

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

Rajna Fani, 24.06.2024, Final Presentation

Lehrstuhl für Software Engineering betrieblicher Informationssysteme (sebis)  
Fakultät für Informatik  
Technische Universität München  
[www.matthes.in.tum.de](http://www.matthes.in.tum.de)

**Goal:** Collaboration with  to develop HR chatbot

Split into **2 guided research projects**

→ **Alex:** Functionalities + Implementation

→ **Rajna:** Evaluation with human-in-the-loop

**Paper:** *Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human-in-the-Loop*

→ Accepted at DaSH Workshop at NAACL 24 in Mexico

→ Best Paper Award

# Agenda

- 1** Motivation
- 2** Research Questions
- 3** Methodology
- 4** Results & Main Findings

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



SAP Employees

QA Chatbot

HR Policies

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

● **Benefit 2:** Automation of Manual tasks

Goal: *Process 30% of HR tickets* with chatbot functionalities

## 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 respond to these challenges

# Challenges in Evaluating Text Generative Models : Example

**Question**

● *Am I entitled to use my company benefits immediately after joining the company*

**Golden Answer**

● *Yes, you will automatically be enrolled in pension plan, health scheme and various insurances.*

**Generated Answer (GPT4)**



● *Yes, you are entitled to use your company benefits immediately after joining the company. You will automatically be enrolled in the pension plan, health scheme, and various insurances. However, for some benefits such as health insurance for spouse and children, you need to apply for them via Mercer. Please raise an HRdirect ticket if you require further assistance.*

**Human Evaluation**

*Readability: 5, Relevance: 5 , Naturalness: 4, Truthfulness: 5*

**Automatic Evaluation**

*BLEU Score: 0.16*

**Bad Correlation**

# Agenda

- 1** Motivation
- 2** Research Questions
- 3** Methodology
- 4** Results & Main Findings

## Research Questions: Problem Statement and Goals



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



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?



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

# Agenda

- 1** Motivation
- 2** Research Questions
- 3** Methodology
- 4** Results & Main Findings



# Architecture Approaches: Baseline vs. our Enhanced RAG Pipeline using LLMs

## Human-in-the-Loop

### Retriever Accuracy

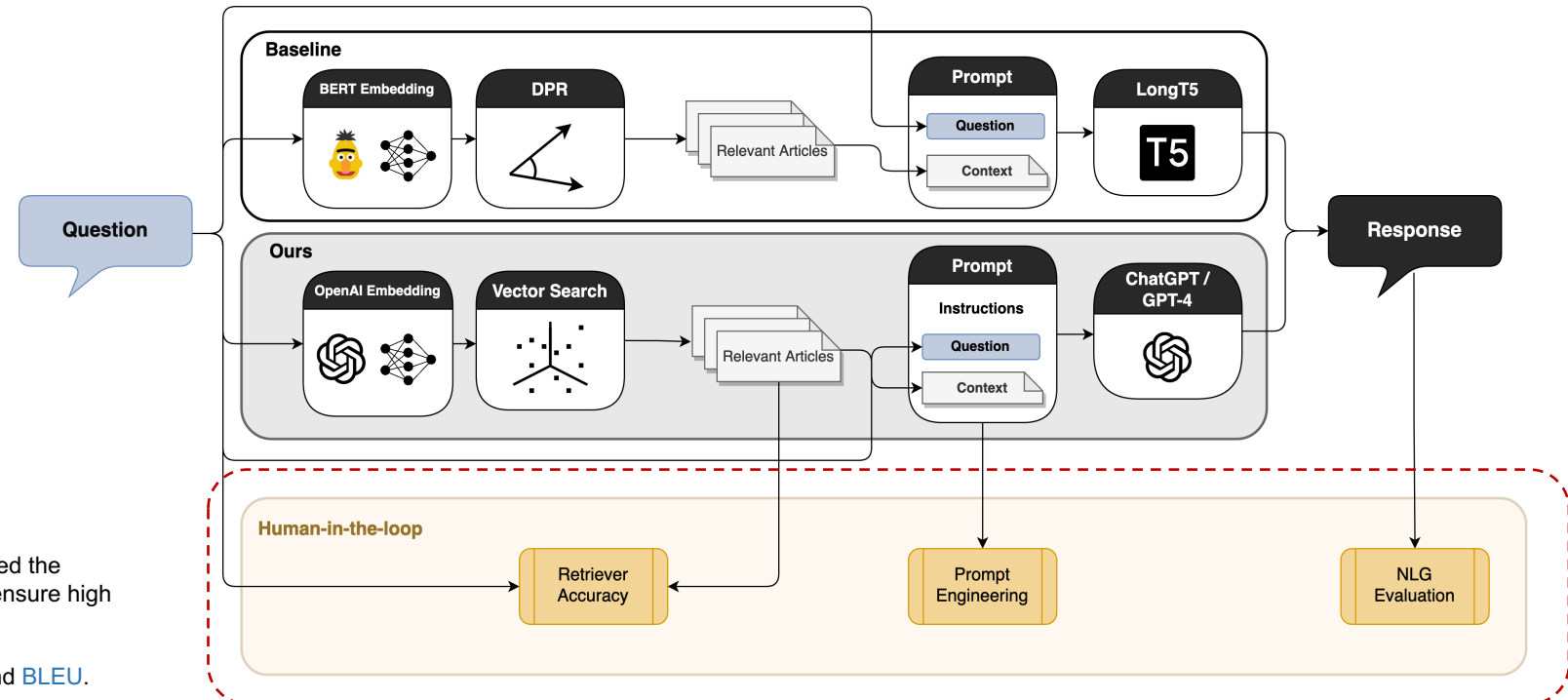
- **Dataset Curation:** Domain experts manually *curated* the dataset and *evaluated* the performance of the retriever by verifying the *accuracy* of matched questions, contextual information (KBA), and correct answers. This ensured the correctness and relevance of the retrieved articles.

