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

Alexander Kowsik, 24.06.2024, Final Presentation

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

General Information

Goal: Collaboration with with SAP 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

Outline

High HR workloads result in extensive manual labor and long delays

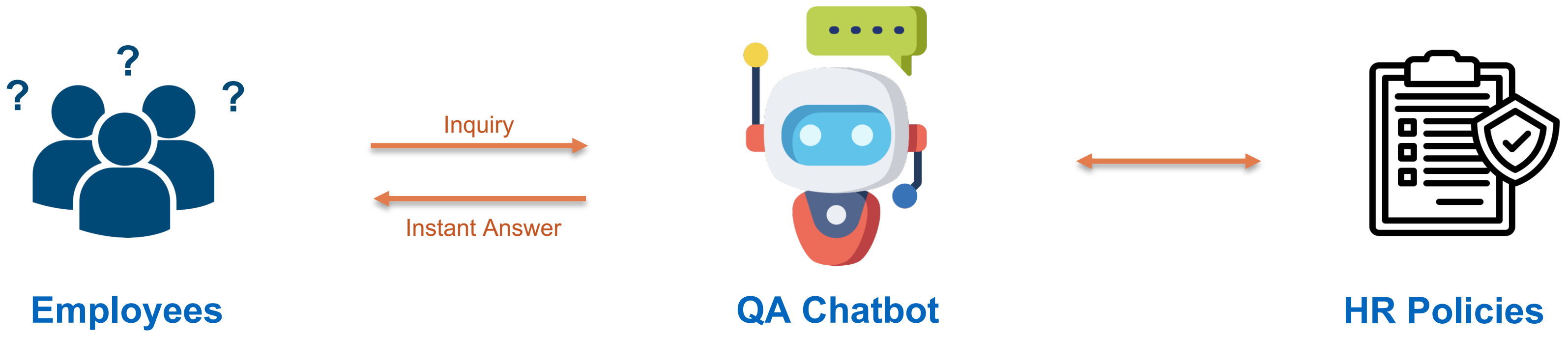
Problem 1: **Large volumes of HR inquiries**

- 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

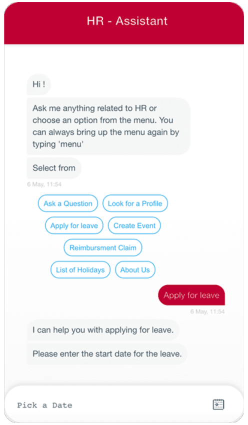


Benefit 1: **Save time for both employees and the HR domain experts**

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

Benefit 2: **Automation of (mundane) manual tasks**

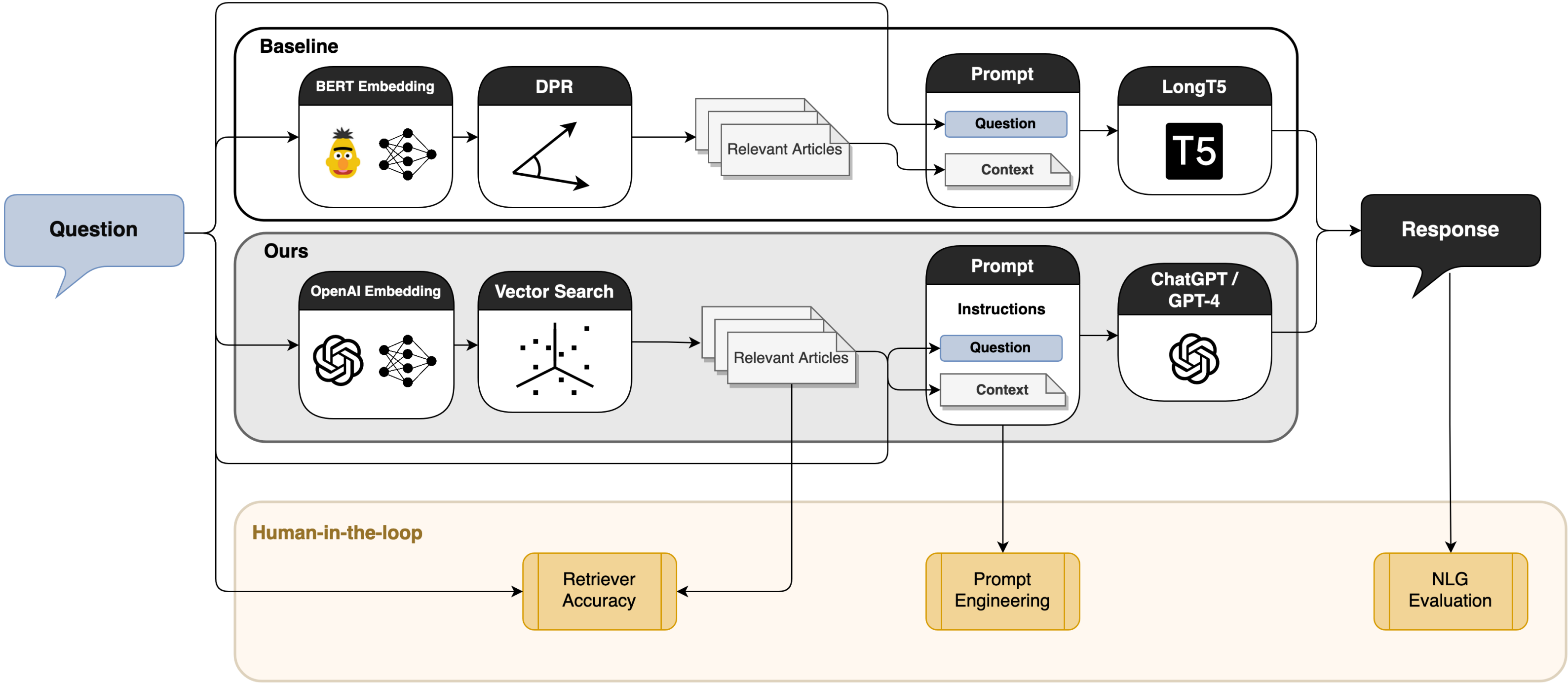
- The process of answering employee questions based on HR policies is *highly automatable using SOTA NLP models* (e.g., LLMs)
- Chatbots utilize the *HR rules and policies as grounding* for their responses



