



SUMMETIX

LLMs for the Public Sector: Applying RAG in Municipal Administrations

OpenRheinMain Conference 2026

summetix GmbH



Erik Kaiser
CEO

Economics and Engineering degree
from TU Darmstadt



Dr. Johannes Daxenberger
COO

PhD in Natural Language Processing
from TU Darmstadt



Benjamin Schiller
CTO

Computer Science degree from TU
Darmstadt

Team in Germany and England



Bilge
Sales Support and
Fulfillment



Tim
IT / Fullstack



Estefania
Marketing Support and
Fulfillment



Michael
IT / Frontend

Partners and Clients (selection):



Digitalization in Public Administration

Everybody wants it – but ...

- ▶ Siloed data
- ▶ Massively heterogeneous landscape (across and within administrations)
- ▶ Manual “Ask-the-expert” loops
- ▶ Compliance (GDPR, freedom-of-information)

Challenges in Citizen Services

Increasing Complexity of Communication

- Growing number of responsibilities for municipalities
- Service guidelines often lag citizens' reality
- Harder to provide precise responses

Securing Knowledge Despite Constant Change

- High staff turnover and retirements
- Institutional knowledge often undocumented
- Long onboarding time for new employees



Operational strain

- Long response times
- High workload for staff
- Inconsistent response quality

AI to the Rescue?



~1 M+ city governments worldwide,
with 10K+ in Germany alone



Several GB of heterogeneous and
legacy documents per city



Citizens expect instant, accurate
answers



LLMs struggle with

- ▶ factual reliability
- ▶ auditability
- ▶ data protection
- ▶ domain specificity

So, where to start?!

- ▶ **Cold-Start Problem:** Training is not an option in most cases

Provide evidence in real-time

So, where to start?!

▶ **Cold-Start Problem:** Training is not an option in most cases

▶ **Lack of Grounding:** Transparency is not the same as reasoning

Provide evidence in real-time

Enforce grounding in addition to reasoning

So, where to start?!

▶ **Cold-Start Problem:** Training is not an option in most cases

▶ **Lack of Grounding:** Transparency is not the same as reasoning

▶ **Non-Determinism:** Unpredictable behavior decreases trust in AI systems

Provide evidence in real-time

Enforce grounding in addition to reasoning

Prevent mistrust through reproducible processes

So, where to start?!

▶ **Cold-Start Problem:** Training is not an option in most cases

▶ **Lack of Grounding:** Transparency is not the same as reasoning

▶ **Non-Determinism:** Unpredictable behavior decreases trust in AI systems

▶ **Distributed Data and Processes:** Opposed to humans, AI struggles to close gaps