### Prompt Engineering

- **Iterative Refinement:** *Refined prompts* based on HR *feedback* and our *qualitative analysis*, ensuring LLM responses met company requirements through continuous adjustments.

### Natural Language Generation (NLG) Evaluation

- **Human Evaluation:** Using a *5-point Likert scale*, domain experts evaluated the responses based on *readability, relevance, truthfulness, and usability* to ensure high linguistic quality and contextual appropriateness.
- **Reference-Based Metrics:** Utilized metrics like *BERTScore, ROUGE, and BLEU*.
- **Reference-Free Metrics:** Explored advanced LLM-based metrics like *G-Eval and Prometheus*.



# Evaluation Framework Approaches: Automatic Evaluation Metrics for NLG Evaluation

## Reference-based Metrics

### N-gram based metrics

- BLEU
- ROUGE

Simplicity and effectiveness in Machine Translation and Automatic Summaries

Based on word-overlappings, no context

### Embedding-based metrics

- BERTScore

Semantic Evaluation

Dependence on Pre-trained Model

## Reference-free Metrics

### Prompt-based Evaluation

- G-Eval

Using GPT-4 chain-of-thought method

High costs since it is not open-source

### Tuning-based metrics

- **Prometheus:** Fine-tuned on Llama-2-Chat-13B

Open-source and cost effective

## Human Evaluation

### Evaluation Criteria

- Readability
- Relevance
- Truthfulness
- Usability

# Agenda

- 1** Motivation
- 2** Research Questions
- 3** Methodology
- 4** Results & Main Findings



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

## Correlation tests:

**Spearman:** Measures how well **two lists of rankings** match each other.

**Kendall:** Measures the **agreement between two lists of rankings** by comparing pairs.

Criteria	LongT5		GPT-3.5		GPT-4	
	Spearman $\rho$	Kendall $\tau$	Spearman $\rho$	Kendall $\tau$	Spearman $\rho$	Kendall $\tau$
BLEU	<b>0.459</b>	0.337	0.345	0.263	0.146	0.116
ROUGE-1	<b>0.435</b>	0.321	0.364	0.284	0.113	0.091
ROUGE-2	<b>0.462</b>	0.341	0.332	0.258	0.056	0.044
ROUGE-L	<b>0.433</b>	0.324	0.353	0.274	0.093	0.075
BERTScore_P	<b>0.457</b>	0.347	0.304	0.234	0.156	0.122
BERTScore_R	<b>0.466</b>	0.305	0.085	0.064	-0.022	-0.018
BERTScore_F1	<b>0.455</b>	0.332	0.246	0.192	0.097	0.077

