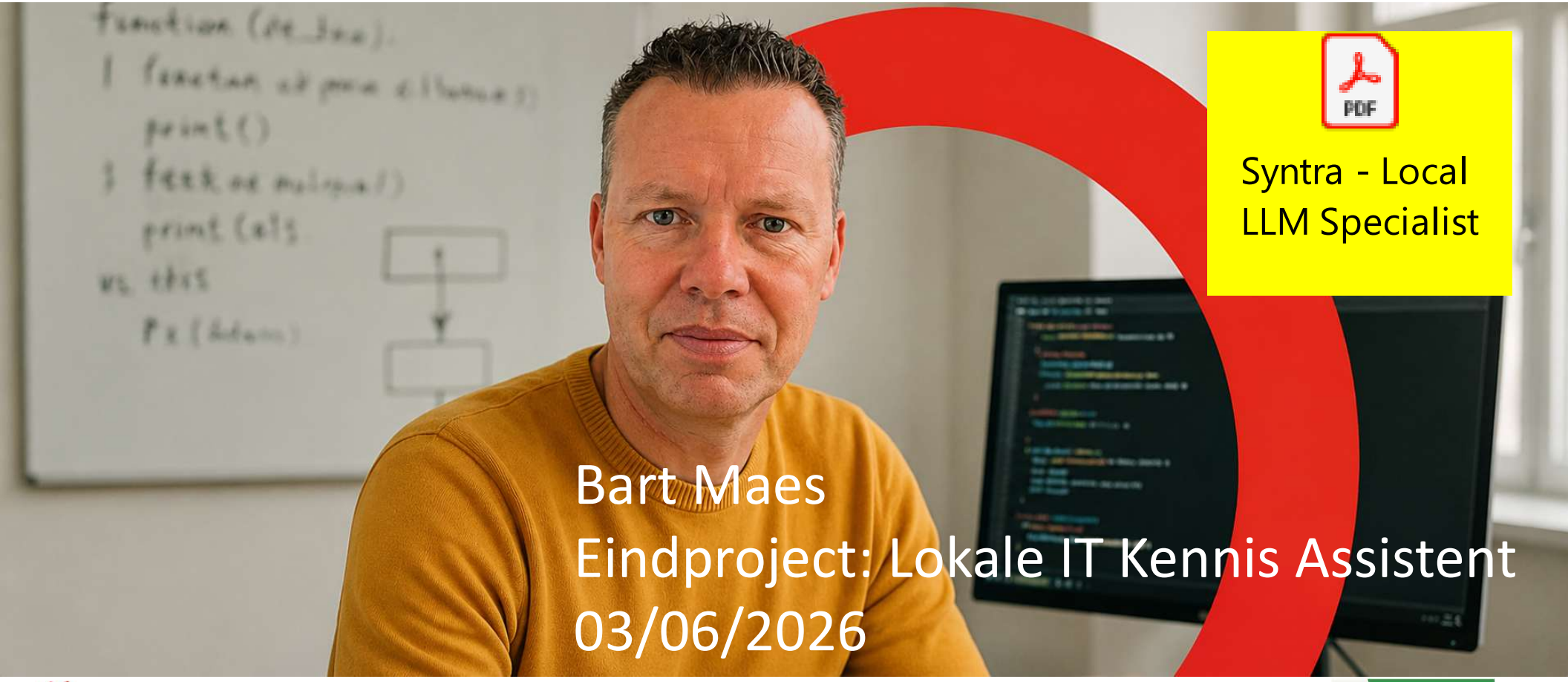
Syntra opleiding
Lokale LLM Specialist / GDPR-Conforme AI Implementatie



Bart Maes

Eindproject: Lokale IT Kennis Assistent

03/06/2026

AI Knowledge Assistant - Executive Summary

- **Probleem**
  - Hoe betrouwbare antwoorden genereren op domeinspecifieke documentatie?
- **Aanpak**
  - Hybrid RAG
  - Local LLM
  - Evaluation-first design
  - RAGAS-inspired benchmarking
- **Resultaat**
  - ~88% answer accuracy
  - HITL (Human In The Loop) is belangrijk
  - Volledig lokaal uitvoerbaar
- GDPR-aware architectuur

- **Lessons Learned**
  - Retrievalkwaliteit en evaluatiemethodologie zijn belangrijker dan modelgrootte alleen.
  - Meten is essentieel om AI betrouwbaar te maken
  - Een AI-project is minstens evenveel een governance-project als een technisch project
  - Productiewaardige AI vereist een andere mindset dan een proof-of-concept

=> Toelichting bij de Lessons Learned in de volgende slide

Lessons learned - explained

1. Retrievalkwaliteit is belangrijker dan modelgrootte

- Tijdens de evaluatie bleek dat de kwaliteit van de opgehaalde context vaak een grotere impact had op de antwoordkwaliteit dan de keuze van het LLM zelf. Een goed retrievalmechanisme compenseert vaak voor een kleiner model, terwijl een krachtig model weinig kan aanvangen met irrelevante of ontbrekende context.
- Tijdens de benchmark bleek een combinatie van vector search en keyword search robuustere resultaten te geven dan een puur semantische aanpak. Vooral bij technische en domeinspecifieke documentatie verhoogde dit de precisie en reproduceerbaarheid van antwoorden.

2. Meten is essentieel om AI betrouwbaar te maken

- Betrouwbaarheid kan niet beoordeeld worden op basis van enkele demo's of subjectieve indrukken. Door meer dan 100 realistische queries systematisch te evalueren met objectieve metrics werd duidelijk welke architectuurkeuzes effectief een verschil maakten.

3. Een AI-project is minstens evenveel een governance-project als een technisch project

- Gebruikers vertrouwen een systeem pas wanneer transparantie, bronvermelding, logging, confidence scoring en menselijke validatie voorzien zijn. AI-governance bleek minstens even belangrijk als modelkeuze of performantie.

4.. Productiewaardige AI vereist een andere mindset dan een proof-of-concept

- Een werkende demo bouwen is relatief eenvoudig. Zaken zoals monitoring, logging, foutafhandeling, schaalbaarheid, auditability en operationele governance bepalen uiteindelijk of een AI-oplossing effectief inzetbaar is binnen een organisatie.

Van versnipperde IT-kennis naar slimme antwoorden

HET PROBLEEM

IT-informatie is versnipperd en moeilijk te vinden

- Informatie staat overal en in verschillende formaten
- Moeilijk om het juiste antwoord te vinden
- Verouderde, dubbele of inconsistente inhoud
- Verlies van tijd en gefrustreerde medewerkers
- Risico op het gebruik van foutieve informatie

DE OPLOSSING

Een lokale, GDPR-conforme AI-assistent (RAG) die via hybride retrieval en een gevalideerd LLM model snel **accurate, context gebaseerde antwoorden** levert op basis van interne IT-documentatie.

