# A Human Assessment of Reference-Free and Reference-Based Evaluation Approaches in the HR Domain

Rajna Fani, 06.10.2023, Kick-Off Presentation

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

# Agenda

1 Motivation

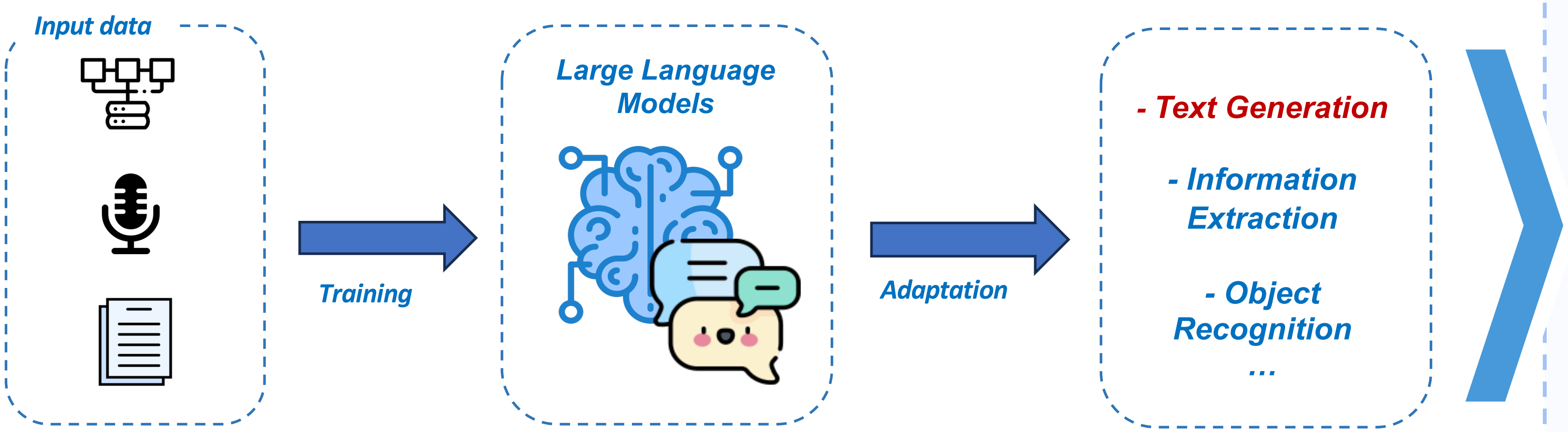
2 Use Case

3 Approaches

4 Research Questions

5 Timeline

# **Challenges in Evaluating Text Generative Models :** Motivation for Exploring Reference-Free Metrics



*Input data*

*Training*

*Large Language Models*

*Adaptation*

*- Text Generation*

*- Information Extraction*

*- Object Recognition*

*…*

## Evaluation

### Motivation

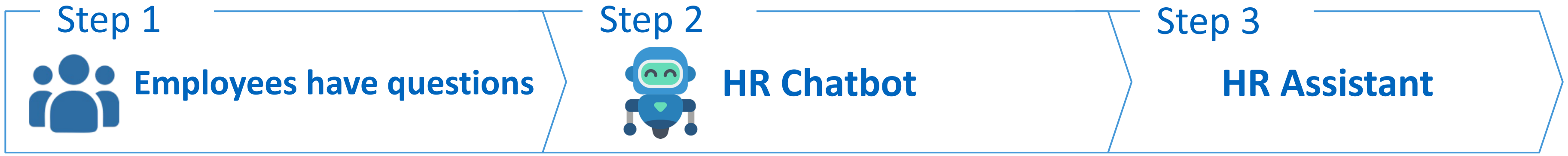
**A** Text Generation with LLMs is a trending task, but evaluating its quality poses a challenge

**B** Automatic Evaluation Metrics show limited effectiveness

### Investigation

- Explore the performance of Reference-Free State-of-the-Art Metrics
- Investigate how Reference-Free metrics respoåçnd to these challenges

# Use Case: SAP HR Chatbot for Employees



Step 1

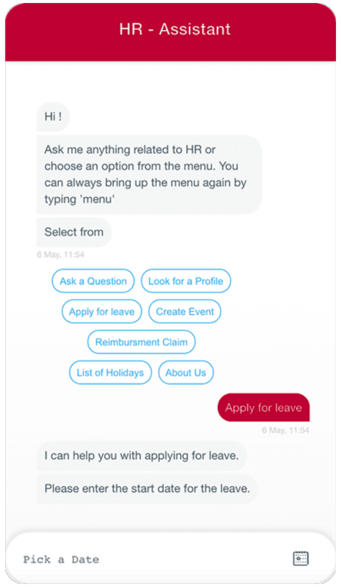
Employees have questions

Step 2

HR Chatbot

Step 3

HR Assistant



Hi !