Provide evidence in real-time

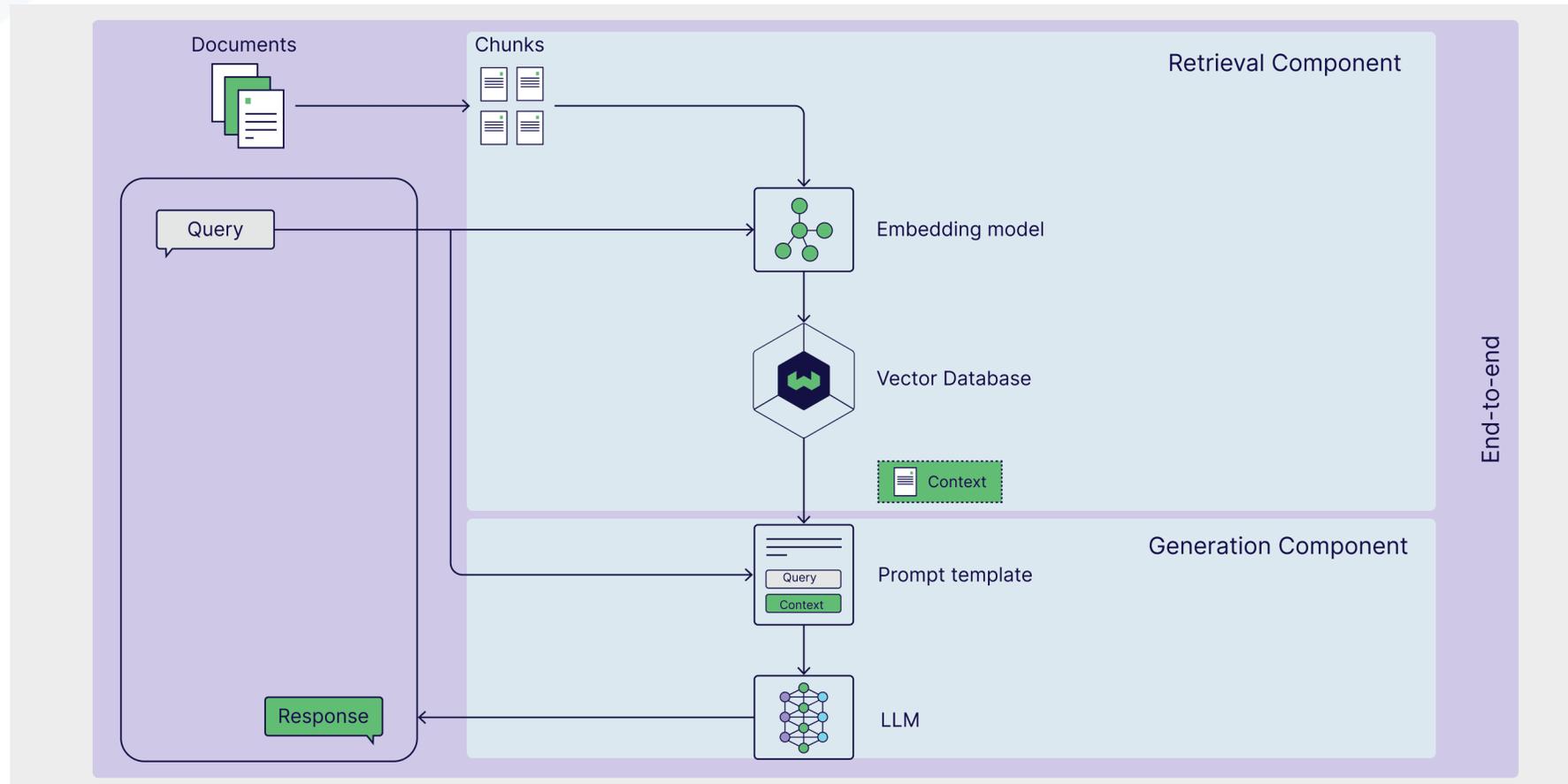
Enforce grounding in addition to reasoning

Prevent mistrust through reproducible processes

Provide document and process-specific pre-processing



Retrieval-Augmented Generation



<https://weaviate.io/blog/introduction-to-rag>

RAG: Practical Advantages in Administration



Soft Introduction

- ▶ No need to onboard all systems and data at once
- ▶ Integrates with other measures for digitalization



On-the-fly Consolidation

- ▶ Knowledge and communication can be added at any time
- ▶ Knowledge can be stored both centrally or decentrally



On-Point Generation

- ▶ Custom generations without prompt engineering exercises
- ▶ Queries are answered in line with previous communication

RAG: Things to Consider

Document Preprocessing

- ▶ **Source identification & collection:** different formats (PDF, Word, images, HTML, DOCX, etc.) need different handling
- ▶ **OCR & text extraction:** PDFs from legacy systems often contain scanned pages
- ▶ **Structural parsing:** Long policy documents have headings, sub-headings, clauses, annexes, and tables.
- ▶ **Segmentation & chunking:** most LLMs cannot deal with infinite context, longer documents need to be segmented in a meaningful way

RAG: Things to Consider

Document Preprocessing (Advanced)

