondersteuning** in de winkel door klanten snel toegang te geven tot productinformatie en basisadvies. Dankzij **real-time toegang** tot **gegevens** en een **gebruiksvriendelijke interface** via scherm of spraak kunnen vragen vlot beantwoord worden, terwijl medewerkers zich op persoonlijkere of complexere interacties kunnen richten. De assistent neemt niet de rol van het winkelpersoneel over, maar **helpt** bij het **stroomlijnen** van de **klantenservice** en het **verlagen** van de **werkdruk**.

Persoonlijkheidskenmerken van de AI-assistent

Hoewel een AI-assistent gemakkelijk een rol kan aannemen die lijkt op die van een menselijke medewerker, is niet elke AI-assistent per definitie succesvol. Het succes hangt sterk af van het **ontwerp** en de **kenmerken**. De **persoonlijkheid** van de AI-assistent speelt een cruciale rol in hoe klanten de interactie ervaren en in hoeverre ze tevreden zijn met de hulp die de AI-assistent biedt.

Bij het kiezen van de persoonlijkheidskenmerken voor een avatar in de retail zijn verschillende factoren belangrijk om de klantbetrokkenheid en -tevredenheid te verhogen. Enkele **essentiële kenmerken** zijn:

Vriendelijk en beleefd

Een AI-assistent moet vriendelijk en beleefd zijn, ongeacht de situatie. Dit helpt om een positieve klantbeleving te creëren. Klanten verwachten een warm welkom en een gevoel van comfort wanneer ze een winkel binnenkomen. De AI-assistent moet een **rustige, aangename toon** hebben, met een **vriendelijke** en **uitnodigende uitstraling**. Beleefdheid zorgt ervoor dat klanten zich gewaardeerd voelen, wat hun algehele winkelervaring verbetert.

Empathisch

Empathie is belangrijk, vooral wanneer de AI-assistent moet omgaan met klanten die gefrustreerd of ontevreden zijn. Bijvoorbeeld, als een klant klaagt over een product of ontevreden is over de winkelindeling, moet de AI-assistent **begrip tonen** ("Het spijt me dat je dat ervaart. Ik help je graag verder.") en **passende oplossingen** aanbieden.

Deskundig en betrouwbaar

De AI-assistent moet een **expert** zijn in de producten en diensten van de winkel. Klanten vertrouwen erop dat de AI-assistent hen voorziet van **betrouwbare en actuele informatie** over producten, beschikbaarheid, prijzen en aanbiedingen. De AI-assistent moet **snel** en **nauwkeurig** kunnen **antwoorden**, zonder onnauwkeurigheden of verwarring. Dit draagt bij aan de klanttevredenheid en versterkt het vertrouwen in de winkel.

Geduldig

Een AI-assistent moet geduldig zijn, zelfs als klanten herhaaldelijk dezelfde vraag stellen of extra tijd nodig hebben. In tegenstelling tot menselijke medewerkers, die wel eens moe of geïrriteerd kunnen raken, kan een AI-assistent in principe **oneindig geduldig** zijn. Dit is belangrijk voor klanten die onzeker zijn of die wat meer tijd nodig hebben om hun vraag of verzoek duidelijk te maken.

Toegankelijk

De AI-assistent is **vlot benaderbaar**, toont **begrip** voor vragen en past zijn communicatie aan het **klantniveau** aan. Klanten verschillen namelijk in hun digitale vaardigheden; een toegankelijke avatar kan eventuele barrières verlagen.



Hulpvaardig

Een AI-assistent moet **gericht** zijn op het **helpen** van de klant. Dit betekent dat de AI-assistent zich snel aanpast aan de situatie van de klant, bijvoorbeeld door het aanbieden van gerichte aanbevelingen, het verstrekken van navigatie-informatie in de winkel of het helpen bij het vinden van producten. **Hulpvaardigheid** moet de **primaire motivator** zijn in de interactie, zodat de AI-assistent als een waardevolle assistent wordt ervaren.



Oplossingsgericht

In lijn met hulpvaardigheid moet de AI-assistent ook **oplossingsgericht** zijn. Dit betekent dat de AI-assistent altijd gericht is op het bieden van **praktische** en **effectieve oplossingen** voor de vragen of problemen van de klant. In plaats van zich te concentreren op problemen of obstakels, moet de AI-assistent snel oplossingen aandragen en ervoor zorgen dat de klant zich geholpen voelt.



Natuurlijk en menselijk

Hoewel de AI-assistent een **digitale entiteit** is, moet de **interactie zo natuurlijk mogelijk** aanvoelen. Dit betekent dat de stem, bewegingen en gezichtsuitdrukkingen op menselijke interacties (**antropomorf**) afgestemd worden. De AI-assistent moet in staat zijn om op een manier te reageren die niet te robotachtig of kunstmatig is, maar die klanten het gevoel geeft met een "echte" persoon te communiceren. Dit verhoogt de **vertrouwensband** en maakt de **ervaring aangenamer**.



Enthousiast en positief

Een **positieve houding** kan de winkelervaring aanzienlijk verbeteren. De AI-assistent moet een positieve en **enthousiaste uitstraling** hebben, iets wat klanten een gevoel van energie en plezier kan geven tijdens hun winkelervaring.



Oprecht en betrokken

Hoewel de AI-assistent geen echte emoties heeft, moet hij een **gevoel** van **oprechtheid** en **betrokkenheid** uitstralen. Dit kan door middel van zijn **lichaamstaal** (bijvoorbeeld knikken) zodat hij echt lijkt te luisteren naar de klant. De AI-assistent moet oprecht interesse tonen in de klant, wat kan helpen om een band op te bouwen en een klantgerichte sfeer te creëren.

Samenvatting

De inzet van AI-assistent in de retail biedt waardevolle kansen om **klantinteracties** te **versterken** en de **winkelervaring** te **personaliseren**. Cruciaal daarbij is de afstemming van de **persoonlijkheidskenmerken** van de AI-assistent op de verwachtingen van de klant. Kenmerken zoals vriendelijkheid, deskundigheid, betrouwbaarheid, toegankelijkheid en – waar passend – speelsheid, blijken essentieel voor het **opbouwen** van **vertrouwen**, het **stimuleren** van **klantbetrokkenheid** en het **ondersteunen** van **conversie**. Retailers die investeren in een AI-assistent met een doordachte persoonlijkheid creëren niet alleen een menselijkere digitale interactie, maar onderscheiden zich ook op het vlak van klantbeleving en merkidentiteit.

persoonlijkheidskenmerken	
vriendelijk en beleefd	natuurlijk
empathisch	menselijk
deskundig	positief
betrouwbaar	oprecht
toegankelijk	betrokken
hulpvaardig	geduldig
oplossingsgericht	enthousiast

Avatars: wie zijn ze, wat doen ze?

Algemeen

Om de persoonlijkheid van de AI-assistent meer tot leven te brengen en de AI-assistent af te stemmen op de branding van de winkel, kan men ervoor opteren om een avatar te gebruiken. De literatuur definieert een avatar als “een algemene **grafische voorstelling** die door middel van **computertechnologie** wordt **gepersonaliseerd**”. Een digitale avatar is een **digitaal gegenereerd personage** – vaak weergegeven op een scherm – dat fungeert als een virtuele winkelmedewerker. Deze avatar wordt aangestuurd door **artificiële intelligentie (AI)** om te **communiceren** met klanten en hen te **assisteren**. Je kan de avatar zien als een vorm van *embodied conversational agent*: AI krijgt een “lichaam” of gezicht op een scherm, waardoor het lijkt alsof men tegen een persoon praat in plaats van een computeralgoritme. De avatar kan o.a. praten (tekst weergeven), luisteren naar vragen en gezichtsuitdrukkingen of handgebaren gebruiken om de communicatie natuurlijker te maken. Moderne avatars maken gebruik van **AI-technologie**, zoals *large language models* (LLM), om te converseren, spraak te synthetiseren en real-time animatie te creëren, waardoor ze bijna menselijke gesprekken kunnen voeren.

Dit type AI-gestuurde avatars is zeer **veelzijdig** en kan op verschillende manieren worden ingezet. Ze kunnen verschijnen als een tweedimensionale figuur op een flatscreen, als een driedimensionaal karakter (bijvoorbeeld op een scherm, via een 3D-hologram of met behulp van AR/VR), of zelfs als een cartoon, afhankelijk van de branding van de winkel. Zo kan een avatar fungeren als een vriendelijke tekenfilmmascotte of net als een levensechte mens. Bedrijven hebben de mogelijkheid om het **uiterlijk**, de **stem** en de **persoonlijkheid** van de avatar af te stemmen op hun **merkidentiteit**. Denk bijvoorbeeld aan een avatar in het bedrijfsuniform of gelijkend op een bekende persoon/medewerker. Het idee hierachter is dat de avatar een herkenbaar “gezicht” van de winkel wordt, die **altijd beschikbaar** is en **consistent** werkt.

In tegenstelling tot een eenvoudige kiosk of chatbot vormt een AI-gestuurde avatar op een scherm een **menselijker** en **interactiever contactpunt**. Klanten kunnen op een natuurlijkere manier communiceren, bijvoorbeeld door een vraag te stellen zoals “Waar kan ik de winterjassen vinden?”. De avatar kan vervolgens een antwoord geven, zowel in gesproken als in geschreven taal, eventueel aangevuld met visuele aanwijzingen. Omdat de avatar beschikt over een menselijk gezicht en stem, voelt de interactie **intuïtiever**. Onze hersenen zijn namelijk geëvolueerd om sterk te reageren op gezichten en stemmen. Zelfs wanneer we weten dat de avatar geen echte persoon is, vertonen we toch vergelijkbare reacties als bij menselijk contact. Emotionele uitdrukkingen van mensachtige avatars kunnen empathie opwekken die doet denken aan echte sociale interacties. Bovendien roepen kunstmatig gegenereerde gezichten neurale reacties op die vergelijkbaar zijn met het zien van een echt menselijk gezicht. Dit betekent dat een goed ontworpen avatar in een winkelomgeving de zogenaamde “*human touch*” deels kan nabootsen. Uiteraard is het contact met een fysieke medewerker onvervangbaar.

Kortom, een AI-avatar is een **virtuele vertegenwoordiger** van de winkel: aangedreven door AI en gevisualiseerd als een karakter op een scherm zodat klanten op een **directe, persoonlijke manier** geholpen worden alsof een **extra medewerker** aanwezig is.

Antropomorfisme en sociale aanwezigheid

Wanneer men de interactie tussen mens en AI onderzoekt, komt de term '**antropomorfisme**' al snel naar boven. Antropomorfisme is het toekennen van **menselijke eigenschappen** (kenmerken, emoties, intenties, motivaties) aan **niet-menselijke wezens** of **voorwerpen**. Denk bijvoorbeeld aan reclamespots waarin dieren spreken of zich als mensen gedragen.

Zelfs als er geen echte persoon in de buurt is, kunnen **mensachtige eigenschappen** een gevoel van **sociale aanwezigheid** (*social presence*) oproepen. Sociale aanwezigheid verwijst naar de mate waarin mensen het gevoel hebben dat ze met een echte persoon communiceren, zelfs in digitale of virtuele omgevingen. Het gaat om de **ervaring** van **nabijheid**, **betrokkenheid** en **verbinding**, ondanks dat de interactie niet fysiek is. Hoe hoger de sociale aanwezigheid, hoe persoonlijker en natuurlijker de communicatie aanvoelt.

Men maakt een onderscheid tussen twee vormen van sociale aanwezigheid:

- **Menselijke sociale aanwezigheid** (*human social presence*), die ontstaat bij interactie met een echte persoon.
- **Geautomatiseerde sociale aanwezigheid** (*automated social presence*), die wordt opgewekt door technologie die sociaal contact nabootst.