Ask me anything related to HR or choose an option from the menu. You can always bring up the menu again by typing 'menu'

Select from

6 May, 11:54

Apply for leave

6 May, 11:54

I can help you with applying for leave.

Please enter the start date for the leave.

Pick a Date

***Benefits***

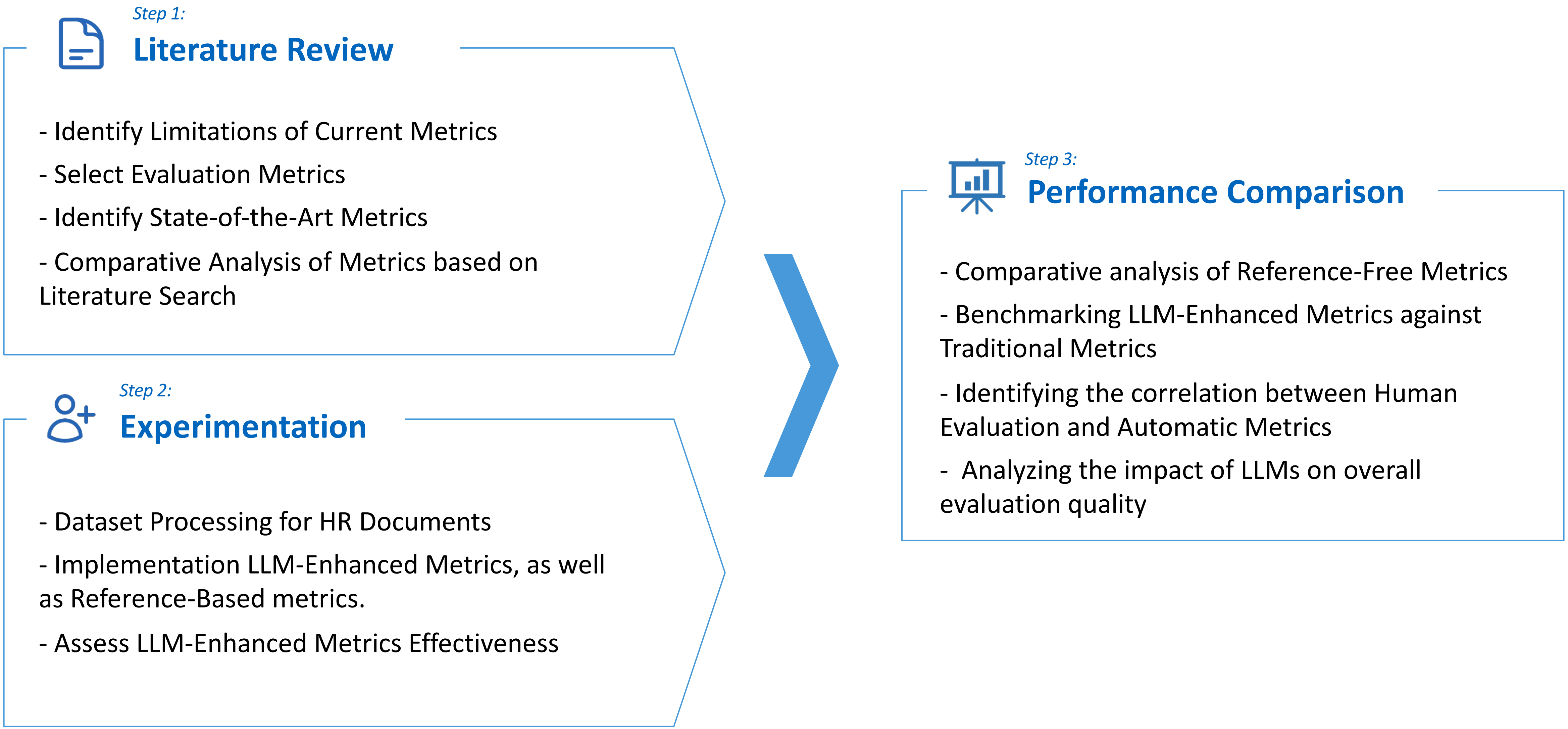
- ***Benefit 1***: Save time for employees and the HR domain experts
- ***Benefit 2***: Automation of Manual tasks

