
# How hard can it be to build a customer support chatbot?
# - Lessons learned

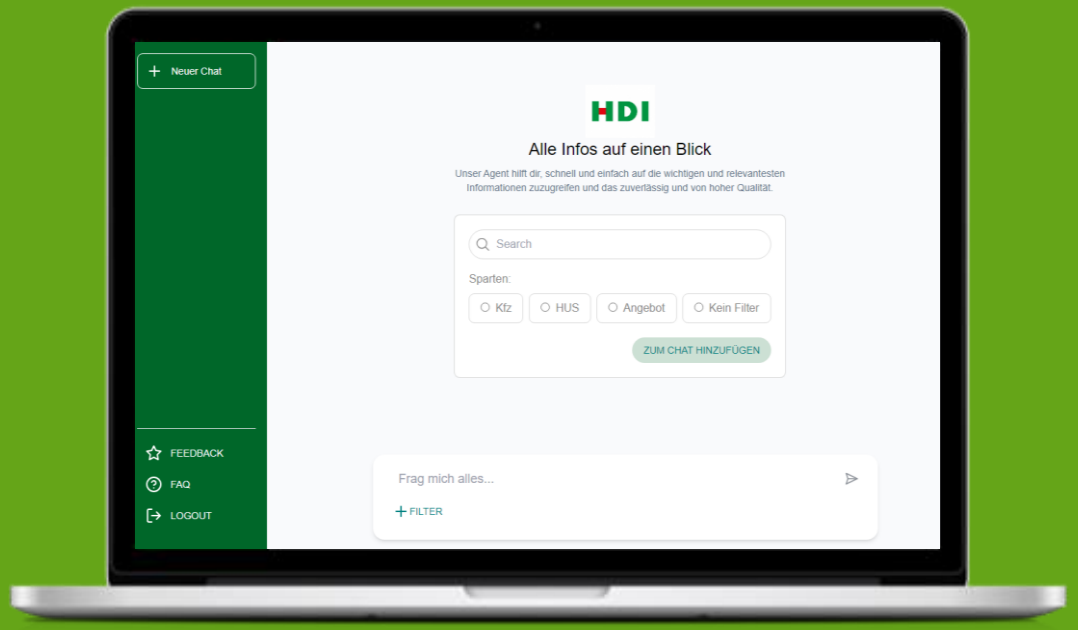
AWS Deep Dive Days – Generative AI

Leonard Zucht (HDI), Elias Stiegelmeier (HDI), Akarsha Sehwag (AWS)

Munich | 20. & 21.02.2024

# The vision – „Alle Infos auf einen Blick“

- **Knowledge-based customer service**
- **Fast and flexible access to information**
- **Chatbot-interface**
- **Start with internal-only scope and scale**



# Retrieval Augmented Generation (RAG)


Question

Information Search

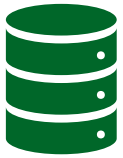
Question & Context

Large Language Model

Answer

# Setting up a first prototype is easy

Langchain

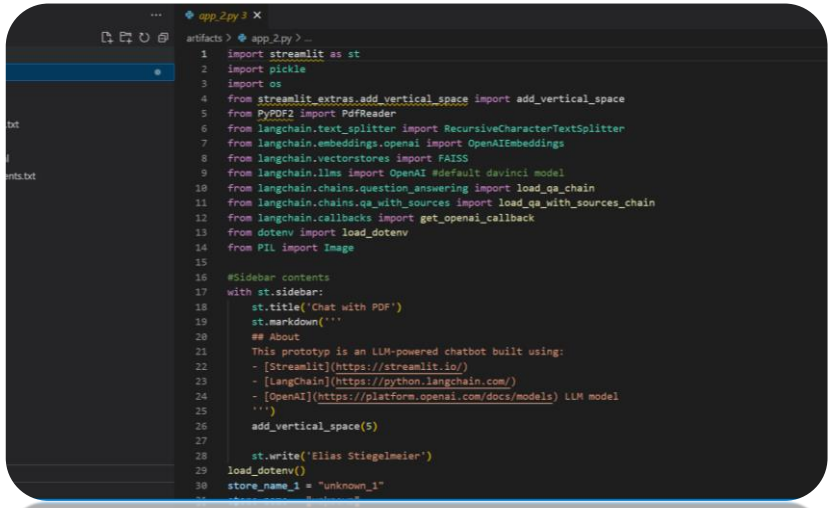


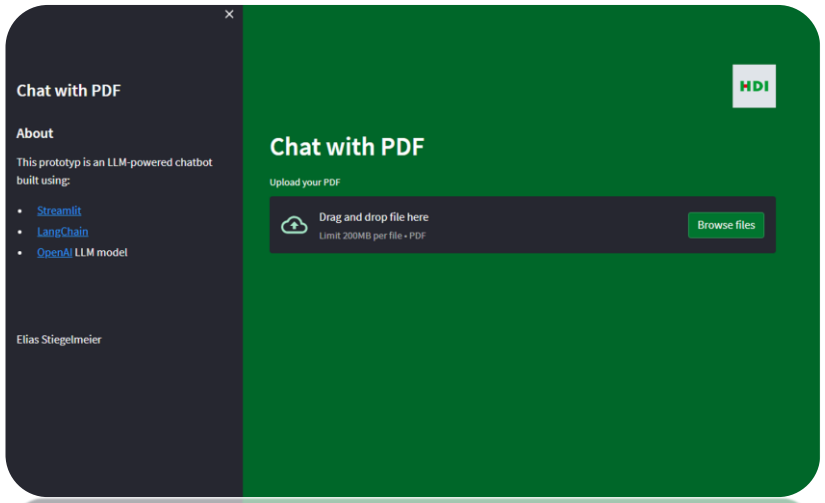
FAISS

Streamlit

https://www.langchain.com/
https://ai.meta.com/tools/faiss/
https://streamlit.io/