Omdat bedrijven vaker, al dan niet uit noodzaak, hun personeelsbestand aanvullen/vervangen met servicerobots wordt de aandacht voor geautomatiseerde sociale aanwezigheid steeds relevanter.



AI-assistent in de fysieke winkel



Meerwaarde voor fysieke winkels

Waarom zou een retailer investeren in een digitale AI-assistent op de winkelvloer? Deze technologie kan aanzienlijke voordelen opleveren die bijdragen aan zowel de klantbeleving als de bedrijfsresultaten. We sommen graag enkele voordelen op.

Altijd beschikbare klantenservice

Een AI-assistent is in principe **altijd beschikbaar** en kan dus **onafgebroken** klanten helpen. In tegenstelling tot personeel heeft een AI-assistent geen pauzes nodig en kan hij op drukke momenten meerdere klanten vlot na elkaar te woord staan. Bijgevolg moeten klanten minder lang wachten en krijgen ze meteen antwoord op hun vragen over bijvoorbeeld productinformatie of garanties. Tijdens piekuren of in vakantieperiodes kan de AI-assistent de druk op het personeel verminderen door routinevragen over te nemen. Bovendien is de AI-assistent **nooit ziek** en komt hij altijd op tijd. De inzet van een AI-assistent vult de **dienstverlening** aan: **medewerkers** kunnen zich op **complexere** of **persoonlijkere service** richten, terwijl de **AI-assistent** de **basisvragen** en **-taken** afhandelt. Klanten kunnen zich zo ook alvast voorbereiden op een gericht gesprek met een medewerker. Het resultaat is een **soepelere totaalervaring** in de winkel, waarin klanten sneller geholpen worden. Studies suggereren dat als de prestaties van zo'n AI-assistent de verwachtingen overtreffen, klanten meer tevreden zijn en de technologie vaker willen gebruiken.

Consistente en hoogwaardige servicekwaliteit

Een goed geprogrammeerde AI-assistent biedt elke klant dezelfde kwaliteitsservice, ongeacht externe factoren. Zoals eerder vermeld heeft een AI-assistent geen last van vermoeidheid of slecht humeur, wat betekent dat **elke interactie professioneel** en **vriendelijk** is. Dit is belangrijk voor ketens die een uniform serviceniveau willen garanderen in hun filialen. Waar menselijke service kan variëren per medewerker of filiaal, zal een AI-assistent dezelfde up-to-date info en behulpzaamheid leveren volgens een **gescript protocol**. Deze consistentie kan leiden tot hogere waargenomen servicekwaliteit. Voor klanten vertaalt dit zich in een consistente ervaring en verhoogde betrouwbaarheid van de informatie. Ook draagt de AI-assistent bij aan meertaligheid en een gepersonaliseerde ervaring in de winkel, wat het serviceaanbod voor diverse klanten vergroot.

Verhoogde klantbetrokkenheid en tevredenheid

Digitale AI-assistent voegen een **interactieve** en **innovatieve dimensie** toe aan het winkelbezoek. In plaats van statische borden of folders krijgen klanten een persoonlijk gesprek. Dit kan de shopervaring leuk en memorabel maken. Nieuwigheid en verrassing spelen ook een rol: een aantrekkelijke AI-assistent **trekt de aandacht** en **nodigt uit** tot uitproberen, vooral bij nieuwsgierige of “*tech-savvy*” klanten (i.e. klanten die goed bekend zijn met technologie en het vlot kunnen gebruiken).

Daarnaast kunnen AI-assistent **drempelverlagend** werken voor klanten die liever geen menselijke medewerker aanspreken. Denk hierbij aan mensen met sociale angst of klanten met een zeer gerichte vraag die ze liever in stilte willen houden. In sommige gevallen voelen consumenten zich comfortabeler om gevoelige of gênante vragen aan een virtuele assistent te stellen dan aan een mens. Een AI-assistent kan vervolgens **anoniem** en **niet-oordelend** hulp bieden. Verder kan het ook taalbarrières verlagen indien je als retailer bijvoorbeeld veel anderstaligen over de vloer krijgt, maar je moeilijkheden hebt om meertalig personeel aan te werven.

Medewerkers ontlasten

Medewerkers kunnen voordelen halen uit de snelle toegang tot productinformatie, procedures en voorraadstatus. Daarnaast kan de AI-assistent dienen **als opleiding- en onboardingtool**, waardoor nieuwe medewerkers sneller productkennis opbouwen.

Tijdens **piekmomenten** fungeert de assistent als een waardevol hulpmiddel, waardoor medewerkers zich kunnen richten op complexere of persoonlijker klantvragen. Zoals een medewerker van Blabloom treffend zegt: “In het weekend krijgen we vaak tien keer dezelfde vraag: een cadeau voor een tweejarige. Het zou fijn zijn als de AI-assistent dat al kan opvangen.” Repetitieve vragen kunnen dus door de AI-assistent worden afgehandeld, waardoor de werkdruk daalt. Voor nieuwe medewerkers werkt de AI-assistent ook als leerinstrument om sneller productkennis op te bouwen.

Hogere verkoop en conversie

Uiteindelijk is het doel van betere service ook om de **omzet te verhogen**. Een AI-assistent kan op subtiële wijze **upselling** en **cross-selling** toepassen, vergelijkbaar met een verkoper die extra producten aanbeveelt. Bijvoorbeeld: bij een klant die schoenen koopt, kan de AI-assistent vragen of de klant bijpassende sokken of verzorgingsproducten nodig heeft. Omdat de AI-assistent over de productdetails beschikt, kan hij de voordelen van aanvullende aankopen vlot uitleggen. Daarnaast kan de AI-assistent **promoties** en **kortingen** aanbieden wanneer dit relevant is, waardoor klanten eerder geneigd zijn iets mee te nemen (“Wist u dat er vandaag 10% korting is op accessoires?”).

Efficiëntie en kostenbesparing

Langs operationele kant kunnen AI-assistenten helpen om **kosten te besparen** of **effectiever** in te zetten. Eén enkele AI-assistent-installatie kan, na de initiële investering, langdurig service verlenen tegen relatief lage operationele kosten (geen salaris, geen training na de initiële setup, lage onderhoudskosten). Dit kan met name interessant zijn voor **routinematige taken**. AI-assistenten zorgen ervoor dat frontline medewerkers tijd kunnen besteden aan het “meer menselijke” werk. De AI-assistent handelt eenvoudige vragen af, zodat medewerkers zich kunnen richten op complexere klantverzoeken of taken als het opruimen van de winkel, persoonlijk stijladvies, etc. Hiermee wordt zowel de klant geholpen (sneller antwoord) als de medewerker ontlast van repetitieve vragen. Verder kan de AI-assistent door de koppeling aan diverse systemen ook het **voorraadbeheer ondersteunen** (bijvoorbeeld melding maken als veel klanten naar een product vragen dat out-of-stock is, zodat het sneller wordt bijbesteld).

Data en inzichten verzamelen

Uniek aan een digitale AI-assistent is het vermogen om data uit klantinteracties vast te leggen. Normaalgezien wordt een gesprek tussen klant en medewerker niet opgenomen, waardoor vragen of feedback verloren gaan. Een AI-assistent kan elke gestelde vraag, elke veelvoorkomende klacht of product waarnaar gevraagd wordt, **registreren** en **analyseren**. Door deze **gegevens** te **verzamelen**, krijgt de retailer meer inzicht in wat klanten willen, waar ze eventueel vastlopen in de winkel of welke producten populair zijn. Je kan het vergelijken met *webanalytics*, maar dan voor de fysieke winkel: men kan bijvoorbeeld ontdekken dat veel klanten vragen naar een bepaald type product dat moeilijk te vinden is in de winkel. Dit geeft het signaal om de bewegwijzering te verbeteren of om het product een prominentere plaats in de winkel te geven. Daarnaast kunnen terugkerende vragen wijzen op informatie die niet duidelijk genoeg gecommuniceerd wordt in de winkel (herhaaldelijk “Waar zijn de paskamers?” impliceert dat de bewegwijzering onvoldoende is). Retailers kunnen op basis van deze inzichten hun **winkelindeling**, **voorraad** en **serviceprocessen optimaliseren**. Bovendien kan de AI-assistent **zelflerend** zijn: de AI-assistent wordt gaandeweg slimmer door te leren van eerdere interacties welke antwoorden het beste werken of welke aanbevelingen effectief zijn, waardoor de **service continu verbetert**. Data-inzichten uit avatar-interacties creëren dus een **feedbackloop** voor continue verbetering van de klantbeleving.

Kortom, een digitale AI-assistent op de winkelvloer draagt bij aan een **hybride winkelervaring** die het beste van twee werelden heeft, namelijk de persoonlijkheid en het advies van een menselijke verkoper, gecombineerd met de onuitputtelijke kennis en consistentie van een computer. Dit helpt fysieke winkels om relevant en competitief te blijven in een tijdperk waarin klantenservice het onderscheid kan maken.

Praktische toepassingen in de winkel

Het is essentieel om de AI-assistent doordacht te gaan toepassen in je winkel. Zonder een strategisch plan over hoe je de AI-assistent in de praktijk zal inzetten, zal deze niet optimaal ingezet kunnen worden. Een aantal zaken waar je zeker rekening mee dient te houden zijn de volgende:

Ten eerste is de **strategische plaatsing** van de AI-assistent essentieel voor een effectieve inzet. De klant moet de assistent op het juiste moment in zijn customer journey tegenkomen om er daadwerkelijk gebruik van te maken. Staat de AI-assistent in voor algemene ondersteuning, dan is een centrale locatie — bijvoorbeeld in de buurt van de helpdesk — aangewezen. Wanneer de assistent specifieke productcategorieën ondersteunt, is een plaatsing dicht bij deze producten het meest effectief. De kern is dat klanten snel en intuïtief toegang hebben tot het gewenste advies.

Verder is het gebruik van **natuurlijke taal** een belangrijk element in de gebruikservaring van de klant. Klanten kunnen hun vraag formuleren zoals ze dat ook bij een medewerker zouden doen. De AI-assistent stelt vervolgens gerichte vervolgvragen om de behoefte van de klant beter te begrijpen en tot een passend advies te komen. Dit maakt de interactie menselijker.

Daarnaast biedt de **integratie met bestaande POS-systemen of apps** een grote meerwaarde. Hierdoor kan de AI-assistent actuele informatie raadplegen over prijzen, promoties, voorraad of klantvoorkeuren. Afhankelijk van de context kan gekozen worden voor een **voice-assistent** of een **tekstinterface** — elk met hun eigen voordelen. Een voice-assistent is natuurlijk en snel in gebruik, terwijl tekstinterfaces vaak beter werken in drukke omgevingen of waar discretie gewenst is.

Interactieve zuilen of tablets: centraal geplaatst in de winkel, waar klanten snel toegang hebben tot advies.

- Natuurlijke taal: klanten kunnen hun vraag stellen zoals ze dit bij een medewerker zouden doen. De AI stelt doorvragen.
- Integratie in bestaande POS-systemen of apps.
- Voice-assistent versus tekstinterface.
- Bij WARA wordt nagedacht over de inzet van kiosken met AI-functionaliteit. Deze moeten klanten ondersteunen bij productkeuzes zonder een volledige webshopervaring te simuleren.

Valkuilen en misverstanden

De inzet van een AI-assistent op de winkelvloer biedt veel potentieel, maar er zijn ook valkuilen en misverstanden die retailers in overweging moeten nemen. Wanneer deze niet tijdig worden aangepakt, bestaat het risico dat de technologie zijn meerwaarde verliest of zelfs weerstand oproept.