***30%*** HR tickets could be replaced by chatbot functionalities[1]

**The goal of this research: Evaluate the Performance of the HR Chatbot**

[1] *Xiang: Generative Question Answering for a Chatbot in the Human Resources Domain* *https://fpt.ai/chatbot-inevitable-trend-human-resources-2022*

# **Approaches:** Evaluating Metrics in Conversational AI



*Step 1:*

## Literature Review

- Identify Limitations of Current Metrics
- Select Evaluation Metrics
- Identify State-of-the-Art Metrics
- Comparative Analysis of Metrics based on Literature Search

*Step 2:*

## Experimentation

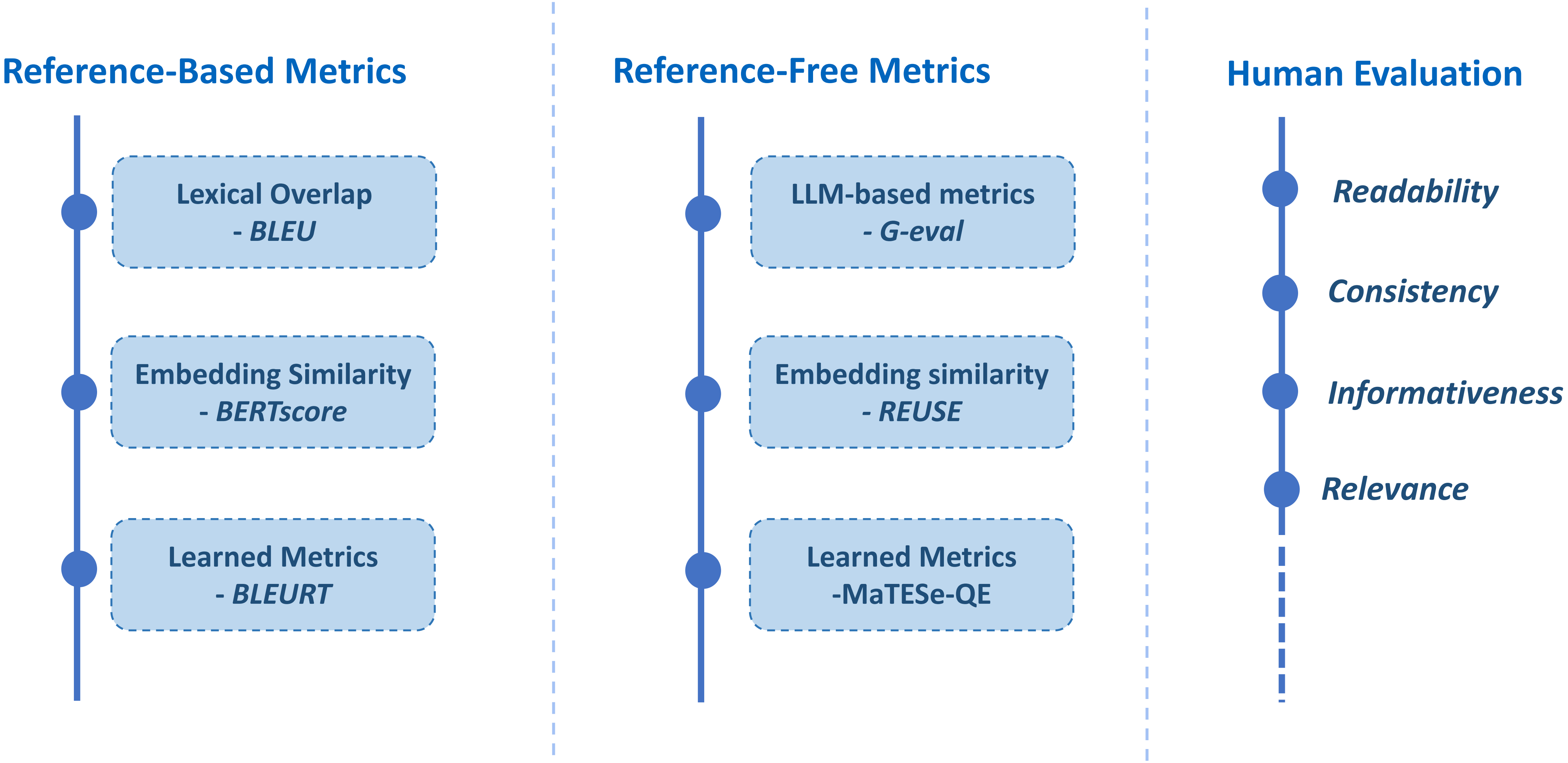
- Dataset Processing for HR Documents
- Implementation LLM-Enhanced Metrics, as well as Reference-Based metrics.
- Assess LLM-Enhanced Metrics Effectiveness

