

Generative Question Answering for a Chatbot in the Human Resources Domain

Tao Xiang, 08.05.2023, Final Presentation

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

Outline

1. Motivation

2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Outline

1. Motivation


2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Motivation



Large Volume of HR Inquiries

- Human Resource departments handle numerous tasks and queries from employees on a daily basis
- More than 330.000 HR tickets per year in SAP SE

Motivation



Large Volume of HR Inquiries

- Human Resource departments handle numerous tasks and queries from employees on a daily basis
- More than 330.000 HR tickets per year in SAP SE

Labor Reduction and Efficiency

- Effective QA system can substantially alleviate the workload of HR staff by handling a higher volume of employee inquiries.
- Employees can receive instant responses without waiting.

Motivation

Large Volume of HR Inquiries

- Human Resource departments handle numerous tasks and queries from employees on a daily basis
- More than 330.000 HR tickets per year in SAP SE

Labor Reduction and Efficiency

- Effective QA system can substantially alleviate the workload of HR staff by handling a higher volume of employee inquiries.
- Employees can receive instant responses without waiting.

Advancements in NLP

- The rapid progress in natural language processing (NLP) technologies offers opportunities to develop more sophisticated and accurate HR chatbots
- Leveraging state-of-the-art NLP techniques allows for better understanding of user intents and improved response generation, resulting in a more seamless and effective user experience

Outline

1. Motivation

2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Introduction

QA Chatbots in HR Domain

In the HR domain, QA chatbots are designed to assist employees by answering their questions and providing support on various topics, such as benefits, policies, and onboarding.

Introduction

QA Chatbots in HR Domain

In the HR domain, QA chatbots are designed to assist employees by answering their questions and providing support on various topics, such as benefits, policies, and onboarding.

Traditional QA

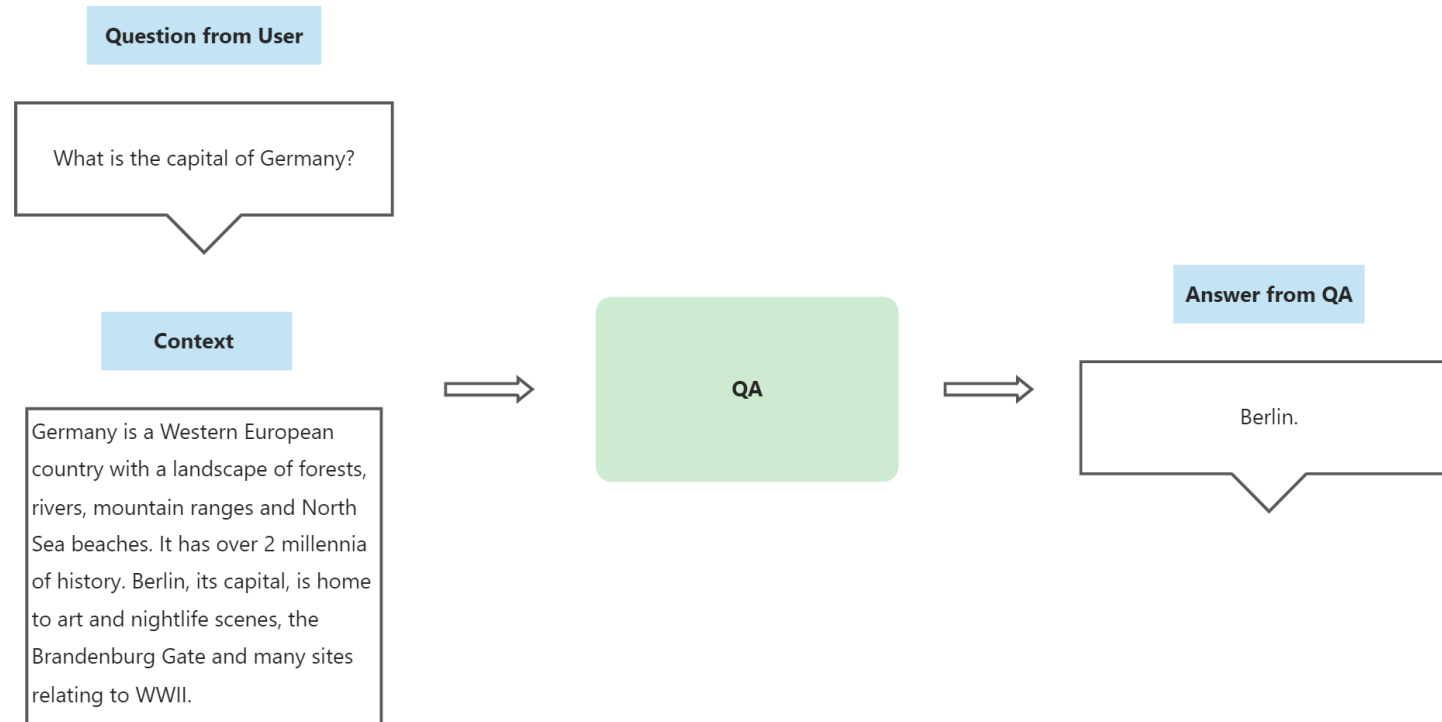
- manually designed intents and predefined responses
- Handle commonly asked questions effectively
- Limited in understanding complex or ambiguous queries
- Not easily scalable due to manual effort required