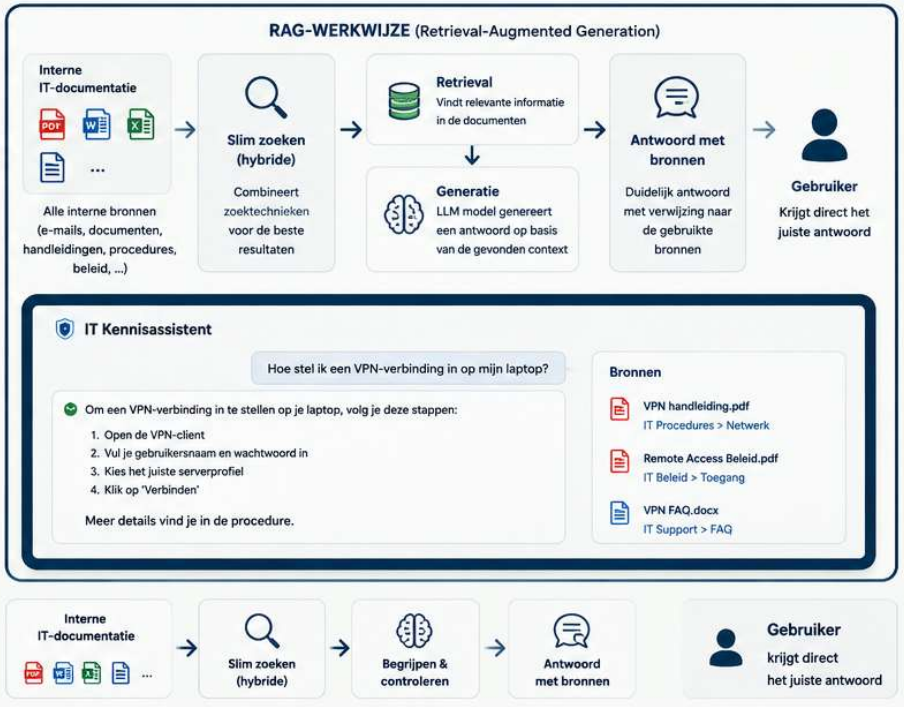
LOKAAL & VEILIG
Jouw data blijft binnen je organisatie

SLIM ZOEKEN
Hybride retrieval vindt de juiste informatie

ACCURATE ANTWOORDEN
LLM met RAG levert direct duidelijke antwoorden met bronverwijzingen

GDPR-CONFORM
Privacy en veiligheid staan centraal

RAG-WERKWIJZE (Retrieval-Augmented Generation)



LOKALE LLM VERSUS CLOUD DESIGN KEUZE

Waarom gekozen werd voor een volledig lokale AI-architectuur

LOKALE ARCHITECTUUR (GEKOZEN OPLOSSING)

Alle componenten draaien binnen de eigen infrastructuur. Data, modellen en resultaten blijven onder lokale controle.

LM Studio
Lokale LLM-inferentie (Qwen3-8B-Instruct)

Qdrant
Vector database voor embeddings en document opslag

n8n
Workflow orchestratie voor ingestion, retrieval en evaluatie

Volledig binnen lokale omgeving
Geen externe datadoorgifte

- Interne IT-documentatie blijft lokaal
- Geen externe datadoorgifte
- Embeddings, prompts en antwoorden onder eigen controle
- GDPR-conforme verwerking
- Reproduceerbare testomgeving
- Geen tokenkosten per query
- Minder afhankelijk van externe AI-platformen

BESLISSINGSMATRIX

Criteria	Lokale LLM	Cloud LLM
Privacy & GDPR	★★★★★	★★
Controleerbaarheid	★★★★★	★★
Reproduceerbaarheid	★★★★★	★★★
Performantie	★★★	★★★★★
Modelkwaliteit	★★★★	★★★★★
Kostcontrole	★★★★★	★★
Schaalbaarheid	★★★	★★★★★

Context
Evaluatie in het kader van een enterprise IT Knowledge Assistant met interne documentatie, procedures en mogelijk gevoelige bedrijfsinformatie.

CLOUDGEBASEERDE LLM-AANPAK

Gebruik van externe AI-platformen via API's voor modelinference en diensten.

Externe AI-platformen
OpenAI, Anthropic, Azure OpenAI, ...

API-gebaseerde verwerking
Verwerking in de cloud

Verbruik per token
Pay-per-use model

Afhankelijkheid van provider
API, beschikbaarheid en wijzigingen

Mogelijke voordelen

- ✓ Hogere schaalbaarheid
- ✓ Sterkere frontier modellen
- ✓ Hogere performantie
- ✓ Minder lokaal beheer

Aandachtspunten

- ⚠ Datadoorgifte buiten organisatie
- ⚠ Logging & retention policies
- ⚠ Vendor lock-in
- ⚠ Contractuele garanties nodig
- ⚠ Variabele operationele kosten

GESELECTEERDE ARCHITECTUUR: LOKALE LLM-ARCHITECTUUR

Beste match
voor enterprise IT-documentatie en interne kennisassistenten

Focus op privacy en controle
Data blijft binnen de eigen organisatie

GDPR-conforme AI-implementatie
Geen onnodige datadoorgifte

Ideaal voor proof-of-concept
Robuust, reproduceerbaar en controleerbaar

> *"Voor enterprise knowledge assistants is controle over data vaak belangrijker dan maximale modelgrootte."*

KMO LEVEL HARDWARE

CPU
Intel Core i9

RAM
32 GB DDR4/DDR5

GPU
NVIDIA GeForce RTX 5070
8 GB VRAM

STORAGE
NVMe SSD (min. 1 TB)

KMO
Standaard werkstation

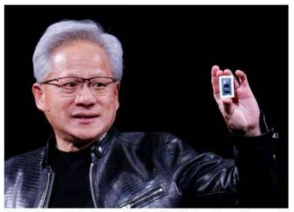
Lokaal
AI-inferentie & RAG-stack

Privacy
100% lokaal GDPR-friendly

Gebruik
RAG, LLM's, Workflows

Geschikt voor:

- Lokale LLM's
- RAG applicaties
- Vector DB (Qdrant)
- Workflows (n8n)



Veel details over zijn nieuwe chips heeft Nvidia-baas Jensen Huang niet vrijgegeven. © reuters

Nvidia lanceert "superchip" voor AI-laptops

Chipmaker Nvidia maakt zijn entree in de markt voor lap- zullen deze herfst laptops op basis van de RTX Spark uitbrengen.

