AI GATEWAY

FROM CHAOS TO CLARITY

MAY 2026

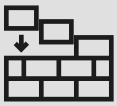
VOORWOORD

Organisaties worden overspoeld door AI-aanbod: tientallen tools, meerdere providers en een groeiende druk om sneller te innoveren. Zonder centrale sturing leidt dit tot datarisico's, oplopende kosten en verlies van controle. Een AI-gateway biedt organisaties een manier om hier grip op te krijgen: een centrale laag die AI-verkeer beheert, beveiligt en inzichtelijk maakt; zonder innovatie onnodig te remmen.

Dit whitepaper helpt technische leiders en beslissers, zoals solution architects, product owners, CDO's, CTO's en CIO's, bij het maken van de belangrijkste architectuurkeuzes rondom de inzet van een AI-gateway. Het behandelt wat een AI-gateway is, hoe deze werkt, welke dilemma's organisaties tegenkomen en hoe je van een eerste gateway groeit naar een volwassen AI-platform.

De drie belangrijkste inzichten uit dit whitepaper

EEN AI-GATEWAY IS GEEN LUXE MAAR EEN FUNDAMENT



Zonder centrale governance groeien risico's rondom kosten, data en compliance sneller dan de baten.

DE BELANGRIJKSTE DESIGN KEUZES ZIJN ERG ORGANISATIE-SPECIFIEK

Denk aan modelkeuze, governance, monitoring en compliance. De juiste keuze hangt af van context, risicoacceptatie en organisatiedoelen.

DE AI-GATEWAY IS HET STARTPUNT VAN EEN BREDERE TRANSFORMATIE

Door een solide basis te leggen en te ontwerpen voor flexibiliteit, bouw je gericht toe naar een AI-native organisatie.

01 PROBLEEM EN AMBITIE

INTRODUCTIE

In steeds meer organisaties is AI geen experiment meer, maar dagelijkse praktijk. De gemiddelde organisatie gebruikt inmiddels meer dan zeven verschillende AI-tools, en dat aantal verdubbelt momenteel elk jaar[01]. Teams en afdelingen schaffen zelfstandig oplossingen aan, experimenteren buiten het zicht van IT en compliance, en bouwen zo ongemerkt een groeiend schaduw-AI landschap op. Dat roept urgente vragen op: waar zitten mijn kosten? Is mijn bedrijfsdata veilig? Voldoe ik aan de AVG, NIS2 en de EU AI Act?

Zoals de voorgaande vragen al doen laten blijken is schaduw-AI niet alleen een technisch probleem, maar vooral het is ook een groot governance vraagstuk. Zonder centraal inzicht en sturing ontstaan risico's zoals datalekken richting externe AI-aanbieders, onverwachte kosten door ongecontroleerd tokengebruik, compliance-overtredingen en een wildgroei aan moeilijk beheerbare AI-integraties. Organisaties die dit niet snel aanpakken, raken controle kwijt over data, kosten en lopen grote risico's. Tegelijkertijd biedt AI een enorme kans: snellere productontwikkeling, hogere productiviteit en nieuwe vormen van dienstverleningsmodellen. De druk om mee te bewegen is groot (er heerst FOMO), en organisaties die AI snel én verantwoord weten toe passen, kunnen een duidelijke voorsprong opbouwen.

Om deze spanning tussen risico en innovatie te managen, kiezen steeds meer organisaties voor een AI-gateway: een centrale laag die al het AI-verkeer binnen een organisatie beheert, beveiligt en inzichtelijk maakt. Dit vormt de basis voor grip op kosten, dataveiligheid en compliance, terwijl teams toch kunnen blijven experimenteren en innoveren. Een sprekend voorbeeld hiervan is de Enexis case[02], waarin een AI-gateway is ingezet om zowel innovatie als compliance te waarborgen binnen een complexe organisatieomgeving. De centrale architectuurvraag die dit whitepaper beantwoord is:

"Hoe kan een AI-gateway worden ontworpen die naadloos integreert met het bestaande architectuurlandschap, de organisatie veilig en compliant houdt, en tegelijkertijd op verantwoorde wijze innovatie versnelt?"

01 https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
02 https://www.teamrockstars.nl/client-stories/versnellen-met-visie-hoe-enexis-een-veilig-en-schaalbaar-ai-platform-bouwde-in-recordtijd/

Een AI-gateway brengt structuur aan door te fungeren als centraal ontsluitings- en beheerpunt — al het AI-verkeer loopt via één gecontroleerde laag.

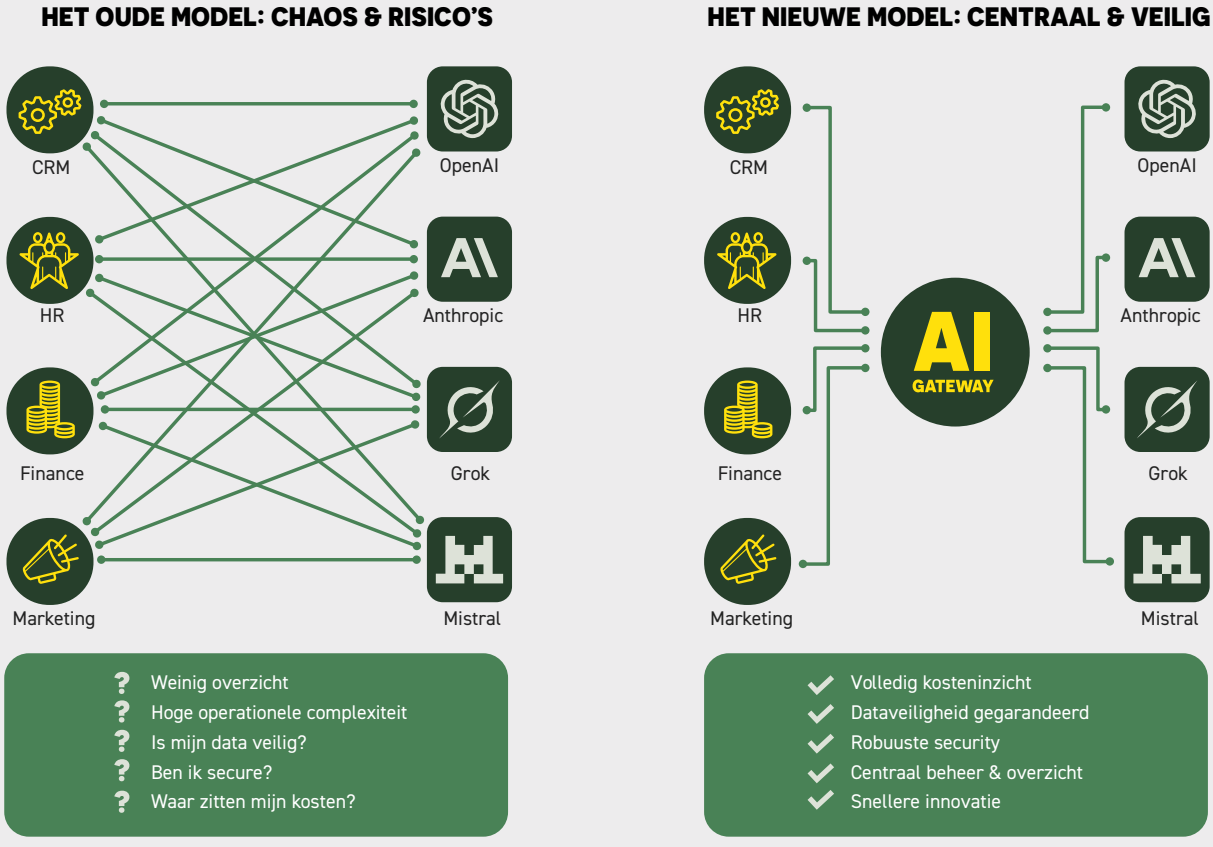
02 WAT IS EEN AI-GATEWAY EN WAAROM GEBRUIK JE DEZE?