Table 3: Correlations between Automated Metrics and Human Evaluation across Models

## Correlation between reference-based and human evaluation:

- **Weak** alignment with human evaluations for GPT-3.5 and GPT-4.
- Fail to assess creative, varied outputs of LLM-based models.
- Prefer less generative, reference-aligned responses like LongT5.

# Correlation Analysis among the metrics



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

## G-Eval:

•**Strengths:** High accuracy in assessing **truthfulness** and steady performance across different models.

•**Weaknesses:** Less effective in **readability** evaluations, needing refinement for subjective nuances.

## Prometheus:

•**Strengths:** Strong in evaluating **usability** assessment.

•**Weaknesses:** Similar to G-Eval, struggles with **readability** assessments. This might come as a conclusion of the HR domain specific vocabulary.

Criteria	LongT5		GPT-3.5		GPT-4	
	Spearman $\rho$	Kendall $\tau$	Spearman $\rho$	Kendall $\tau$	Spearman $\rho$	Kendall $\tau$
BLEU	<b>0.459</b>	0.337	0.345	0.263	0.146	0.116
ROUGE-1	<b>0.435</b>	0.321	0.364	0.284	0.113	0.091
ROUGE-2	<b>0.462</b>	0.341	0.332	0.258	0.056	0.044
ROUGE-L	<b>0.433</b>	0.324	0.353	0.274	0.093	0.075
BERTScore_P	<b>0.457</b>	0.347	0.304	0.234	0.156	0.122
BERTScore_R	<b>0.466</b>	0.305	0.085	0.064	-0.022	-0.018
BERTScore_F1	<b>0.455</b>	0.332	0.246	0.192	0.097	0.077
<b>G-Eval</b>						
Usability	0.675	0.584	0.217	0.198	0.346	0.327
Relevance	0.569	0.499	0.339	0.304	0.325	0.306
Readability	0.208	0.181	0.395	0.373	0.139	0.137
Truthfulness	<b>0.726</b>	0.651	0.694	0.667	0.452	0.432
<b>Prometheus</b>						
Usability	<b>0.723</b>	0.675	0.386	0.351	0.516	0.495
Relevance	0.467	0.439	0.419	0.371	0.382	0.357
Readability	0.493	0.468	0.378	0.358	0.225	0.213
Truthfulness	0.541	0.521	0.439	0.402	0.454	0.427

Table 3: Correlations between Automated Metrics and Human Evaluation across Models



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?

### Traditional Metrics (BLEU, ROUGE, BERTScore):

- **Favor Less Generative Models:** Higher scores for LongT5 due to less generative outputs.
- **BERTScore:** Provides a more nuanced evaluation, better suited for assessing modern generative models like GPT-4 and GPT-3.5.

### LLM-based metrics (G-Eval, Prometheus):

- **Better Align with Human Judgment:** Better reflection of relevance, readability, truthfulness, and usability.
- **Higher Scores for Advanced Models:** Favor GPT-4 and GPT-3.5, showcasing their generative strengths.
- **Comprehensive Evaluation:** Capture nuanced text quality aspects.

⇒ State-of-the-art Metrics offer a more accurate, human-aligned evaluation of generative models.

⇒ Traditional Metrics are effective for simpler models but inadequate for advanced generative models.