# Setting up a first prototype is easy



# Setting up a first prototype is easy

## The next steps

- Scalability
- Professional frontend
- Improved document preprocessing for tables and layout
- Modularize solution for optimization

# We need a product team

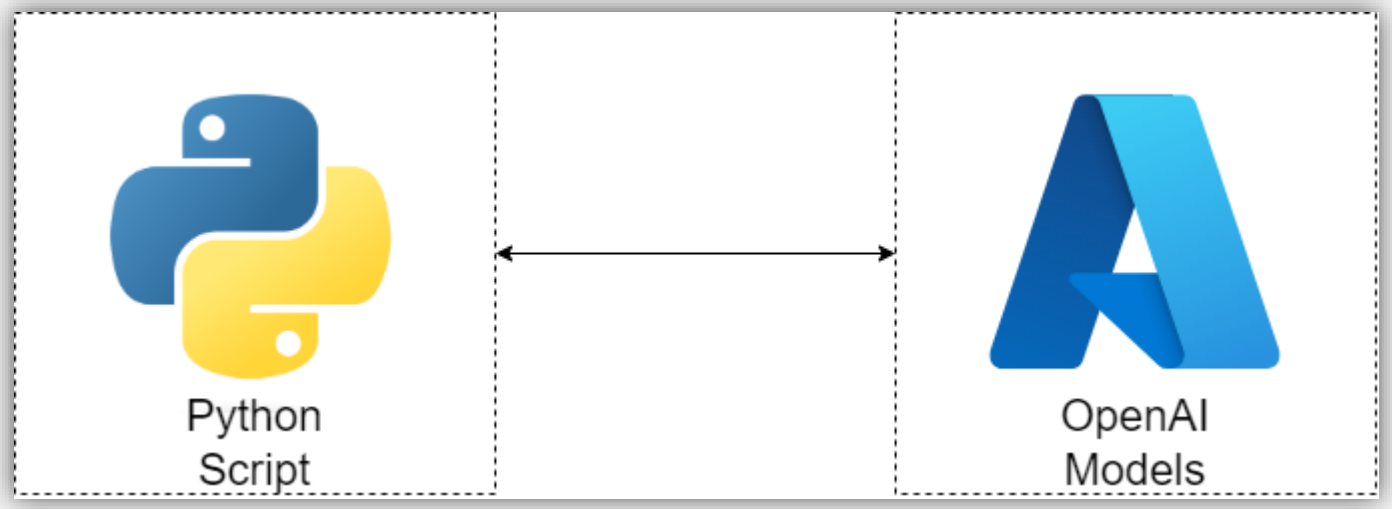


Build a scalable solution which solves the limitiations and can be adapted in a modular manner

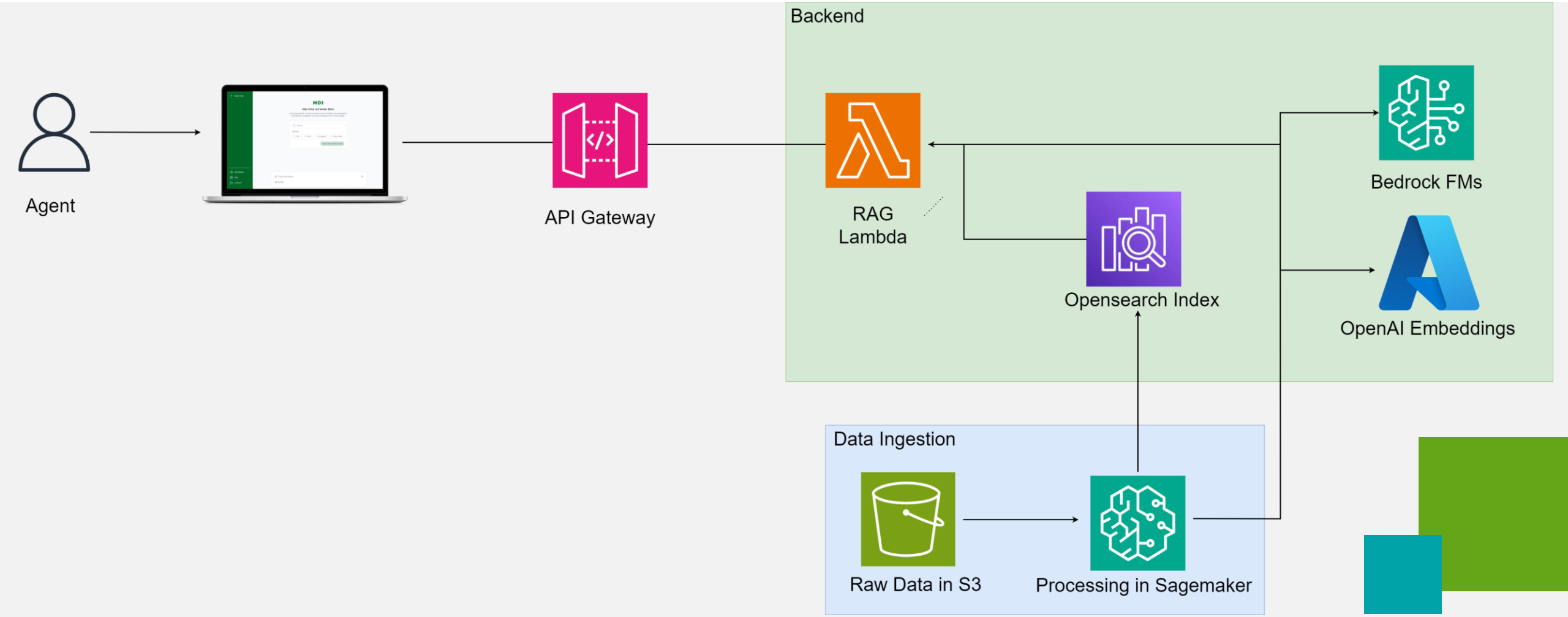
# The journey

| | Streamlit | Blogpost | AWS artifact | Own application |
|---|---|---|---|---|
| ***Backend*** | | | | |
| | Simple ingest | Simple ingest | OCR-based ingest | Adapted OCR-based ingest |
| | In-memory retrieval | OpenSearch-based retrieval | OpenSearch-based retrieval | Adapted OpenSearch -based retrieval |
| | Local | Public-facing | Public-facing | HDI-internal VPC |
| ***Frontend*** | | | | |
| | Simple design (Streamlit) | Simple design (Gradio) | Simple design (Streamlit) | Professional design (React) Streaming enabled |