*Step 3:*

## Performance Comparison

- Comparative analysis of Reference-Free Metrics
- Benchmarking LLM-Enhanced Metrics against Traditional Metrics
- Identifying the correlation between Human Evaluation and Automatic Metrics
- Analyzing the impact of LLMs on overall evaluation quality

# **Approaches:** Literature Review on Traditional and State-of-the-Art Evaluation Metrics



## Reference-Based Metrics

- Lexical Overlap *- BLEU*
- Embedding Similarity *- BERTscore*
- Learned Metrics *- BLEURT*

## Reference-Free Metrics

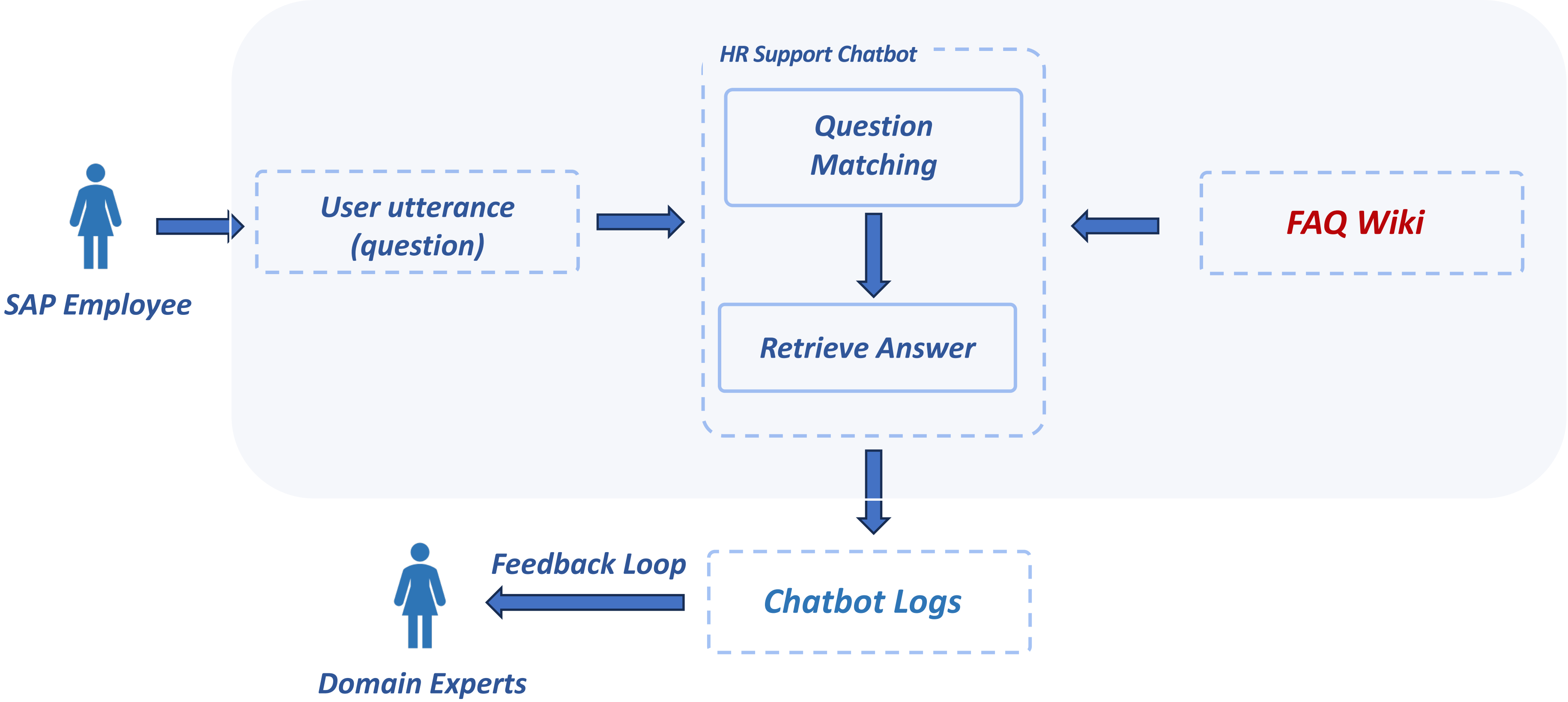
- LLM-based metrics *- G-eval*
- Embedding similarity *- REUSE*
- Learned Metrics -MaTESe-QE