Generative QA

- Leverage advancements in NLP for improved accuracy and user experience
- Automatically understand user intents and find answers in context
- Can handle a wider range of questions, including complex and ambiguous queries
- More scalable and cost-effective solution for handling high volume of HR inquiries

Introduction

Generative QA: An example



Research Questions

- 1 How to effectively address the issue of lengthy input (context) in generative QA systems?
- 2 Which analytical scores would be ideal to evaluate the performance of the models?
- 3 How to accurately assess the performance of generative QA models in real-world scenarios?

Outline

1. Motivation

2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Methodology - Datasets

Domain experts internally prepared two datasets:

Dataset 1

T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	
NE	COMP	EXCL	EXC	COMPAN	KEYWORD	ARTICLE DATA TEXT	LANG	FAQ DATA	DESCRIPTION	QUESTION	ANSWER	SWER_RA	FA
0262, _	00	FALSE	FALSE	['0262', '_	, my_learnin	If employees have used all annual leave entitlen	EN	<tr><td><h2><span ["		How can I request Unpaid Leave?/ How	To apply for Unpaid Leave, please follow the	To apply fo	['Unpaid Leave (Kf
0262, _	00	FALSE	FALSE	['0262', '_	, my_learnin	If employees have used all annual leave entitlen	EN	<tr><td><h2><span ["		How is my salary calculated in case I us	The salary will be prorated and paid out for ti	The salary [['Unpaid Leave (Kf
0262, _	00	FALSE	FALSE	['0262', '_	, my_learnin	If employees have used all annual leave entitlen	EN	<tr><td><h2><span ["		Can Intern/Temporary Contractor use	Be kindly informed that only permanent emp	Be kindly ii	['Unpaid Leave (Kf
0262, _	00	FALSE	FALSE	['0262', '_	, my_learnin	If employees have used all annual leave entitlen	EN	<tr><td><h2><span ["		How can I get the approval from HRBP	You can obtain approval emails from the resp	You can ob	['Unpaid Leave (Kf
0262, _	00	FALSE	FALSE	['0262', '_	, my_learnin	If employees have used all annual leave entitlen	EN	<tr><td><h2><span ["		What local benefits can I claim during	You are only eligible for Flexible Benefits in S	You are on	['Unpaid Leave (Kf
0450, 0800	FALSE	FALSE	FALSE	['0450', '08	counseling, The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron ["		Where can I get help regarding a pers	The Employee Assistance Program (EAP) help	The Emplo	['Employee Assista	
0450, 0800	FALSE	FALSE	FALSE	['0450', '08	counseling, The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron ["		What is the Employee Assistance Progr	The Employee Assistance Program (EAP) help	The Emplo	['Employee Assista	
0450, 0800	FALSE	FALSE	FALSE	['0450', '08	counseling, The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron ["		Can I ask the Employee Assistance Prog	The Employee Assistance Program (EAP) help	The Emplo	['Employee Assista	
0450, 0800	FALSE	FALSE	FALSE	['0450', '08	counseling, The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron ["		Who is eligible to use the Employee Ass	All SAP employees and their immediate famili	All SAP emil	['Employee Assista	

- Contains (question, context, answer) tuples
- The questions are very standardized and highly structured

Methodology - Datasets

Dataset 1 Example

Question	Context	Answer
What local benefits can I claim during Long-Term Leave?	Q: How can I request Unpaid Leave?/ How can I apply for Unpaid Leave?/I would like to understand the application process for Unpaid Leave.</p><p>A: To apply for Unpaid Leave, please ...	You are only eligible for Flexible Benefits in Success Map.