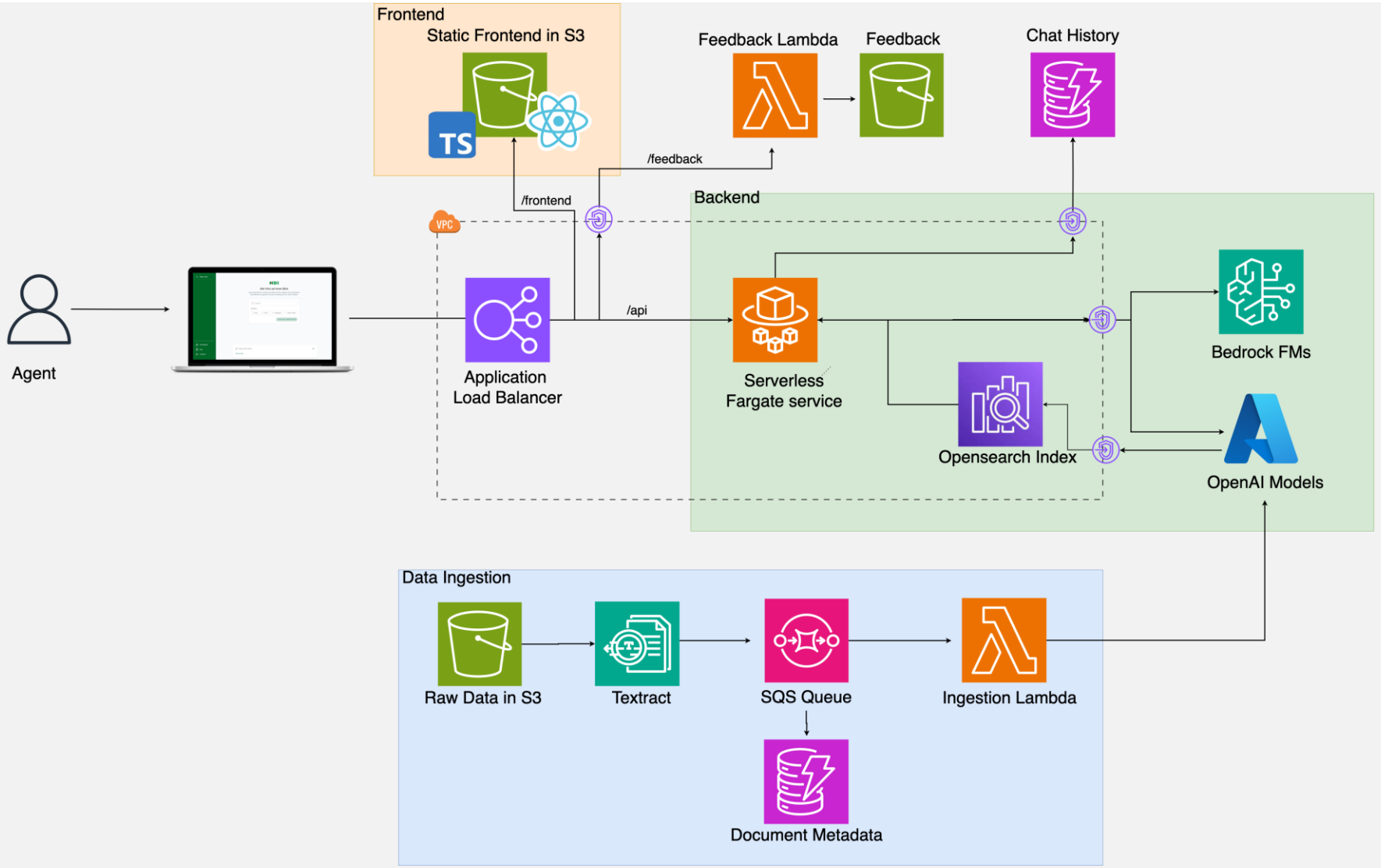
# The journey - Streamlit



# The journey - blogpost



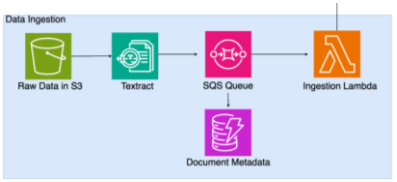
# The journey – own application



# There are many moving parts

| Challenges | Contents | Approaches |
|---|---|---|
| Complex document parsing | Tables, layout, pictures | OCR, ML-Parser |
| Chunking | Columns, paragraphs, lists, tables | Layout |
| Domain knowledge | Domain language, abbreviations, conventions | Prompt engineering, look-ups, finetuning |
| Variable parameters | Top-k, chunk-size, relevance, overlap, temperature, … | Performance test |
| Document similarity | Retrieval, embedding, sources | Filter, hybrid-search, reranking |
| Performance | Recall, answer quality | Performance test, advanced prompting |
| User intent | Information search, summarization, follow-up questions | Prompt engineering, routing, agents / tools |