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

Outline

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



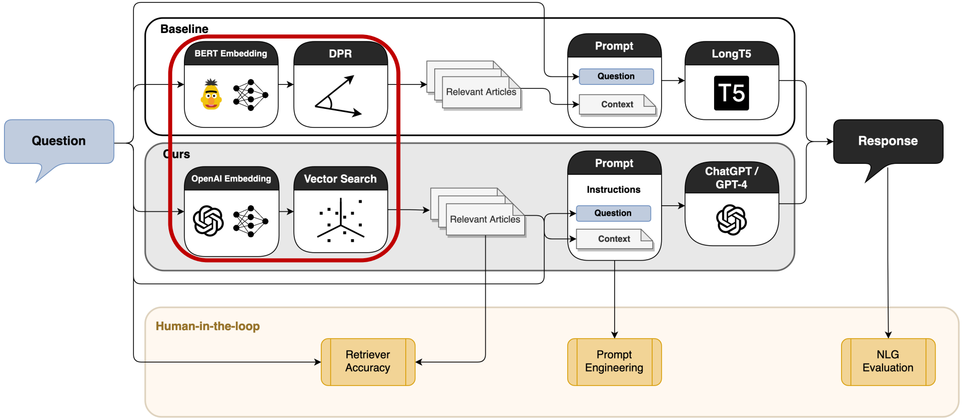
Retrieval-augmented Generation using LLMs – Most flexible and least limiting solution

Document Retriever

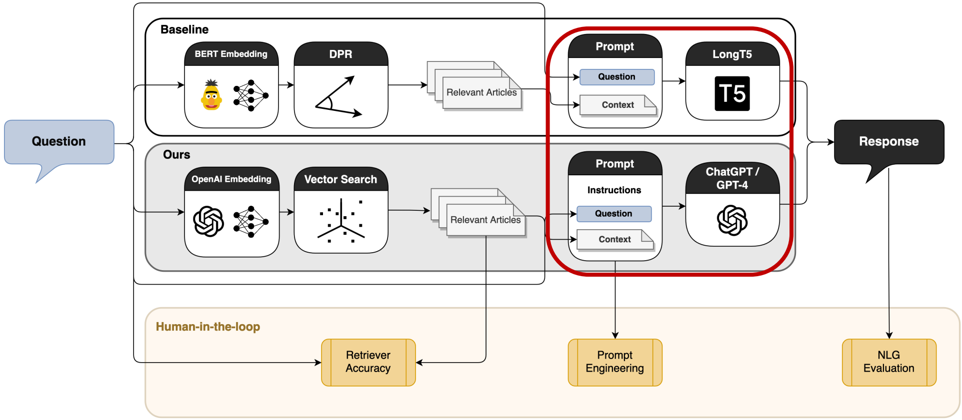
Goal: Higher retrieval accuracy + better performance