AI als vervanging van medewerkers

Een van de meest voorkomende misverstanden is dat een AI-assistent medewerkers zou vervangen. In realiteit is de bedoeling net het tegenovergestelde: de **technologie** vangt **repetitieve** en **eenvoudige** vragen op, zodat **medewerkers** tijd krijgen voor **complexere** en **persoonlijkere** klantvragen. Zoals een medewerker bij Blabloom het verwoordde: “Onze kennis blijft de meerwaarde, maar de AI-assistent kan repetitieve vragen al opvangen.”

Toch kan dit idee van vervanging leiden tot onzekerheid of weerstand bij het team. Een interview met WARA benadrukt dit punt: medewerkers vrezen dat hun rol onder druk komt te staan. Deze angst kan enkel worden weggenomen door duidelijke communicatie, herhaalde training en het actief betrekken van personeel bij de introductie van de technologie.

Te complexe interacties

Een ander risico is dat de AI-assistent te lange of ingewikkelde interacties vraagt van de klant. Retailbezoekers hebben meestal weinig geduld en verwachten directe, overzichtelijke antwoorden. Wanneer de technologie omslachtig is of onnodige stappen vraagt – zoals inlogprocedures of te technische vragen – haken zowel klanten als medewerkers snel af. **Eenvoud, snelheid** en **gebruiksgemak** zijn dus cruciaal voor succes.

Overschatting van de technologie

AI wordt soms gepresenteerd als een wondermiddel dat alle problemen oplost. In werkelijkheid is de effectiviteit sterk afhankelijk van de context. Een AI-assistent kan alleen goed functioneren als de onderliggende productdata volledig en correct is. Onvoldoende of foutieve informatie leidt tot frustratie en kan de geloofwaardigheid van zowel de technologie als de retailer ondermijnen.

Daarnaast is het belangrijk om onderscheid te maken tussen algemene AI-modellen en retail-specifieke tuning. Zonder **afstemming op de sector** en het **specifieke productaanbod**, blijft de AI-assistent te algemeen en mist hij relevantie in de winkelpraktijk.

Onderschatting van de retailer zijn input

Een **AI-assistent is dus maar zo slim als de data waarmee hij gevoed wordt**. Om echt waardevolle ondersteuning te bieden, moet de retailer zorgen voor goede, rijke en begrijpelijke inputdata. Dat betekent dat productinformatie niet beperkt mag blijven tot een droge lijst met specificaties of technische kenmerken, maar ook context en betekenis moet bevatten die aansluiten bij hoe klanten denken en spreken. Wanneer een AI-assistent bijvoorbeeld productaanbevelingen doet, moet hij begrijpen wat een klant bedoelt met “een zachte trui voor de winter” of “een stevige buggy voor dagelijks gebruik”, en dat kan alleen als de onderliggende productdata dit soort beschrijvende, mensgerichte informatie bevat. Met andere woorden: de taal van de data moet vertaald worden naar de taal van de klant. Pas dan kan de AI-assistent niet enkel correcte, maar ook relevante en menselijk aanvoelende antwoorden geven. Retailers die investeren in goed gestructureerde en begrijpelijke productinformatie, leggen daarmee de basis voor een AI-ervaring die echt werkt op de winkelvloer.

AI als gimmick

Nog een valkuil is dat de AI-assistent wordt ingezet als marketingstunt of gimmick, zonder echte functionele waarde. Klanten prikken daar snel doorheen. Om effectief te zijn, moet de AI-assistent **daadwerkelijk bijdragen aan het winkelproces**, bijvoorbeeld door klanten te helpen keuzes te maken, productinformatie te geven of wachttijden te verkorten.

Onvoldoende betrokkenheid en training van personeel

Het succes van een AI-assistent hangt niet alleen af van de technologie, maar ook van de medewerkers die ermee werken. Zoals eerder aangehaald, kan gebrek aan training leiden tot terughoudend gebruik en zelfs wantrouwen. Training en herhaling zijn essentieel om **medewerkers vertrouwd te maken met de tool**, zodat zij zelfzeker blijven en de technologie actief durven inzetten.

De ervaring bij WARA toont aan dat weerstand vaak ontstaat uit onbekendheid. Duidelijke communicatie over de meerwaarde – dat AI er is om te ondersteunen en niet om te vervangen – kan dit wegnemen. Betrokkenheid van medewerkers bij de uitrol vergroot bovendien de kans dat de technologie in de praktijk omarmd wordt. Het zijn namelijk de medewerkers die tijdens klantenbezoeken kunnen doorverwijzen naar de AI-assistent, bijvoorbeeld op piekmomenten.

AI-assistenten kunnen dus een waardevolle aanvulling zijn in retail, maar enkel wanneer ze eenvoudig, betrouwbaar en functioneel worden ingezet, mét de actieve betrokkenheid van medewerkers en de juiste input van de retailer. Het vermijden van de valkuilen hierboven is cruciaal om de technologie echt impactvol te maken.

Van data tot dialoog: de technische fundering



Hoe maak je interne informatie beschikbaar?

Om echt bruikbaar te zijn als AI-assistent, moet het **toegang** hebben tot **relevante** en **actuele informatiebronnen** binnen de **organisatie**. Dit gaat bijvoorbeeld over de productcatalogus, maar ook over voorraadinformatie, technische specificaties, handleidingen, prijslijsten en andere winkelgerelateerde documenten. Deze gegevens zitten typisch verspreid over verschillende systemen, zoals databases, interne documentbeheerplatformen of zelfs Excel- of PDF-bestanden. Het is meestal niet de bedoeling om deze gegevens publiek toegankelijk te maken, bijvoorbeeld door ze online te zetten. Een manier om deze **gegevens** toch **toegankelijk** te maken voor het systeem, **zonder** ze **publiek** te delen, is **Retrieval-Augmented Generation**.

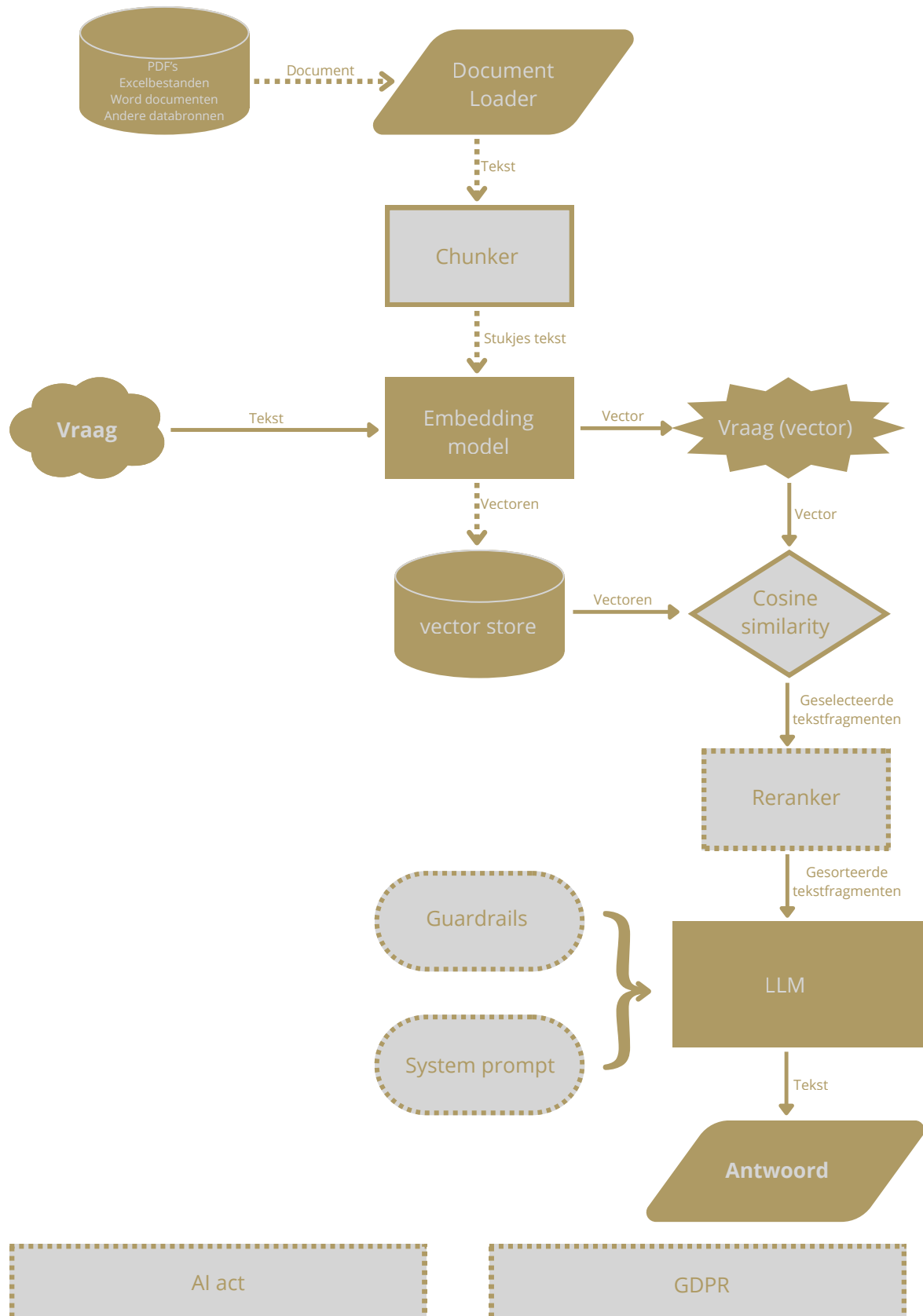
Wat is Retrieval-Augmented Generation ?

Retrieval-Augmented Generation (RAG) is een techniek waarbij een AI-assistent niet alleen antwoorden genereert op basis van de informatie waarop het getraind is, maar ook in **real-time relevante informatie ophaalt uit eigen bronnen**. Een AI-assistent met RAG kan werken met uiteenlopende informatiebronnen die door de gebruiker zelf worden aangeleverd, zoals PDF-documenten, Excel-tabellen of interne webpagina's. Al deze data worden omgezet naar een uniforme structuur en opgeslagen in een database die geoptimaliseerd is voor snelle opvraging. Wanneer een klant een vraag stelt, doorzoekt de AI-assistent deze database efficiënt en haalt het de meest relevante informatie op om een nauwkeurig antwoord te formuleren. Zo kan de AI-assistent **niet alleen snel** informatie terugvinden, maar **ook** inspelen op **wijzigingen in beschikbare gegevens** zonder dat het model zelf opnieuw getraind hoeft te worden.

Omdat de AI-assistent altijd werkt met **eigen, up-to-date gegevens**, kan het inspelen op uiteenlopende situaties waarin juiste informatie cruciaal is. In een winkelcontext betekent dit dat het niet alleen algemene productkenmerken kent, maar ook gegevens kan opvragen zoals de **actuele voorraad, promoties** of de **nieuwste producthandleidingen**. Wanneer een klant vraagt of een bepaald model op voorraad is, zal de AI-assistent niet gokken of een verouderd antwoord geven, maar rechtstreeks in het voorraadstelsel kijken en een actueel antwoord formuleren. Dit verhoogt niet alleen de betrouwbaarheid van de AI-assistent, maar maakt hem ook flexibeler en waardevoller als hulpmiddel voor klanten.