# There are many moving parts

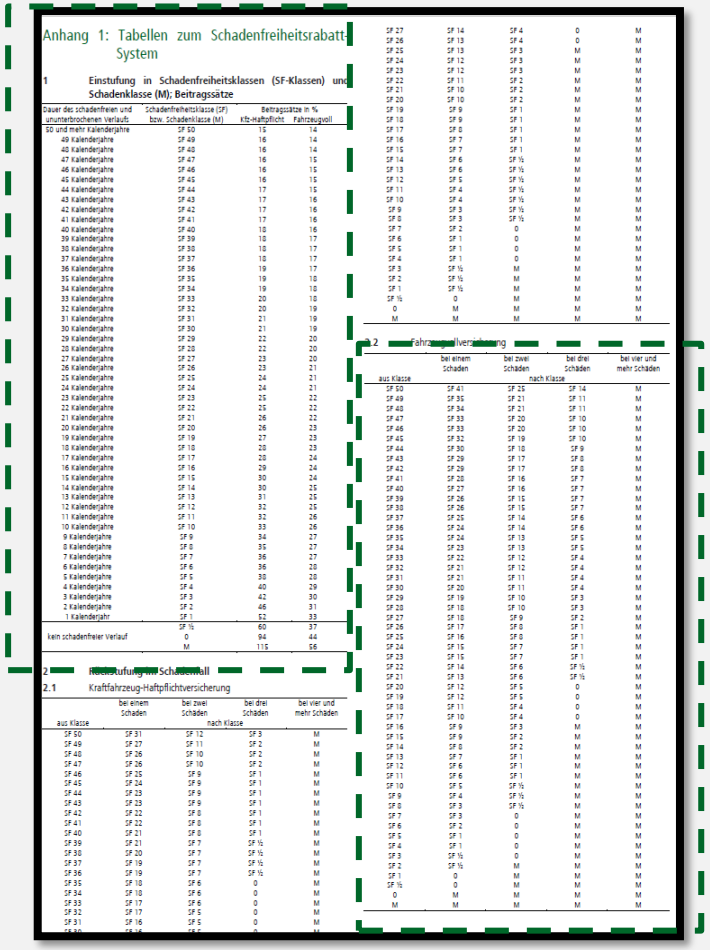



Pre-processing

Retrieval

| PREPROCESSING | | | | | | | | | | | RETRIEVAL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PARSING | | | | CHUNKING | | | | | | | RETRIEVAL | | | | | | | | | |
| use_layout | process_tables | duplicate_text_in_merged_cells | copy_headers_in_each_cell | fix_slicing | add_metadata | splitter | target_chunk_size | chunk_overlap | max_chunk_size | embeddings_model_id | index | num_docs | doc_length | rewrite_query | rerank_docs | rerank_service | rerank_service | u_shaped | hide_met |

# Spotlight – advanced document processing



Anhang 1: Tabellen zum Schadenfreiheitsrabatt-System

1 Einstufung in Schadenfreiheitsklassen (SF-Klassen) und Schadenklasse (M); Beitragssätze

| Dauer des schadenfreien und ununterbrochenen Verlaufs | Schadenfreiheitsklasse (SF) bzw. Schadenklasse (M) | Beitragssätze in % Kfz-Haftpflicht | Fahrzeugvoll |
|---|---|---|---|
| 50 und mehr Kalenderjahre | SF 50 | 15 | 14 |
| 49 Kalenderjahre | SF 49 | 16 | 14 |
| 48 Kalenderjahre | SF 48 | 16 | 14 |
| 47 Kalenderjahre | SF 47 | 16 | 15 |
| 46 Kalenderjahre | SF 46 | 16 | 15 |
| 45 Kalenderjahre | SF 45 | 16 | 15 |
| 44 Kalenderjahre | SF 44 | 17 | 15 |
| 43 Kalenderjahre | SF 43 | 17 | 16 |
| 42 Kalenderjahre | SF 42 | 17 | 16 |
| 41 Kalenderjahre | SF 41 | 17 | 16 |
| 40 Kalenderjahre | SF 40 | 18 | 16 |
| 39 Kalenderjahre | SF 39 | 18 | 17 |
| 38 Kalenderjahre | SF 38 | 18 | 17 |
| 37 Kalenderjahre | SF 37 | 18 | 17 |
| 36 Kalenderjahre | SF 36 | 19 | 17 |
| 35 Kalenderjahre | SF 35 | 19 | 18 |
| 34 Kalenderjahre | SF 34 | 19 | 18 |
| 33 Kalenderjahre | SF 33 | 20 | 18 |
| 32 Kalenderjahre | SF 32 | 20 | 19 |
| 31 Kalenderjahre | SF 31 | 21 | 19 |
| 30 Kalenderjahre | SF 30 | 21 | 19 |
| 29 Kalenderjahre | SF 29 | 22 | 20 |
| 28 Kalenderjahre | SF 28 | 22 | 20 |
| 27 Kalenderjahre | SF 27 | 23 | 20 |
| 26 Kalenderjahre | SF 26 | 23 | 21 |
| 25 Kalenderjahre | SF 25 | 24 | 21 |
| 24 Kalenderjahre | SF 24 | 24 | 21 |
| 23 Kalenderjahre | SF 23 | 25 | 22 |
| 22 Kalenderjahre | SF 22 | 25 | 22 |
| 21 Kalenderjahre | SF 21 | 26 | 22 |
| 20 Kalenderjahre | SF 20 | 26 | 23 |
| 19 Kalenderjahre | SF 19 | 27 | 23 |
| 18 Kalenderjahre | SF 18 | 28 | 23 |
| 17 Kalenderjahre | SF 17 | 28 | 24 |
| 16 Kalenderjahre | SF 16 | 29 | 24 |
| 15 Kalenderjahre | SF 15 | 30 | 24 |
| 14 Kalenderjahre | SF 14 | 30 | 25 |
| 13 Kalenderjahre | SF 13 | 31 | 25 |
| 12 Kalenderjahre | SF 12 | 32 | 25 |
| 11 Kalenderjahre | SF 11 | 32 | 26 |
| 10 Kalenderjahre | SF 10 | 33 | 26 |
| 9 Kalenderjahre | SF 9 | 34 | 27 |
| 8 Kalenderjahre | SF 8 | 35 | 27 |
| 7 Kalenderjahre | SF 7 | 36 | 27 |
| 6 Kalenderjahre | SF 6 | 36 | 28 |
| 5 Kalenderjahre | SF 5 | 38 | 28 |
| 4 Kalenderjahre | SF 4 | 40 | 29 |
| 3 Kalenderjahre | SF 3 | 42 | 30 |
| 2 Kalenderjahre | SF 2 | 46 | 31 |
| 1 Kalenderjahr | SF 1 | 52 | 33 |
| | SF ½ | 60 | 37 |
| kein schadenfreier Verlauf | 0 | 94 | 44 |
| | M | 115 | 56 |