- ▶ **Images/Audio/Video:** Dealing with multimodal information, within and in addition to text documents (multimodal embeddings or image-to-text)
- ▶ **PII Filtering:** GDPR compliance and privacy requirements might enforce PII detection and removal
- ▶ **Version control & change tracking:** Updates (changes, removal and additions) should be easy/fast

RAG: Things to Consider

Embedding model selection and evaluation

- ▶ **Language coverage:** multilingual models vs. language-specific ones
- ▶ **Long-term support:** make sure you have control over the model (APIs might change at any time!)
- ▶ **Latency & compute cost:** how much does it cost to reindex your whole dataset?
- ▶ **Vector Storage:** where and how to store? Consider retrieval capabilities and support for metadata
- ▶ **Evaluation on *Gold-Standard Test Set*:** qualitative: Policy-Q&A pairs, quantitative: Recall@k, Mean Reciprocal Rank (MRR)

RAG: Things to Consider

Retrieval and Grounding

- ▶ **Hybrid search:** always make sure to use dense *and* lexical retrieval
- ▶ **Chunking strategy:** important to chose wisely and hard to change later on
- ▶ **Reranking:** helps to account for relevance and recency
- ▶ **Grounding** (after generation): Provide evidence after generation, e.g. using structured LLM output

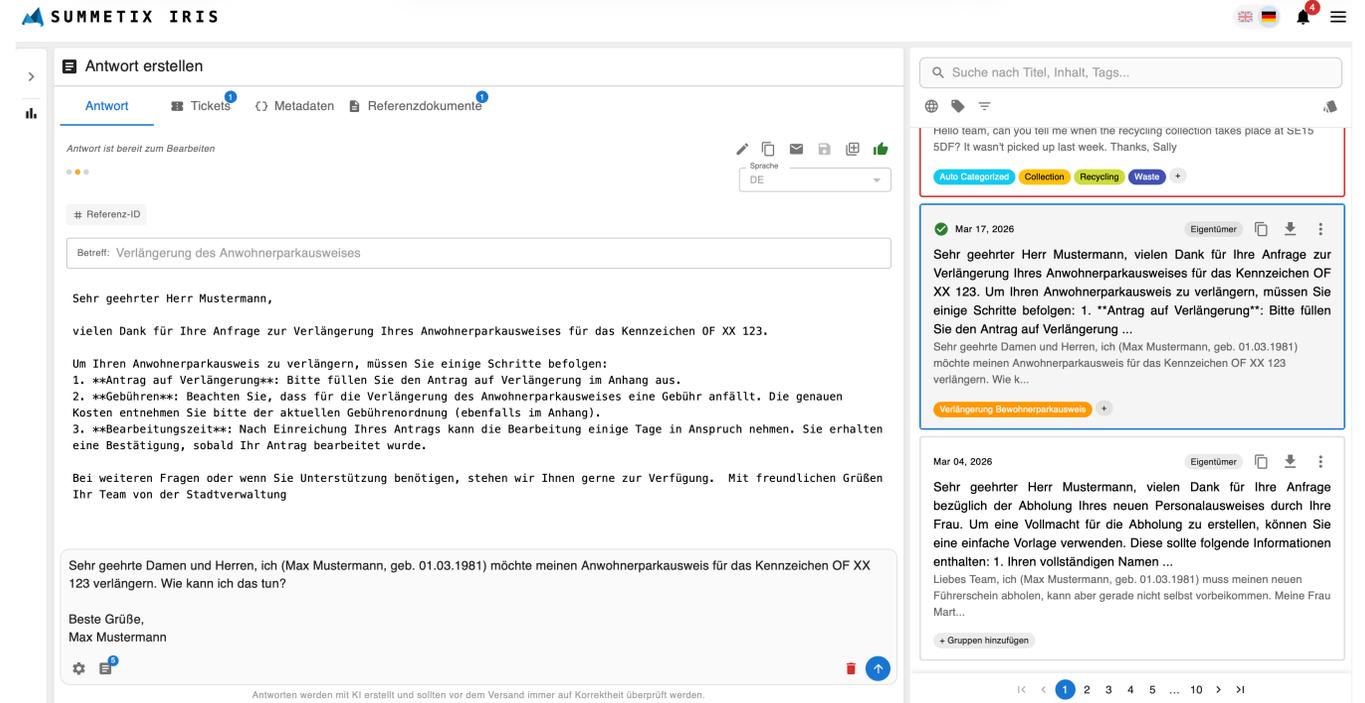
RAG: Things to Consider

Metadata and Access Control

- ▶ **Source:** keep track of original document sources, keep real-time access for displaying (file links)
- ▶ **Temporal and geographical relevance:** timestamps and locations to be considered at retrieval time
- ▶ **Confidentiality Level:** relevant for access control and compliance in document sources
- ▶ **Role-Based Access Control (RBAC):** make sure to keep permission levels on internal documents (public ones are easier to manage)

SUMMETIX IRIS

Case Study



SUMMETIX IRIS

Antwort erstellen

Antwort Tickets Metadaten Referenzdokumente

Antwort ist bereit zum Bearbeiten

Sprache DE

Referenz-ID

Betreff: Verlängerung des Anwohnerparkausweises