In veel organisaties ontstaat het AI-landschap niet vanuit een centraal ontwerp, maar groeit het organisch. Een AI-gateway brengt structuur aan in de tools en modellen die op verschillende plekken in de organisatie gebruikt worden door te fungeren als centraal ontsluitings- en beheerpunt: al het AI-verkeer loopt via één gecontroleerde laag. Conceptueel lijkt een AI-gateway op een API management-laag, maar deze is specifiek ontworpen voor de unieke eigenschappen van AI-verkeer. Denk aan grote, ongestructureerde payloads (zoals prompts en completions), sterk variërende tokenkosten en de noodzaak om inhoud vooraf te controleren op veiligheid, compliance en mogelijk tal van andere guardrails; zowel vóórdat data naar een model wordt gestuurd als voordat een antwoord wordt teruggegeven.

Door al het AI-verkeer via één centrale laag te laten lopen, ontstaat realtime inzicht in gebruik, kosten en prestaties van AI. Dit maakt gerichte optimalisatie en consistent beveiligingsbeleid mogelijk. Tegelijkertijd wordt het eenvoudiger om nieuwe modellen te beheren en sneller te adopteren (zie afbeelding 1).

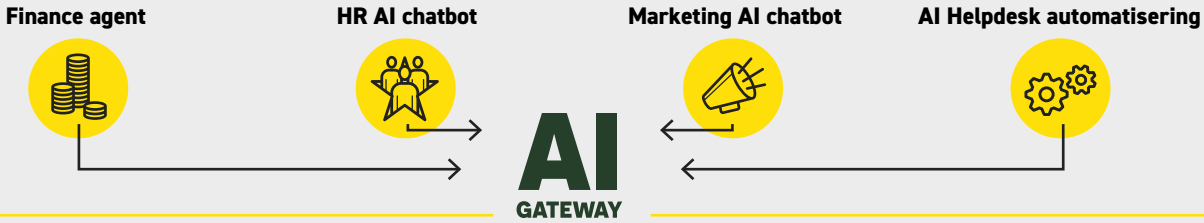
Realtime inzicht in gebruik, kosten en prestaties maakt gerichte optimalisatie en consistent beveiligingsbeleid mogelijk.

Hoewel een AI-gateway een initiële investering vraagt, levert deze vaak kostenbesparing op door efficiënter modelgebruik, beter quotabeheer en technieken zoals caching. De AI-gateway fungeert daarmee niet alleen als controlemechanisme, maar ook als enabler voor schaalbare AI-innovatie binnen de organisatie.



Afb. 1: Van chaos naar controle — de AI-gateway als centraal beheerpunt

03 HOE WERKT EEN AI-GATEWAY



Een AI-gateway vormt de centrale schakel tussen applicaties en AI-modellen. In plaats van directe integraties met verschillende modelproviders, loopt al het AI-verkeer via één gecontroleerd toegangspunt. Hierdoor kunnen beleid, beveiliging en monitoring consistent worden toegepast op alle AI-interacties binnen de organisatie. Conceptueel lijkt een AI-gateway op een klassieke API-gateway, maar de aard van AI-verkeer stelt andere eisen. Prompts en responses zijn vaak groot, ongestructureerd en inhoudelijk relevant. Kosten variëren sterk per verzoek en risico's zitten niet alleen in techniek, maar ook in de inhoud die wordt verstuurd en gegenereerd. De AI-gateway is specifiek ontworpen om met deze eigenschappen om te gaan (zie afbeelding 2).

Om deze uitdagingen te adresseren, biedt een AI-gateway een set aan gerichte functionaliteiten. Deze zorgen ervoor dat AI-verkeer niet alleen technisch wordt afgehandeld, maar ook beheersbaar, veilig en kosten efficiënt blijft. De belangrijkste functies worden hieronder toegelicht.

Functionaliteiten van een AI Gateway

CENTRALE TOEGANG EN GOVERNANCE

Doordat alle AI-verzoeken via één centraal punt verlopen, ontstaat een logisch moment om governance af te dwingen. Authenticatie en autorisatie zorgen ervoor dat alleen geautoriseerde gebruikers, applicaties of agents toegang krijgen tot AI-modellen. Met rollen, API-keys en identity-integraties kan onderscheid gemaakt worden tussen teams, omgevingen en use cases.

Daarnaast maakt de gateway het mogelijk om beleid af te dwingen over het gebruik van AI. Denk aan policies die bepalen welke tools agents mogen aanroepen, welke acties expliciet verboden zijn en hoe lang sessies mogen bestaan. Dit voorkomt dat agentic AI zich ontwikkelt tot een onbeheersbare black box in productieomgevingen.

VEILIGHEID, COMPLIANCE EN DATABEHEERSING

Een kernfunctie van de AI-gateway is het beschermen van data en het beperken van risico's. Omdat de gateway zowel prompts als responses verwerkt, kan inhoud worden gecontroleerd vóórdat deze een model bereikt en voordat output bij de gebruiker terugkomt.

Content safety-mechanismen filteren schadelijke of ongepaste inhoud en voorkomen dat deze wordt verwerkt of terug geleverd. Daarnaast kan de gateway automatisch persoonsgegevens (PII) detecteren en maskeren, zodat gevoelige data niet onbedoeld bij externe modelproviders terecht komt.
Alle interacties kunnen worden vastgelegd in auditlogs. Deze maken inzichtelijk welk model is gebruikt, met welke input en welke output. Dit is essentieel voor audits, incidentanalyse en compliance met regelgeving zoals de AVG, NIS2 en de EU AI act.

FLEXIBILITEIT, KWALITEIT EN CONTINUÏTEIT

Doordat de gateway fungeert als abstractielaag boven AI-modellen, ontstaat flexibiliteit in modelkeuze. Verzoeken kunnen dynamisch worden gerouteerd op basis van kosten, prestaties, beschikbaarheid of functionele geschiktheid. Dit maakt A/B testing mogelijk en verkleint de afhankelijkheid van één provider.