Veel details over zijn RTX Spark-chips heeft Nvidia-baas Jensen Huang maandag niet vrijgegeven, tijdens een speech op de Computex-beurs in Taiwan. Hij grossierde vooral in superlatieven en sprak van de "heruitvinding van de computer", waarbij hij dit moment "even ingrijpend" noemde als de overgang van de klassieke gsm op een smartphone.

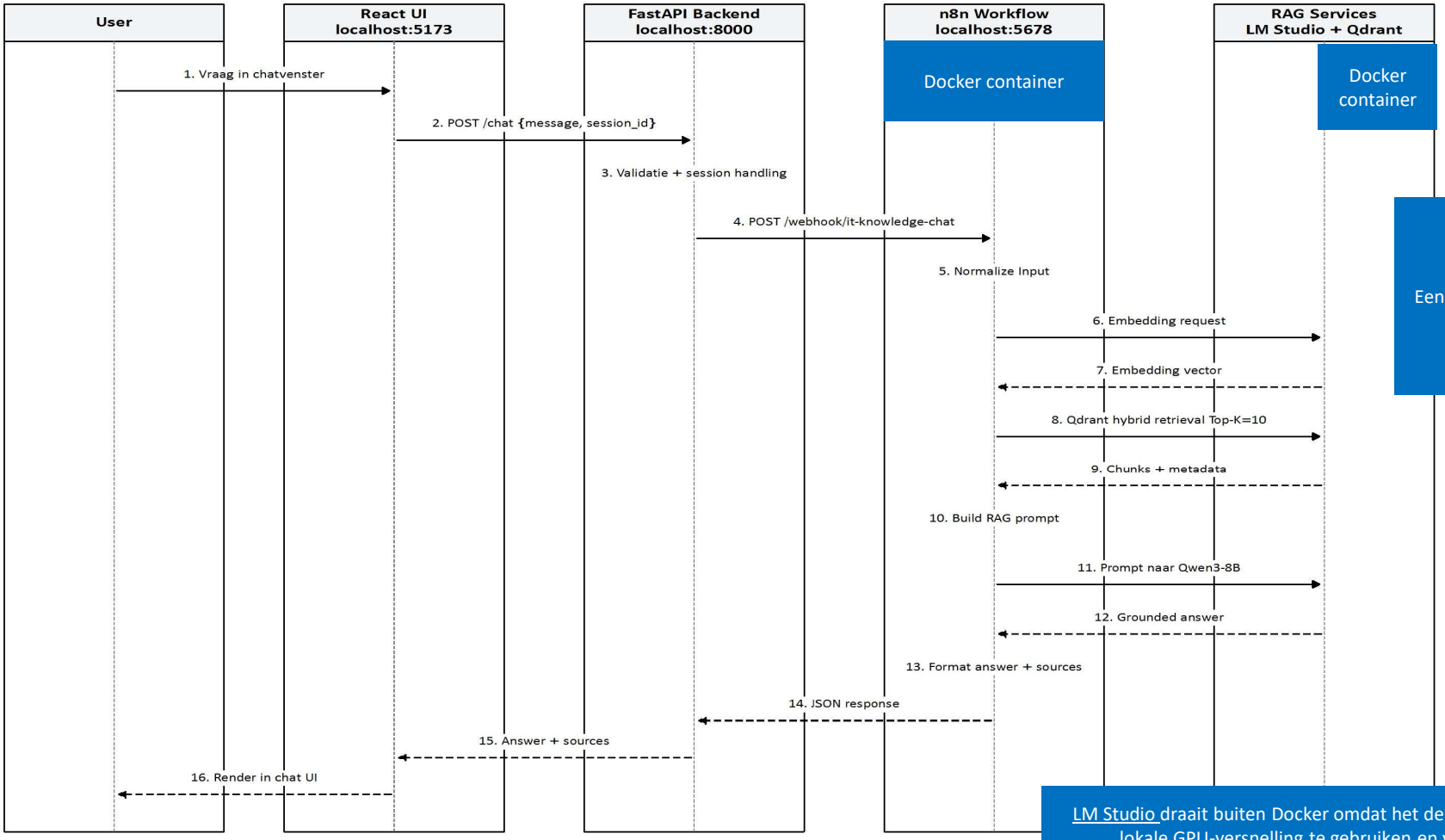
brikanten, waarond
vo, Dell, Asus en M

Gisteren in De Standaard

Nvidia omschrijft de RTX Spark als een "superchip", met zowel een forse GPU met 6.144 kernen aan boord als een door Nvidia ontworpen Grace CPU met tot twintig processorkernen. Ze delen tot 128 gigabytes aan geheugen. Die combinatie moet het toelaten om een AI-agent te draaien met een plaatselijk *large language model* (LLM) van 120 miljard parameters, wat naar de huidige normen een middelgrote LLM is. Maar daarnaast

"De toekomst van AI = hybride model waarbij gevoelige bedrijfsinformatie lokaal verwerkt wordt en cloudcapaciteit selectief wordt ingezet??"

Architecture



Docker voordelen
Eenvoudige installatie
Reproduceerbaarheid
Eenvoudiger beheer (start, stop, updates)
Portabiliteit
Isolatie

LM Studio draait buiten Docker omdat het de eenvoudigste manier bood om lokale GPU-versnelling te gebruiken en verschillende modellen te benchmarken. Voor deze proof-of-concept was ontwikkelsnelheid belangrijker dan volledige containerisatie."

Architecture – RAG Data Ingestie Flow

Van interne IT-documentatie naar Qdrant Vector Database

1 Bronnen (Interne IT-documentatie)

PDF

- PDF documenten (7 documenten)

2 Inname (Upload)

- Handmatige upload
- Batch upload
- Scheduling (optioneel)

3 Extractie & Parsing

- Text extractie (OCR indien nodig)
- Structuurbehoud
- Metadata extractie (titel, auteur, datum, ...)

4 Opschonen & Normalisatie

- White-space / layout cleanup
- Normalisatie
- Taal detectie
- Duplicaten detectie

Docker

5 Chunking

Vaste chunking strategie (Fixed Size with Overlap)

chunk_length = 500
chunk_overlap = 75

- Consistente chunk grootte
- Overlap voor context behoud
- Chunk metadata toevoegen (bron, pagina, categorie, ...)

