Optimizing a Retrieval-Augmented QA Chatbot for HR Support using LLMs

Alexander Kowsik, 20.11.2023, Kick-off Presentation

Lehrstuhl für Software Engineering betrieblicher Informationssysteme (sebis)
Fakultät für Informatik
Technische Universität München
wwwmatthes.in.tum.de

Outline

High HR workloads result in extensive manual labor and long delays

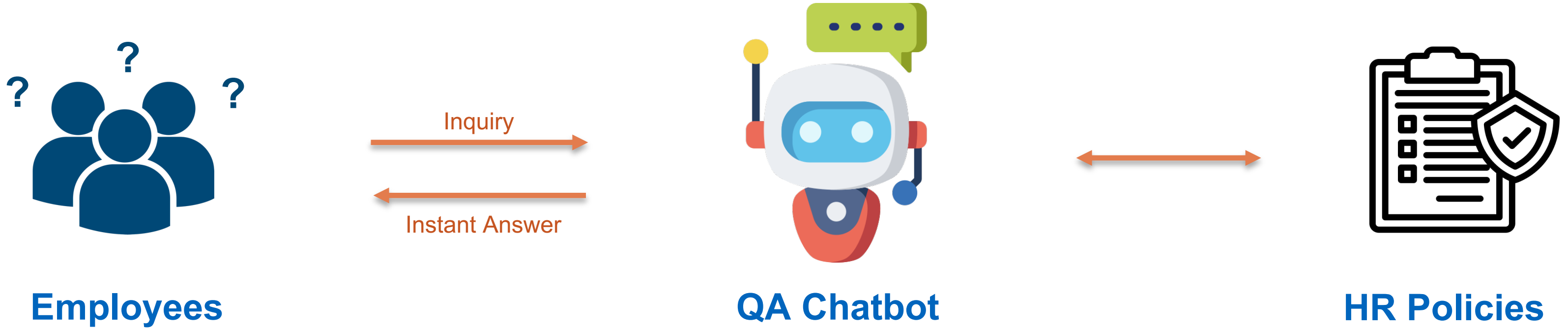
Problem 1: **Large volumes of HR inquiries**

- HR departments deal with *large quantities of daily tasks* and queries from employees
- *More than **330.000** HR tickets* per year are created at SAP
- To effectively manage this, *a substantial number of HR experts* are needed

Problem 2: **Manual and time-consuming process**

- Employee *questions must be manually processed and responded to* in accordance with HR rules and policies
- The result is *long waiting times*, ranging from hours to days or even weeks

QA chatbots reduce HR workloads by processing inquiries significantly more efficiently

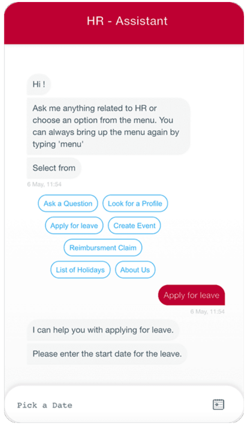


Benefit 1: Save time for both employees and the HR domain experts

- *QA chatbots provide immediate responses*, effectively eliminating any answer delays
- *Reduction of HR workload* allows the HR experts to focus on more important tasks
- Goal: *Replace 30% of HR tickets* with chatbot functionalities