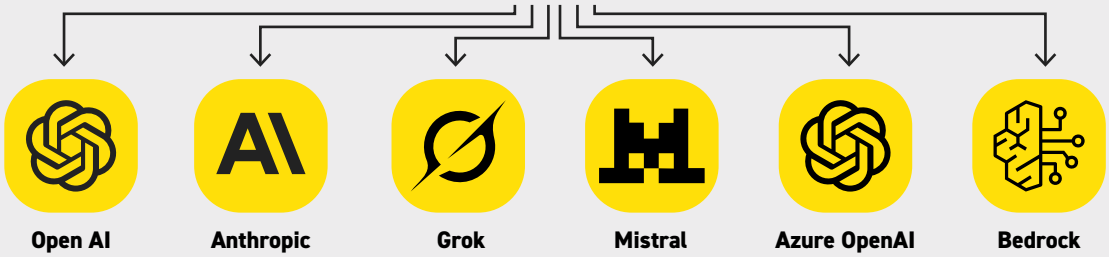
Voor productieomgevingen is continuïteit essentieel. De AI-gateway kan fallback en failover-mechanismen bevatten. Bij uitval van een model of provider wordt automatisch overgeschakeld naar een alternatief, waardoor continuïteit gewaarborgd blijft. Hierdoor blijft AI-functionaliteit robuuster en beter bestand tegen incidenten bij individuele leveranciers.

Tot slot zorgt observability voor inzicht in gebruik, prestaties, kosten en kwaliteit. Deze inzichten vormen de basis om AI-toepassingen continu te verbeteren en governancebeleid gericht bij te sturen.

KOSTENBEHEERSING EN OPERATIONELE CONTROLE

AI-gebruik brengt variabele kosten met zich mee, vaak gebaseerd op tokengebruik en modelkeuze. Zonder centrale sturing worden deze kosten pas achteraf zichtbaar.
Een AI-gateway maakt kosten inzichtelijk en beheersbaar door gebruik te meten per applicatie, team of gebruiker. Met quota, budgetten en alerts kan vooraf worden gestuurd in plaats van achteraf bijgestuurd.

Daarnaast kan semantic caching herhaalde of vergelijkbare vragen opvangen. Dit verlaagt kosten en verbetert tegelijkertijd de responsetijd. Rate limiting en throttling beschermen modellen tegen overbelasting en misbruik, en zorgen voor voorspelbare prestaties bij bedrijfskritische toepassingen.



CENTRALE TOEGANG & GOVERNANCE

- Authenticatie & autorisatie
- Agent policies

VEILIGHEID, COMPLIANCE & DATABEHEERSING

- Content safety
- PII detectie
- Compliance logging

KOSTENBEHEERSING & OPERATIONELE CONTROLE

- Token beheer & quota
- Semantic caching
- Rate limiting

FLEXIBILITEIT, KWALITEIT & CONTINUITEIT

- Multi-model routing
- Fallover & resilience
- A/B testing & observability

Afb. 2: De functionaliteiten van een AI Gateway

04 KEUZES: DE BELANGRIJKSTE ARCHITECTUURDILEMMA'S

De inrichting van een AI-gateway is geen one-size-fits-all beslissing. De keuzes die je maakt, bepalen in grote mate de wendbaarheid, veiligheid en kosten van je AI-landschap.

Elke organisatie heeft andere eisen op het gebied van compliance, performance en innovatie. Daarom bestaat er geen standaardoplossing. In dit hoofdstuk bespreken we de belangrijkste architectuurdilemma's, inclusief de bijbehorende trade-offs, zodat je bewuste keuzes kunt maken die passen bij jouw context.

4.1 WAAR HAAL IK MIJN AI-MODELLEN (LLM'S) VANDAAN?

De keuze voor LLM's en modelproviders is voor veel organisaties een strategisch vraagstuk. Het gaat niet alleen om kwaliteit en kosten, maar ook om compliance, data residency en afhankelijkheid van leveranciers.

Een belangrijk architectuurdilemma ligt tussen het gebruik van hyperscalers en het inzetten self-hosted open source modellen. Hyperscalers bieden snelheid, schaalbaarheid en geïntegreerde cloudfunctionaliteit. Met AI-modellen met een pay-as-you-go constructie kunnen organisaties snel opschalen zonder zware operationele last. Dit ontzorgt interne teams, maar creëert ook afhankelijkheid van één of meerdere leveranciers. Het is dan ook aan te raden om pas te kijken naar vooraf inkopen van gereserveerde capaciteit (bijvoorbeeld via provisioned throughput units ofwel PTU's) zodra voor langere tijd ingezet wordt op een specifieke provider, de meeste organisaties doen dit meestal zodra ze een goede baseline hebben wat betreft gebruik. Daarnaast kan gereserveerde capaciteit ook aantrekkelijk zijn wanneer er een specifieke eis voor performance of snelheid geldt, aangezien de meeste pay-as-you-go-modellen geen SLA's bieden.

Open source oplossingen bieden juist maximale controle en flexibiliteit. Organisaties kunnen modellen zelf hosten en aanpassen, zonder vendor lock-in. Daar staat tegenover dat dit vraagt om meer technische expertise, infrastructuur en doorlopend beheer en onderhoud. Zonder centrale regie leidt het direct integreren met verschillende model-API's per team of applicatie al snel tot versnippering en verlies van overzicht. Een AI-gateway biedt hier een oplossing door deze integraties te centraliseren. In plaats van directe koppelingen met individuele providers, worden modelaanroepen via de gateway gerouteerd. Hierdoor kan centraal worden bepaald welke modellen zijn toegestaan, onder welke voorwaarden ze gebruikt mogen worden en hoe het gebruik wordt gemonitord. Dit zorgt voor grip op compliance, kosten en performance, terwijl de flexibiliteit behouden blijft om meerdere modellen naast elkaar te gebruiken en per use case de beste keuze te maken.

DILEMMA

KIES JE VOOR GEMAK EN SNELHEID, OF VOOR EIGENAARSCHAP EN FLEXIBILITEIT?

Bij een klant van Team Rockstars IT in de energiesector is ervoor gekozen om vooral commerciële pay-as-you-go modellen te gebruiken, omdat de meeste use cases beperkte performance-eisen hebben en weinig verkeer ontvangen. Voor kritische use cases met eisen aan snelheid en beschikbaarheid worden commerciële modellen via PTU ingezet, zodat performancegaranties en state-of-the-art modellen beschikbaar zijn.

MATHIJS VAN BREE

4.2 COMMERCIAL GATEWAY OF OPEN SOURCE?

Vaak zijn er al één of meerdere modelproviders en integraties gekozen voordat een organisatie met een AI Gateway aan de slag gaat. Dit leidt tot een nieuw dilemma: kies je voor een commerciële AI-gateway of bouw je zelf een open source variant?

Deze keuze draait om de afweging tussen snelheid en gemak enerzijds, en controle en flexibiliteit anderzijds. Commerciële platforms zoals Kong Konnect, Portkey, Apigee en Azure API Management (met AI-extensies) bieden een snelle start en ondersteuning voor de meeste leveranciers. Ze bevatten vaak ingebouwde functionaliteit voor beveiliging, monitoring en compliance, en nemen een groot deel van het operationele beheer uit handen. Dit versnelt de time-to-value, maar gaat gepaard met licentiekosten en vendor lock-in.