6 Embedding Generatie (LM Studio)

- Embedding model: tekst-embedding-bge-m3
- Batch verwerking
- Vector generatie

Docker

7 Opslag in Qdrant Vector DB

- Upsert vector + payload
- Metadata: bron, pagina, categorie, tags, datum, ...
- Hybrid search ready
- Index optimalisatie

Data Overview – Interne PDF Documenten

#	Document	Pagina's	#chunks	Categorie(s)
1	ITIL 4 Foundation Knowledge Base	46	162	general, ITIL
2	IT Onboarding Guide	22	74	general, onboarding
3	Incident Management Procedure	20	76	general, incident
4	Change Management Policy	19	55	general, change
5	Acceptable Use Policy	21	63	general, usepolicy
6	Network- en VPN Handleiding	20	61	general, network
7	Software Request Procedure	19	59	general, software
	Totaal	**167**	**550**	

Chunking Strategy

Strategie: 'fixed_size_with_overlap'
chunk_length = 500
chunk_overlap = 75

Voordeel:

- Behoud van context over chunk grenzen
- Betere retrieval kwaliteit
- Consistente chunk grootte voor embedding

Resultaat in Qdrant

Vector + Payload (metadata)

```
[ 0.12, -0.08, 0.33, .., 0.41 ]  {  bron": "ITIL 4 Foundation Knowledge Base",
                                    pagina": 12,
                                    categorie": "ITIL",
                                    chunk_id": "ITIL4_12_1",
                                    datum": "2024-05-10",
                                 }  tags": ["itil", "foundation", "principles"]
[ -0.21, 0.37, -0.11, .., 0.09 ]  {  ...  }
[ 0.05, -0.31, 0.22, .., -0.14 ]  {  ...  }
...                                ...
```

- **Hybrid Search** (dichtheid + vector)
- **Top-K Retrieval** (bijv. K=10)
- **Metadata Filtering** (categorie, bron, datum, ...)
- **Snel & schaalbaar** Klaar voor productie

Omgeving

- Dockerized services (n8n, Qdrant)
- Lokaal (niet in Docker) (LM Studio)

n8n
Workflow Orchestratie
(Ingestie pipeline)

LM Studio
Embeddings
(tekst-embedding-bge-m3)

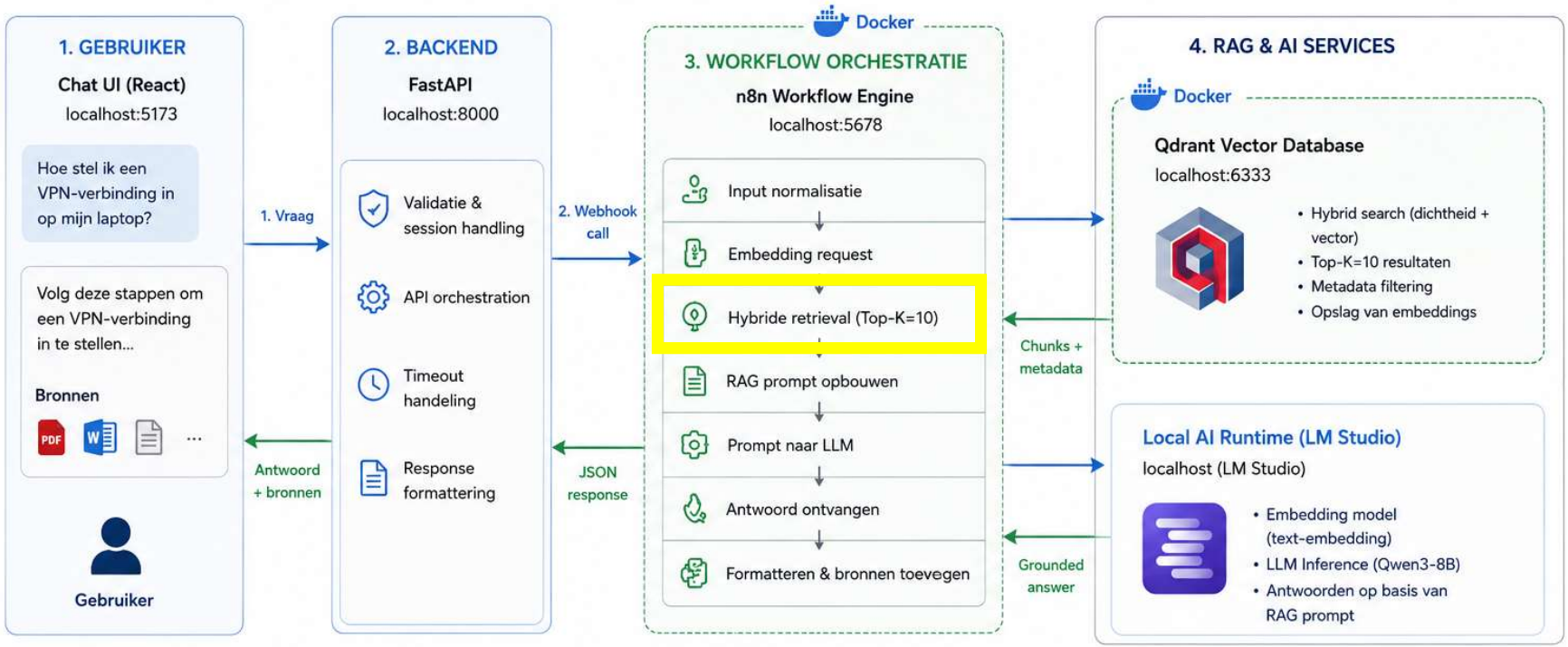
Qdrant
Vector Database
(Opslag & Retrieval)

Robuust & Betrouwbaar

- Foutafhandeling & logging
- Retry & fallback strategieën
- Idempotente verwerking
- Monitoring & validatie

Architecture – RAG Retrieval Flow

Van vraag tot antwoord via hybride retrieval (RAG)



GDPR-conform & privacy by design
Data blijft binnen de organisatie

Accurate & context gebaseerd
RAG zorgt voor onderbouwde antwoorden

Betrouwbaar & robuust
Timeouts, validatie en fallback handling

Snel & efficiënt
Hybride retrieval voor relevante resultaten

Architecture – RAG Retrieval Evaluation & Optimization

100 IT-queries geëvalueerd met RAGAS-geïnspireerde LLM-as-a-judge evaluatie

1. EVALUATIE-AANPAK

Testopzet
- 100 representatieve IT-queries
- 4 query types

- Definition – korte definities
- Numerical – numerieke waarden
- List – opsommingen
- Explanation – conceptuele uitleg

Volledige evaluatie binnen de RAG pipeline

Query → Retrieval → Context → LLM → Antwoord

Evaluatie (LLM-as-a-judge)