Metric	GPT-3.5	GPT-4	LongT5
<i>Reference-based Evaluation</i>			
BLEU Score	0.27	0.28	<b>0.41</b>
ROUGE-1	0.48	<b>0.52</b>	0.51
ROUGE-2	0.36	0.35	<b>0.43</b>
ROUGE-L	0.46	<b>0.50</b>	0.49
BERTScore_P	0.88	0.90	<b>0.91</b>
BERTScore_R	<b>0.96</b>	0.93	0.91
BERTScore_F1	0.90	<b>0.91</b>	0.90
<i>Reference-free Evaluation (LLM-based)</i>			
G-Eval: Relevance	4.03	<b>4.51</b>	3.17
G-Eval: Readability	4.26	<b>4.49</b>	3.52
G-Eval: Truthfulness	4.12	<b>4.80</b>	3.36
G-Eval: Usability	4.67	<b>4.79</b>	3.29
Prometheus: Relevance	3.25	<b>3.70</b>	2.83
Prometheus: Readability	3.07	<b>4.22</b>	3.73
Prometheus: Truthfulness	3.20	<b>3.75</b>	3.32
Prometheus: Usability	3.98	<b>4.32</b>	2.83
<i>Domain Expert Evaluation</i>			
Human Eval: Readability	4.31	<b>4.76</b>	4.02
Human Eval: Relevance	4.31	<b>4.67</b>	3.46
Human Eval: Truthfulness	4.09	<b>4.41</b>	3.67
Human Eval: Usability	3.32	<b>4.11</b>	2.59

Table 2: Average Evaluation Scores. BLEU (0 to 1), ROUGE (0 to 1) and BERTScore (-1 to +1 ) were computed on 200 samples, Prometheus (1 to 5) on 60 samples, and Domain Expert Evaluation (1 to 5) & G-Eval (1 - 5) on 100 samples.

# Conclusion

## Superiority of GPT-4:

- GPT-4 **outperforms** GPT-3.5 and LongT5 in generating accurate, relevant responses, ideal for HR chatbots.

## Ineffectiveness of Traditional Metrics:

- BLEU and ROUGE are **less effective** for evaluating complex outputs from advanced models.

## Effectiveness of Reference-free Metrics:

- G-Eval and Prometheus align **closely** with human judgment, providing more reliable NLG assessments.

## Future of NLG Evaluation:

- Advanced LLM-powered metrics come very close to human evaluation on average. Our findings highlight the continued importance of human judgment, especially for domain-specific use cases.

# Future Work

## *Refine Reference-free Metrics:*

- Enhance accuracy and integrate organizational knowledge (fine-tune the model for evaluation specifically on the HR domain).

## *Explore with new Models:*

- With the new Models and Metrics, there is still space for future research to find a more suitable metric.

## *Improve Human Evaluation:*

- Use multiple domain experts for unbiased results.



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  
[www.matthes.in.tum.de](http://www.matthes.in.tum.de)





# G-Eval Prompt

---

## SYSTEM PROMPT

You will be given a generated answer for a given question. Your task is to act as an evaluator and compare the generated answer with a reference answer on one metric. The reference answer is the fact-based benchmark and shall be assumed as the perfect answer for your evaluation. Please make sure you read and understand these instructions very carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria: {criteria}

Evaluation Steps: {steps}

---

## USER PROMPT

Example: {example}

Question: {question}

Generated Answer: {generated\_answer}

Reference Answer: {reference\_answer}

Evaluation Form: Please provide your output in two parts separate as a Python dictionary with keys rating and explanation. First the rating in an integer followed by the explanation of the rating.

{metric\_name}

---

## METRIC SCORE CRITERIA

{The degree to which the generated answer matches the reference answer based on the metric description.}

Readability(1-5) - Please rate the readability of each chatbot response. This criterion assesses how easily the response can be understood. A response with high readability should be clear, concise, and straightforward, making it easy for the reader to comprehend the information presented. Complex sentences, jargon, or convoluted explanations should result in a lower readability score.

---

## METRIC SCORE STEPS

{Readability Score Steps}

1. Read the chatbot response carefully.
2. Assess how easily the response can be understood. Consider the clarity and conciseness of the response.
3. Consider the complexity of the sentences, the use of jargon, and how straightforward the explanation is.
4. Assign a readability score from 1 to 5 based on these criteria, where 1 is the lowest (hard to understand) and 5 is the highest (very easy to understand).

Table 6: G-Eval Prompt Example for Readability Criteria

# Prometheus Prompt

## SYSTEM PROMPT

---

**Task Description:** An instruction (might include an input inside it), a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing an evaluation criterion is given.

