Generative Question Answering for a Chatbot in the Human Resources Domain

Tao Xiang
Technical University of Munich
tao.xiang@tum.de

Abstract

Traditional question-answering (QA) systems for the Human Resource domain face challenges such as the need to manually design user intents, which is not scalable. To address this challenge, we opt for generative QA systems that can automatically understand user intent and find the correct answer within the context. However, generative QA systems often encounter the problem of lengthy input texts in user questions and contexts. In this work, we aim to tackle this issue by exploring the use of an efficient transformer, LongT5, and comparing its performance to that of a conventional transformer, T5. To assess the impact of different training resources (data distribution) on the performance of QA systems in practice, we create three distinct datasets: (1) a dataset containing only highly structured questions, context, and answers; (2) a dataset with real user questions, context, and answers, where the context is retrieved by a retriever (possibly yielding inaccurate results); and (3) a modified version of the second dataset, with additional manual corrections for erroneous retrievals made by domain experts. These datasets are designed to simulate different situations, ranging from highly structured and controlled data to more realistic situations. The results reveal that T5 models outperform LongT5 in short-input scenarios, while LongT5 exhibits promising potential for handling longer inputs. Moreover, models trained on the third dataset yield the best performance in real-world scenarios. These results emphasize the importance of high-quality, realistic training data and selecting appropriate model architectures based on input length in the context of generative QA system development.

1 Introduction

The rapid advancement of artificial intelligence and natural language processing techniques has led to a growing interest in developing intelligent chatbots for a wide range of applications. One such application is the human resource (HR) domain, where chatbots can help answer employee queries, provide information about company policies, and assist with various HR-related tasks. Recognizing the potential of chatbots in the HR sector, organizations are seeking to leverage this technology to enhance employee experience and streamline HR processes.

Traditional question-answering (QA) chatbots have played a significant role in addressing user queries. However, these QA chatbots often need manual design of user intents, which might limit their capacity to handle complex and diverse questions. Consequently, there is a growing interest in generative question-answering chatbots, which can generate more natural and coherent responses by understanding the context and user queries more effectively.

In this study, we focus on creating a generative question-answering (GQA) chatbot tailored for HR departments. The primary goal of the chatbot is to understand and effectively respond to employee questions by leveraging the context available in the dataset. A key challenge in developing such a chatbot is dealing with lengthy input texts, which include user queries and contextual information. Lengthy texts can lead to increased computational complexity and may affect the chatbot's performance.

To address this issue, we investigate the use of efficient transformers, specifically the LongT5 model (Guo et al., 2021), which is designed to handle longer sequences more effectively than conventional transformer models, such as the T5 (Raffel et al., 2019).

Moreover, the quality of the dataset plays a crucial role in training an effective chatbot. In real-world scenarios, user questions can differ significantly from those found in pre-existing datasets. To address this challenge and ensure the chatbot's relevance and effectiveness, real user questions were collected during actual user consultations. These

real questions were then paired with appropriate context using a retrieval model. However, there were instances where the retrieval model provided inaccurate context. To resolve this issue, domain experts manually corrected the erroneous context, ensuring a higher-quality dataset. With this in mind, we constructed three distinct datasets for our experiments to understand the impact of data quality on the chatbot's performance in practice. These datasets are as follows:

1. A dataset containing only highly structured questions, context, and answers, representing a controlled environment for model training and evaluation.

2. A dataset comprising real user questions, context, and answers, where the context was retrieved by a retrieval model. This dataset presents a more realistic scenario, with the possibility of inaccuracies in the retrieved context.

3. A modified version of the second dataset, where additional manual corrections were applied by domain experts to the erroneous context retrieved by the model. This dataset demonstrates the potential improvements that can be achieved through manual intervention.

We train both LongT5 and T5 models on each of these three datasets, allowing us to compare the performance of different models under various data quality conditions. This approach provides valuable insights into the chatbot's effectiveness and adaptability when faced with diverse real-world scenarios and helps us determine the most suitable model configuration for an HR chatbot application.

The structure of this paper is organized as follows: section 2 delves into the background and relevant literature in the domain of generative question-answering chatbots and efficient transformers. In section 3, we outline the methodology employed in our research, with a primary focus on the datasets and their processing. section 4 presents the results obtained from our experiments. In section 5, we discuss the implications of the experimental outcomes and current limitations. Lastly, section 6 offers concluding remarks and suggests potential avenues for future research.