LLM-as-a-judge evaluatie
- Evaluatiemodel: openai/gpt-oss-20B
- Elke metric met specifieke prompt
- Orthogonale evaluatie voor objectiviteit

2. RAGAS-STYLE EVALUATIEMETRICS (5)

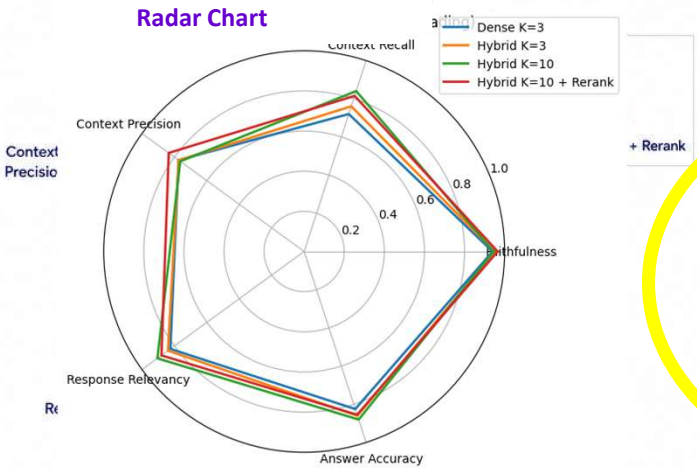
Faithfulness	Bevat het antwoord alleen informatie die ondersteund wordt door de context? Geen hallucinaties.
Context Recall	Werd alle relevante informatie die nodig is voor het antwoord opgehaald?
Context Precision	Hoeveel irrelevante of ruis-achtige informatie bevat de opgehaalde context?
Response Relevancy	Beantwoordt het gegenereerde antwoord de vraag effectief?
Answer Accuracy	Hoe correct is het antwoord inhoudelijk (t.o.v. de ground truth)?

Gebaseerd op afzonderlijke evaluatieprompts per metric om orthogonaliteit te garanderen.

3. GEËVALUEERDE RETRIEVAL APPROACHES

1. **Dense K=3** – Pure vector retrieval (baseline)
2. **Hybrid K=3** – Dense + BM25
3. **Hybrid K=10** – High recall retrieval (meer context)
4. **Hybrid K=10 + Rerank** – Cross-encoder reranking (context filtering)
5. **Adaptive (A5)** – Query-type gebaseerde routing (optimaal per vraagtype)

Radar Chart



4. BELANGRIJKSTE INZICHTEN

- ✅ **Hybrid retrieval > Dense retrieval**
  Dense + BM25 geeft systematisch betere answer quality.
- ✅ **Context recall is de belangrijkste driver**
  Meer relevante context bleek belangrijker dan perfecte filtering.
- ✅ **Hogere Top_K verbeterde answer accuracy**
  Top_K 3 → 10:
  - Recall: 76% → 84% (+8pp)
  - Accuracy: 86% → 88% (+2pp)
- ⚠ **Reranking verhoogt precision, maar niet accuracy**
  Perfect gefilterde context leidde niet noodzakelijk tot betere antwoorden.

5. FINALE SELECTIE

Geselecteerde retrievalstrategie:
HYBRID RETRIEVAL + TOP_K = 10

Waarom deze keuze?
- ✓ Hoogste globale answer accuracy (88%)
- ✓ Beste response relevancy (90%)
- ✓ Sterkste context recall (84%)
- ✓ Beste balans tussen recall en precision
- ✓ Robuust over verschillende query types
- ✓ Relatief eenvoudige implementatie

Beste balans tussen prestaties, robuustheid en implementatiecomplexiteit.

6. NUANCERING: ADAPTIVE RETRIEVAL (APPROACH 5)

Adaptive Retrieval (per query type optimale strategie)
→ Hoogste theoretische answer accuracy: **90%**

Slechts +2% winst t.o.v. Hybrid K=10 (88%)

Hogere complexiteit (routing / classificatie nodig)

Interessante optimalisatie voor een latere fase.

MODELKEUZEPROCES – STAP 1

Standalone LLM Evaluatie (zonder RAG)

Edge Case & Robuustheid Testing op LLM's (zonder RAG)

Geselecteerde modellen draaien op mijn laptop (32 GB RAM, NVIDIA GPUs met 8 GB VRAM).

1. Testmethodologie

Edge Case Types

- Length
- Domain mismatch
- Ambiguity
- Malformed input

Failure Classificatie

F = Format
C = Completeness
A = Accuracy
L = Language
R = Refusal
H = Hallucination

Meetmethode

- 32 Edge Case prompts
- Identieke inference settings
- LLM-as-a-judge evaluatie
- GPT-OSS-20B als evaluator-model

Bron: Local LLM Model Selection Rapport (§3.2 – Context Position Testing & §3.3 – Test Resultaten)

2. Kwalitatieve vergelijking van de geselecteerde modellen

Criteria	Qwen3-8B	Ministral-3-8B (Reasoning)	Llama-3.2-3B
Accuracy	★★★★★	★★★★☆	★★★☆☆
Reasoning	★★★★★	★★★★☆	★★☆☆☆
Hallucinatie-controle	★★★★★	★★★★★	★★★☆☆
Taal (NL/EN)	★★★★★	★★★★☆	★★★☆☆
Completeness	★★★★☆	★★★☆☆	★★★☆☆
Snelheid	★★☆☆☆	★★★☆☆	★★★★☆
RAG-geschiktheid	Beste	Goed	Beperkt

Eerste voorlopige keuze: Qwen3-8B

- ✓ Sterkste robuustheid (edge cases)
- ✓ Beste reasoning-capaciteiten
- ✓ Zeer goede hallucinatie-controle
- ✓ Beste RAG-geschiktheid

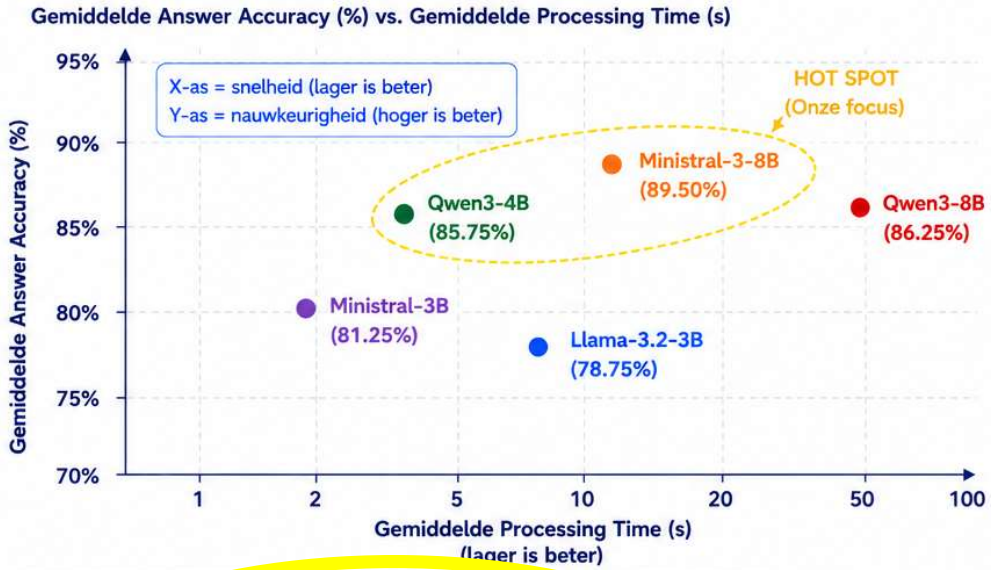
Snelheid werd initieel als secundair criterium beschouwd.