Open-source gateways zoals LiteLLM, Agent Gateway, One API of HelixML bieden juist maximale controle en aanpasbaarheid. Organisaties kunnen de gateway volledig afstemmen op hun eigen architectuur en use cases, zonder al te veel afhankelijkheid van één leverancier. Daar staat tegenover dat implementatie, beheer en eventueel doorontwikkeling intern georganiseerd moeten worden, wat vraagt om voldoende expertise en capaciteit. De rol van de AI-gateway blijft in beide gevallen hetzelfde: het centraliseren van AI-verkeer, afdwingen van beleid en het bieden van inzicht. De keuze bepaalt vooral waar de complexiteit ligt; bij de leverancier of binnen de organisatie; en hoeveel verantwoordelijkheid je zelf wilt dragen.

Om risico's in de supply-chain te beheersen, is het essentieel om de betrouwbaarheid van de leverancier of open-source community te beoordelen. Let daarbij op de frequentie van updates, transparantie over kwetsbaarheden en de beschikbaarheid van support. Bij commerciële oplossingen kunnen security- en compliance-eisen contractueel worden vastgelegd. Bij open-source alternatieven is het belangrijk om zelf verantwoordelijkheid te nemen, bijvoorbeeld door actieve betrokkenheid bij de community, het uitvoeren van security-audits en het inrichten van processen voor snelle patching. Incidenten zoals de supply-chain aanval[03] op o.a. Trivy en LiteLLM laten zien dat zelfs securitytools zelf een aanvalsvector kunnen worden en dat continue due diligence noodzakelijk is.

DILEMMA

KIES JE VOOR HET COMFORT VAN EEN LEVERANCIER BIJ DE KEUZE VAN EEN AI-GATEWAY, OF GA JE VOOR ZELFSTANDIG EIGENAARSCHAP, MET MEER VRIJHEID MAAR OOK MEER VERANTWOORDELIJKHEID?

> **"Bij deze specifieke klant van Team Rockstars IT staat veiligheid op 1. Comfort van een leverancier is verleidelijk, maar alleen zelfstandig eigenaarschap geeft ons de transparantie en controle die wij nodig hebben over ons LLM-verkeer."**
>
> **ROBBIE KOUWENBERG**

4.3 HOE GEEF JE TEAMS TOEGANG TOT DE GATEWAY?

Wanneer een organisatie een AI-gateway in gebruik neemt, is het essentieel om teams op een veilige en gecontroleerde manier toegang te geven. Dit gebeurt via identiteiten, rollen en beleid, waarmee wordt bepaald wie welke modellen en functionaliteiten mag gebruiken

.

Een belangrijk architectuurdilemma hierbij is de keuze voor het governance-model. In een centraal model voert het platformteam de regie en worden standaarden uniform afgedwongen. Dit zorgt voor consistente beveiliging en een betrouwbare audit trial. Het alternatief is een meer self-service georiënteerd model, bijvoorbeeld door een Internal Developer Portal zoals Backstage. Teams krijgen hierin meer autonomie om zelf AI-toepassingen te ontwikkelen en configureren. Dit versnelt innovatie en vermindert afhankelijkheid van een centraal team, maar vraagt om duidelijke kaders en voldoende volwassenheid binnen teams.

Beide benaderingen hebben risico's. Te veel centrale controle kan innovatie vertragen en ertoe leiden dat teams buiten de gateway om oplossingen bouwen. Te weinig sturing kan juist resulteren in ongecontroleerde kosten, versnippering en risico's rondom datagebruik. In beide gevallen is het aan te raden om ook op het gebied van je cloud infrastructuur te werken met policies om te garanderen dat Teams geen shadow AI gaan afnemen buiten de AI Gateway om, bijvoorbeeld met behulp van policies om het aanmaken van AI-resources te beperken. Daarnaast is een goede samenwerking met je inkoopafdeling aan te raden om inkoop van ongewenste schaduw-AI af te buigen naar je eigen AI Gateway.

De uitdaging ligt in het vinden van de juiste balans. In de praktijk werkt een gelaagde aanpak vaak het beste: centrale regels voor security, compliance en kosten, gecombineerd met ruimte voor teams om binnen die kaders zelfstandig keuzes te maken.

De gateway speelt hierin een sleutelrol. Doordat alle AI-interacties via deze laag verlopen, ontstaat één centrale plek waar gebruik, modellen en gedrag worden geregistreerd en gemonitord. Daarmee fungeert de gateway niet alleen als toegangspunt, maar ook als fundament voor een breder AI-governance- en registrymodel.

DILEMMA

CENTRALISEER JE DE TOEGANG TOT DE AI GATEWAY VOOR MAXIMALE CONTROLE, OF GEEF JE TEAMS SELF SERVICE OM INNOVATIE TE VERSNELLEN?

[03] https://www.kaspersky.com/blog/critical-supply-chain-attack-trivy-litellm-checkmarx-teampcp/55510/

> Volledige controle verkleint het risico op datalekken, maar kan de adoptie van AI remmen. Meer vrijheid versnelt innovatie, maar brengt grotere risico's met zich mee.

4.4 HOE GA JE OM MET PII EN GEVOELIGE DATA?

AI-gateways spelen een cruciale rol in het beschermen van gevoelige data. Doordat alle prompts en responses via deze centrale laag lopen, kan data worden geanalyseerd, verrijkt en aangepast voordat deze naar een extern model wordt gestuurd of terugkomt bij de gebruiker. Typische functionaliteiten zijn automatische detectie van persoonsgegevens (PII), masking of pseudonimisering, en het blokkeren van gevoelige input. Hiermee wordt gecontroleerd AI-gebruik mogelijk binnen de kaders van de AVG en andere regelgeving.
Het belangrijkste dilemma hierbij is de balans tussen compliance en bruikbaarheid. Enerzijds wil je risico's minimaliseren door gevoelige data strikt te controleren of blokkeren. Anderzijds kan te veel restrictie het gebruik frustreren en innovatie vertragen.

Organisaties staan daarmee voor een belangrijke keuze: monitor je datagebruik en geef je teams ruimte om flexibel te werken, of grijp je actief in door requests met gevoelige informatie te blokkeren? Volledige controle verkleint het risico op datalekken, maar kan de adoptie van AI remmen. Meer vrijheid versnelt innovatie, maar brengt grotere risico's met zich mee.

In de praktijk werkt een gefaseerde aanpak vaak het beste. Start met monitoren en signaleren om inzicht te krijgen in datastromen en gebruikspatronen. Op basis van deze inzichten kan gericht beleid worden opgesteld en waar nodig worden aangescherpt met actieve maatregelen zoals masking of blocking. Zo blijft het beleid zowel veilig als werkbaar.

DILEMMA

BLOKKEER JE GEVOELIGE DATA VROEGTIJDIG VOOR MAXIMALE COMPLIANCE EN VEILIGHEID, OF MONITOR JE EERST OM TEAMS SNELHEID EN FLEXIBILITEIT TE GEVEN, MET EEN GROTER RISICO OP DATALEKKEN?