Dense Passage Retriever → Vector search + optimizations
BERT embeddings → OpenAI embeddings

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



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



NLG Module

Challenge: Get outputs conforming with requirements

LongT5 → ChatGPT / GPT-4 (+ open source LLMs)
Fine-tuning → Prompt Engineering

- Removes need for fine-tuning custom model
- Instruct LLM to return desired output
  → Prompt engineering / In-context learning
  → More flexible w.r.t. changing requirements
  → (Muti-turn) conversational capabilities
- Attach relevant context to ground responses
  → Prevent hallucinations

{Instructions}

{Question}

{
Relevant Context
…
}

Prompt

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 **vector search** in comparison to the current **DPR model**?

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

Outline

4

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

- 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, …
- Improved quality of training data with LLMs: summarized unnecessary long responses, improved grammar / output structure, …

1

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

Evaluation Results

1. **Automatic Evaluation:** reference-based and reference-free metrics

2. **Human Evaluation:** *Readability*, *Relevance*, *Truthfulness*, *Usability*

→ ChatGPT/GPT-4 achieve **very high scores on all metrics**

→ Responses are **factually correct** due to **grounding in HR documents**

→ **Significant improvements** compared to the baseline

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

Table 4: 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.

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**?

SAP Requirements

- **Grounding:** responses must be based on HR articles
- **No hallucinations:** LLM should not make up things + instruct the user to open HR ticket when uncertain
- **Appropriate length:** <150 words
- **User friendliness**
- …

Iterative process

- Qualitative evaluation
- Back and forth with SAP using evaluation samples

→ Worked well, based on qualitative and quantitative evaluation

SYSTEM PROMPT

You are an HR chatbot for Large multi-national company and you provide truthful and concise answers to employee questions based on provided relevant HR articles.
1. Stay very concise and keep your answer below 150 words.
2. Do not include too much irrelevant information unrelated to the posed question.
3. Keep your response brief and on point.
4. Include URLs from the relevant article if it is important to answer the question.
5. If the answer applies to specific labs/countries/companies, include this information in your response.
6. Refer to the employee directly as "you" and not indirectly as "the employee".
7. If the provided HR article does not include the answer to the question, tell the employee to create an HRdirect ticket.
8. Answer in a polite, personal, user-friendly, and actionable way.
9. Never make up your response! If you do not know the answer to the question, just say so and ask the user to create an HRdirect ticket!

USER PROMPT

Question: {question}
Relevant Article: {article}

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

Approach: Query Transformation Techniques

→ transform user query → embed transformed query → use in vector search
Idea: transformed queries are closer aligned to desired article in the embedding space

Intended Topics

"How to request request a parental leave?" → **LLM** →
- *parental leave*
- *childcare leave*
- *maternity leave*

HyDE (Hypothetical Document Embeddings)

"How to request request a parental leave?" → **LLM** →
- *To request parental leave, please submit...*
- *If you wish to request...*

Multi-Query + Re-ranking

"How to request request a parental leave?" → **LLM** →
- *How can I request a childcare leave?*
- *Where can I apply for parental leave?*
- + IT & HyDE queries

1. Perform vector search with each query
2. Re-rank results using Reciprocal Rank Fusion
   → Adjust scores (cosine distance) by taking into account the retrieval rank of other queries
3. Return top articles

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**?

Metric: *Accuracy*, since we only retrieve the top-1 article for a given question

	HR Test Dataset				Stackexchange English
Method	**top-1**	**top-2**	**top-3**	**top-5**	**top-1**
Old DPR	**22.24%**	30.03%	35.08%	40.06%	-
Basic	**11.12%**	15.06%	16.82%	18.53%	69.5%
Intended Topics	9.33%	-	-	-	57.25%
HyDE	10.01%	-	-	-	65.91%
Multi-Query	10.92%	-	-	-	**71.31%**

Table 3: Retriever Accuracy on the HR test data and the Stackexchange benchmark dataset for various retriever methods and top-k retrieved articles

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**?