Sehr geehrter Herr Mustermann,

vielen Dank für Ihre Anfrage zur Verlängerung Ihres Anwohnerparkausweises für das Kennzeichen OF XX 123.

Um Ihren Anwohnerparkausweis zu verlängern, müssen Sie einige Schritte befolgen:

- Antrag auf Verlängerung:** Bitte füllen Sie den Antrag auf Verlängerung im Anhang aus.
- Gebühren:** Beachten Sie, dass für die Verlängerung des Anwohnerparkausweises eine Gebühr anfällt. Die genauen Kosten entnehmen Sie bitte der aktuellen Gebührenordnung (ebenfalls im Anhang).
- Bearbeitungszeit:** Nach Einreichung Ihres Antrags kann die Bearbeitung einige Tage in Anspruch nehmen. Sie erhalten eine Bestätigung, sobald Ihr Antrag bearbeitet wurde.

Bei weiteren Fragen oder wenn Sie Unterstützung benötigen, stehen wir Ihnen gerne zur Verfügung. Mit freundlichen Grüßen
Ihr Team von der Stadtverwaltung

Sehr geehrte Damen und Herren, ich (Max Mustermann, geb. 01.03.1981) möchte meinen Anwohnerparkausweis für das Kennzeichen OF XX 123 verlängern. Wie kann ich das tun?

Beste Grüße,
Max Mustermann

Antworten werden mit KI erstellt und sollten vor dem Versand immer auf Korrektheit überprüft werden.

Suche nach Titel, Inhalt, Tags...

Auto Categorized Collection Recycling Waste

Mar 17, 2026 Eigentümer

Sehr geehrter Herr Mustermann, vielen Dank für Ihre Anfrage zur Verlängerung Ihres Anwohnerparkausweises für das Kennzeichen OF XX 123. Um Ihren Anwohnerparkausweis zu verlängern, müssen Sie einige Schritte befolgen: 1. **Antrag auf Verlängerung**: Bitte füllen Sie den Antrag auf Verlängerung ...

Sehr geehrte Damen und Herren, ich (Max Mustermann, geb. 01.03.1981) möchte meinen Anwohnerparkausweis für das Kennzeichen OF XX 123 verlängern. Wie k...

Verlängerung Bewohnerparkausweis

Mar 04, 2026 Eigentümer

Sehr geehrter Herr Mustermann, vielen Dank für Ihre Anfrage bezüglich der Abholung Ihres neuen Personalausweises durch Ihre Frau. Um eine Vollmacht für die Abholung zu erstellen, können Sie eine einfache Vorlage verwenden. Diese sollte folgende Informationen enthalten: 1. Ihren vollständigen Namen ...