Methodology - Datasets

Domain experts internally prepared two datasets:

Dataset 2

	B	C	D	E	F	G	H	I	J	K	P	Q	T	U	V	W	X
1	COMPANY	SIMILARIT	W2VDIST	DISTANCE	ENSEMBL	ISCORRECT	ISMANAG	CONTENT	EMPLOYEE	QUESTION	DATE	CORRECT	RESPONSEQUESTION	RESPONSEANSWER			
2	0413, 0700	0.777757	0.290638	0.527216	0.408927	FALSE	FALSE	1.58E+09	internal	I have misplaced my	35:24.0	1.58E+09	Can you explain the CVS R	CVS Rx Maintenance Medication service is convenient, wi			
3	0413, 0700	0.777194	0.290638	0.529289	0.409963	FALSE	FALSE	1.58E+09	internal	I have misplaced my	35:24.0	1.58E+09	For CVS, what drugs are co	For information regarding the drugs covered under your m			
4	0251, 0413	0.783853	0.506104	0.289315	0.39771	FALSE	FALSE	2.07E+09	internal	how do I get an empl	06:14.0	1.58E+09	I have a corporate credit ca	First, make sure all expenses have been processed and all			
5	0251, 0413	0.785821	0.496457	0.29172	0.394089	FALSE	FALSE	2.07E+09	internal	how do I get a corpor	03:52.0	1.58E+09	I have a corporate credit ca	First, make sure all expenses have been processed and all			
6	0251, 0413	0.687193	0.583979	0.567151	0.575565	FALSE	FALSE	2.07E+09	internal	employee credit card	06:01.0	1.58E+09	I have a corporate credit ca	First, make sure all expenses have been processed and all			
7	0063, 0265	0.785463	0.33916	0.450337	0.394748	TRUE	FALSE	1.86E+09	internal	I am trying to log into	18:08.0		How can I access the Bene	You can access the BenefitFocus tool via corporate portal:			
8	0800, 0063	0.739866	0.509911	0.447383	0.478647	TRUE	FALSE	1.58E+09	internal	insurance id card	24:20.0		Where can I get my Aetna	You will receive an Aetna ID card for yourself, which may I			
9	0063, 0265	0.74557	0.452506	0.483795	0.468151	FALSE	FALSE	1.86E+09	internal	Where do I go to log	54:37.0	1.86E+09	What will I do, if I can't log	Capture the error screenshot. Attach a copy your system is			
10	0063, 0265	0.735181	0.458695	0.51584	0.487268	FALSE	FALSE	1.58E+09	internal	Where do I go to log	54:37.0	1.86E+09	How can I access my Aetna	Once your account has been created in the Aetna™'s sys			
11	0063, 0265	0.67305	0.501949	0.701227	0.601588	FALSE	FALSE	1.58E+09	internal	Logging into Benefit	17:00.0	1.86E+09	How can I access mv Aetna	Once your account has been created in the Aetna™'s sys			

- Contains (question, context, answer) tuples
- The questions are from real users (improved randomness)
- A question matching model is applied to retrieve the most similar question in Dataset 1
 - 60% accuracy
- For wrong retrievals: domain experts annotate correct questions