## Human Evaluation

- ***Readability***
- ***Consistency***
- ***Informativeness***
- ***Relevance***

# **Approaches:** SAP Q&A Dataset Structure

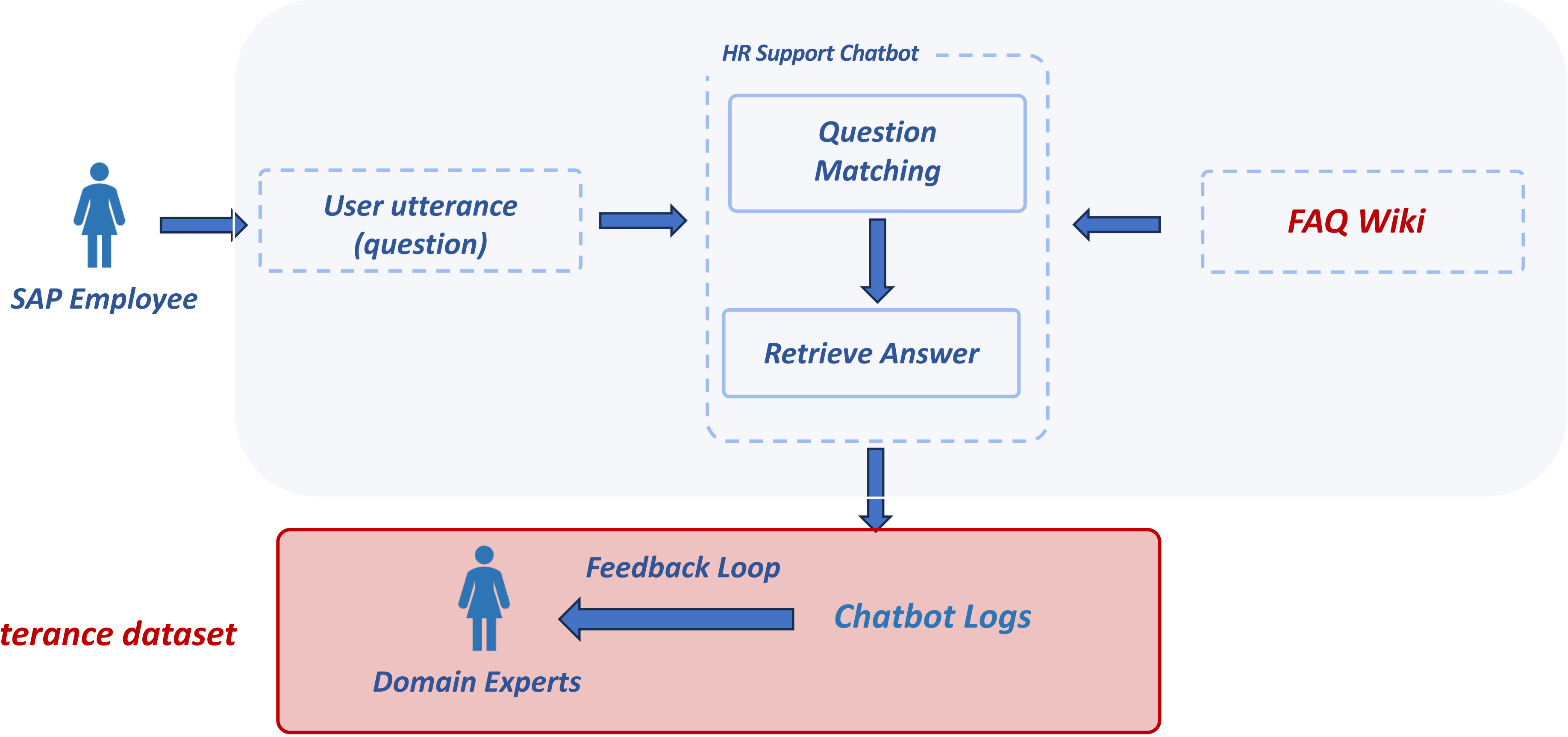
***First Approach: Question Matching***



**Dataset Overview: Selected Questions and Answers as Reference**

# **Approaches:** SAP Q&A Dataset Structure

*First Approach: Question Matching*

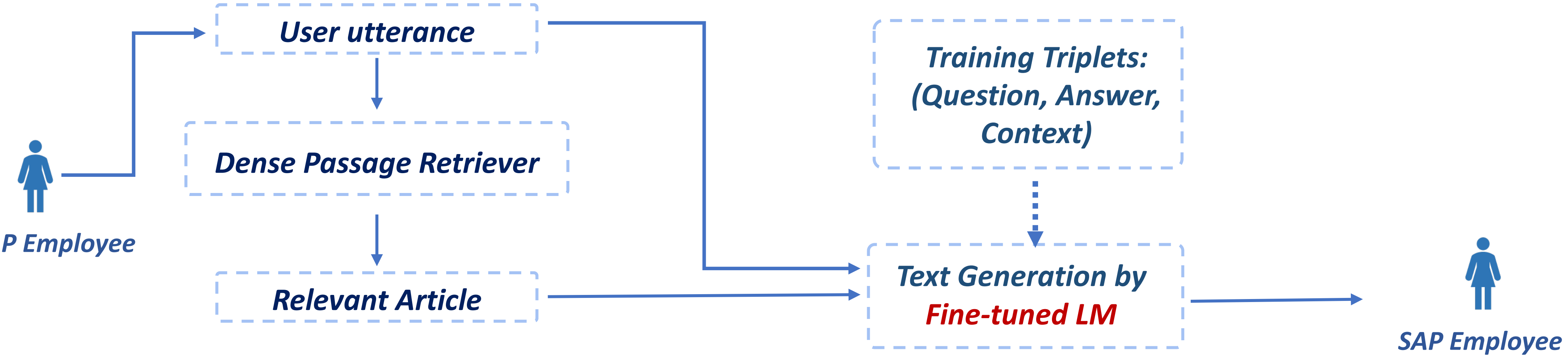


*User utterance dataset*

Dataset Overview: Selected Questions and Answers as Reference

# **Approaches:** SAP Q&A Dataset Structure

*Fine-tuned LM Approach*



*User utterance*

*Dense Passage Retriever*

*Relevant Article*

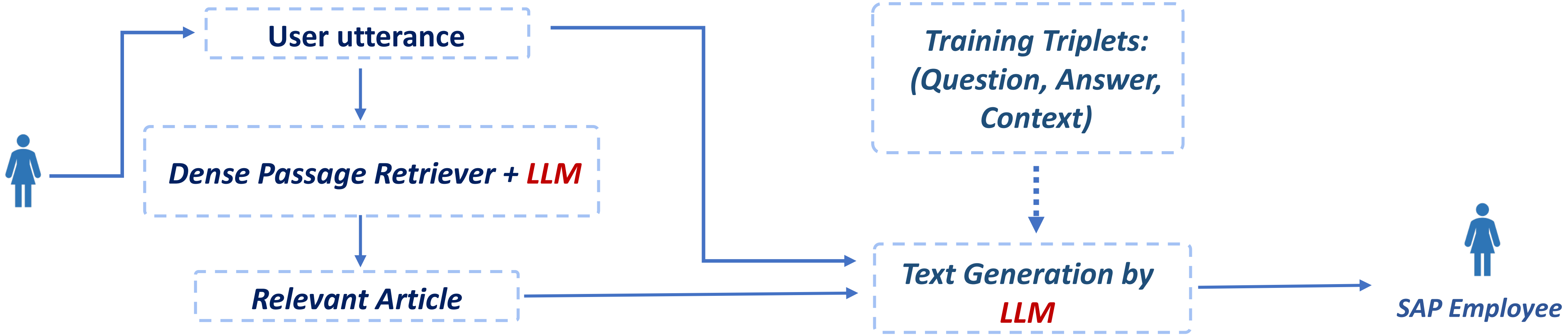