Liebes Team, ich (Max Mustermann, geb. 01.03.1981) muss meinen neuen Führerschein abholen, kann aber gerade nicht selbst vorbeikommen. Meine Frau Mart...

Gruppen hinzufügen

1 2 3 4 5 ... 10 >

The IRIS – Process

A continuously improving process



Incoming Request



Identification of already answered queries.

Combination with other data sources such as:

- past Emails
- manuals
- websites

Responses and queries are added to the database.

This way, the system learns and keeps improving the generated answers.



Creation of a **new text proposal** by SUMMETIX IRIS within seconds.



The **user edits and saves the text** and forwards it to the citizens.



Approval processes (e.g., between manager and department director) can be implemented.



Document Search with IRIS

Why not use an arbitrary LLM?

Security

- ▶ Only **defined content** is searched
- ▶ No AI **Training**
- ▶ 100% hosted in **Germany**

Quality

- ▶ Document **parsers** adjusted to sources (website, PDF, E-Mail etc.)
- ▶ Fast and continuous **updates of data**
- ▶ **Specialized language models** for search/retrieval

Process Integration

- ▶ Link to **defined processes** in the background
- ▶ **Human-in-the-loop approach** to prevent losing control

Business Case

Citizen Services in Public Administration

Case studies showed that with IRIS, emails can be answered much faster and more comprehensively, reducing processing time **from 10–15 minutes to just one minute.**

Business case

- ➔ approx. **94% cost saving** on this proportion of work
- ➔ Net effect: 5-digit savings per licence annually, **approx. €300k+ per year** with 30 users (mid size city)

Assumptions

~30% time savings, 20% of an employee's time is spent with responding to requests

Quotes

“With IRIS, response times can be significantly reduced.
Emails can be answered within one minute instead of 10–15 minutes — while improving quality and consistency.”