Benefit 2: Automation of (mundane) manual tasks

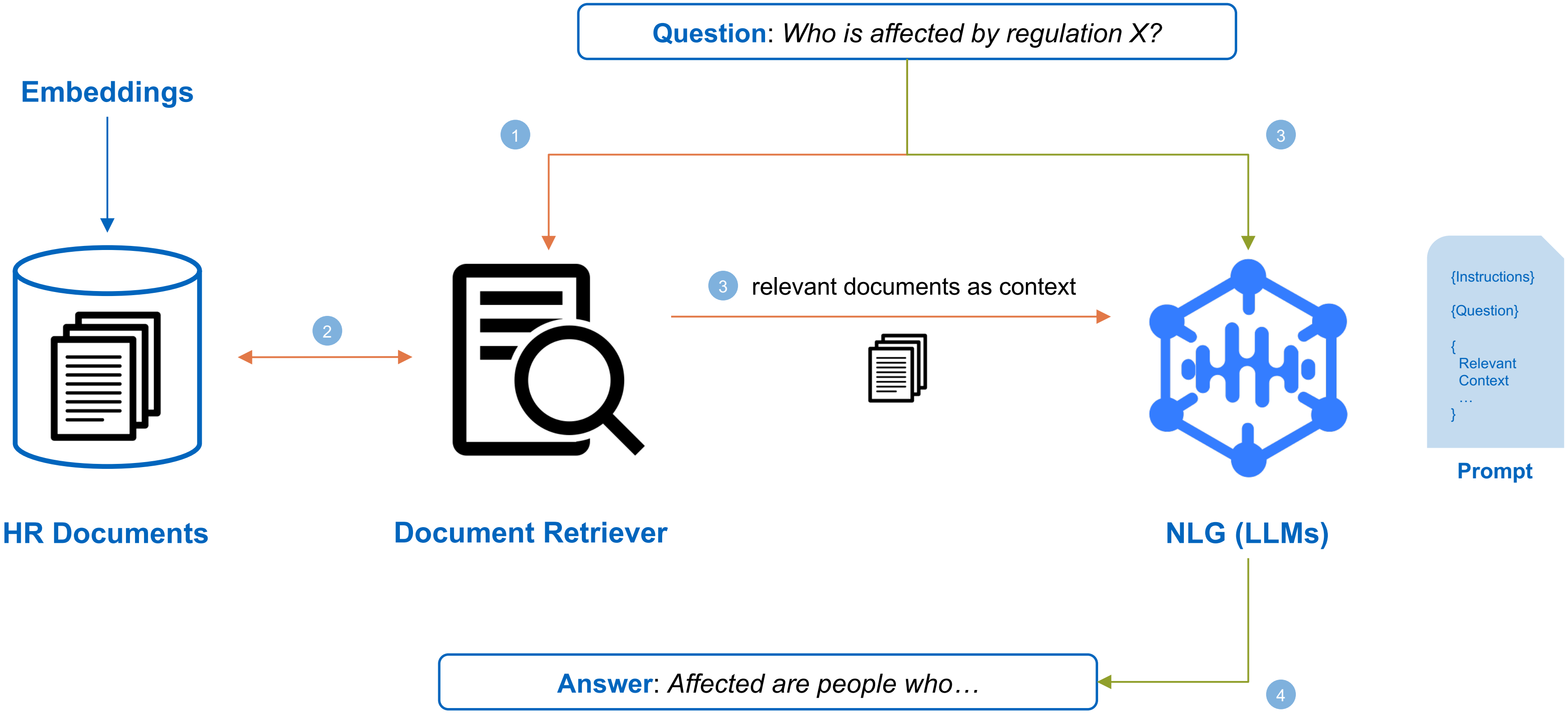
- The process of answering employee questions based on HR policies is *highly automatable using SOTA NLP models* (e.g., LLMs)
- Chatbots utilize the *HR rules and policies as grounding* for their responses



https://fpt.ai/chatbot-inevitable-trend-human-resources-2022

Outline

Retrieval-augmented Generation using LLMs – Most flexible and least limiting solution



Retrieval-augmented Generation using LLMs – Most flexible and least limiting solution

Document Retriever

Current solution: haystack by deepset — **Haystack DPR**
Fine-tuned

Proposed solution: **OpenAI Embeddings + Vector Search**
Better embeddings → better performance

Goal: Higher retrieval accuracy + better performance

- Removes need for fine-tuning DPR
- Embed new documents → include in vector search
- Better embeddings lead to better context retrieval
- Vector Database: Scalability, Hybrid Search
- Advanced retrieval methods: Query Transformations (Intended Topics, HyDE), Reranking, …

NLG (LLMs)

Current solution: T5 — **T5 / LongT5**
Fine-tuned

Proposed solution: **LLMs**: APIs (OpenAI) / Open-source
Prompting → flexible, more powerful
Conversational capabilities

Challenge: Get desired outputs

- Removes need for fine-tuning custom model
- Instruct LLM to return desired output
  → Prompt engineering / tuning
  → In-context learning
- Attach relevant context to ground responses
  → Prevent hallucinations

{Instructions}

{Question}

{
Relevant
Context
…
}

Prompt

A look at the **data**

User Question

“How can I request an unpaid leave?”