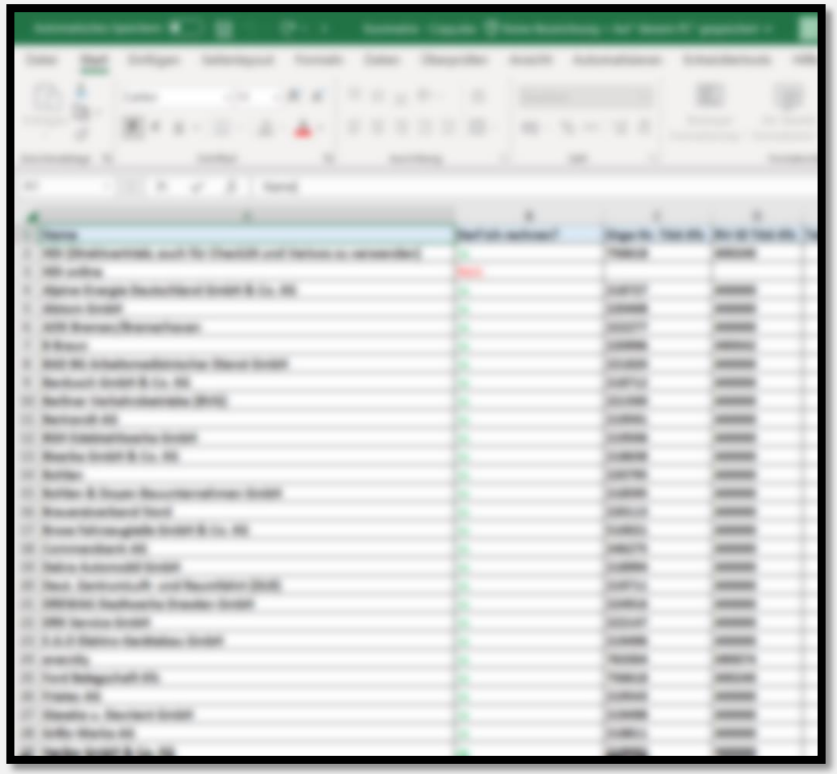
2 Rückstufung im Schadenfall

2.1 Kraftfahrzeug-Haftpflichtversicherung

| aus Klasse | bei einem Schaden | bei zwei Schäden | bei drei Schäden | bei vier und mehr Schäden |
|---|---|---|---|---|
| | nach Klasse | | | |
| SF 50 | SF 31 | SF 12 | SF 3 | M |
| SF 49 | SF 27 | SF 11 | SF 2 | M |
| SF 48 | SF 26 | SF 10 | SF 2 | M |
| SF 47 | SF 26 | SF 10 | SF 2 | M |
| SF 46 | SF 25 | SF 9 | SF 1 | M |
| SF 45 | SF 24 | SF 9 | SF 1 | M |
| SF 44 | SF 23 | SF 9 | SF 1 | M |
| SF 43 | SF 23 | SF 9 | SF 1 | M |
| SF 42 | SF 22 | SF 8 | SF 1 | M |
| SF 41 | SF 22 | SF 8 | SF 1 | M |
| SF 40 | SF 21 | SF 8 | SF 1 | M |
| SF 39 | SF 21 | SF 7 | SF ½ | M |
| SF 38 | SF 20 | SF 7 | SF ½ | M |
| SF 37 | SF 19 | SF 7 | SF ½ | M |
| SF 36 | SF 19 | SF 7 | SF ½ | M |
| SF 35 | SF 18 | SF 6 | 0 | M |
| SF 34 | SF 18 | SF 6 | 0 | M |
| SF 33 | SF 17 | SF 6 | 0 | M |
| SF 32 | SF 17 | SF 5 | 0 | M |
| SF 31 | SF 16 | SF 5 | 0 | M |
| SF 30 | SF 15 | SF 5 | 0 | M |
| SF 29 | SF 15 | SF 5 | 0 | M |
| SF 28 | SF 14 | SF 5 | 0 | M |
| SF 27 | SF 14 | SF 4 | 0 | M |
| SF 26 | SF 13 | SF 4 | 0 | M |
| SF 25 | SF 13 | SF 3 | M | M |
| SF 24 | SF 12 | SF 3 | M | M |
| SF 23 | SF 12 | SF 3 | M | M |
| SF 22 | SF 11 | SF 2 | M | M |
| SF 21 | SF 10 | SF 2 | M | M |
| SF 20 | SF 10 | SF 2 | M | M |
| SF 19 | SF 9 | SF 1 | M | M |
| SF 18 | SF 9 | SF 1 | M | M |
| SF 17 | SF 8 | SF 1 | M | M |
| SF 16 | SF 7 | SF 1 | M | M |
| SF 15 | SF 7 | SF 1 | M | M |
| SF 14 | SF 6 | SF ½ | M | M |
| SF 13 | SF 6 | SF ½ | M | M |
| SF 12 | SF 5 | SF ½ | M | M |
| SF 11 | SF 4 | SF ½ | M | M |
| SF 10 | SF 4 | SF ½ | M | M |
| SF 9 | SF 3 | SF ½ | M | M |
| SF 8 | SF 3 | SF ½ | M | M |
| SF 7 | SF 2 | 0 | M | M |
| SF 6 | SF 1 | 0 | M | M |
| SF 5 | SF 1 | 0 | M | M |
| SF 4 | SF 1 | 0 | M | M |
| SF 3 | SF ½ | M | M | M |
| SF 2 | SF ½ | M | M | M |
| SF 1 | SF ½ | M | M | M |
| SF ½ | 0 | M | M | M |
| 0 | M | M | M | M |
| M | M | M | M | M |