*SAP Employee*

*Training Triplets: (Question, Answer, Context)*

*Text Generation by* ***Fine-tuned LM***

*SAP Employee*

*LLM-Powered Approach*



User utterance

*Dense Passage Retriever + LLM*

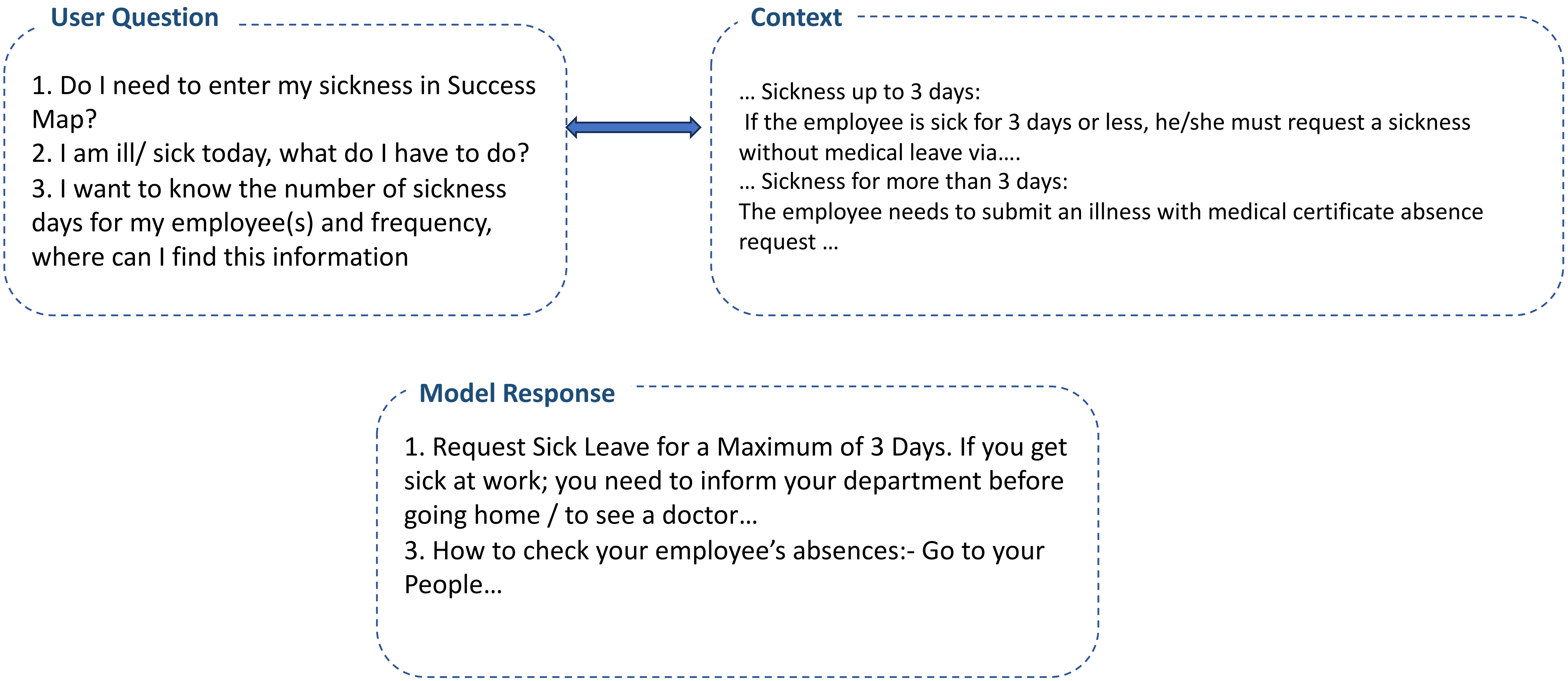
*Relevant Article*

*Training Triplets: (Question, Answer, Context)*

*Text Generation by* ***LLM***

*SAP Employee*

# **Approaches:** Illustrative SAP Q&A Dataset Structure



### User Question

1. Do I need to enter my sickness in Success Map?
2. I am ill/ sick today, what do I have to do?
3. I want to know the number of sickness days for my employee(s) and frequency, where can I find this information

### Context

… Sickness up to 3 days:
If the employee is sick for 3 days or less, he/she must request a sickness without medical leave via….
… Sickness for more than 3 days:
The employee needs to submit an illness with medical certificate absence request …

### Model Response

1. Request Sick Leave for a Maximum of 3 Days. If you get sick at work; you need to inform your department before going home / to see a doctor…
3. How to check your employee's absences:- Go to your People…

# **Research Questions:** Problem Statement and Goals

## Research Question 1

What are the emerging **state-of-the-art metrics** in the evaluation of generative conversational agents, and how do they **compare** to **traditional metrics**?