2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.

3. The output format should look as follows: Feedback: [write a feedback for criteria] [RESULT] [an integer number between 1 and 5].

4. Please do not generate any other opening, closing, and explanations.

---

**Question to Evaluate:** {instruction}

**Response to Evaluate:** {response}

**Reference Answer (Score 5):** {reference answer}

**Score Rubrics:** {criteria description}

Score 1: {Very Low correlation with the criteria description}

Score 2: {Low correlation with the criteria description}

Score 3: {Acceptable correlation with the criteria description}

Score 4: {Good correlation with the criteria description}

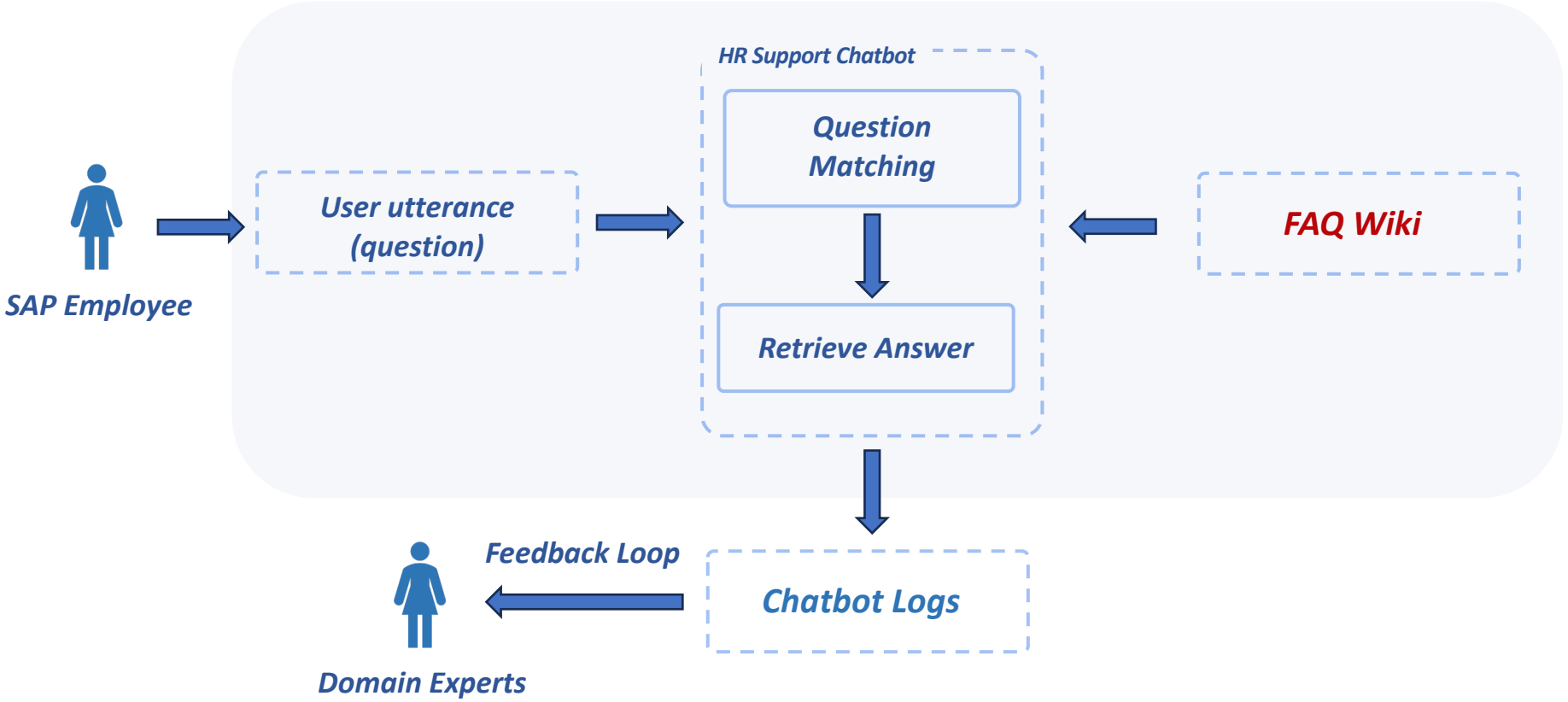
Score 5: {Excellent correlation with the criteria description}


**{criteria description}:** Readability(1-5) - Please rate the readability of each chatbot response. This criterion assesses how easily the response can be understood. A response with high readability should be clear, concise, and straightforward. Complex sentences, jargon, or convoluted explanations should result in a lower readability score.