2.2 Fahrzeugvollversicherung

| aus Klasse | bei einem Schaden | bei zwei Schäden | bei drei Schäden | bei vier und mehr Schäden |
|---|---|---|---|---|
| | nach Klasse | | | |
| SF 50 | SF 41 | SF 25 | SF 14 | M |
| SF 49 | SF 35 | SF 21 | SF 11 | M |
| SF 48 | SF 34 | SF 21 | SF 11 | M |
| SF 47 | SF 33 | SF 20 | SF 10 | M |
| SF 46 | SF 33 | SF 20 | SF 10 | M |
| SF 45 | SF 32 | SF 19 | SF 10 | M |
| SF 44 | SF 30 | SF 18 | SF 9 | M |
| SF 43 | SF 29 | SF 17 | SF 8 | M |
| SF 42 | SF 29 | SF 17 | SF 8 | M |
| SF 41 | SF 28 | SF 16 | SF 7 | M |
| SF 40 | SF 27 | SF 16 | SF 7 | M |
| SF 39 | SF 26 | SF 15 | SF 7 | M |
| SF 38 | SF 26 | SF 15 | SF 7 | M |
| SF 37 | SF 25 | SF 14 | SF 6 | M |
| SF 36 | SF 24 | SF 14 | SF 6 | M |
| SF 35 | SF 24 | SF 13 | SF 5 | M |
| SF 34 | SF 23 | SF 13 | SF 5 | M |
| SF 33 | SF 22 | SF 12 | SF 4 | M |
| SF 32 | SF 21 | SF 12 | SF 4 | M |
| SF 31 | SF 21 | SF 11 | SF 4 | M |
| SF 30 | SF 20 | SF 11 | SF 4 | M |
| SF 29 | SF 19 | SF 10 | SF 3 | M |
| SF 28 | SF 18 | SF 10 | SF 3 | M |
| SF 27 | SF 18 | SF 9 | SF 2 | M |
| SF 26 | SF 17 | SF 8 | SF 1 | M |
| SF 25 | SF 16 | SF 8 | SF 1 | M |
| SF 24 | SF 15 | SF 7 | SF 1 | M |
| SF 23 | SF 15 | SF 7 | SF 1 | M |
| SF 22 | SF 14 | SF 6 | SF ½ | M |
| SF 21 | SF 13 | SF 6 | SF ½ | M |
| SF 20 | SF 12 | SF 5 | 0 | M |
| SF 19 | SF 12 | SF 5 | 0 | M |
| SF 18 | SF 11 | SF 4 | 0 | M |
| SF 17 | SF 10 | SF 4 | 0 | M |
| SF 16 | SF 9 | SF 3 | M | M |
| SF 15 | SF 9 | SF 2 | M | M |
| SF 14 | SF 8 | SF 2 | M | M |
| SF 13 | SF 7 | SF 1 | M | M |
| SF 12 | SF 6 | SF 1 | M | M |
| SF 11 | SF 6 | SF 1 | M | M |
| SF 10 | SF 5 | SF ½ | M | M |
| SF 9 | SF 4 | SF ½ | M | M |
| SF 8 | SF 3 | SF ½ | M | M |
| SF 7 | SF 3 | 0 | M | M |
| SF 6 | SF 2 | 0 | M | M |
| SF 5 | SF 1 | 0 | M | M |
| SF 4 | SF 1 | 0 | M | M |
| SF 3 | SF ½ | 0 | M | M |
| SF 2 | SF ½ | M | M | M |
| SF 1 | 0 | M | M | M |
| SF ½ | 0 | M | M | M |
| 0 | M | M | M | M |
| M | M | M | M | M |