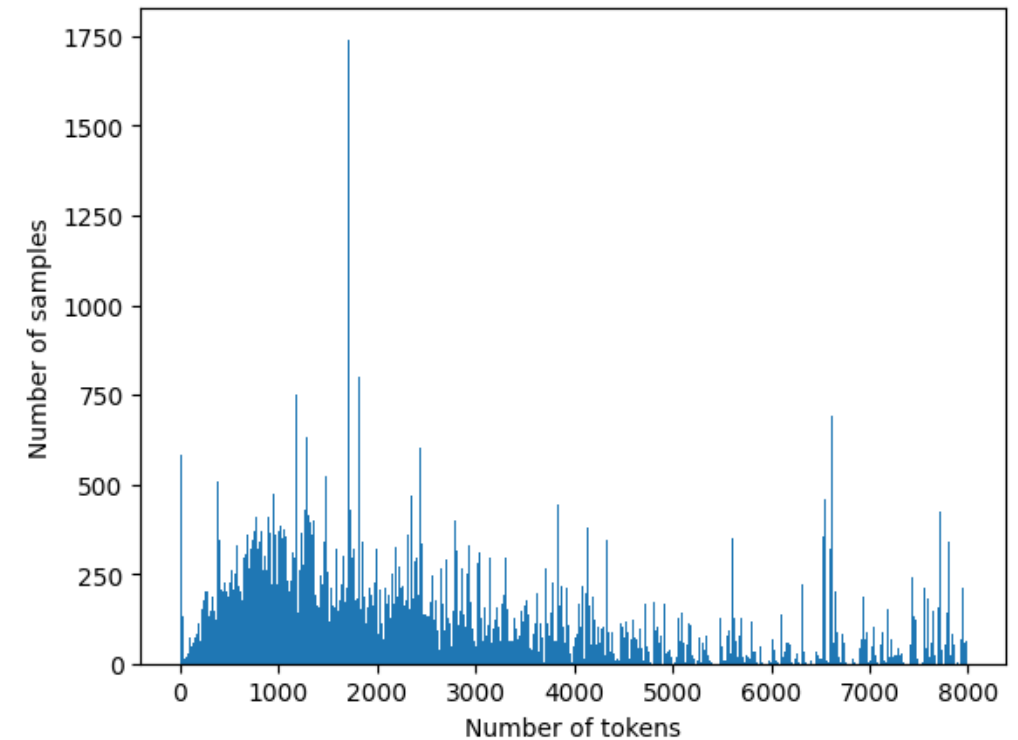
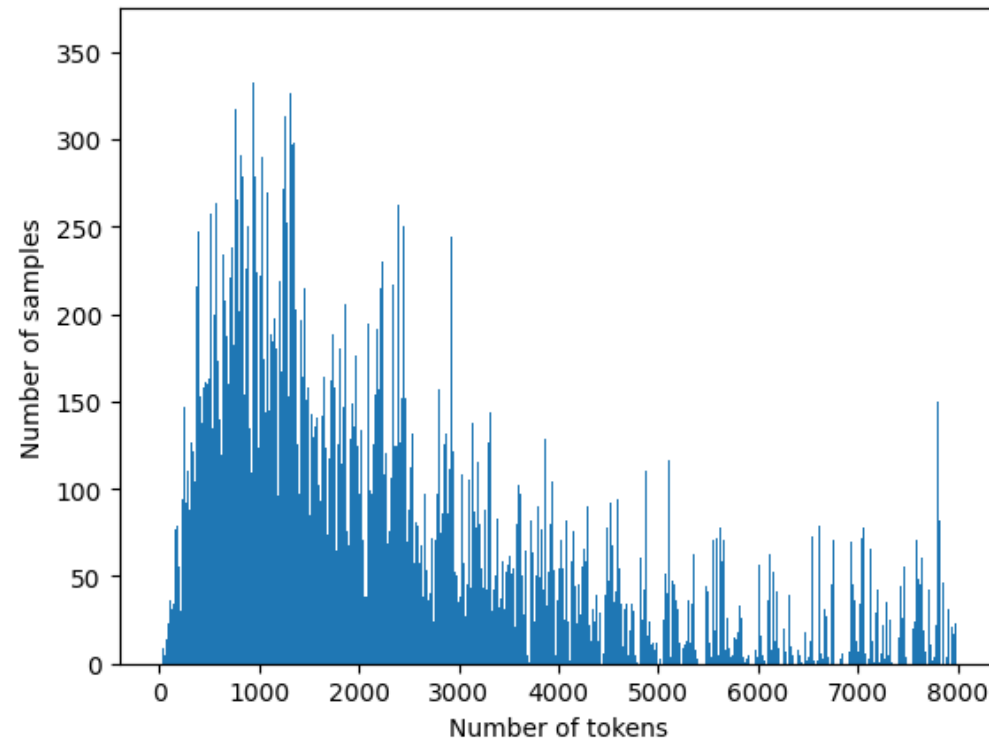
Methodology - Datasets

Dataset 2 Example

User Question	Matched Question	Context	Answer
insurance id card	Where can I get my Aetna ID card	The Aetna Medical Plan name is Aetna Choice POS II and the Group Number is 12345 (includes Fieldglass EEs) followed by your company code...	You will receive an Aetna ID card for yourself, which may list up to three dependents. If you have more than three dependents, you will receive an additional card showing those dependents (note that dependents info should be complete in their BenefitFocus profile for either SSN/DOB for Aetna to recognize the dependent and be included)...