MODELKEUZEPROCES – STAP 2

Evaluatie binnen volledige RAG Pipeline

100 realistische IT-vragen in de RAG architectuur

1. Resultaten – 100 IT queries in de RAG architectuur



2. Overzicht resultaten (100 IT queries)

Model	Gem. Processing Time (s) (lager is beter)	Gem. Answer Accuracy (%) (hoger is beter)
Qwen3-8B	50.21	86.25%
Qwen3-4B	6.10	85.75%
Ministral-3-8B	11.27	89.50%
Ministral-3B	4.31	81.25%
Llama-3.2-3B	8.85	78.75%

Definitieve modelkeuze: Qwen3-4B

- ✓ Quasi identieke accuracy als Qwen3-8B (-0,50%)
- ✓ ~8× sneller dan Qwen3-8B (6,10 s vs. 50,21 s)
- ✓ Beste balans tussen snelheid en nauwkeurigheid
- ✓ Betere responsiveness voor eindgebruikers
- ✓ Efficiënt lokaal inzetbaar en schaalbaar

Beste [illegible]nelheid/accuracy verhouding voor onze productie[illegible]geving.

Beste balans voor productie

Fallback model: Ministral-3-8B

- Hoogste pure accuracy (89,50%)
- Sterk alternatief voor complexe queries

Maar:

- bijna dubbel zo traag als Qwen3-4B (11,27 s vs. 6,10 s)
- kortere antwoorden
- minder consistente procedurele uitleg

Conclusie: binnen een productie-RAG-systeem bleek retrievalkwaliteit belangrijker dan pure modelgrootte.

Kosten-Baten Analyse – Lokale Oplossing

Business case voor een kleine KMO (10 personen) op basis van tijdsbesparing

1. ASSUMPTIES

Lokale oplossing

• Eenmalige ontwikkelkost (160u à €80)	€12.800
• Eenmalige HW + installatiekost	€3.500
Totale initiële investering	**€16.300**

Jaarlijkse operationele kost (OPEX)

• Stroom, onderhoud, updates (12 mnd x €1.280)	€15.360 / jaar

Gebruikersprofiel (Kleine KMO)

• Aantal gebruikers	10
• Queries per gebruiker per dag	2
• Werkdagen per jaar	240
• **Totaal queries per jaar**	**4.800**

Tijdsbesparing per query

• Zonder IT Assistant:	10 minuten
• Met IT Assistant:	2 minuten

Netto tijdswinst: 8 minuten per query

Kost werknemer

- Gemiddelde kost: €80 / uur

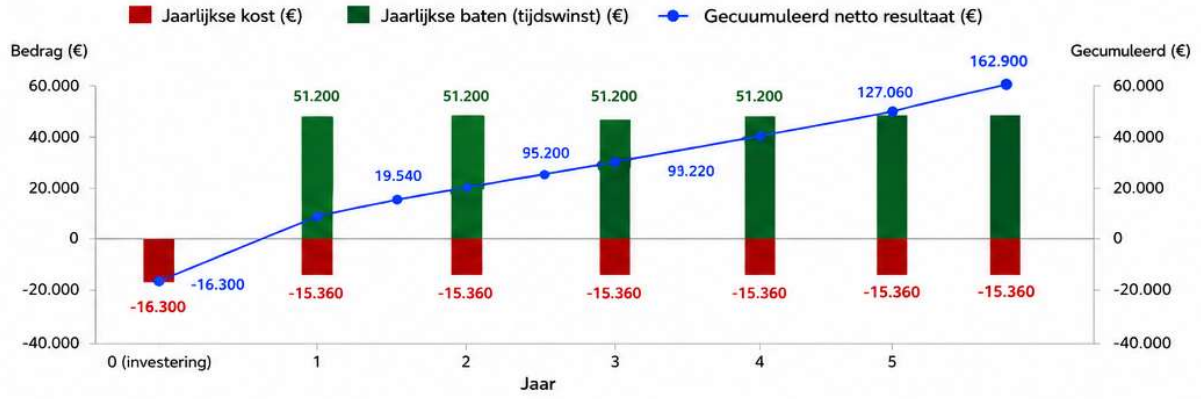
Formule tijdsbesparing

Bespaarde tijd per query = 10 min – 2 min = 8 min
Bespaarde uren per jaar = 4.800 x 8 / 60 = 640 uur
Financiële besparing per jaar = 640 x €80 = €51.200

2. KOSTEN EN BATEN OVER 5 JAAR (Kleine KMO – 10 personen)

Jaar	Jaarlijkse kost (€)	Jaarlijkse baten (tijdswinst) (€)	Netto resultaat per jaar (€)	Gecusmuleerd netto resultaat (€)
0 (investering)	-16.300	0	-16.300	-16.300
1	-15.360	51.200	35.840	19.540
2	-15.360	51.200	35.840	55.380
3	-15.360	51.200	35.840	91.220
4	-15.360	51.200	35.840	127.060
5	-15.360	51.200	35.840	162.900

3. KOSTEN EN BATEN OVER 5 JAAR – VISUEEL OVERZICHT



Break-even bereikt tussen maand 5 en 6.