4.5 HOE EN WAT WIL JE MONITOREN?

AI-gateways bieden organisaties inzicht in het gebruik, de prestaties en de kwaliteit van AI-interacties. Dit maakt het mogelijk om gericht te sturen op kosten, performance en betrouwbaarheid.

Een belangrijke ontwerpkeuze is het niveau waarop je monitort. Richt je je op individuele gebruikers en applicaties voor maximale accountability, of kies je voor geaggregeerde inzichten om privacy te waarborgen? Daarnaast speelt de vraag of je inzet op real-time signalering van afwijkingen, of vooral op passieve monitoring.

Te weinig inzicht belemmert effectieve sturing en vertraagt incidentrespons. Te veel detail leidt tot privacyrisico's en een hogere operationele last.

Het onderliggende dilemma draait om de balans tussen observability en privacy, zonder dat de complexiteit en beheerlast te groot worden. Te weinig inzicht belemmert effectieve sturing en vertraagt incidentrespons. Tegelijkertijd kan te gedetailleerde monitoring leiden tot privacyrisico's en een hogere operationele last. Regelgeving zoals de EU AI Act stelt bovendien aanvullende eisen aan logging en transparantie.

Een pragmatische aanpak is om vooraf duidelijke keuzes te maken in wat je meet en bewaart. Denk aan het definieren van relevante KPI's, het inrichten van real-time alerts op kritieke afwijkingen en het vastleggen van een helder retentiebeleid voor logs. Zo ontstaat voldoende inzicht om te sturen, zonder onnodige data te verzamelen.

Tot slot verschilt de mate van ondersteuning per platform. Commerciële oplossingen bieden vaak out-of-the-box functionaliteit voor monitoring en compliance, terwijl open-source alternatieven meer flexibiliteit bieden, maar ook vragen om eigen inrichting en beheer.

DILEMMA

GA JE VOOR FIJNMAZIGE MONITORING OM TE KUNNEN STUREN EN VERANTWOORDEN, OF VOOR BEPERKT INZICHT OM PRIVACY EN EENVOUD TE BORGEN?

4.6 HOE ZORG JE VOOR CONTINUE BESCHIKBAARHEID?

AI-gateways maken het mogelijk om fallback- en multi-providerstrategieën te implementeren, waardoor de continuiteit van AI-diensten beter kan worden gewaarborgd. Naarmate organisaties afhankelijker worden van AI, neemt het belang van beschikbaarheid toe. Uitval van een model of provider kan directe impact hebben op bedrijfsprocessen.

Door gebruik te maken van automatische failover en provider-onafhankelijkheid kan de gateway verzoeken dynamisch routeren naar alternatieve modellen of leveranciers. Dit verhoogt de betrouwbaarheid en verkleint de afhankelijkheid van één partij.

Het belangrijkste dilemma hierbij is de balans tussen maximale zekerheid en een werkbare, betaalbare oplossing. Meer redundantie verhoogt de beschikbaarheid, maar brengt ook extra complexiteit, beheerlast en kosten met zich mee. Als een fallback model niet enkel ingericht is als failover maar als graceful degradation presteert het fallback-model wellicht niet op hetzelfde niveau als het primaire model, wat impact kan hebben op de kwaliteit van de output.

De juiste aanpak begint bij het bepalen hoe bedrijfkritisch de AI-toepassing is. Niet elke use case vraagt om dezelfde mate van resilience. Voor bedrijfskritische processen kan een hogere mate van redundantie en strengere SLA's noodzakelijk zijn, terwijl voor minder kritische toepassingen een eenvoudiger opzet volstaat.

In de praktijk betekent dit dat organisaties per use case expliciet bepalen wat een acceptabele downtime is en welk serviceniveau nodig is. Op basis daarvan kan de architectuur van de AI-gateway worden ingericht, inclusief fallback mechanismen, failoverstrategieën en testprocedures. Zo wordt resilience afgestemd op risico en impact, in plaats van standaard gemaximaliseerd.

DILEMMA

GA JE VOOR FIJNMAZIGE MONITORING OM TE KUNNEN STUREN EN VERANTWOORDEN, OF VOOR BEPERKT INZICHT OM PRIVACY EN EENVOUD TE BORGEN?

4.7 HOE HOUD JE KOSTEN ONDER CONTROLE?

AI-gateways bieden organisaties inzicht in het gebruik en de kosten van AI. Door gebruik te meten per team, applicatie en model ontstaat transparantie, waardoor gerichte sturing mogelijk wordt. Functionaliteiten zoals quota's, semantische caching en slimme modelselectie helpen om kosten actief te beheersen.

Caching kan hierbij aanzienlijke besparingen opleveren door herhaalde of vergelijkbare requests op te vangen. Daarnaast maakt monitoring inzichtelijk welke use cases de meeste kosten veroorzaken en waar optimalisatie mogelijk is, bijvoorbeeld door alternatieve modellen in te zetten.

Het belangrijkste dilemma ligt in de balans tussen kostenoptimalisatie en het waarborgen van performance en kwaliteit. Strikte quota of een te sterke focus op kosten kunnen de snelheid en kwaliteit van AI-toepassingen negatief beïnvloeden. Tegelijkertijd leidt te weinig sturing al snel tot onverwachte kosten en gebrek aan controle.

Een pragmatische aanpak is om gefaseerd te werk te gaan. Start met het creëren van inzicht in gebruik en kosten. Denk aan het creëren van een overzicht van de verschillende AI-modellen per kostencategorie (low, medium, high, ultra), en het rapporteren van gebruik en kosten aan de verschillende teams. Stel vervolgens budgetten en quota vast per team of use case, en pas caching toe als eerste optimalisatiemechanisme. Op basis van deze inzichten kan verder worden gestuurd op modelkeuze en gebruikspatronen. Ook is het belangrijk de tokenkosten goed in de gaten te houden en actief met teams in gesprek te blijven bij wijzigingen in prijsstelling van de leverancier. Voor organisaties welke AI assisted development of Agentic Coding practices hebben kan het zelfs interessant zijn om de AI-gateway te gebruiken om een stukje kosten mitigatie te doen daar waar developers buiten hun bestaande quota's van bijvoorbeeld Claude Code of Github Copilot raken.

Zo blijft het beleid zowel kostenefficiënt als werkbaar, en behoudt de organisatie grip op zowel uitgaven als kwaliteit.

DILEMMA

GRIJP JE CENTRAAL IN OM AI KOSTEN ACTIEF TERUG TE DRINGEN, OF BEPERK JE JE TOT KOSTENINZICHT EN LAAT JE DE VERANTWOORDELIJKHEID BIJ TEAMS?

4.8 HOE ADOPTEER JE DE AI-GATEWAY BINNEN DE ORGANISATIE?

Je kunt technisch een uitstekende AI-gateway bouwen, maar de echte uitdaging begint daarna: hoe zorg je ervoor dat teams deze ook daadwerkelijk gaan gebruiken? Voor veel teams voelt een gateway in eerste instantie als extra regels, beperkingen en monitoring, terwijl de directe voordelen niet altijd zichtbaar zijn. Zonder draagvlak bestaat het risico dat teams alternatieve routes blijven gebruiken en de gateway omzeilen.