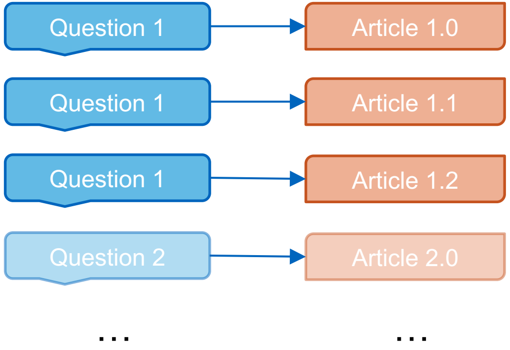
Evaluation Challenges

For most articles, **multiple versions exist** in the evaluation dataset

→ Same/similar questions mapped to slightly different articles

→ Leads to low accuracy, even with *correct* retrievals

→ DPR mostly chooses most frequent version → higher accuracy



Qualitative evaluation confirms correctness
→ also confirmed by human evaluations (**high correctness scores**)
→ LLM can detect if the article is irrelevant → refers to HR ticket

→ LLMs can **reliably reason through inconsistencies in the data**, due to internal reasoning capabilities + domain knowledge

Conclusion

Improvement over Baseline chatbot

- Vector search with query transformations + LLMs is effective in providing grounded answers to HR questions

GPT-4 yields the best results overall

- GPT-4 beats ChatGPT and LongT5 models on all metrics
- Is able reason through inconsistencies in the data and provide useful responses

High retriever performance

- Despite evaluation difficulties, the retriever was able to retrieves relevant articles in virtually all tested cases

Future Work

- Multi-turn conversational capabilities for follow-up questions
- Develop better ways of evaluating retriever quality than accuracy
- Implement actions to allow the chatbot to access/use HR resources (e.g. retrieve paychecks, …)

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

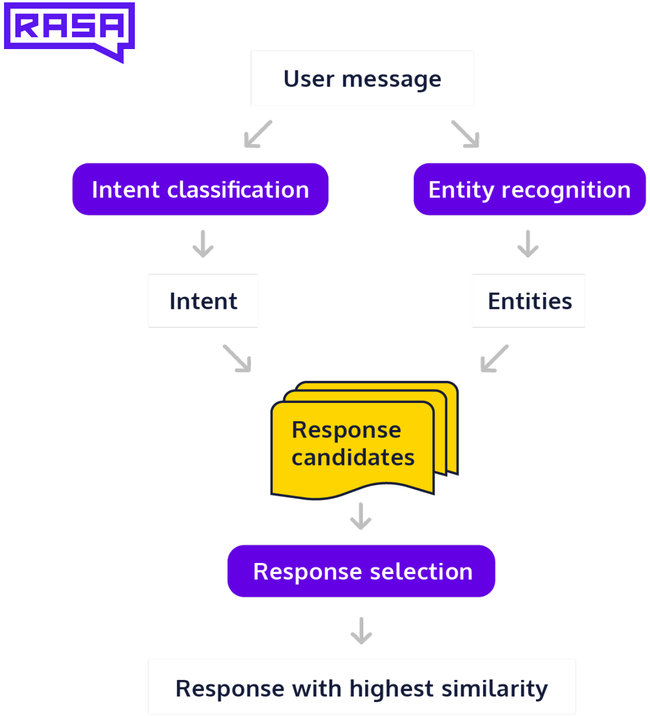
Thank you!

Let’s discuss!

Paper

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

Dataset Statistics

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

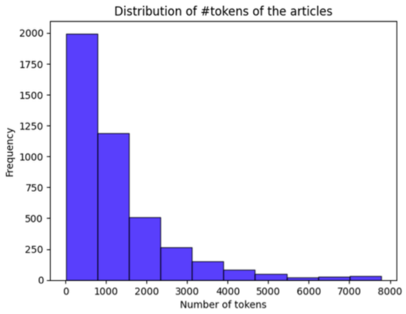


Figure 1: Distribution over the number of tokens of all unique articles in our HR dataset.

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

Correlation Analysis among the metrics

Criteria	LongT5		ChatGPT		GPT-4	
	Spearman ρ	Kendall τ	Spearman ρ	Kendall τ	Spearman ρ	Kendall τ
BLEU	**0.459**	0.337	0.345	0.263	0.146	0.116
ROUGE-1	**0.435**	0.321	0.364	0.284	0.113	0.091
ROUGE-2	**0.462**	0.341	0.332	0.258	0.056	0.044
ROUGE-L	**0.433**	0.324	0.353	0.274	0.093	0.075
BERTScore_P	**0.457**	0.347	0.304	0.234	0.156	0.122
BERTScore_R	**0.466**	0.305	0.085	0.064	−0.022	−0.018
BERTScore_F1	**0.455**	0.332	0.246	0.192	0.097	0.077
G-Eval						
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	**0.726**	0.651	0.694	0.667	0.452	0.432
Prometheus						
Usability	**0.723**	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 9: Correlations between Automated Metrics and Human Evaluation across Models

Initial Project Timeline