Jörn Büttner



Head of Traffic and Mobility Planning
City of Aschaffenburg

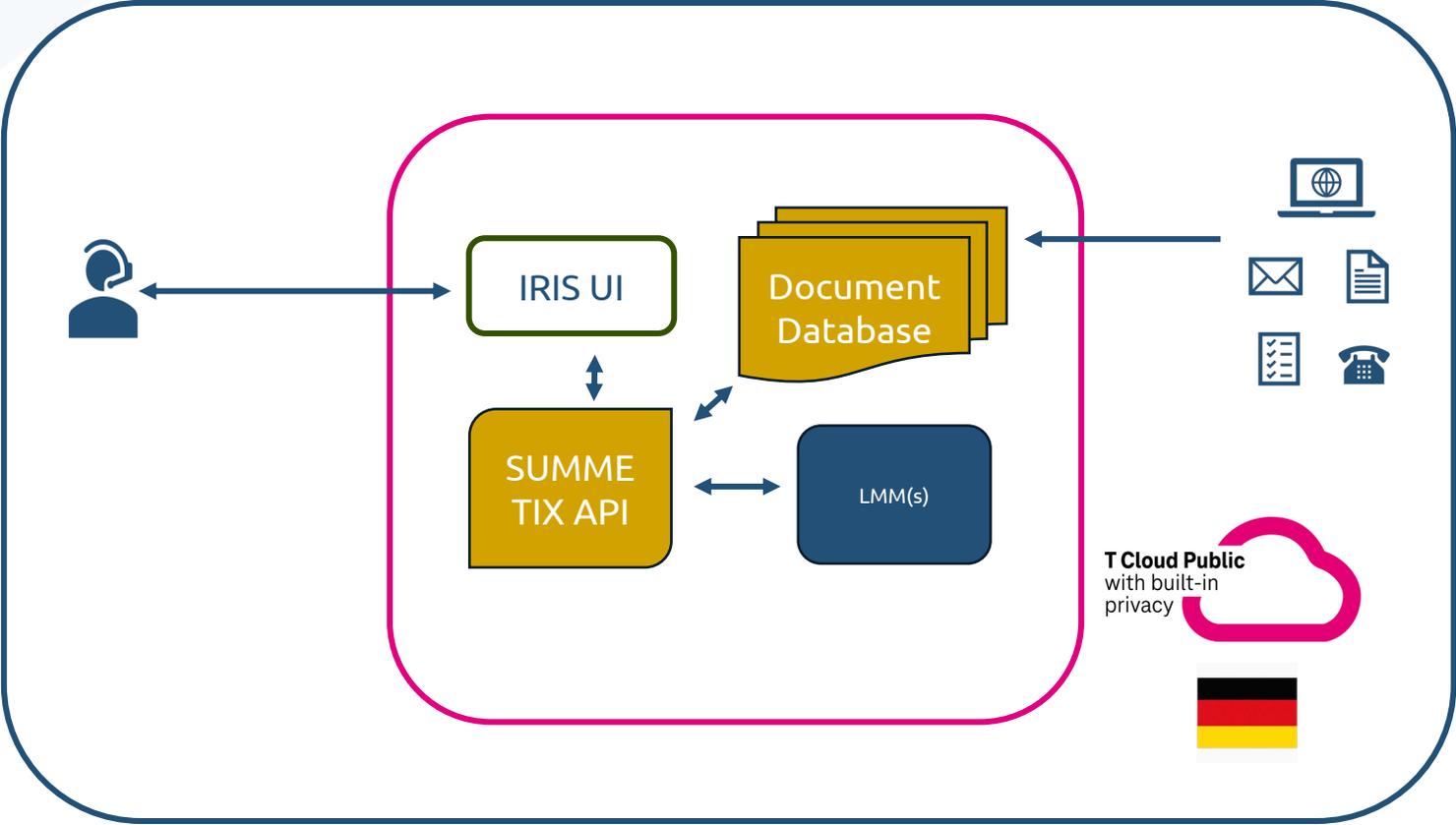
“IRIS shows how digitalization can be implemented responsibly and in a practical way — for the benefit of our employees and the entire community.”

Ludwig Horn



Mayor
City of Tutzing

Deploying IRIS



- ✓ Physically hosted LLMs in **Germany**
- ✓ On servers of **German companies** (Telekom)
- ✓ **No external** communication
- ✓ **No training** of AI systems with the entered data

Integrating IRIS



Forward requests to IRIS from your inbox

Admin staff answers requests with the help of IRIS

Tracking of issues in the CRM

“Deterministic Agents”

Connecting RAG and AI-based Automation

- ▶ Integrating RAG with AI-supported processes
- ▶ Like an agent, but predictable and only for defined processes
- ▶ Connects defined internal processes to external communication

Integrating IRIS with AI-supported processing



SUMMETIX IRIS



Citizen: Request

Text example:

Dear Ladies and Gentlemen, I (John Doe, born 01.03.1981) would like to extend my residents' parking permit for the license plate OF XX 123. How can I do this?

Best regards,
John Doe

IRIS: Communication, Process Detection

IRIS analyzes and assigns emails, extracts metadata, and determines relevant processes:

Person: Max Mustermann born 01.03.1981
Process: Renewal Residents' parking permit

Execution of Workflows

Receives the data from IRIS and triggers the appropriate process.

Hands over the finalized documents or missing information to IRIS

Generates feedback to the citizen based on external actions





Erik Kaiser / Johannes Daxenberger

Managing Directors

summetix GmbH | Frohsinnstraße 32, 63739 Aschaffenburg
Amtsgericht Aschaffenburg; HRB 16040

Bildnachweis: Icons erstellt von Freepik - Flaticon

- <https://www.flaticon.com/de/kostenlose-icons/feder>
- <https://www.flaticon.com/de/kostenlose-icons/flug>
- <https://www.flaticon.com/de/kostenlose-icons/ziel>

info@summetix.com



www.summetix.com