De lokale IT Knowledge Assistant levert voor een kleine KMO (10 personen) een jaarlijkse tijdswinst op van **€51.200**.
Ondanks de initiële investering wordt de oplossing reeds binnen het eerste jaar terugverdiend.
Over 5 jaar resulteert dit in een gecumuleerd netto voordeel van: **€162.900**

ROI OVER 5 JAAR
175%
(€162.900 netto voordeel t.o.v. totale kost)

Lokale kost vs Cloud kost vergelijking

Vergelijking van jaarlijkse en gecumuleerde kosten voor een KMO van 10 personeelsleden

1. ASSUMPTIES (volgens Input sheet)

Gebruikersprofiel (KMO)

- Aantal gebruikers: 10
- Queries per gebruiker per dag: 2
- Werkdagen per jaar: 200
- Totaal queries per jaar: 4.000

Lokale oplossing

- Eenmalige ontwikkelkost (160u à €80): €12.800
- Eenmalige HW + installatiekost: €3.500
- Jaarlijkse operationele kost (OPEX): €15.360
  (Stroom, onderhoud, updates) (12 mnd x €1.280)

Cloud oplossing (API-gebaseerde LLM)

- Eenmalige ontwikkelkost (opstart): €12.800
- Maandelijkse user support kost: €640
  (= €7.680 per jaar)
- Gemiddelde tokens per query
  - Input tokens (incl. context): 1.500
  - Output tokens (antwoord): 250
- Prijzen per 1K tokens
  - Input tokens: €0,003
  - Output tokens: €0,012

Kost per query (totaal): €0,00755

Jaarlijkse cloud kost = 4.000 queries × €0,00755
= €30,20 per jaar

- Totale jaarlijkse kost (query + support): **€7.710**
  (€30,20 + €7.680 = €7.710)

Werkgeverskost

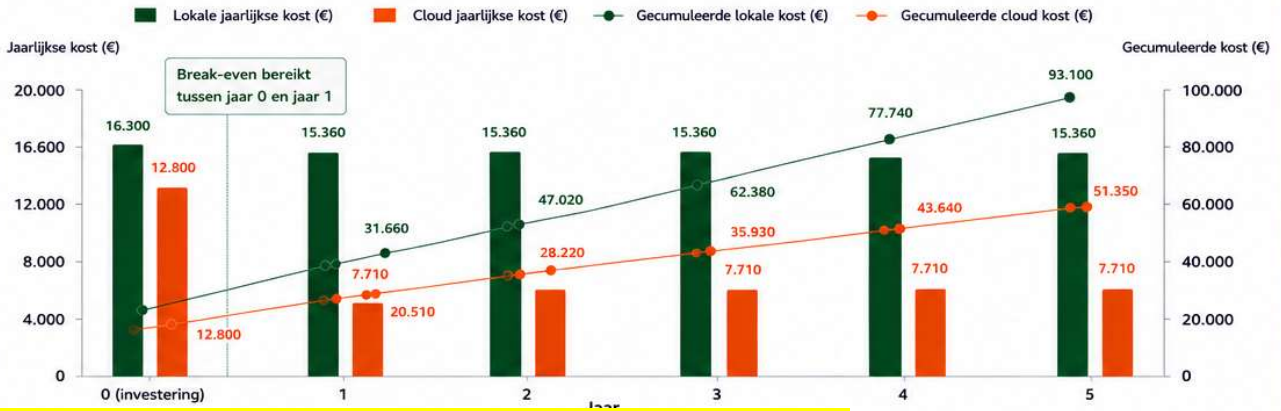
- Niet van toepassing in deze vergelijking

2. KOSTENOVERZICHT PER JAAR (KMO – 10 PERSONEELSLEDEN)

Jaar	Lokale oplossing		Cloud oplossing		Verschil (Cloud – Lokaal)
	Jaarlijkse kost (€)	Gecumuleerde kost (€)	Jaarlijkse kost (€)	Gecumuleerde kost (€)	Gecumuleerde verschil (€)
0 (investering)	16.300	16.300	12.800	12.800	-3.500
1	15.360	31.660	7.710	20.510	-11.150
2	15.360	47.020	7.710	28.220	-18.800
3	15.360	62.380	7.710	35.930	-26.450
4	15.360	77.740	7.710	43.640	-34.100
5	15.360	93.100	7.710	51.350	-41.750

Lokale oplossing is vanaf jaar 1 duurder dan cloud oplossing en blijft elk jaar duurder.

3. KOSTEN OVER 5 JAAR – LOKAAL VS CLOUD



CUMULATIEF VOORDEEL NA 5 JAAR
€41.750
(cloud oplossing t.o.v. lokale oplossing)

Voor een KMO van 10 personeelsleden en 4.800 queries per jaar blijft de cloudoplossing financieel goedkoper dan de lokale oplossing, zelfs over een horizon van 5 jaar. Dit wordt voornamelijk verklaard door het relatief beperkte queryvolume van deze IT Knowledge Assistant toepassing .Toch werd bewust gekozen voor een lokaal geïmplementeerde LLM-architectuur, omdat voor enterprise knowledge assistants met interne data privacy en controle belangrijker wordt geacht dan louter minimale operationele kost.

Production Readiness & Operational Maturity

Huidige status (POC bereikt)

- Werkende end-to-end RAG architectuur
- Lokale LLM-integratie via LM Studio
- React GUI + FastAPI backend
- Dockerized infrastructuur (n8n + Qdrant)
- Error Handling (retries, reporting)
- Logging van queries, resultaten en metrics
- Evaluatieframework met RAGAS & LLM-as-a-Judge
- GDPR-bewuste lokale architectuur

Mogelijke Vervolgstappen

- Multi-user concurrency (>4 gebruikers)
- Load- & stress-testing
- Monitoring & alerting dashboards
- Automatische backup & recovery procedures
- Security hardening & role-based access
- High-availability inferencing (bv. vLLM)
- Centrale deployment & update procedures

Het systeem bereikt vandaag een sterke proof-of-concept maturiteit, maar bijkomende operationalisatie en schaalbaarheidstesten zijn nodig voor enterprise productiegebruik.