De productcatalogus van de retailers interactief maken

Onderstaande schematische voorstelling helpt als leidraad om de samenhang van de volgende secties beter te begrijpen;



Een AI-assistent die productinformatie op een natuurlijke manier kan presenteren, heeft meer nodig dan een eenvoudige zoekfunctie. **Klanten** stellen vaak **vragen** die **niet** zomaar met een **lijst van zoekresultaten beantwoord** kunnen worden. Ze willen weten welk product het beste bij hun **situatie** past, waarom een bepaald model beter is dan een ander, of hoe een product precies werkt. Om dit mogelijk te maken, is een systeem nodig dat niet alleen relevante informatie kan opzoeken, maar ook kan interpreteren en op een begrijpelijke manier kan uitleggen. In deze sectie bekijken we welke componenten hiervoor nodig zijn

Document Loaders

De **informatie** die een AI-assistent nodig heeft, staat vaak **verspreid** over verschillende documenten en in uiteenlopende formaten. Een prijslijst kan bijvoorbeeld in een Excel-bestand staan, terwijl productspecificaties in een PDF zijn opgeslagen en aanvullende uitleg in een Word-document staat. Om al deze documentvormen te verwerken, gebruiken we een **Document Loader** die documenten kan uitlezen en de inhoud **omzet in gestructureerde tekst**.

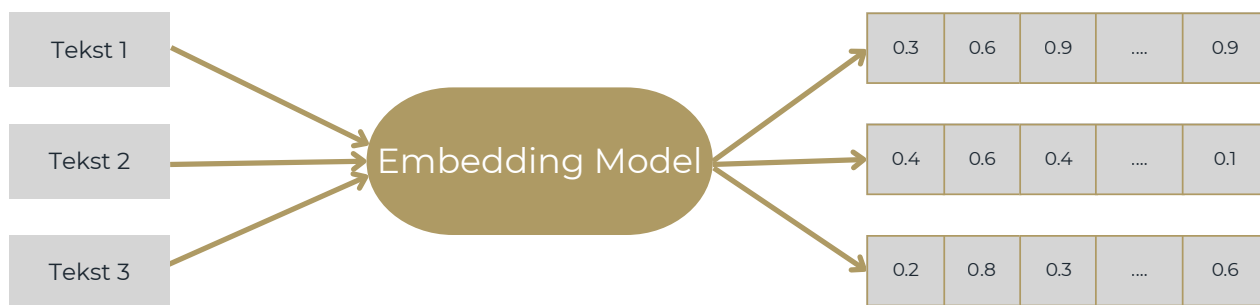
Wanneer een document wordt ingeladen, is het de taak van de Document Loader om de tekst uit het document te halen. Voor sommige formaten zoals txt-bestanden is dit heel eenvoudig, maar voor andere, zoals PDF's, kan dit behoorlijk complex zijn. Tekst kan verspreid staan over meerdere kolommen, opgesplitst zijn in afzonderlijke blokken of zelfs als afbeelding zijn opgeslagen, waardoor optische tekenherkenning (OCR) of andere technieken nodig zijn om de inhoud correct op te halen.

Als de tekst correct uit het document is gehaald, wordt dit meestal doorgegeven aan een **embedding model**, dat de tekst omzet naar **vectoren** die in de vector store opgeslagen kunnen worden.

Embedding modellen

Een embedding model zet **tekst** om in **numerieke representaties**, oftewel vectoren, zodat de **betekenis** van woorden en zinnen **wiskundig kan worden vastgelegd**. In plaats van een reeks losse woorden te beschouwen, plaatst een embedding model tekst in een hogedimensionale ruimte, waar woorden die inhoudelijk bij elkaar horen dicht bij elkaar liggen. Dit maakt het mogelijk om niet alleen exacte overeenkomsten te vinden, maar ook **semantisch gerelateerde** concepten te herkennen.

Door deze vectorrepresentatie kunnen embedding models teksten efficiënt vergelijken en snel bepalen welke stukken informatie het meest relevant zijn bij een zoekopdracht. In een **RAG**-oplossing wordt een **embedding model** gebruikt om **zowel klantvragen als** beschikbare **documenten** te verwerken, zodat de AI-assistent niet afhankelijk is van exacte trefwoorden, maar op basis van betekenis de juiste informatie kan terugvinden. Dit is essentieel voor het goed functioneren van een AI-assistent, omdat het zorgt voor flexibelere en nauwkeurigere zoekresultaten.



Vector stores

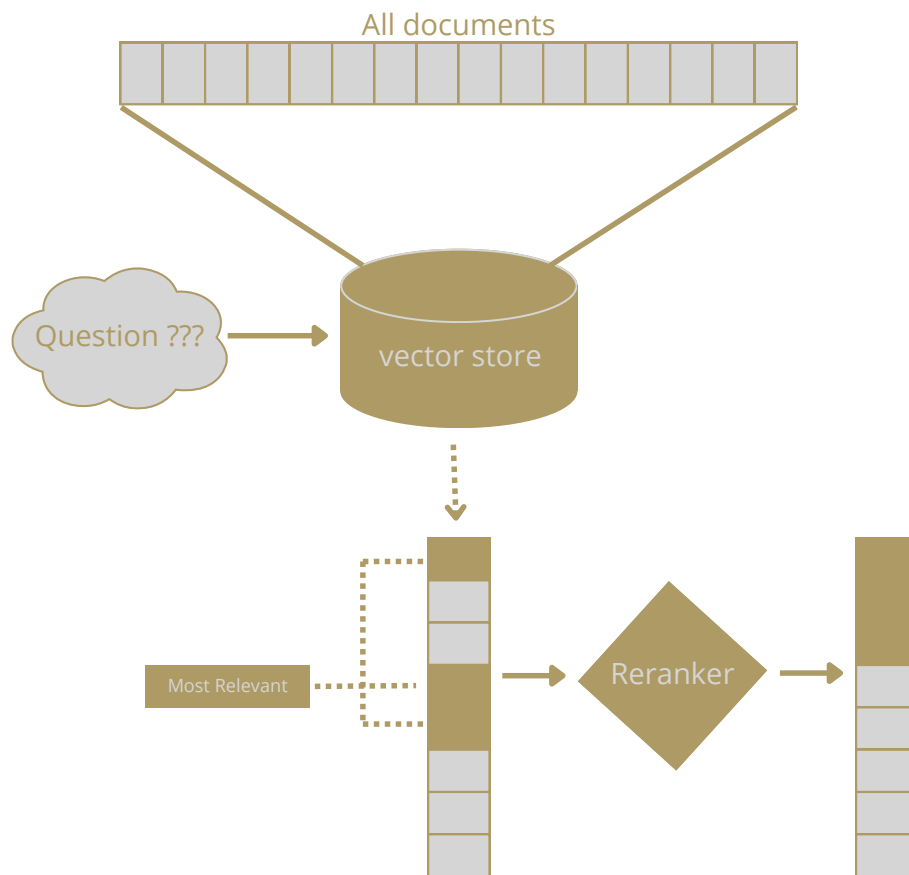
Een **vector store** is een **gespecialiseerde database** waarin de vectoren die door een embedding model zijn gegenereerd, worden opgeslagen en efficiënt kunnen worden doorzocht. Wanneer een AI-assistent informatie nodig heeft, **vergelijkt het de vraag** met de **opgeslagen teksten** in de vector store en haalt hij de meest relevante fragmenten op. Dit gebeurt niet door exacte woorden te matchen, maar met een wiskundige methode zoals cosine similarity.

Cosine similarity is een techniek om te **meten** hoe sterk **twee stukken** tekst **op elkaar lijken**, niet op basis van exacte woorden, maar **op basis van betekenis**. De AI zet de vraag van een gebruiker en de opgeslagen teksten om in vectoren en berekent vervolgens hoe dicht ze bij elkaar liggen in de vectorruimte. Hoe kleiner de hoek tussen twee vectoren, hoe groter de overeenkomst. Hierdoor kan een AI-assistent ook relevante informatie vinden, zelfs als de exacte zoekwoorden niet overeenkomen. Dankzij deze methode kan een vector store duizenden of zelfs miljoenen tekstfragmenten efficiënt doorzoeken en binnen milliseconden de meest relevante stukken teruggeven.

Rerankers

Wanneer een AI-assistent informatie ophaalt uit een documentensysteem of productcatalogus, zijn er vaak **meerdere stukken** tekst of productbeschrijvingen die de vector store zal **teruggeven op basis van de cosine similarity**. Het probleem is dat de verschillen tussen deze topresultaten vaak klein zijn, waardoor de standaard sortering niet altijd de meest nuttige informatie als eerste geeft.

Een **reranker** zorgt ervoor dat de AI niet alleen de meest relevante tekstfragmenten terugvindt, maar ze ook **beter beoordeelt op inhoud**. In plaats van enkel te vertrouwen op de numerieke afstand tot de vraag, analyseert een reranker de **werkelijke betekenis** van de gevonden stukken en herschikt ze op basis van hoe goed ze aansluiten bij wat de gebruiker echt zoekt. Dit maakt de resultaten niet alleen nauwkeuriger, maar zorgt er ook voor dat de AI-assistent alleen de meest relevante stukken gebruikt om een antwoord te formuleren.



Large Language Models

Een AI-assistent moet niet alleen relevante informatie opzoeken, **maar deze ook in een begrijpelijke en samenhangende vorm presenteren**. De teksten die uit een vector store worden opgehaald, zijn vaak losse fragmenten die niet altijd direct een volledig antwoord vormen. Daarom wordt een **Large Language Model (LLM)** gebruikt om de opgehaalde informatie te verwerken en om te zetten in een vloeiende en natuurlijke tekst. Een LLM begrijpt de vraag van de gebruiker, combineert deze met de opgehaalde informatie en genereert een antwoord dat logisch opgebouwd is en aansluit bij de context van de vraag.

Niet alle LLM's zijn gelijk. Er bestaat een groot verschil tussen modellen die lokaal draaien en modellen die via een API worden geraadpleegd. **Lokale** modellen bieden **volledige controle** en zijn ideaal voor situaties waar **privacy of offline** verwerking een belangrijke rol speelt. Ze vereisen echter **aanzienlijke rekenkracht** en moeten **zelf worden onderhouden**. **API**-gebaseerde modellen, zoals die van OpenAI, Anthropic of Google, worden **extern** gehost en zijn vaak **krachtiger**, omdat ze continu worden bijgewerkt en profiteren van **schaalbare infrastructuur**. Het **nadeel** is dat deze afhankelijk zijn van een externe aanbieder en dat er **kosten** verbonden zijn aan het gebruik.

Naast de keuze tussen lokaal of API-gebaseerd gebruik, zijn er ook verschillen in **modelgrootte**, kosten en **toegankelijkheid**. Kleinere modellen kunnen efficiënter en goedkoper zijn, maar leveren minder accurate antwoorden op dan grote modellen met miljarden parameters. Open-source modellen zoals LLaMA of Mistral kunnen gratis gebruikt worden, terwijl commerciële modellen vaak rekenen per aantal verwerkte tokens. Daarnaast zijn sommige modellen gemakkelijker toegankelijk binnen specifieke ecosystemen, zoals OpenAI-modellen binnen Microsoft-producten of gespecialiseerde modellen die optimaal werken binnen een bepaalde cloudomgeving. De **keuze** voor een **LLM** hangt dus af van factoren zoals **prestatie-eisen, kosten, privacy** en **integratiemogelijkheden**.

Overige

In de vorige paragrafen hebben we gefocust op de grote onderdelen die nodig zijn om zelf een RAG te bouwen. Daarnaast zijn er nog een aantal andere componenten die een rol spelen bij het optimaliseren van het systeem.

Systeem prompts

Een systeem prompt is een **instructie** die een **LLM vooraf** krijgt om zijn **gedrag en toon** te bepalen bij het beantwoorden van een vraag. Dit kan gaan van algemene richtlijnen zoals “Wees beknopt en professioneel” tot specifieke instructies zoals “Baseer je antwoorden uitsluitend op de gegeven documenten”. Daarnaast kan een systeem prompt worden gebruikt om de AI een persona mee te geven, zoals een behulpzame klantenservice-medewerker, een technische expert of een formele adviseur. Dit helpt om de interactie consistent en beter afgestemd te maken op de doelgroep. De systeem prompt is niet zichtbaar voor de gebruiker.