Methodology - Datasets

Token count distributions of Dataset 1 & 2



Methodology - Data Preprocessing

After several preprocessing steps:

- remove invalid samples (NaN, numeric)
- discard irrelevant metadata

Methodology - Data Preprocessing

After several preprocessing steps:

- remove invalid samples (NaN, numeric)
- discard irrelevant metadata

We create 3 datasets to simulate different data environments:

FAQ (N≈48k)

- derived from Dataset 1
- standard questions
- represents a controlled environment for model training

User utterance dataset (N≈89k)

- combining Dataset 1 and 2
- **excluding** the manual corrections made by do-main experts
- represents a more realistic environment with inaccuracies

User utterance dataset with human in the loop (N≈89k)

- combining Dataset 1 and 2
- **Manual annotated contexts and answers** are used for wrong samples
- represents a more realistic environment enhanced by domain experts

Methodology - RQ1

How to effectively address the issue of lengthy input (context) in generative QA systems?

Methodology - RQ1

How to effectively address the issue of lengthy input (context) in generative QA systems?

We investigate the use of **efficient Transformers**

Methodology - RQ1

How to effectively address the issue of lengthy input (context) in generative QA systems?

We investigate the use of **efficient Transformers**

Efficient Transformers

- variant of Transformer models that aim to improve limitations of traditional Transformer models
- Enhanced computational and memory efficiency for handling lengthy inputs
- Examples: LongT5

Methodology - RQ1

Due to limited training resources, in this project we choose **LongT5** model and **T5** model for experiments.

LongT5

- `google/long-t5-local-base`
- $O(l)$
- Up to 16,384 tokens
- 296M trainable parameter

T5

- `t5-base`
- $O(l^2)$
- Up to 512 tokens
- 220M trainable parameters

Methodology - RQ2

Which analytical scores would be ideal to evaluate the performance of the models?

Methodology - RQ2

Which analytical scores would be ideal to evaluate the performance of the models?

Rouge Score

- comparing the overlap of n-grams
- lexical similarity

BERTScore

- cosine similarities between the contextual embeddings
- semantic similarity

Methodology - RQ3

How to accurately assess the performance of generative QA models in real-world scenarios?

Methodology - RQ3

How to accurately assess the performance of generative QA models in real-world scenarios?

We build a “Real User Question Only” dataset (N≈4k) for evaluation

- This dataset comprises only real user questions with correct context and answer pairs.
- It was constructed by filtering the test set of the “User utterance dataset with human in the loop” dataset to include only the real user question

Outline

1. Motivation

2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Evaluation & Results

We have a total of 9 configurations in our experiments:

ID	Model	Training Dataset	Max Tokens
1	T5	FAQ	512
2	T5	User utterance dataset	512
3	T5	User utterance dataset with human in the loop	512
4	LongT5	FAQ	512
5	LongT5	User utterance dataset	512
6	LongT5	User utterance dataset with human in the loop	512
7	LongT5	FAQ	5120
8	LongT5	User utterance dataset	5120
9	LongT5	User utterance dataset with human in the loop	5120

Evaluation & Results

We have a total of 9 configurations in our experiments:

ID	Model	Training Dataset	Max Tokens
1	T5	FAQ	512
2	T5	User utterance dataset	512
3	T5	User utterance dataset with human in the loop	512
4	LongT5	FAQ	512
5	LongT5	User utterance dataset	512
6	LongT5	User utterance dataset with human in the loop	512
7	LongT5	FAQ	5120
8	LongT5	User utterance dataset	5120
9	LongT5	User utterance dataset with human in the loop	5120

- Comparing the performance of T5 and LongT5 models trained on the three different datasets
- Assessing the influence of varying input lengths on the quality of the generated answers

Evaluation & Results

We have a total of 9 configurations in our experiments:

ID	Model	Training Dataset	Max Tokens
1	T5	FAQ	512
2	T5	User utterance dataset	512
3	T5	User utterance dataset with human in the loop	512
4	LongT5	FAQ	512
5	LongT5	User utterance dataset	512
6	LongT5	User utterance dataset with human in the loop	512
7	LongT5	FAQ	5120
8	LongT5	User utterance dataset	5120
9	LongT5	User utterance dataset with human in the loop	5120

- Comparing the performance of T5 and LongT5 models trained on the three different datasets
- Assessing the influence of varying input lengths on the quality of the generated answers

For each configuration, we choose the model checkpoint with the highest BERTScore score on the validation set, obtained at a specific training epoch
Then evaluate them on the Real User Question Only dataset.