# Spotlight – third party dependencies

## Challenges

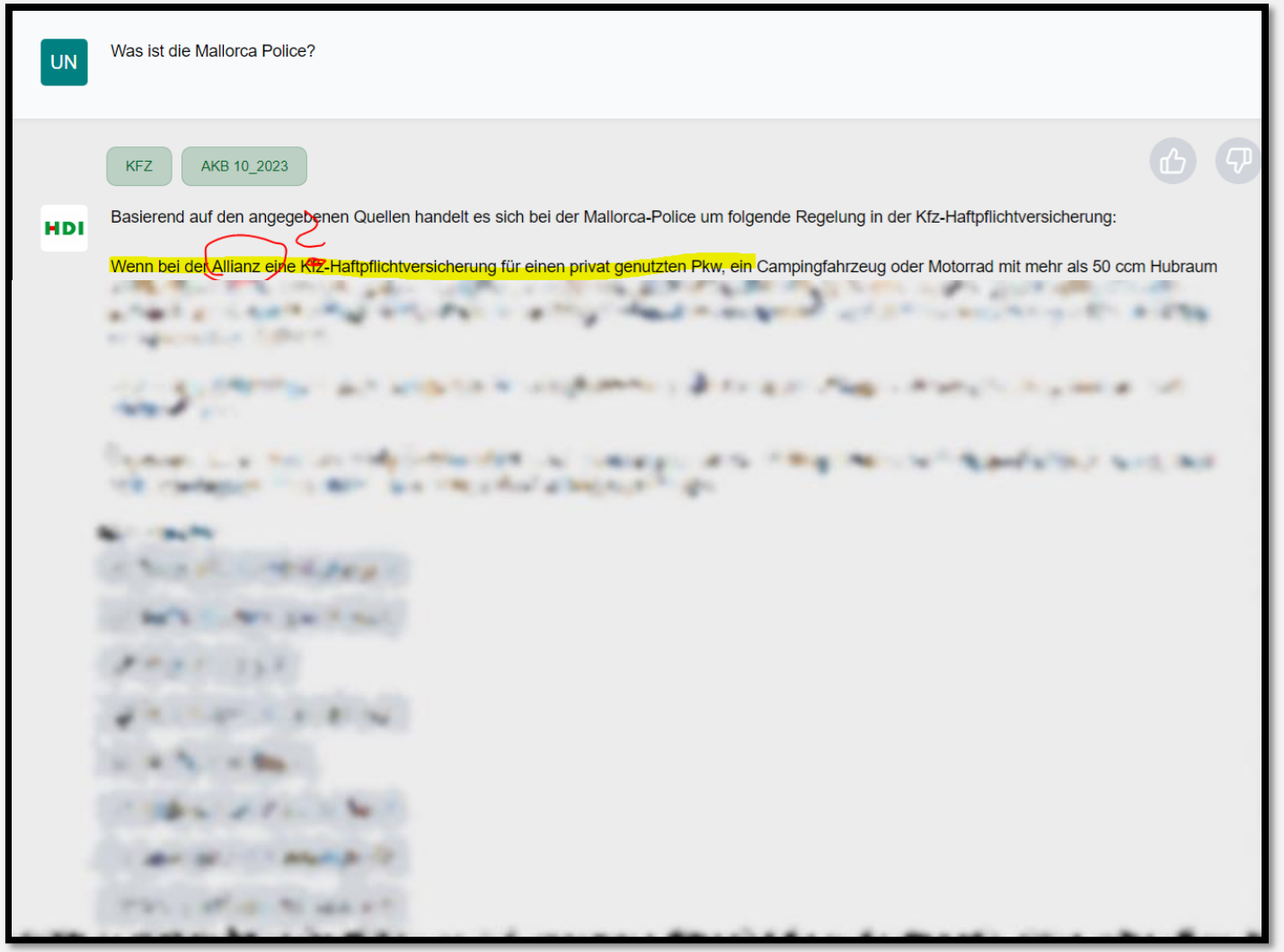
- Inconsistent
- Incomplete documentation
- Little transparency e.g. for timeouts

## Learnings

- Not everything popular is good.
- Rather prefer low level code

(customizations and flexibility as needed)

# Spotlight - one prompt to rule them all?



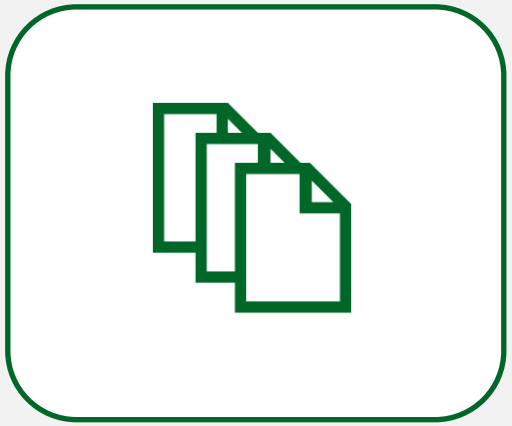
# Spotlight - one prompt to rule them all?

- Different system prompt can be useful for questions with differing complexity
- System prompt sometimes does not work as assumed
- How do you want your answers to look like?
- What should be included in the system prompt?
- Do you need to prevent prompt injection?

# Spotlight - one prompt to rule them all?

- Different system prompt can be useful for questions with differing complexity
- System prompt sometimes does not work as assumed
- How do you want your **Compromise**
- What should be included in the system prompt?
- Do you need to prevent prompt injection?

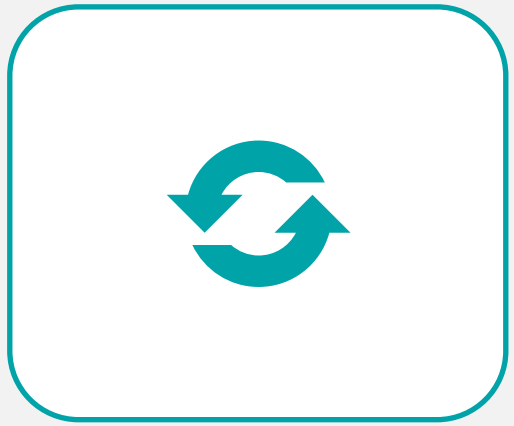
# How to test our solution? - The setup



**Ground truth**

(Questions, Sources, Answers)