Chunkers

Een chunker is een tool die **lange documenten opsplijt** in kleinere, behapbare stukken voordat ze naar een vector store worden gestuurd. Dit is noodzakelijk omdat LLM's een beperkt context window hebben en omdat kleinere tekstfragmenten efficiënter kunnen worden doorzocht en vergeleken met een zoekopdracht. Een goede chunking-strategie houdt rekening met de structuur van een document—bijvoorbeeld door tekst op te splitsen per alinea, sectie of op basis van semantische samenhang.

Opslag

In sommige gevallen kan het handig zijn om de **originele documenten**, afbeeldingen of andere media **bij de hand te hebben**, zodat we ernaar kunnen verwijzen in het antwoord, bijvoorbeeld om de volledige context te kunnen bekijken of meer visuele achtergrond te geven. Deze bestanden kunnen worden **opgeslagen** in verschillende systemen, zoals een cloudgebaseerde opslagdienst of een lokale database. De keuze hangt af van factoren zoals schaalbaarheid, kosten en toegangsbeheer.

Guardrailing

Guardrailing verwijst naar de maatregelen die worden genomen om de AI-assistent **binnen veilige en gewenste grenzen te houden**. Dit omvat het filteren van ongepaste vragen, het voorkomen van foutieve of misleidende antwoorden en het beperken van de AI tot betrouwbare bronnen. Een guardrailing-systeem kan verschillende technieken gebruiken, zoals inhoudsfilters om gevoelige of schadelijke verzoeken te blokkeren, query-modificatie om vragen te herformuleren binnen de juiste context, en responsvalidatie om te controleren of een gegenereerd antwoord voldoet aan vooraf bepaalde richtlijnen.

AI Act

De AI Act, een **wetgevend kader van de Europese Unie**, legt regels op voor het ontwikkelen en inzetten van AI-systemen om transparantie, veiligheid en ethiek te waarborgen. AI-systemen worden ingedeeld in verschillende risiconiveaus, waarbij strengere eisen gelden voor toepassingen met een hoog risico, zoals geautomatiseerde besluitvorming met grote impact op individuen. Voor een AI-assistent in een RAG-systeem betekent dit dat ontwikkelaars rekening moeten houden met gegevensbescherming, uitlegbaarheid en verantwoord gebruik. Dit kan onder andere inhouden dat de herkomst van informatie duidelijk wordt vermeld, dat de AI geen ongefundeerde antwoorden genereert en dat er mechanismen zijn om menselijke controle en correctie mogelijk te maken. Het naleven van de AI Act is essentieel om zowel juridische risico's als reputatieschade te vermijden.

GDPR

De General Data Protection Regulation (GDPR) **stelt strikte eisen aan hoe persoonlijke gegevens worden verwerkt**, opgeslagen en gedeeld binnen AI-systemen. In een RAG-oplossing betekent dit dat er zorgvuldig moet worden omgegaan met gebruikersgegevens en de informatiebronnen die de AI-assistent raadpleegt. Dit omvat het minimaliseren van verzamelde data, het anonimiseren of pseudonimiseren van persoonsgegevens en het voorzien van transparantie over hoe gegevens worden gebruikt. Daarnaast moeten gebruikers het recht hebben om hun gegevens in te zien, te laten corrigeren of te laten verwijderen.

Voldoen aan de noden van de retailer

Een AI-assistent is natuurlijk meer dan enkel het aan elkaar koppelen van enkele componenten. Om succesvol te zijn, moet hij ook voldoende aan de noden van de klant. In dit hoofdstuk overlopen we zes basisbouwstenen die samen bepalen of een AI-oplossing praktisch inzetbaar en duurzaam is:

- **presentatie**
- **vraag van de consument**
- **data**
- **infrastructuur**
- **veiligheid**
- **regelgeving**

Elk van deze bouwstenen vertegenwoordigt een essentieel aspect van het systeem, van hoe de AI-assistent met de gebruiker communiceert tot hoe gegevens worden verwerkt en beveiligd.

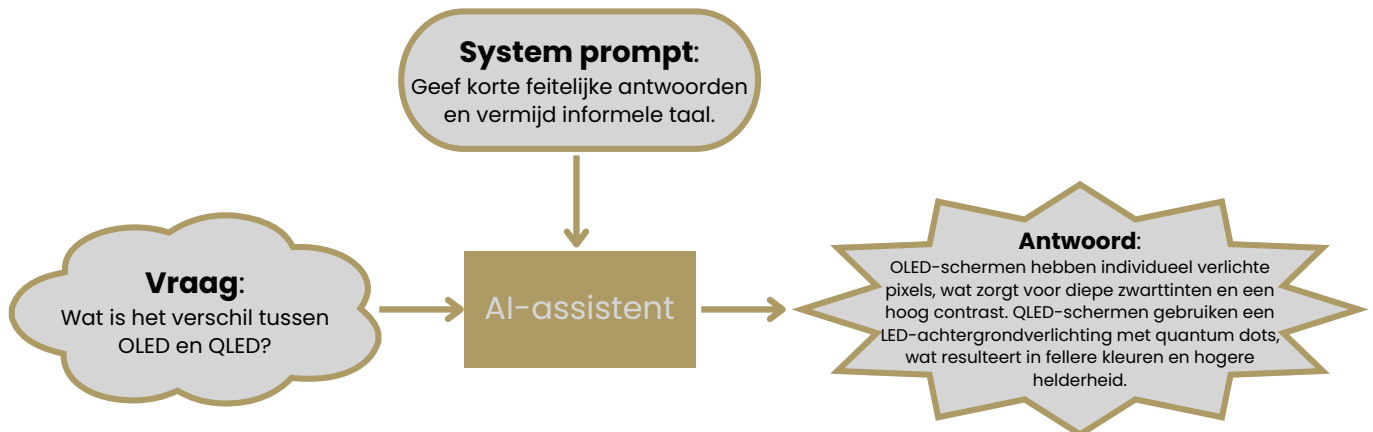
In het vorige hoofdstuk bespraken we de belangrijkste componenten van een Retrieval-Augmented Generation (RAG)-systeem, zoals de document loader, embedding models, de vector store en het LLM. In dit hoofdstuk bekijken we **hoe deze componenten samenwerken om de zes basisbouwstenen af te vinken**. Dit biedt een praktisch inzicht in hoe een AI-assistent niet alleen technisch, maar ook functioneel en betrouwbaar kan worden geïmplementeerd.

Presentatie

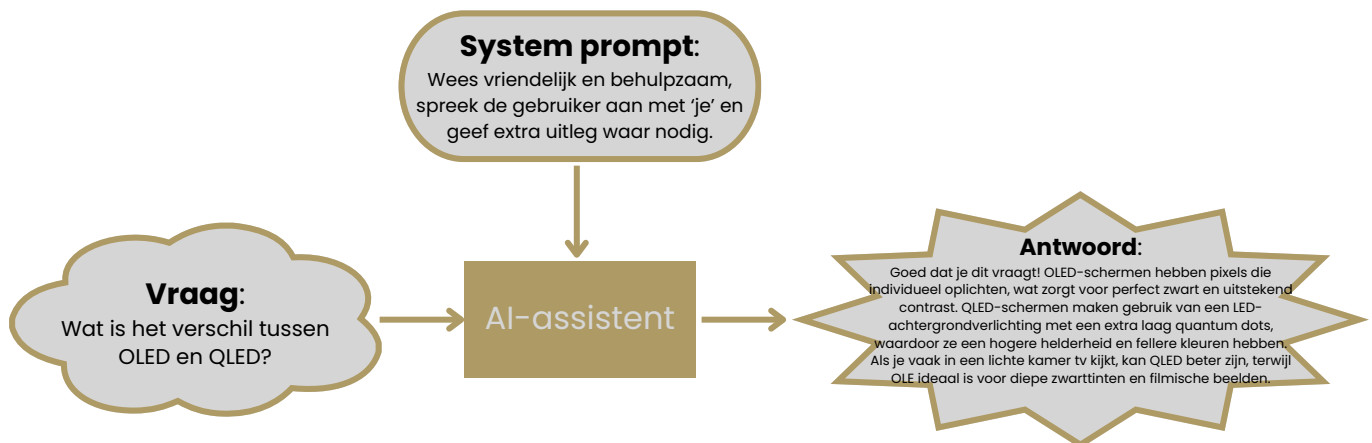
De manier waarop een AI-assistent informatie presenteert, bepaalt in grote mate hoe gebruiksvriendelijk en effectief de interactie met de gebruiker is. Presentatie gaat niet alleen over hoe je antwoorden visueel weergeeft, maar ook over de toon, stijl en duidelijkheid van de communicatie. Het is niet voldoende dat de informatie die de AI-assistent geeft factueel juist is, maar hij moet dit ook op een manier doen die aansluit bij de verwachtingen van de gebruiker. Dit betekent dat hij zakelijk en beknopt kan antwoorden in een professionele setting, terwijl een informele en vriendelijke toon beter past in een interactie met klanten.

Hoe de AI-assistent zijn antwoorden formuleert, kan je sturen met behulp van de **system prompt**. Deze bepaalt niet alleen wát er wordt gezegd, maar ook hóé het wordt gebracht. Een kleine aanpassing in de prompt kan ervoor zorgen dat de assistent formeel en beknopt reageert of juist informeel en uitnodigend communiceert. Dit verschil wordt geïllustreerd in de figuren op de volgende bladzijde, waar dezelfde vraag over het verschil tussen OLED en QLED op twee verschillende manieren wordt beantwoord. In de eerste figuur is de prompt gericht op korte, feitelijke antwoorden zonder informele taal, waardoor de respons direct en zakelijk is. In de tweede figuur is de system prompt ingesteld op een vriendelijke en behulpzame toon, wat resulteert in een uitgebreid, toegankelijk antwoord met extra uitleg.

Voorbeeldprompt voor formele communicatie:



Voorbeeldprompt voor informele communicatie:



Vraag van de consument

De **manier waarop** een klant een **vraag** stelt, kan sterk variëren en heeft een directe impact op hoe de AI-assistent deze moet verwerken. Sommige klanten stellen korte en bondige vragen met enkele kernwoorden, terwijl anderen een uitgebreide beschrijving zullen geven, met de nodige mogelijk irrelevante details. De AI-assistent moet in staat zijn om beide typen vragen correct te interpreteren en de essentie eruit te halen.

Naast de stijl van de vraag is ook de **taal** een bepalende factor. Is de AI-assistent ontworpen om uitsluitend in één taal te functioneren, of moet hij ondersteuning kunnen bieden in meerdere talen? Moet hij kunnen antwoorden in de taal van de klant?

Het **embedding model** is een cruciale component om correct om te gaan met de vraag van de klant, omdat het ervoor zorgt dat varianten van dezelfde vraag op dezelfde manier geïnterpreteerd worden. Door elke vraag om te zetten in een vectorrepresentatie, worden inhoudelijk gelijkaardige vragen dicht bij elkaar geplaatst in de vectorruimte. Hierdoor maakt het niet uit of een klant vraagt “Waar staan de OLED tv’s” of “Ik zoek een OLED scherm”. De assistent zal beide formuleringen als inhoudelijk hetzelfde beschouwen en dezelfde relevante informatie terugvinden. Dit voorkomt dat kleine verschillen in woordkeuze of zinsopbouw invloed hebben op de nauwkeurigheid van het antwoord en zorgt ervoor dat de assistent flexibel omgaat met uiteenlopende vraagstijlen.