it_knowledge_query_log

- Success/Failure
- Failure Reason

- response time
- aantal bronnen
- confidence level

- Faithfulness
- Context Recall
- Context Precision
- Response Relevancy

Operationele evaluatie

Debug Mode evaluatie

id	ses	user_query	answer	status	failure_	response_time_	sources_count	top_score	avg_top3_score	#	confidence_level	debug_mode	score_faithfulness	score_context_recall	score_response_relevancy	score_context_precisic
21	user-1	hoe reset ik vpn?	Om een VPN-verbinding te resetten, vc	success	Empty	50 047	10	0.5	0.45	0.15(	Null	Null	Null	Null	Null	Null
22	user-1	Hoe reset ik VPN?	Om een VPN-verbinding te resetten of	success	Empty	57 673	10	0.5	0.4444444466666666	0.16(	Null	Null	Null	Null	Null	Null
23	user-1	wat betekent ITIL?	ITIL staat voor **Information Technolo	success	Empty	19 815	10	0.6111111	0.5148148133333333	0.17;	low	Null	Null	Null	Null	Null
24	user-1	What are the three types of changes in Change Enablemer	The three types of changes in Change	success	Empty	26 692	10	0.7	0.6222222333333333	0.19(	medium	Null	Null	Null	Null	Null
25	user-1	Welke persoonlijke cloudopslag is toegestaan voor werkbe	Persoonlijke cloudopslag voor werkbes	success	Empty	30 166	10	0.64285713	0.5809523766666667	0.14;	medium	Null	Null	Null	Null	Null
26	user-1	Wanneer zijn freezeperiodes van toepassing?	Freezeperiodes zijn van toepassing in (	success	Empty	59 387	10	0.8333334	0.6111111333333333	0.33	medium	Null	Null	Null	Null	Null
27	user-1	What is the cost of ITIL 4 Foundation certification?	I could not find this information in the a	success	Empty	16 676	10	0.75	0.6666666666666666	0.25	high	Null	Null	Null	Null	Null
28	user-1	Welke acties moet je nemen wanneer je een verdachte phi	Wanneer je een verdachte phishing-ma	success	Empty	48 935	10	0.8333334	0.6944445	0.41(	medium	Null	Null	Null	Null	Null
29	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Op je eerste werkdag moet je na inlogg	success	Empty	44 702	10	0.8333334	0.6388889333333333	0.33	medium	Null	Null	Null	Null	Null
30	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Op je eerste werkdag moet je na het inl	success	Empty	46 825	10	0.8333334	0.6388889333333333	0.33	medium	Null	Null	Null	Null	Null
31	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Op je eerste werkdag moet je na inlogg	success	Empty	64 950	10	0.8333334	0.6388889333333333	0.33	medium	0	Null	Null	Null	Null
32	user-1	python	Op je eerste werkdag moet je na inlogg	success	Empty	44 434	10	0.8333334	0.6388889333333333	0.33	medium	0	Null	Null	Null	Null
33	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Op je eerste werkdag moet je na inlogg	success	Empty	41 526	10	0.8333334	0.6388889333333333	0.33	high	0	Null	Null	Null	Null
34	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Op je eerste werkdag moet je na het inl	success	Empty	93 376	10	0.8333334	0.6388889333333333	0.33	high	1	0	0	0	0
35	user-1	Wat moet je doen op je eerste werkdag nadat je met je tijd	Empty	failure	llm_error	337 547	0	Null	Null	Null	low	Null	Null	Null	Null	Null
36	user-1	Wat betekent ITIL?	ITIL staat voor **Information Technolo	success	Empty	163 912	10	0.6111111	0.5117845133333333	0.18(	medium	0	Null	Null	Null	Null
37	user-1	What are the 4 dimensions of service management in ITIL	The four dimensions of service manage	success	Empty	50 590	10	1	0.6722222333333333	0.65	high	0	Null	Null	Null	Null
38	user-1	What are the 4 dimensions of service management in ITIL	The four dimensions of service manage	success	Empty	94 106	10	1	0.6722222333333333	0.65	high	1	0	0	0	0
39	user-1	What are the 4 dimensions of service management in ITIL	The four dimensions of service manage	success	Empty	84 723	10	1	0.6722222333333333	0.65	high	1	0	0	1	0.1
40	user-1	What are the 4 dimensions of service management in ITIL	The four dimensions of service manage	success	Empty	64 503	10	1	0.6722222333333333	0.65	high	1	0	0	1	0

+

SYNTRA ANTWERPEN & VLAAMS-BRABANT

VLAIO

GDPR & AI Act Compliance Evaluatie

GDPR Sterke Punten

- ✅ Volledig lokale verwerking van data
- ✅ Geen externe cloud-LLM's vereist
- ✅ Interne documenten blijven lokaal
- ✅ Chunk-traceability via metadata
- ✅ Logging & auditability voorzien
- ✅ Human-in-the-loop ondersteuning
- ✅ Dataminimalisatie via retrieval
- ✅ DPIA uitgevoerd

AI Act & Governance

- ✅ Transparantie over AI-gebruik
- ✅ Bronvermelding in antwoorden
- ✅ Explainability via retrieval chunks
- ✅ Mogelijkheid tot menselijke verificatie
- ✅ Evaluatie via objectieve metrics

Mogelijke Verdere Verbeteringen

- Fine-grained toegangscontrole
- Security & retention policies
- Formele incident response procedures
- Bias- en hallucination monitoring

De gekozen lokale architectuur ondersteunt een GDPR-conforme aanpak en vormt een sterke basis voor toekomstige AI Act-compliance.

Evaluatie van het Project & Toekomstige Evolutie

Belangrijkste Resultaten

- Werkende lokale enterprise RAG-stack
- Hybride retrieval verhoogt answer accuracy
- Qwen3-4B biedt beste snelheid/kwaliteit verhouding
- Volledig lokaal draaibaar op KMO-hardware
- Sterke reproduceerbaarheid & controle

Mogelijke Vervolgstappen

- Pilot met echte eindgebruikers
- Query type monitoring & analytics
- Adaptive retrieval routing per querytype
- Multi-user scalability testing en implementatie
- vLLM voor hogere concurrency
- Uitbreiding naar multimodale documenten

Dit project toont aan dat lokale, GDPR-conforme AI-oplossingen vandaag technisch haalbaar zijn voor KMO's, mits een doordachte architectuur en evaluatie-aanpak.

Samenvatting: Lokale IT Kennis Assistent

PROBLEEM	Bedrijven beschikken over veel interne IT-kennis, maar vinden en gebruiken die vandaag te traag, inconsistent en onvoldoende betrouwbaar om efficiënt beslissingen en acties te ondersteunen.
OPLOSSING	Een lokale, GDPR-conforme AI-assistent die via hybride retrieval en een gevalideerd LLM model snel accurate, context gebaseerde antwoorden levert op basis van interne IT-documentatie.
BEWIJS	Evaluatie van 100 realistische queries (van verschillende types) toont dat de gekozen hybride RAG-aanpak tot ~88% Antwoord Nauwkeurigheid en ~90% Relevantie behaalt
KOSTEN	160 uur ontwikkeling (12,800 €) + 3,500 € hardware/installatie = **16,300 €** eenmalig
BATEN	💼 **Baten per bedrijf** 🟢 **Klein (10 users) → €51,200/jaar** 🟡 **Middelgroot (50 users) → €256,000/jaar** 🔵 **Groot (200 users) → €1,024,000/jaar**
VOLGENDE STAP	Piloot 2 maanden — monitoring -> definieer optimalisaties – Schaalbaarheid Impact Analyse

Dank voor je aandacht