Organisaties staan daarbij voor een fundamenteel dilemma. Je kunt het gebruik van de AI-gateway afdwingen via beleid en technische controles, of je kiest voor een aanpak waarbij je samenwerkt met teams: hun behoeften begrijpt en de gateway zo inricht dat deze hen daadwerkelijk helpt sneller, veiliger en efficiënter te werken.

In de praktijk vraagt succesvolle adoptie om een combinatie van beide. Alleen afdwingen leidt vaak tot weerstand en schaduwoplossingen, terwijl volledige vrijheid kan resulteren in gebrek aan controle. Het is daarom essentieel om duidelijke kaders te stellen, terwijl je tegelijkertijd investeert in begeleiding, communicatie en het zichtbaar maken van de toegevoegde waarde.

Een effectieve aanpak houdt de drempel laag voor teams, bijvoorbeeld door eenvoudige onboarding, goede documentatie en herbruikbare templates. Tegelijkertijd worden niet-onderhandelbare basisregels voor security, compliance en kosten centraal geborgd.

Op die manier wordt de AI-gateway niet gezien als een rem op innovatie, maar als een platform dat teams ondersteunt en versnelt. Dit vergroot de kans op brede adoptie en zorgt ervoor dat de voordelen van centrale governance daadwerkelijk worden gerealiseerd.

DILEMMA

VERPLICHT JE HET GEBRUIK VAN DE AI GATEWAY, OF OVERTUIG JE TEAMS DOOR TE LATEN ZIEN DAT HET HEN DAADWERKELIJK HELPT SNELLER EN BETER TE WERKEN?

4.9 BESLISWIJZER: QUICK REFERENCE

Onderstaande tabel biedt een beknopt overzicht van de belangrijkste architectuurkeuzes, inclusief de centrale afweging en signalen die helpen bij het maken van een passende keuze.

	CENTRALE OVERWEGING	SIGNAAL VOOR KEUZE
Model sourcing	Hyperscaler vs. open source vs. eigen	Risicotolerantie en dataclassificatie
Gateway type	Commercieel vs. open source	Time-to-value versus behoefte aan controle
Leveringsmodel	Centraal vs. self-service	Team-volwassenheid en governancebehoefte
Securitymodel	Centrale identity provider vs. API-KEY	Security posture versus gewenste snelheid van adoptie
PII / data	Blokkeren vs. monitoren	Compliance-eisen en criticiteit aan use case
Monitoring	Detailniveau vs. privacy	Accountability-eisen, opslag en AI Act
Resilience	Maximale redundantie vs. eenvoud	SLA-vereisten en criticiteit van gebruik
Kosten	Strikte quota vs. flexibiliteit	Budget en kwaliteitseisen
Adoptie	Nadruk op top-down guardrails of bottom-up innovatie	Organisatiecultuur en gewenste snelheid van adoptie

05 TOEKOMSTIGE ONTWIKKELINGEN

5.1 VENDOR LOCK-IN ALS STRATEGISCH RISICO

Hoewel een AI-gateway organisaties meer controle geeft over hun AI-landschap, introduceert het tegelijkertijd een nieuw strategisch risico: vendor lock-in. Niet alleen de keuze voor modelproviders, maar ook de gekozen gateway-oplossing zelf kan een afhankelijkheid creëren die toekomstige flexibiliteit beperkt.

Wanneer organisaties sterk leunen op specifieke functionaliteiten of proprietary integraties van een leverancier, kan dit migraties complex en kostbaar maken. Dit geldt zowel voor commerciële platforms als voor specifieke model-API's die moeilijk uitwisselbaar zijn.

Om dit risico te beperken, kiezen steeds meer organisaties voor een architectuur gebaseerd op standaarden, interoperabele interfaces en een multi-vendor strategie. Hiermee blijft de mogelijkheid bestaan om modellen of gateways te vervangen of te combineren, zonder ingrijpende wijzigingen in het landschap.

De kern van het dilemma ligt in de afweging tussen het maximaal benutten van specifieke functionaliteiten van een leverancier en het behouden van strategische wendbaarheid op de lange termijn. Organisaties die hier bewust mee omgaan, voorkomen dat hun AI-architectuur een rem wordt op toekomstige innovatie.

5.2 INSPELEN OP DE VERANDERENDE AI-WERELD

Het AI-landschap ontwikkelt zich in hoog tempo. Nieuwe modellen, providers en capabilities volgen elkaar snel op. Een AI-gateway stelt organisaties in staat om hier flexibel op in te spelen, doordat nieuwe modellen eenvoudig kunnen worden toegevoegd zonder ingrijpende wijzigingen in applicaties.

> **Niet alleen de keuze voor modelproviders, maar ook de gateway-oplossing zelf kan een afhankelijkheid creëren die toekomstige flexibiliteit beperkt.**

Functionaliteiten zoals routing en A/B-testing maken gecontroleerde experimenten mogelijk, waardoor organisaties innovatie kunnen benutten zonder direct grote risico's te nemen.

De kernuitdaging ligt in de balans tussen stabiliteit en vernieuwing. Te snel experimenteren kan de betrouwbaarheid van productieomgevingen ondermijnen, terwijl een te afwachtende houding kan leiden tot een achterstand ten opzichte van concurrenten.

Nieuwe modellen, providers en capabilities volgen elkaar snel op — een AI-gateway stelt organisaties in staat om hier flexibel op in te spelen zonder ingrijpende wijzigingen in applicaties.

5.3 AGENT GATEWAYS EN EDGE GATEWAYS

Nieuwe architectuurpatronen, zoals agent gateways en edge gateways, winnen snel aan belang.

Agent gateways richten zich op het orkestreren van autonome AI-agents en maken het mogelijk om complexe, multi-step processen te beheren. Edge gateways brengen AI dichter bij de bron, bijvoorbeeld lokaal of op device-niveau, wat voordelen biedt op het gebied van datasoevereiniteit, latency en offline beschikbaarheid.

Deze ontwikkelingen brengen echter ook nieuwe uitdagingen met zich mee. Ze vragen om extra investeringen en verhogen de complexiteit rondom beveiliging, beheer en governance.