Veel embedding modellen zijn getraind voor slechts één taal, vaak een grote taal zoals Engels of Chinees. In sommige gevallen is dit voldoende, bijvoorbeeld wanneer zowel de vragen als de beschikbare informatie zich binnen die ene taal bevinden. In een meertalige omgeving, zoals België, is dit echter minder praktisch. Een model dat alleen voor het Nederlands getraind is, komt in de praktijk weinig voor, waardoor een model nodig is dat meerdere talen ondersteunt. Het voordeel van zo’n meertalig embedding model is niet alleen dat het vragen in verschillende talen kan verwerken, maar ook dat het relaties legt tussen die talen. Dit werkt doordat vragen, in welke taal ze ook gesteld worden, naar een gedeelde vectorruimte worden omgezet. Hierdoor kan een klant bijvoorbeeld zijn vraag in het Nederlands formuleren en een antwoord krijgen uit een Engelstalige productcatalogus of technische documentatie.

Wanneer de AI-assistent de vraag heeft geïnterpreteerd en de relevante informatie heeft opgehaald, is het de taak van het **LLM** om een begrijpelijk antwoord te formuleren. Maar in welke taal zal het antwoorden? Sommige modellen zijn getraind in slechts één taal en kunnen alleen in die taal antwoorden, zelfs als de vraag in een andere taal gesteld wordt. Andere modellen zijn wel meertalig, herkennen automatisch de taal van de vraag en kunnen ook in die taal antwoorden. Zo kan een klant dus een vraag in het Nederlands stellen en een begrijpelijk antwoord in het Nederlands terugkrijgen, zelfs als de technische documentatie in het Engels (of Chinees) staat.

Tot slot speelt ook de **reranker** hier een mogelijke rol. Bij algemene vragen vindt het embedding model vaak veel mogelijke antwoorden die in de vectorruimte ongeveer even dicht bij elkaar liggen. De reranker kan dan helpen om de stukken tekst uit te kiezen die het best aansluiten op de vraag. In plaats van simpelweg de eerste overeenkomsten terug te geven, analyseert hij de inhoud en geeft hij voorrang aan de meest bruikbare informatie.

Data

Om een AI-assistent effectief te laten werken, is niet alleen de verwerking van vragen belangrijk, maar ook **hoe de onderliggende data wordt beheerd**. Informatie moet op een manier worden opgeslagen die zowel efficiënte raadpleging als nauwkeurige interpretatie mogelijk maakt. Dit betekent dat gegevens niet alleen eenvoudig doorzoekbaar moeten zijn, maar ook in een **formaat** moeten worden bewaard dat **geschikt** is voor verdere verwerking door de AI. Daarnaast is het in veel gevallen nuttig om de originele documenten te behouden, zodat de assistent niet alleen antwoorden kan genereren, maar ook verwijzingen naar de brondocumenten kan geven.

Bij het verwerken van documenten vormt het PDF-formaat vaak de grootste uitdaging voor een **document loader**. Afhankelijk van de inhoud en opmaak kunnen eenvoudige extractiemethoden voldoende zijn, maar complexere documenten vereisen krachtigere technieken. Kiezen we voor lokale extractie of gaan we voor een cloudgebaseerde oplossing? Ingebouwde bibliotheken kunnen tekst uit eenvoudige PDF's halen, maar lopen vaak vast bij ingescande documenten of tabellen, waar OCR nodig is. Cloudoplossingen kunnen deze complexiteit doorgaans beter aan, maar zijn meestal betalend en vereisen een externe API-call, wat impact kan hebben op snelheid en vertrouwelijkheid.

Een ander bepalend aspect bij het kiezen van een document loader is het gewenste outputformaat. De eenvoudigste loaders geven tekst zonder opmaak terug, wat in veel gevallen volstaat. Krachtigere loaders bieden echter de mogelijkheid om documenten in Markdown te exporteren, waardoor titels, lijsten en andere opmaak behouden blijven. Dit helpt bij het chunken van documenten, omdat secties duidelijker afgebakend zijn en de structuur behouden blijft.

De keuze van een **vector store** hangt af van hoe groot de dataset is en hoe de opslag en zoekopdrachten moeten verlopen. Een eerste optie zijn kleine vector stores die volledig in RAM worden geladen. Deze zijn extreem snel, maar de data verdwijnt bij een herstart, waardoor de vectoren telkens opnieuw moeten worden opgebouwd of ingeladen. Dit maakt ze vooral geschikt voor prototyping of toepassingen met beperkte datasets. Een tweede optie zijn lokaal gehoste vector stores, die op een eigen server draaien en vectoren ook op schijf kunnen opslaan. Dit zorgt ervoor dat data persistent blijft en de vector store niet telkens opnieuw opgebouwd moet worden. Deze aanpak is ideaal wanneer controle over de infrastructuur en privacy van de data belangrijk zijn. De derde categorie zijn remote, cloudgebaseerde vector stores, die ontworpen zijn voor zeer grote toepassingen met miljoenen tot miljarden vectoren. Ze zijn schaalbaar en altijd beschikbaar, maar voor een AI-assistent in een retailsetting is dat meestal niet nodig. De dataset blijft relatief klein, terwijl snelheid en controle zwaarder doorwegen dan een onbeperkte opslagcapaciteit.

Soms is het handig om **opslag** voor de originele documenten te voorzien, zodat de AI-assistent niet alleen antwoorden kan geven, maar ook kan verwijzen naar de bron. Dit kan lokaal, wat volledige controle geeft, maar betekent dat de bestanden enkel op dat toestel beschikbaar zijn. Een alternatief is cloudopslag, wat bredere toegankelijkheid biedt en vaak eenvoudiger te integreren is met andere systemen. Hoewel cloudopslag meestal niet gratis is, blijven de kosten doorgaans beperkt.

Infrastructuur

De infrastructuur van een AI-assistent omvat alle technische middelen die nodig zijn om de verschillende componenten correct te laten functioneren. Dit gaat niet alleen over de hardware waarop de AI draait, maar ook over de manier waarop rekenkracht, opslag en netwerkverbindingen worden beheerd. Een belangrijke keuze hierbij is of de componenten lokaal worden gehost, in de cloud draaien of een combinatie van beide gebruiken.

De **LLM** is hierbij vaak de bepalende factor. Wanneer de AI-assistent ook offline beschikbaar moet zijn, is lokaal draaien de enige mogelijkheid, bijvoorbeeld voor afgelegen locaties of plaatsen met weinig of geen internet toegang. Daarnaast spelen prestatieverschillen een rol. Hoewel cloudgebaseerde modellen vaak sneller zijn door de beschikbaarheid van gespecialiseerde hardware, is de reactietijd afhankelijk van serverbelasting en netwerklatentie. Lokale modellen hebben geen last van externe vertragingen, maar de inferentiesnelheid wordt beperkt door de beschikbare rekenkracht. Vooral bij grotere modellen kan een lokaal gehost systeem trager zijn door een beperkt aantal tokens per seconde, terwijl kleinere modellen vaak vlot kunnen draaien op een krachtige GPU.

De grootte van een LLM bepaalt dus in grote mate of een eigen infrastructuur een haalbare optie is. Kleinere modellen, tot ongeveer 7 miljard parameters (7B), kunnen nog relatief efficiënt draaien op een krachtige consumenten-GPU met voldoende VRAM. Met quantization-technieken kan zelfs een meer bescheiden workstation geschikt zijn, zij het met een mogelijke impact op de modelkwaliteit. Modellen tussen 13B en 30B parameters vormen een grensgebied: ze kunnen nog lokaal draaien op high-end workstations met meerdere GPU's, maar dit vereist aanzienlijke hardware-investeringen en optimalisaties om inference-tijden acceptabel te houden.

Boven 30B wordt lokaal draaien in de meeste gevallen onpraktisch, tenzij er gespecialiseerde AI-hardware beschikbaar is, zoals meerdere professionele GPU's of een clusteropstelling. De huidige state-of-the-art modellen bevinden zich echter uitsluitend in de 70B+ categorie en vereisen dermate veel geheugen en rekenkracht dat een lokale implementatie buiten bereik ligt van standaard hardware. Dit betekent dat, naarmate de modelgrootte toeneemt, een lokale opstelling steeds minder realistisch wordt en een hybride of volledige cloudgebaseerde aanpak noodzakelijk is.

Hoewel de keuze van een **vector store** en **document loader** invloed heeft op de infrastructuur, is deze impact doorgaans minder groot dan bij een LLM. Een vector store vereist opslagruimte en snelle zoekfunctionaliteit, maar kan vaak draaien op standaard hardware zonder extreme rekenkracht. Voor kleinere toepassingen volstaat een lokaal gehoste vector store, terwijl grootschalige implementaties baat hebben bij een cloudgebaseerde oplossing voor betere schaalbaarheid.

Een document loader stelt nauwelijks specifieke infrastructuureisen, tenzij grote hoeveelheden documenten in real-time verwerkt moeten worden. In dat geval kan een krachtige server of cloudoplossing nodig zijn om zware OCR- en extractietaken efficiënt uit te voeren.

Veiligheid

Een AI-assistent moet niet alleen nuttige en correcte antwoorden geven, maar ook veilig functioneren binnen de grenzen van de beoogde toepassing. Veiligheid draait in deze context om het voorkomen van ongepaste, misleidende of schadelijke output en het beschermen van vertrouwelijke informatie. Dit wordt gerealiseerd via **guardrailing**, een reeks technieken die ervoor zorgen dat de AI zich aan vooraf bepaalde richtlijnen houdt.

Veel grote **LLM**'s hebben al ingebouwde guardrailing om te voorkomen dat ze ongepaste of gevaarlijke inhoud genereren. Dit betekent dat ze automatisch weigeren om vragen over bijvoorbeeld haatspraak, illegale activiteiten of geweldsinstructies te beantwoorden. Deze beveiligingen zijn een standaardonderdeel van de meeste cloudgebaseerde modellen en beperken het risico op misbruik.

Daarnaast kan de **system prompt** worden ingezet om extra guardrails op maat te implementeren. Hierin kunnen expliciete instructies worden meegegeven, zoals welke onderwerpen vermeden moeten worden, hoe gevoeligere vragen moeten worden beantwoord en welke toon de assistent moet hanteren. Bijvoorbeeld, een system prompt kan het model instrueren om altijd neutraal en feitelijk te blijven of om gebruikers door te verwijzen naar een menselijke medewerker bij juridische of medische vragen.

Naast het voorkomen van ongepaste output is het ook belangrijk om de AI-assistent te beschermen tegen pogingen om het systeem te misleiden. Gebruikers kunnen proberen de guardrails te omzeilen via jailbreaking-technieken, zoals het vragen om een "hypothetisch scenario" waarin het model toch vertrouwelijke informatie onthult. Ook kunnen ze proberen de AI te laten redeneren in meerdere stappen om alsnog verboden antwoorden te genereren. Een goed geformuleerde system prompt kan hier tegen beschermen door het model expliciet te instrueren om altijd de oorspronkelijke richtlijnen te volgen, ongeacht de context van de vraag.

Regelgeving

Bij de ontwikkeling en inzet van een AI-assistent komen niet alleen technische en functionele keuzes kijken, maar ook naleving van de geldende regelgeving. Wetgeving rond AI en gegevensbescherming evolueert snel en bepaalt in welke mate een AI-systeem mag functioneren. Twee belangrijke kaders zijn de **AI Act**, die risico's en verantwoordelijkheden van AI-systemen vastlegt, en de **GDPR** wetgeving, die regelt hoe persoonsgegevens verwerkt mogen worden. Afhankelijk van de toepassing kunnen er aanvullende eisen gelden, zoals transparantie over hoe de AI tot beslissingen komt of het recht van gebruikers om niet uitsluitend door AI beoordeeld te worden.