Relevant Context

…

If employees have used all annual leave entitlement and still need time of leave, they can request unpaid leave under below condition.

Eligibility

Permanent employees are eligible for unpaid leave or the company may order an
employee to take a leave of absence due to the following reasons:

* An employee cannot work for one week or more due to illness unrelated to work.

…

Model Response

To apply for Unpaid Leave, please follow the steps below:

1. Fill out the Unpaid Leave Application Form.
2. You need to obtain an approval from your manager and HR Business Partner.
3. Once you get the approval, please send the application form with the approval to HRdirect via ticket for the creation of unpaid leave quota in the system.
4. You need to provide an address and contact number so that you can be contacted during your leave of absence.

Description

- Data contains (1) curated FAQ training data and (2) cleaned actual user data from previous iterations
- Final dataset: merge of FAQ + user data
- >**89.000 data samples** (48k FAQ, rest user data)

Data improvements

- Data cleaning / correction, …
- Improve quality of training data with LLMs: summarize unnecessary long responses, improving grammar / output structure, …

Outline

Research Questions

1. For domain-specific use-cases, are **LLM Chatbot Systems** able to address the user queries as effectively as humans?

2. Can direct inference yield adequate results *without the need for fine-tuning*, and what prompt-tuning techniques can be used to **improve the quality of the responses**?

3. What methods can be used to optimize the **retrieval** when using **LLM embeddings** and a **vector database** in comparison to the current **DPR module**?

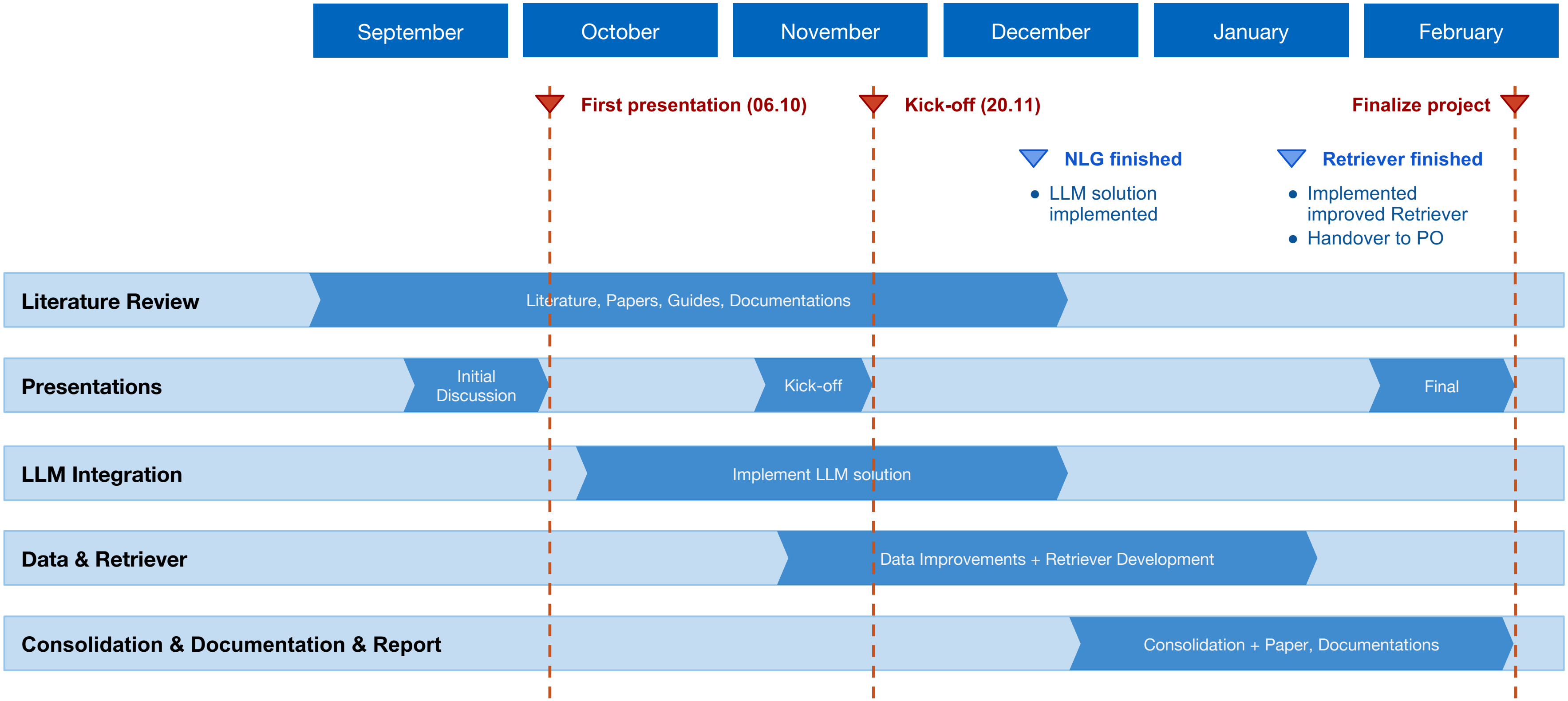
4. How can **LLMs** be utilized to improve the quality of the training data?