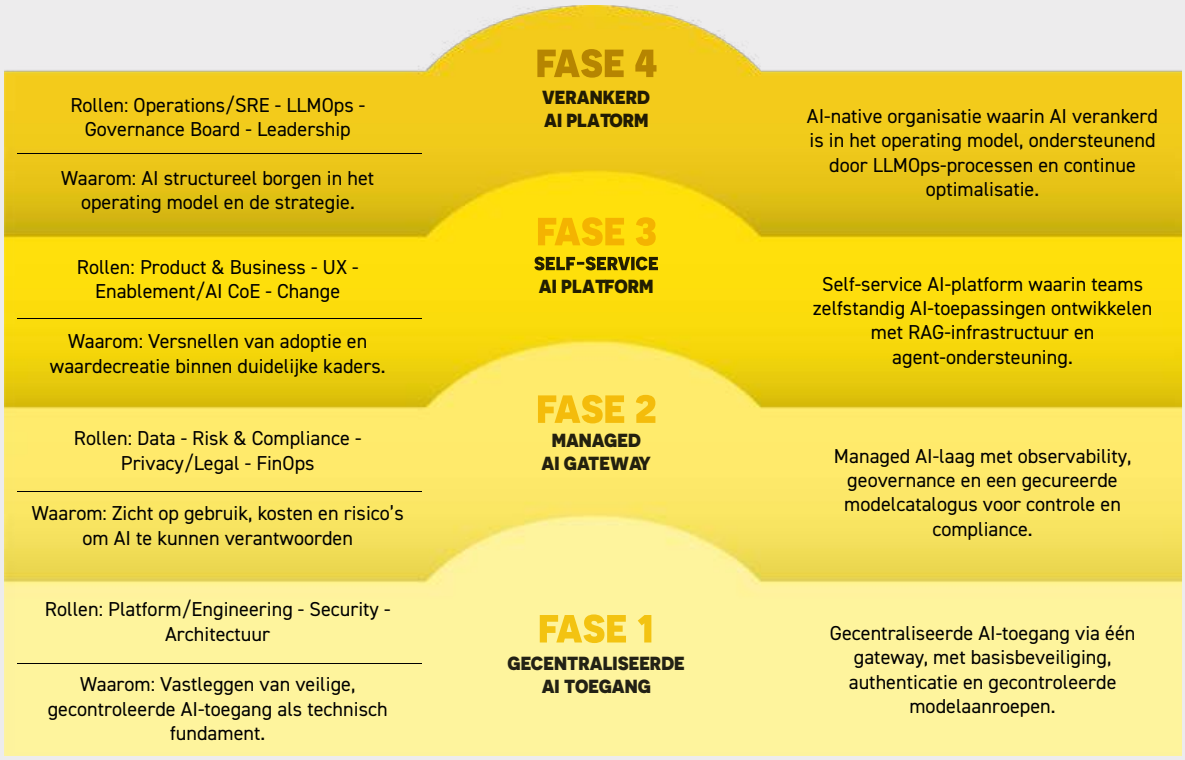
De belangrijkste afweging ligt in de timing van adoptie: te vroeg investeren in onvolwassen technologie brengt risico's met zich mee, terwijl te laat instappen kan leiden tot een strategische achterstand.

5.4 DOORGROEI NAAR EEN AI-PLATFORM

De AI-gateway vormt vaak het startpunt van een bredere ontwikkeling naar een volwassen AI-platform. Dit groeipad bestaat uit meerdere fasen, waarin organisaties stapsgewijs meer controle, functionaliteit en volwassenheid opbouwen (zie afbeelding 3).

Dit groeipad maakt incrementele adoptie mogelijk, met optimaal hergebruik van eerdere investeringen. Tegelijkertijd vraagt het om actieve sturing om te voorkomen dat complexiteit en bottlenecks ontstaan naarmate het platform groeit.

Afb. 3: AI Gateway Maturity Model — van gateway naar AI-native organisatie



Niet al het AI-gebruik binnen een organisatie loopt via zelfgebouwde toepassingen — en precies daar ontstaan de blinde vlekken.

06 WAT DOE JE MET NIET IN-HOUSE ONTWIKKELDE AI?

Niet al het AI-gebruik binnen een organisatie loopt via zelfgebouwde toepassingen. In de praktijk zijn er twee veelvoorkomende patronen waarbij de governance-uitdaging wezenlijk anders is en een AI-gateway niet vanzelfsprekend onderdeel is van de controleketen:

1. EMBEDDED AI
Dit betreft AI-functionaliteit die geïntegreerd is in SaaS-oplossingen van leveranciers, zoals Salesforce Einstein, Microsoft Copilot en ServiceNow AI. Deze toepassingen communiceren direct met AI-diensten van de leverancier en omzeilen daarmee de AI-gateway.

Governance-aanpak: richt je op contractuele afspraken rondom dataverwerking, privacy en security. Denk aan dataverwerkingsovereenkomsten, vendor risk/supply chain risk assessments en het gebruik van tooling zoals CASB voor extra zichtbaarheid.

2. BYOM (BRING-YOUR-OWN-MODEL)
Sommige platformen bieden de mogelijkheid om eigen AI-modellen te integreren. Hoewel dit meer controle geeft, kan er alsnog fragmentatie ontstaan wanneer deze integraties buiten de AI-gateway plaatsvinden.

Governance-aanpak: stel BYOM-ondersteuning als expliciete eis bij procurement en zorg dat modelverkeer via de AI-gateway blijft lopen. Dit maakt consistente logging, security en kostenbeheer mogelijk. Houd er rekening mee dat in sommige gevallen additionele kosten bovenop licenties bij komen, bijvoorbeeld voor het aansluiten van externe modellen of het verwerken van gebruik via eigen AI-endpoints.

Praktisch advies
Leg in het AI-sourcingbeleid van jouw bedrijf expliciet vast hoe wordt omgegaan met Embedded AI en BYOM. Zonder duidelijke keuzes ontstaan al snel blinde vlekken in het governance-model. Hoewel deze patronen essentieel zijn binnen AI-governance, vallen ze buiten de scope van dit whitepaper.

07 COMPLIANCE EN DE WET

Voor Nederlandse en Europese organisaties is de AI-gateway niet alleen een architectuurkeuze, maar ook een belangrijk compliance-instrument. De EU AI Act , die gefaseerd van kracht wordt, stelt eisen aan transparantie, menselijk toezicht, auditbaarheid en risicobeheersing van AI-systemen. Maar niet te vergeten de AVG én de NIS2 zijn beiden van toepassing naast de EU AI Act en stellen soms overlappende eisen. Een goed ingerichte AI-gateway helpt om hier structureel aan te voldoen.

LEES DE BELANGRIJKSTE BIJDRAGEN VAN EEN AI-GATEWAY AAN COMPLIANCE OP DE VOLGENDE PAGINA.

Organisaties die auditlogging via een AI-gateway inrichten vanaf dag 1, bouwen direct een voorsprong op.

https://digital-strategy.ec.europa.eu/nl/policies/regulatory-framework-ai [04]

DE BELANGRIJKSTE BIJDRAGEN VAN EEN AI-GATEWAY AAN COMPLIANCE ZIJN

- **Auditlogging**
  De gateway registreert alle AI-interacties, waardoor inzichtelijk wordt welke input, modellen en output zijn gebruikt. Dit is essentieel voor audits en voor hoog-risico AI-systemen onder de EU AI Act.
- **Human-in-the-loop**
  Via policies kan worden afgedwongen dat bepaalde AI-output eerst door een mens wordt beoordeeld voordat deze wordt gebruikt of doorgezet.
- **Transparantie**
  Logging maakt zichtbaar welk model, welke versie en welke parameters zijn toegepast. Dit is cruciaal voor reproduceerbaarheid en incidentanalyse.
- **Datasoevereiniteit**
  Door gebruik van edge gateways of restricties op regio's kan worden geborgd dat data binnen de EU blijft, in lijn met AVG-vereisten rondom data residency.
- **PII-bescherming**
  Automatische detectie en masking van persoonsgegevens vóór verzending naar AI-modellen helpt om gevoelige data te beschermen en compliant te blijven met privacyweting.
- **Risicobeheersing**
  Content Safety en policy-enforcement beperken de kans op schadelijke, onjuiste of discriminerende output, en ondersteunen daarmee verantwoord AI-gebruik.