Evaluation & Results

Rouge Score results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.568	0.331	0.410
User utterance dataset	0.581	0.409	0.432
User utterance dataset with HIL	0.677	0.506	0.601

BERTScore results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.879	0.798	0.838
User utterance dataset	0.883	0.827	0.849
User utterance dataset with HIL	0.913	0.859	0.906

Evaluation & Results

Rouge Score results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.568	0.331	0.410
User utterance dataset	0.581	0.409	0.432
User utterance dataset with HIL	0.677	0.506	0.601

BERTScore results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.879	0.798	0.838
User utterance dataset	0.883	0.827	0.849
User utterance dataset with HIL	0.913	0.859	0.906

Observations

- For both T5 and LongT5 models, the highest Rouge Score and BERTScores are achieved when trained on the **User utterance dataset with human in the loop**.

Evaluation & Results

Rouge Score results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.568	0.331	0.410
User utterance dataset	0.581	0.409	0.432
User utterance dataset with HIL	0.677	0.506	0.601

BERTScore results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.879	0.798	0.838
User utterance dataset	0.883	0.827	0.849
User utterance dataset with HIL	0.913	0.859	0.906

Observations

- For both T5 and LongT5 models, the highest Rouge Score and BERTScores are achieved when trained on the **User utterance dataset with human in the loop**.
- For 512 tokens when trained on the same dataset, the T5 models **consistently outperform** the LongT5 models

Evaluation & Results

Rouge Score results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.568	0.331	0.410
User utterance dataset	0.581	0.409	0.432
User utterance dataset with HIL	0.677	0.506	0.601

BERTScore results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.879	0.798	0.838
User utterance dataset	0.883	0.827	0.849
User utterance dataset with HIL	0.913	0.859	0.906

Observations

- For both T5 and LongT5 models, the highest Rouge Score and BERTScores are achieved when trained on the **User utterance dataset with human in the loop**.
- For 512 tokens when trained on the same dataset, the T5 models **consistently outperform** the LongT5 models
- LongT5 has better performance on **long input data**

Evaluation & Results

Rouge Score results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.568	0.331	0.410
User utterance dataset	0.581	0.409	0.432
User utterance dataset with HIL	0.677	0.506	0.601

BERTScore results

	Token = 512		Token = 5120
	T5	LongT5	LongT5
FAQ	0.879	0.798	0.838
User utterance dataset	0.883	0.827	0.849
User utterance dataset with HIL	0.913	0.859	0.906

Observations

- For both T5 and LongT5 models, the highest Rouge Score and BERTScores are achieved when trained on the **User utterance dataset with human in the loop**.
- For 512 tokens when trained on the same dataset, the T5 models **consistently outperform** the LongT5 models
- LongT5 has better performance on **long input data**
- For all configurations, BERTScores are **substantially higher** than the Rouge Score scores

Implications

- 1 The choice of training data has a significant impact on the performance of the models in practice

Implications

- 1 The choice of training data has a significant impact on the performance of the models in practice
- 2 Traditional transformers like T5 may have better performance on short-input data than efficient transformers

Implications

- 1 The choice of training data has a significant impact on the performance of the models in practice
- 2 Traditional transformers like T5 may have better performance on short-input data than efficient transformers
- 3 LLMs like T5 and LongT5 are capable of generating answers that may not have a close token-level match with the gold answers but still convey similar semantics

Outline

1. Motivation

2. Introduction

3. Methodology

4. Evaluation & Results

5. Conclusion & Future Work

Conclusion & Future Work

Conclusion

- Developed a generative QA chatbot tailored for HR domain
- Investigated T5 and LongT5 language models
- Impact of different training data distributions and input lengths
- Best performance: Real user questions with human intervention
- T5: Better for short inputs, LongT5: Promising for longer inputs

Future Work

- Employ more models for comparison
- Collect more realistic datasets for training and evaluation
- Use a broader range of evaluation metrics
- Model selection based on input length (T5 for short inputs, LongT5 for longer inputs)

Acknowledgement

We would like to express our sincere gratitude to SAP SE for their invaluable assistance and support throughout this project. Their generous provision of resources, including computational resources and datasets, has been instrumental in the successful completion of our work.

Thanks for Attention! Any Questions?