**Business Testers**

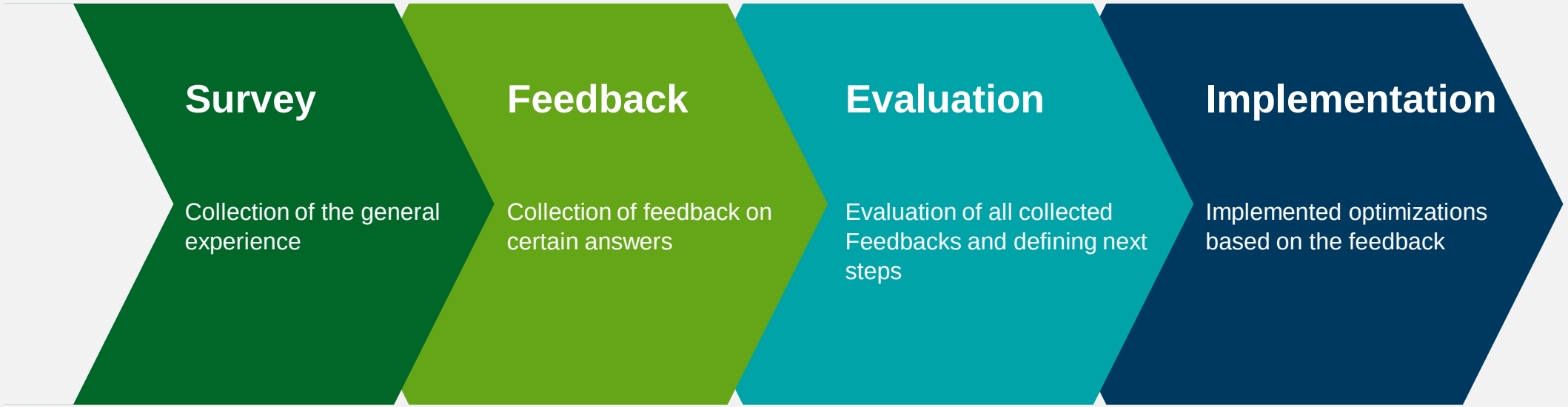


**Feedback Loop**

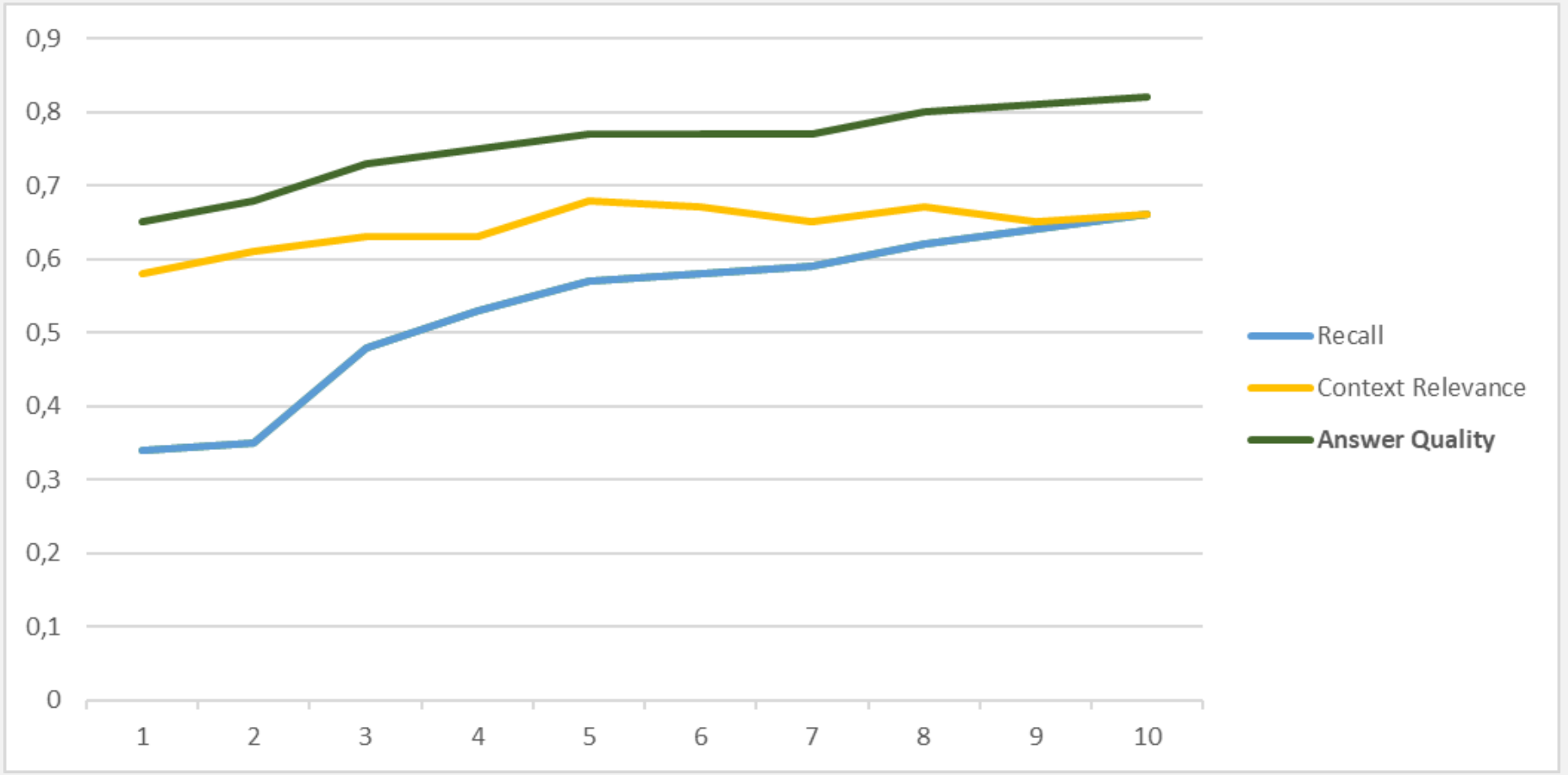
**&**

**Incremental Improvements**

# How to test our solution? - Feedback loop



# Our results (so far)

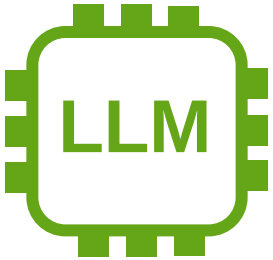


# Key take-aways

**It is not easy, it is a journey!**

- User feedback and evaluation matters
- Testing pipeline and experiments are helpful
- Start small and think big
- Modular approach for a fast-changing landscape
- Hallucinations need to be monitored to some extent.

# Outlook

| Improve data quality | Connection to further data | Finetuning and hosting of own models |
| --- | --- | --- |
|  |  |  |
| Database is refined | Relevant data | Models tailored to use-case |

# Thank you for your attention!

# Questions?