Let op: Het is belangrijk om te beseffen dat de EU AI Act, met name voor hoog-risico AI-systemen, expliciete eisen stelt aan logging en traceerbaarheid. Organisaties die deze capabiliteit vanaf het begin via een AI-gateway inrichten, bouwen direct een voorsprong op. Wie dit pas later toevoegt loopt mogelijk het risico op boetes (tot €35 miljoen of 7% van de jaaromzet), of complexe en kostbare aanpassingen achteraf.

08 CONCLUSIE

De AI-revolutie is niet te stoppen, maar de complexiteit die ermee gepaard gaat, is wel beheersbaar. Een AI-gateway vormt het architectuurfundament waarmee organisaties snel en veilig kunnen innoveren, zonder de controle te verliezen over kosten, data en compliance.

Dit whitepaper laat zien dat de implementatie van een AI-gateway meer is dan een technische exercitie. Het is een strategische keuze die raakt aan governance, organisatiecultuur, leveranciersstrategie en de lange termijn ambitie om AI te verankeren als kerncompetentie van de organisatie.

DE BELANGRIJKSTE BESLISSINGEN OM MEE TE STARTEN

Bepaal je risicobereidheid en definieer hierop je governance-model: maak expliciet welke risico's je accepteert per dataclassificatie en use case (bijv. PII, bedrijfsgeheimen, kritieke processen) en vertaal dit naar heldere spelregels zoals centraal vs. self-service, verplichte guardrails (logging, human-in-the-loop, content safety), en een escalatiepad voor uitzonderingen.

Bepaal wie toegang heeft tot welke modellen en welke data nooit een extern model of geolocatie buiten de EU mag bereiken. Zonder dit fundament blijft elke technische oplossing kwetsbaar. Kies je eerste gateway bewust, maar ontwerp voor vervanging.

Maak gebruik van open standaarden, beperk afhankelijkheden van specifieke leveranciers en evalueer periodiek of de gekozen oplossing nog past bij de behoefte. Richt observability in vanaf dag 1, je kunt alleen sturen wat je meet. Logging, kostenoverzicht en kwaliteitsmonitoring zijn geen nice-to-haves, maar essentiële bouwstenen.

Organisaties die nu investeren in een solide AI-gateway leggen de basis voor een duurzaam concurrentievoordeel: snellere innovatie, beter beheersbare kosten en robuuste compliance. Tegelijkertijd behouden zij de flexibiliteit om de beste modellen te blijven gebruiken, ongeacht de provider.

[05] https://artificialintelligenceact.eu/article/99/

BENIEUWD HOE EEN AI-GATEWAY JOUW ORGANISATIE VERDER KAN BRENGEN?

Wil je sparren over het opzetten, verbeteren of implementeren van een AI-gateway en hoe je hiermee grip krijgt op innovatie, kosten en compliance? Of ben je benieuwd hoe andere organisaties dit aanpakken en welke lessen zij hebben geleerd? Neem gerust contact op.

mathijs.vanbree@teamrockstars.nl

simone.vanerp@teamrockstars.nl

Op www.teamrockstars.nl vind je meer informatie en praktijkvoorbeelden. We denken graag met je mee!

De AI-gateway is daarmee geen eindpunt, maar een startpunt. Het vormt de basis voor de ontwikkeling naar een AI-native organisatie: veilig schaalbaar en voorbereid op de volgende golf van innovatie.

OVER DE AUTEURS

MATHIJS VAN BREE

Mathijs van Bree is Principal Consultant AI bij Team Rockstars IT en hij houdt zich sinds 2017 actief bezig met AI. Hij helpt teams om AI niet alleen technisch werkend te krijgen, maar vooral om het te vertalen naar oplossingen die processen verbeteren, meetbaar renderen en verantwoord in productie kunnen draaien.

In de afgelopen jaren werkte hij aan AI-initiatieven binnen de energiesector, overheid en financiële dienstverlening, onder meer bij organisaties als Essent, Rabobank, ING en nu Jumbo. Zijn focus ligt op de vraag hoe je van pilots naar structurele waarde komt: van use case selectie en evaluatie tot data, governance, adoptie en het inrichten van herhaalbare bouwblokken om op schaal te kunnen leveren.

ROBBIE KOUWENBERG

Robbie Kouwenberg is een strategisch AI & Cloud Architect met een diepgewortelde passie voor het snijvlak tussen software architectuur en kunstmatige intelligentie. Met bijna twee decennia aan ervaring in het bouwen van complexe enterprise-systemen, heeft Robbie zich de afgelopen jaren gespecialiseerd in de architectuur van Generatieve AI-platformen.

Zijn expertise ligt in het overbruggen van de kloof tussen ambitieuze AI-concepten en veilige, schaalbare implementaties. Door technische diepgang te combineren met een pragmatische visie, helpt hij organisaties om AI niet alleen als experiment te zien, maar als een fundamenteel en betrouwbaar onderdeel van hun digitale ecosysteem.

STAN RUESSINK

Stan Ruessink is ML Engineer binnen Team Rockstars IT, gespecialiseerd in MLOps, data engineering en (Gen)AI. Met een achtergrond in data science en cloud engineering helpt hij organisaties om machine learning- en GenAI-oplossingen gestructureerd, schaalbaar en reproduceerbaar naar productie te brengen. Bij TenneT houdt hij zich de afgelopen anderhalf jaar bezig met het neerzetten en implementeren van de GenAI-propositie en bijbehorende cloudinfrastructuur. Daarbij ligt de focus op het versnellen van GenAI-projecten op schaal, onder andere via de realisatie van een centrale AI-gateway, zodat teams sneller kunnen leveren zonder beheersbaarheid, security en compliance uit het oog te verliezen.

INZET AI VOOR DIT RAPPORT

Voor het tot stand komen van dit rapport hebben we gebruikgemaakt van een aantal ondersteunende AI-tools, waaronder Microsoft Copilot, ChatGPT, Claude en Le Chat. Deze tools zijn ingezet voor het ondersteunen bij vertalingen, het maken van samenvattingen en het verbeteren van spelling en grammatica. Ze hebben bijgedragen aan de redactionele kwaliteit en consistentie van de teksten.

MEEWERKEN AAN EEN VOLGENDE EDITIE?

Wil je meedenken over onze toekomstige AI whitepapers? Elke 8 weken publiceren we onze inzichten, waarbij we samenwerken met partners uit het hele AI werkveld. Wanneer je wil bijdragen of sparren, neem contact op met:

@ simone.vanerp@teamrockstars.nl
@ mathijs.vanbree@teamrockstars.nl

Met dank aan o.a.:

Simone van Erp, Mirjam van Olst, Serge Don, Roel van Bergen en Zita Janssen.