De AI Act classificeert AI-systemen op basis van risico, waarbij strengere regels gelden voor toepassingen met een hogere impact op mensenrechten en veiligheid. Een AI-assistent in een winkelomgeving valt doorgaans in een lagere risicocategorie, maar moet nog steeds voldoen aan basisprincipes zoals transparantie en naleving van ethische richtlijnen. De GDPR wetgeving speelt een rol zodra de AI persoonsgegevens verwerkt, zoals gebruikersvragen die kunnen worden opgeslagen of gekoppeld aan klantprofielen. Dit betekent dat er een juridische basis nodig is voor de verwerking, gebruikers duidelijk geïnformeerd moeten worden over wat er met hun data gebeurt, en er mechanismen moeten zijn om gegevens te corrigeren of te verwijderen indien gewenst.

All you need is RAG?

RAG is een krachtige methode om een AI-assistent te voorzien van actuele en relevante informatie uit een documentencollectie. Toch heeft deze aanpak beperkingen. RAG werkt uitsluitend met de documenten die vooraf zijn ingeladen, wat betekent dat informatie die daarbuiten valt niet beschikbaar is voor de AI. Dit kan problematisch zijn wanneer de gebruiker een vraag stelt over real-time data, veranderlijke informatie of domeinen die niet volledig gedekt worden door de beschikbare documenten.

Om de beperkingen van RAG op te vangen, kan een AI-assistent worden uitgebreid met andere tools. Een voorbeeld hiervan zijn **API-integraties**. Dit stelt de AI in staat om niet alleen antwoorden te genereren op basis van statische documenten, maar ook om real-time informatie op te halen of acties uit te voeren. Bijvoorbeeld, wanneer een klant vraagt of een product op voorraad is, volstaat het niet om enkel documentatie te doorzoeken; de assistent moet direct een voorraad-API kunnen raadplegen. Hetzelfde geldt voor prijsinformatie, orderstatussen, of afsprakenplanning, waarbij een directe koppeling met een extern systeem noodzakelijk is.

Naast API-integraties kan het in sommige gevallen ook nuttig zijn als de AI-assistent ook informatie kan ophalen van **websites**. Misschien zijn bepaalde gegevens niet beschikbaar in de interne documentatie, maar wel op de website van de fabrikant. Wel moeten we opletten dat de assistent geen informatie opvraagt van onbetrouwbare bronnen of rechtstreekse concurrenten. We zouden immers niet willen dat de assistent aanbevelingen doet uit de catalogus van een concurrent. Daarom is het goed om toegankelijke websites in te perken via **whitelisting**.

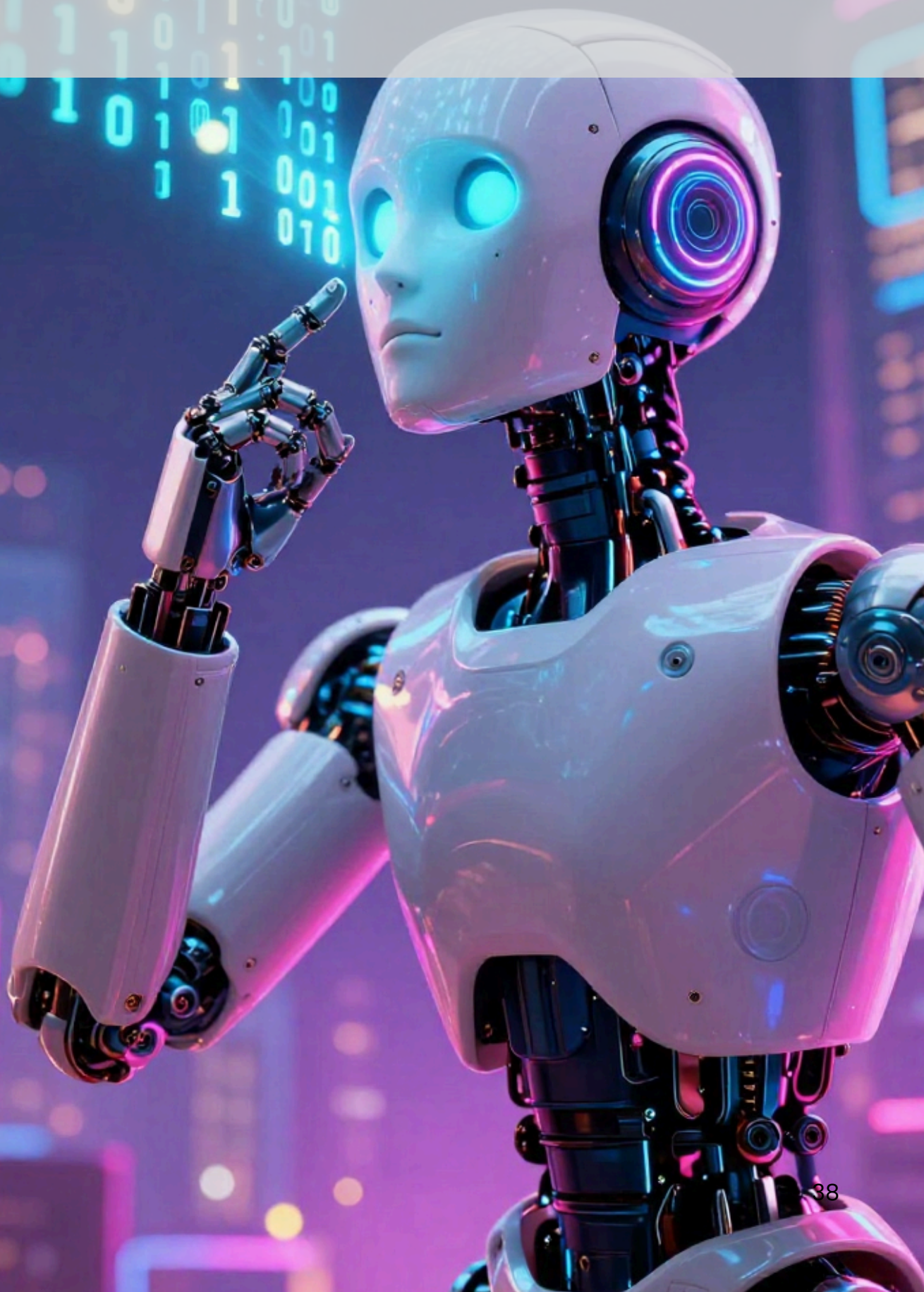
Idealiter willen we niet zelf van tevoren bepalen uit welke bron de assistent zijn informatie moet ophalen. We willen dat de assistent zelf kan bepalen wanneer het best zijn eigen documenten raadpleegt, wanneer het online naar informatie zoekt of wanneer het een externe database aanspreekt. Om dit te doen, heeft de assistent een beslissingsmechanisme nodig dat kan inschatten welke bron het meest geschikt is op basis van de vraag van de gebruiker. Dit kan door middel van **tool calling**, waarbij de AI-assistent toegang heeft tot verschillende tools en, afhankelijk van de context, bepaalt welke tool wordt aangesproken.

Bijvoorbeeld, wanneer een klant vraagt naar een specifieke productspecificatie, kan de assistent eerst de interne documentatie raadplegen. Als het antwoord daar niet gevonden wordt, kan het een vooraf goedgekeurde externe bron, zoals de website van de fabrikant, doorzoeken. Voor vragen over real-time gegevens, zoals de beschikbaarheid van een product in de winkel, kan de AI direct een API-aanroep doen naar het voorraadsysteem. Door deze technieken te combineren, kan de assistent flexibel en intelligent omgaan met informatiebronnen, terwijl er controle blijft over welke externe data hij mag gebruiken.

Naast het correct ophalen van informatie, is ook de manier waarop deze aan de gebruiker wordt gepresenteerd cruciaal. De interactie tussen de klant en de AI-assistent bepaalt in grote mate hoe bruikbaar en toegankelijk het systeem aanvoelt. Een eenvoudige en veelgebruikte aanpak is een interface waarin klanten hun vragen via tekst invoeren en het antwoord ook tekstueel gepresenteerd krijgen. Dit werkt intuïtief en vereist geen bijzondere hardware, maar het beperkt zich tot lezen en typen, wat niet voor elke gebruiker even toegankelijk is.

In sommige situaties kan het wenselijk zijn om een meer natuurlijke vorm van interactie aan te bieden, bijvoorbeeld via **spraakherkenning** en gesproken antwoorden. Dit maakt het gebruik van de assistent toegankelijker, bijvoorbeeld in situaties waar typen moeilijk of onpraktisch is. Ook een visuele representatie van de AI-assistent in de vorm van een **avatar** kan een meer persoonlijke en intuïtieve interactie bieden. Zo'n avatar kan niet alleen het gebruiksgemak verhogen, maar ook bijdragen aan een positieve gebruikservaring doordat klanten het gevoel hebben persoonlijk geholpen te worden.

Aandachtspunten



Aandachtspunten bij implementatie

Gebruiksvriendelijkheid en toegankelijkheid

Een klant wilt een aangename winkelbeleving en wil op een vlotte manier geholpen worden. De interface van de AI-assistent moet daarom intuïtief zijn en werken met korte, duidelijke stappen. Lange of omslachtige processen schrikken klanten af en leiden vaak tot afhaken. Houd het dus laagdrempelig voor de klant. Verder is het belangrijk voor de gebruiksvriendelijkheid dat de AI-assistent de potentiële vragen van een klant kan beantwoorden. Is het bijvoorbeeld belangrijk voor jouw klanten om snel zicht te hebben op de voorraad? Zorg dan dat de AI-assistent dit kan door de juiste data aan te leveren. Ook meertaligheid kan een belangrijke meerwaarde bieden om de AI-assistent toegankelijker te maken voor een breder publiek.

Integratie met bestaande systemen

Een AI-assistent is slechts zo sterk als de data waarop hij draait. Het is essentieel dat de technologie wordt gekoppeld aan bestaande systemen zoals de productdatabases, live stock etc. Zo blijft de informatie actueel, volledig en betrouwbaar. Een AI-assistent die geen correcte of recente productinformatie kan geven, verliest snel zijn geloofwaardigheid bij klanten. Stel jezelf dus de vraag met welke data je welke consumentenvragen kan beantwoorden.

Training en rol van medewerkers

Succesvolle implementatie valt of staat met de betrokkenheid van medewerkers. Training is noodzakelijk, bijvoorbeeld via een praktisch draaiboek, korte video's of hands-on workshops. Dit geeft medewerkers het vertrouwen om de AI-assistent actief aan te raden en zelf te gebruiken om zo klanten correct te begeleiden. Een menselijke fallback moet altijd mogelijk zijn: wanneer de AI-assistent een vraag niet kan beantwoorden, moet hij eenvoudig kunnen doorverwijzen naar een medewerker.

Communicatie naar klanten

De lancering van een AI-assistent vraagt ook duidelijke communicatie richting klanten, want zelfs de beste AI-assistent kan niet helpen wanneer de klant deze niet opneemt in zijn winkelproces. Denk aan visuele aanduidingen in de winkel, instructies bij de kiosk en promotie via social media. Een andere mogelijkheid is het inschakelen van een sociale robot die zich kan voortbewegen en die zelf naar klanten toestapt. Op die manier worden klanten actief gestimuleerd om de technologie te omarmen. Communicatie draagt er ook toe bij dat klanten begrijpen waarvoor de AI-assistent bedoeld is en welke voordelen ze kunnen verwachten.