Table 7: Prometheus Prompt Example for Readability Criteria

# Approaches: SAP Q&A Dataset Structure

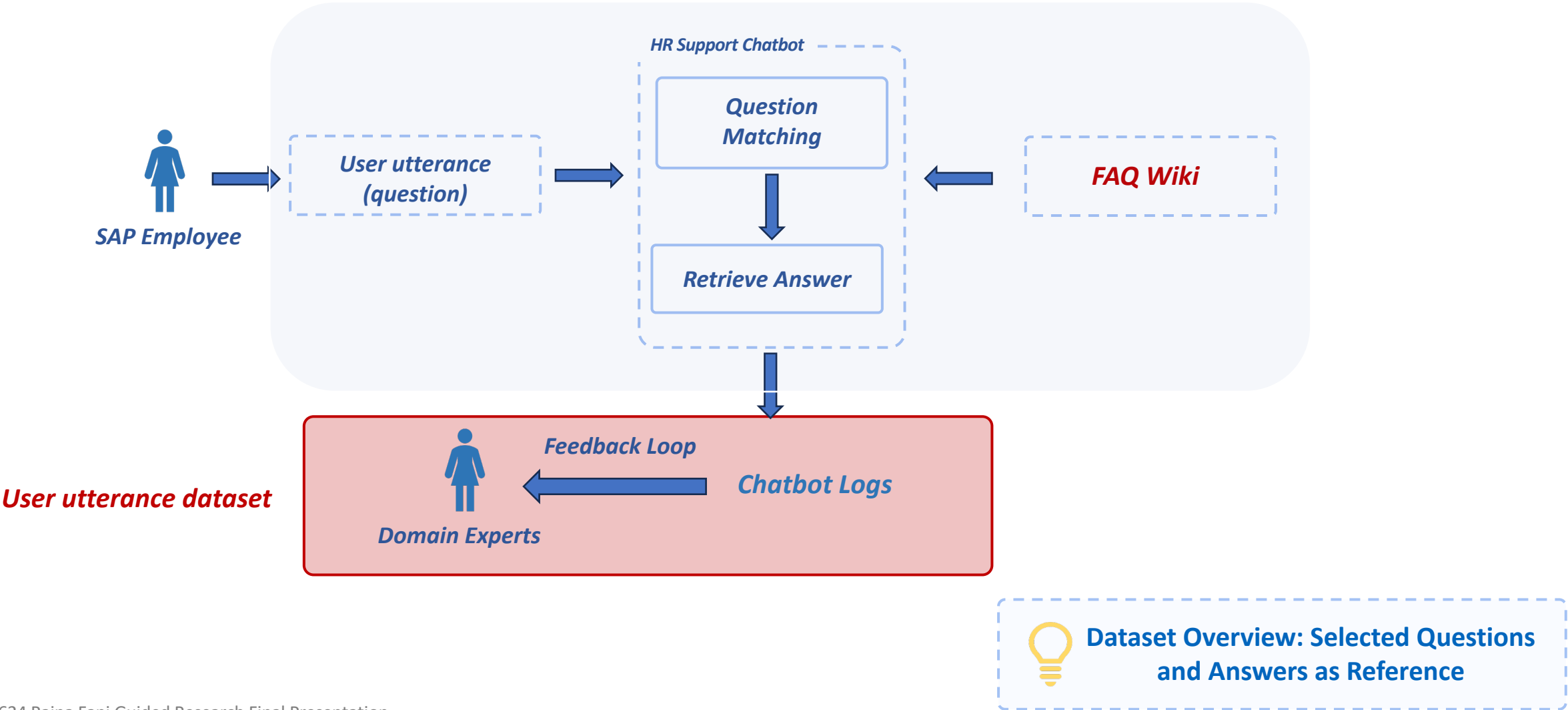
## First Approach: Question Matching



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

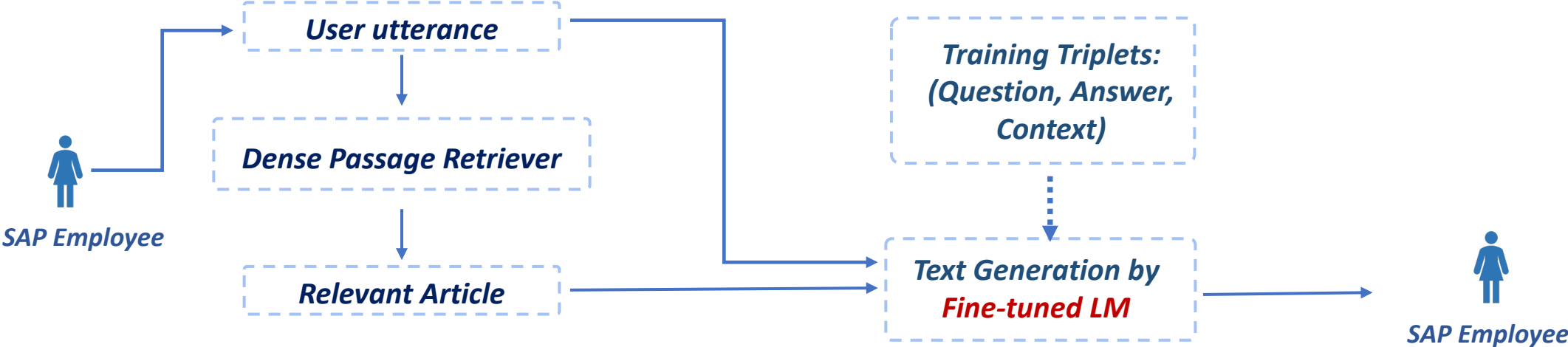
# Approaches: SAP Q&A Dataset Structure

## First Approach: Question Matching

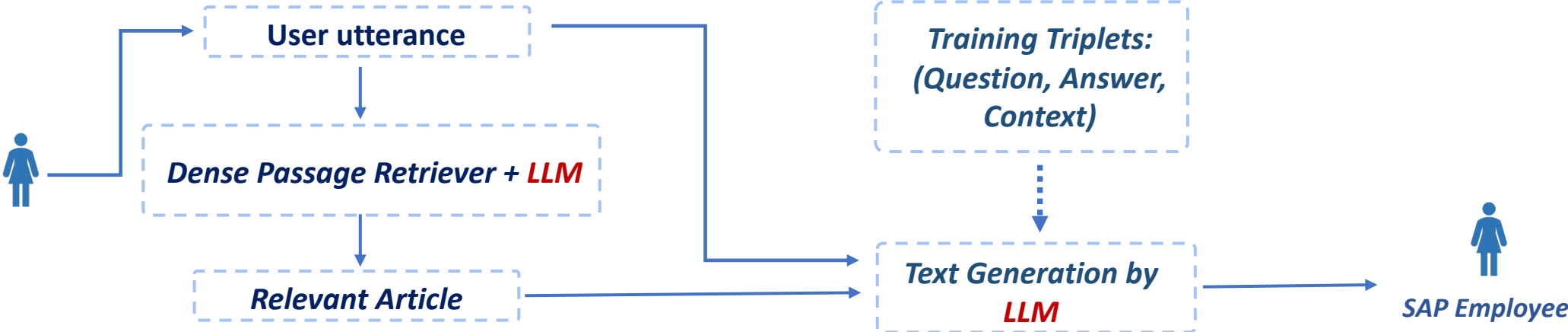


# Approaches: SAP Q&A Dataset Structure

## Fine-tuned LM Approach



## LLM-Powered Approach



# 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...

# Dataset

## DATA TRIPLET

---

**Question:** How can I apply for half a day of holiday?

**Answer:** Unfortunately, vacation days in your country can only be taken as full days.

**Context:** {Relevant Article}

## META DATA

---

**User Role:** Employee

**Name of KBA:** Vacation

**Company Name:** {Company Name}

**Company Code:** {Company Code}

**Region:** {Region}

**Country Code:** {Country Code}

**FAQ Category:** {FAQ Category}

**Process ID:** {Process ID}

**Service ID:** {Process ID}

Table 1: HR Dataset Sample

## 10 most frequent user queries

---

How can I change my approver?