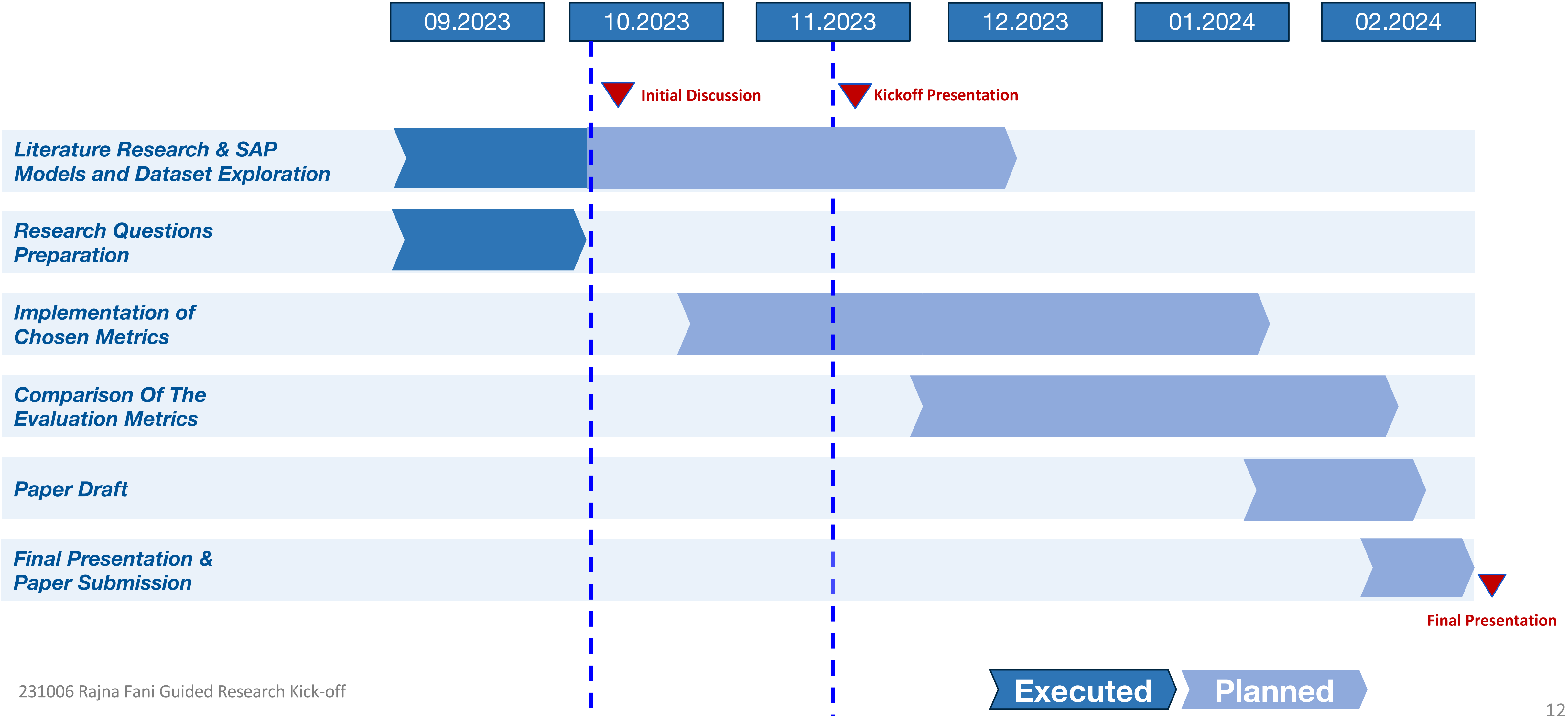
## Research Question 2

Are **reference-free evaluation metrics**, especially those leveraging advanced language models, a more **reliable** indicator of a generative model's performance compared to **traditional reference-based** metrics?

## Research Question 3

How effectively do **automatic metrics** perform in assessing generative model performance when subjected to **human evaluation** by domain experts?

# Timeline



09.2023 | 10.2023 | 11.2023 | 12.2023 | 01.2024 | 02.2024

Initial Discussion

Kickoff Presentation

*Literature Research & SAP Models and Dataset Exploration*

*Research Questions Preparation*

*Implementation of Chosen Metrics*

*Comparison Of The Evaluation Metrics*

*Paper Draft*

*Final Presentation & Paper Submission*

Final Presentation

Executed

Planned

Prof. Dr.

**Florian Matthes**

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de