Outline

Approaches & Evaluation

1	**Literature Review**	**Literature, Papers**, Guides, Documentations, …
2	**Improve data** with LLMs	**Summarizations**, …
3	Replace **DPR** System	OpenAI **Embeddings**, **Vector Search**, Optimizations, …
4	Implement the **LLM-NLG** module	**Prompt-Engineering/-tuning**, Model / APIs Benchmarks, …
5	**Evaluation** (w/ Rajna)	Evaluate **components** and the **whole pipeline *end-to-end***
6	**Consolidation**	Combine all components into a **deployable system**, **Documentation, Handover, Report Paper,** …

Project Timeline



Prof. Dr.
Florian Matthes
Inhaber des Lehrstuhls

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Software Engineering
betrieblicher Informationssysteme

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289.17132
Fax +49.89.289.17136

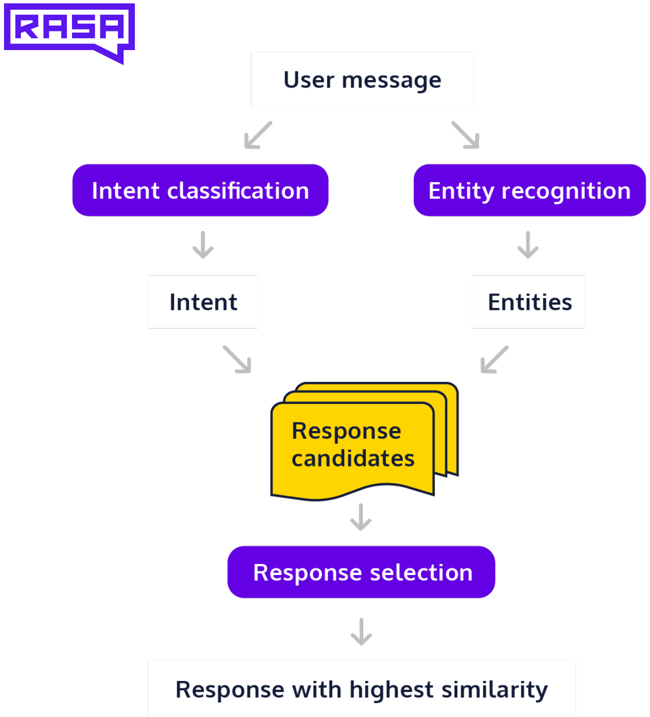
matthes@in.tum.de
wwwmatthes.in.tum.de

Thank you!

Any questions?

Backup

Traditional/Retrieval-based chatbots are limited and require extensive manual effort



https://www.codecademy.com/learn/retrieval-based-chatbots/modules/retrieval-based-chatbots/cheatsheet

Goal: Answer questions based on documents

Functionalities = Intents

Problems

- Manually designed, limited intents
- Not manageable and scalable
- Limited understanding of complex queries

Response Generation

Problems

- Rule-based and pre-defined responses
- No abstractive summarization of results