Where do I see how much leave I have left?

How can I view my payslip online?

Am I paid during maternity leave?

If I am sick whilst on holiday, can I claim my holiday back?

Can I cancel a leave request?

I have a question about my payslip, who do I contact?

Where can I find information about my payslip?

Do I receive sick pay?

How can I have an overview of my leave?

Table 2: Top 10 most frequent user queries

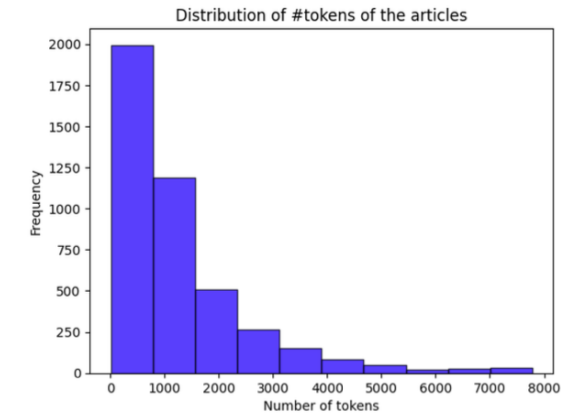


Figure 1: Distribution over the number of tokens of all unique articles in our HR dataset.

# TigerScore Output

Question: how can i change my leave request's approver

Chatbot Answer: You can change your approver, if you Choose Replace in order to change the approver of the request Enter the name or the user ID of the approver and select Find. A list of Possible approvers is displayed. Click on the name you want or select the row and press Choose. You will then be returned to the request form. You can enter a note in the Note for approver field.

TigerScore Results:

```
[
  {
    "num_errors": 2,
    "score": -8.0,
    "errors": {
      "error_0": {
        "location": "The entire generated answer",
        "aspect": "Informativeness",
        "explanation": "The generated answer is too long and detailed, which makes it less likely to be found in a search engine. It could be improved by summarizing the information more concisely. For example, 'To become eligible for the Service Excellence Award, you must be a permanent SAP employee who has made contributions that exceed the expectations of internal and external customers and partners.'",
        "severity": "Major",
        "score_reduction": "4"
      },
      "error_1": {
        "location": "'The LoB manager will make the final decision about the recipients of the award. You will receive the award during the quarterly staff meeting, and the award will be credited to your monthly payroll.'",
        "aspect": "Comprehension",
        "explanation": "The generated answer fails to understand the context that the award is related to SAP employees only. This information should have been included in the summary to avoid confusion for users who are not familiar with the company's policies. For example, 'To become eligible for the Service Excellence Award, you must be a permanent SAP employee...'",
        "severity": "Major",
        "score_reduction": "4"
      }
    },
    "raw_output": " The model-generated output contains 2 errors, with a total score reduction of 8.\nError location 1: The entire generated answer\nError aspect 1: Informativeness\nExplanation 1: The generated answer is too long and detailed, which makes it less likely to be found in a search engine. It could be improved by summarizing the information more concisely. For example, 'To become eligible for the Service Excellence Award, you must be a permanent SAP employee who has made contributions that exceed the expectations of internal and external customers and partners.'\nSeverity 1: Major\nScore reduction 1: 4\nError location 2: 'The LoB manager will make the final decision about the recipients of the award. You will receive the award during the quarterly staff meeting, and the award will be credited to your monthly payroll.'\nError aspect 2: Comprehension\nExplanation 2: The generated answer fails to understand the context that the award is related to SAP employees only. This information should have been included in the summary to avoid confusion for users who are not familiar with the company's policies. For example, 'To become eligible for the Service Excellence Award, you must be a permanent SAP employee...'\nSeverity 2: Major\nScore reduction 2: 4"
```