Dataveiligheid en privacy

Naast een duidelijke communicatie naar klanten is vertrouwen ook een belangrijke succesfactor tot gebruik. Klanten willen geholpen worden zonder dat hun persoonlijke gegevens meteen worden opgevraagd. Het is daarom aan te raden om anonieme interactie als standaard te hanteren, in lijn met GDPR-regelgeving. Dit verlaagt de drempel om de AI-assistent te gebruiken en geeft klanten de zekerheid dat hun privacy wordt gerespecteerd. Denk dus na over het type AI-assistent dat bij jouw type winkel past. Eerder een klein, privé scherm voor winkels met producten of services

Training en personalisatie van het model

Naast de opleiding van medewerkers moet ook de AI-assistent zelf regelmatig getraind en geoptimaliseerd worden. Een retail-specifieke fine-tuning van het model zorgt ervoor dat antwoorden relevant en waardevol zijn. Dit kan bijvoorbeeld door veelgestelde vragen uit de winkelpraktijk toe te voegen of door integratie met de interne kennisbank.

Kosten-batenanalyse

Tot slot moet de implementatie altijd gepaard gaan met een grondige kosten-batenanalyse. De investering in technologie, training en onderhoud moet worden afgewogen tegen de verwachte voordelen zoals kortere wachttijden, hogere klanttevredenheid en efficiënter winkelbeheer. Een realistische inschatting helpt om draagvlak te creëren bij het management én het personeel.

Conclusie en aanbevelingen

De case van Blabloom toont dat een Instore AI Assistent een belangrijke meerwaarde kan bieden in een fysieke winkelomgeving. Vooral in situaties met veel repetitieve vragen en complexe productassortimenten kan AI het verschil maken. Het verkort interactietijden, verlaagt werkdruk en verhoogt klanttevredenheid.

De inzichten uit het WARA-interview bevestigen daarenboven dat een AI-assistent in de winkel een waardevol instrument kan zijn, zowel voor klanten als voor personeel. Het verkorten van wachttijden, het ondersteunen van productkeuzeprocessen en het helpen bij opleiding van nieuwe medewerkers zijn concrete voordelen. WARA benadrukt bovendien het belang van evenwicht: de digitale assistent moet de winkelervaring versterken zonder deze te reduceren tot een webshop in fysieke vorm.

Aanbevelingen:

1. Start met een duidelijke use case: identificeer repetitieve klantvragen.
2. Zorg voor een rijke en gestructureerde productdatabase en denk na hoe klanten vragen stellen.
3. Positioneer de kiosk strategisch in de winkel.
4. Voorzie opleiding via key users en ondersteunend materiaal (draaiboek, video's).
5. Communiceer duidelijk naar klanten dat de AI een hulpmiddel is, niet een vervanging van persoonlijk advies.

Een AI-assistent kan zowel klant als medewerker ondersteunen en past binnen de bredere trend van phygital retail. Het is een praktische stap richting de winkel van morgen

Wil je meer hands-on aan de slag gaan? Check dan onze website waar je praktijkgerichte workshops en e-learnings kan terugvinden. Daarmee kan je toegang krijgen tot verschillende checklists, draaiboeken en implementatietips die zowel voor de retailer als technologieaanbieder relevant zijn. [Instore AI-assistent PXL](#)

Bronnen

Amazon Web Services. (z.d.). *What are Embeddings in Machine Learning?* Amazon Web Services, Inc. <https://aws.amazon.com/what-is/embeddings-in-machine-learning/>

Blanca-Sancho, C. (2025, 17 september). *Guardrails for Responsible AI. Academia and Government.* <https://clarivate.com/academia-government/blog/guardrails-for-responsible-ai/>

Bogue, R. (2019). *Strong prospects for robots in retail.* Industrial Robot: the international journal of robotics research and application, 46(3), 326-331.

Brown, D. (2023, 29 juni). *AI, 3D avatars revolutionize customer service, drive-thru.* Digital Signage Today. <https://www.digitalsignagetoday.com/articles/ai-3d-avatars-revolutionize-customer-service-drive-thru/>

Chatbase. (2024, September 29). *Reranking explained: Why it matters for RAG systems.* <https://www.chatbase.co/blog/reranking>

Dashboard - Aalst. (z.d.-b). <https://provincies.incijfers.be/>

Deferme, S. (2022, August 2). *Winkels moeten steeds vaker tijdelijk sluiten door tekort aan personeel.* Het Belang van Limburg. https://m.hbvl.be/cnt/dmf20220801_96465453

European Commission. (2026). *AI Act – regulatory framework for artificial intelligence.* <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

Grewal, D., Benoit, S., Noble, S. M., Guha, A., Ahlbom, C. P., & Nordfält, J. (2023). *Leveraging in-store technology and AI: Increasing customer and employee efficiency and enhancing their experiences.* Journal of Retailing, 99(4), 487-504.

Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). *The influence of avatars on online consumer shopping behavior.* Journal of marketing, 70(4), 19-36.

Hudlicka, E. (2008, August). *Affective computing for game design.* In Proceedings of the 4th Intl. North American Conference on Intelligent Games and Simulation (pp. 5-12). McGill University Montreal.

Bronnen

Kershaw, M. (2024b, februari 8). *The New Face of E-Commerce Storefronts: Digital Human Avatars*. Total Retail. <https://www.mytotalretail.com/article/the-new-face-of-e-commerce-storefronts-digital-human-avatars/>

Kim, S. S., Kim, J., Badu-Baiden, F., Giroux, M., & Choi, Y. (2021). *Preference for robot service or human service in hotels? Impacts of the COVID-19 pandemic*. International Journal of Hospitality Management, 93, 102795.

Krämer, N. C., Tietz, B., & Bente, G. (2011). *Effects of embodied conversational agent's warmth and competence on user perception*. Interacting with Computers, 23(3), 317-323.

McKinsey & Company. (2024, October 30). *What is retrieval-augmented generation (RAG)?* McKinsey & Company. <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-retrieval-augmented-generation-rag>.

McLean, G., & Osei-Frimpong, K. (2019). *Hey Alexa... examine the variables influencing the use of artificial intelligent in-home voice assistants*. Computers in Human Behavior, 99, 28–37.

Microsoft. (n.d.). *Text split skill example – strategies for chunking documents*. Microsoft Learn. <https://learn.microsoft.com/en-us/azure/search/vector-search-how-to-chunk-documents>

Nass, C., & Moon, Y. (2000). *Machines and mindlessness: Social responses to computers*. Journal of Social Issues, 56(1), 81–103.

OneBonsai. (2025). Digital assistant. <https://onebonsai.com/digital-assistant>

Oracle. (2020, 7 oktober). *What Is a Digital Assistant?* <https://www.oracle.com/chatbots/what-is-a-digital-assistant/>

Pan, J. J., Wang, J., & Li, G. (2023, 21 oktober). *Survey of Vector Database Management Systems*. arXiv.org. <https://arxiv.org/abs/2310.14021>

Ramlochan, S. (2024, 8 maart). *System prompts in large language models*. Prompt Engineering Institute. <https://promptengineering.org/system-prompts-in-large-language-models/>

Reich, I. (2024, July 20). *Generative AI digital assistants in the enterprise – design guide*. <https://itzikr.wordpress.com/2024/07/21/generative-ai-digital-assistants-in-the-enterprise-design-guide/>

Retailinnovatie PXL. (z.d.). *Events – Digitale transformatie fysieke winkel*. Hogeschool PXL. <https://retailinnovatie.pxl.be/digitale-transformatie-fysieke-winkel/events/>

Roozen, I., Raedts, M., & Yanycheva, A. (2023). *Are retail customers ready for service robot assistants?* International Journal of Social Robotics, 15(1), 15-25.

Bronnen

Schuler, M. (2025, June 18). *How the EU AI Act supplements GDPR in the protection of personal data*. International Trademark Association. <https://www.inta.org/perspectives/features/how-the-eu-ai-act-supplements-gdpr-in-the-protection-of-personal-data/>

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). *Customer service chatbots: Anthropomorphism and adoption*. Journal of Business Research, 115, 14-24.

Sirius Insight. (2021, 15 januari). *Meer of minder fysieke verkooppunten? - SIRIUS Insight*. SIRIUS Insight. <https://siriusinsight.be/article-blog/meer-of-minder-verkooppunten/?lang=nl>

Slob, G. (2024, 17 januari). *Leegstand België 2024 - locatus*. Locatus. <https://locatus.com/blog/leegstand-belgie-2024>

Stuurgroep ESF Opleidingen van de toekomst: retail. 19 december 2023

Techstrong.ai. (2024, October 15). *AI-powered interactive kiosks: Transforming customer engagement in retail*. <https://techstrong.ai/retail/ai-powered-interactive-kiosks-transforming-customer-engagement-in-retail/>

Van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., & Petersen, J. A. (2017). *Domo arigato Mr. Roboto: Emergence of automated social presence in organizational frontlines and customers' service experiences*. Journal of service research, 20(1), 43-58.

VDAB. (2024). *Knelpuntberoepen in Vlaanderen 2022*. <https://www.vdab.be/sites/default/files/media/files/Knelpuntberoepen2024.pdf>

VLAIO. (z.d.). *Waarom met artificiële intelligentie aan de slag?* <https://www.vlaio.be/nl/begeleiding-advies/digitalisering/artificiele-intelligentie/waarom-met-artificiele-intelligentie-aan-de-slag>

Databricks. (z.d.). *What are Large Language Models (LLMs)?* Databricks. <https://www.databricks.com/glossary/large-language-models-llm>

Wijaya, C. Y. (2024, April 22). *Semantic search with vector databases*. KDnuggets. <https://www.kdnuggets.com/semantic-search-with-vector-databases>

Zilliz. (2024, August 13). *What are rerankers and how they enhance information retrieval*. <https://zilliz.com/learn/what-are-rerankers-enhance-information-retrieval>

Over ons

Aan Hogeschool PXL doen we praktijkgericht onderzoek naar innovatieve toepassingen van AI-technologie. Eén van onze concrete realisaties is een volledig werkende implementatie van een AI-assistent genaamd PiXie, ontworpen en ontwikkeld door ons eigen onderzoeksteam. PiXie, gebaseerd op de nieuwste inzichten rond Retrieval-Augmented Generation (RAG) en Large Language Models (LLM's), kan klanten ondersteunen bij vragen in fysieke winkelomgevingen en levert informatie uit diverse bronnen op een toegankelijke en gebruiksvriendelijke manier.

Het bouwen van PiXie vindt plaats binnen het kader van het **TETRA-project "AI-Driven Retail Excellence: Optimalisering van klantenervaring en efficiëntie in de fysieke winkel"**. Dit project is een interdisciplinaire samenwerking tussen drie expertisecentra van Hogeschool PXL:

- **Business Research**
- **Smart ICT**
- **Centrum Digitaal Leren**

Onze AI-assistent PiXie combineert kracht en flexibiliteit. Door gebruik te maken van geavanceerde technieken zoals meertalige embedding modellen, tool calling, en integratie met API's en databases, kan PiXie efficiënt schakelen tussen statische documentatie en dynamische, real-time data. Bovendien besteden we bijzondere aandacht aan guardrailing en veiligheid, waardoor onze AI-assistent niet alleen accuraat en snel, maar ook veilig en betrouwbaar is.

Onderzoek aan Hogeschool PXL betekent altijd samenwerking met en voor het werkveld. Onze expertise ligt niet alleen in de technische realisatie, maar ook in het begeleiden van bedrijven bij de implementatie en integratie van AI-oplossingen in hun eigen processen. Door actief mee te denken over infrastructuur, regelgeving en gebruikerservaring, zorgen we ervoor dat AI-technologie concreet toepasbaar wordt en waarde toevoegt binnen elke specifieke bedrijfscontext.

Heb je interesse of wil je meer weten over hoe PiXie en onze AI-oplossingen jouw organisatie kunnen versterken? Neem dan zeker contact met ons op!





All rights reserved. No part of this publication may be reproduced, stored in an automated database, or disclosed in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.



pixie.pxl.be
robin.schrijvers@pxl.be



centrumdigitaalleren.be
hanne.westhovens@pxl.be



retailinnovatie.pxl.be
hermien.raedts@pxl.be