2 Related Work

2.1 Generated Question-Answering Chatbots

The development of generative question-answering chatbots has been an active area of research in recent years.

The introduction of deep learning techniques, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), enabled more sophisticated GQA systems. These models were capable of processing and understanding natural language input more effectively. Sutskever et al. (2014) demonstrated the effectiveness of RNNs for sequence-to-sequence tasks, which paved the way for the development of more advanced GQA systems.

The emergence of the Transformer architecture (Vaswani et al., 2017) has significantly improved the performance of GQA models. Transformers leverage self-attention mechanisms, allowing for greater parallelization and scalability. Notable examples of Transformer-based models include BERT (Devlin et al., 2018), GPT (Radford et al., 2018) and T5 (Raffel et al., 2019), which have achieved state-of-the-art performance on a variety of natural language processing tasks, including GQA.

2.2 Efficient Transformers

While the Transformer architecture has greatly improved the performance of GQA models, it also introduced increased computational and memory requirements, especially for handling long sequences. To address these limitations, several efficient Transformer variants have been proposed.

The Reformer (Kitaev et al., 2020) combines locality-sensitive hashing with reversible layers, enabling efficient processing of long sequences with limited memory consumption. Longformer (Beltagy et al., 2020) introduces a sliding window self-attention mechanism, allowing the model to handle longer text while maintaining the benefits of the Transformer architecture. The Linformer (Wang and Li, 2020) introduces low-rank approximations to reduce the complexity of the self-attention mechanism, making it more efficient for long sequences. The Linear Transformer (Katharopoulos et al., 2020) employs kernel-based self-attention, achieving linear complexity with respect to the sequence length. Big Bird (Zaheer et al., 2020) employs sparse attention patterns to

reduce computational complexity while preserving global context information. Lastly, Pegasus-x (Zhang et al., 2021) incorporates cross-attention, which allows the model to selectively focus on relevant input tokens, reducing computational costs.

Among these efficient Transformer variants, LongT5 (Guo et al., 2021) is particularly relevant to our work. LongT5 is a new model that integrates attention ideas from long-input transformers (ETC) (Liu et al., 2020) and pre-training strategies from summarization pre-training (PEGASUS) (Zhang et al., 2021) into the scalable T5 architecture. LongT5 is an extension of the T5 model that handles long sequence inputs more efficiently. LongT5 achieves state-of-the-art performance on several summarization benchmarks that required longer context or multi-document understanding.

We chose LongT5, specifically the local attention variant, over the transient-global (tglobal) attention variant for the following reasons: Firstly, the local attention-based LongT5 exhibits linear complexity with respect to input sequence length, making it more efficient for processing lengthy input texts. This is achieved through the sparse sliding-window local attention operation, which allows a given token to attend to only r tokens to its left and right, with r=127 by default. This yields a linear complexity of $O(l \cdot r)$ where l is the input sequence length (Guo et al., 2021). Secondly, the local attention mechanism does not introduce any new parameters to the model, maintaining its simplicity. With these techniques, the local attention-based LongT5 is capable of handling input sequences of a length up to 16,384 tokens, which is particularly beneficial for our task involving long input texts.

3 Methodology

3.1 Problem Statement & Notation

In this section, we introduce the primary notations used in this paper and formulate the task briefly.

Let $Q = q_1, q_2, ..., q_n$ denote a set of employee questions, and $C = c_1, c_2, ..., c_m$ represent a set of contextual documents containing information relevant to the HR domain. Each question q_i is associated with a context c_j that provides the necessary information to answer the question. The ground truth answer to question q_i is denoted by a_i. Our goal is to develop a generative question-answering chatbot that, given a question q_i and its corresponding context c_j, can generate a response $\hat{a}_i$ that is close to the ground truth answer a_i.

With the problem statement and notation introduced, we now proceed with the methodology, including the description of models, dataset preparation, training, and evaluation methods.

3.2 Models

In this experiment, we explore the performance of two models, T5 and LongT5. The specific model variants chosen for this study are as follows:

- For T5, we use the T5-base variant [1], comprising 220 million trainable parameters. This model represents a well-established transformer architecture, serving as a suitable baseline for evaluating the performance of LongT5 in the HR chatbot context.

- For LongT5, we employ the local-attention-based variant [2], which consists of 296 million trainable parameters. This model incorporates local attention mechanisms, allowing it to efficiently process long input sequences and potentially surpass the performance of the baseline T5 model.

Both models are of a similar scale in terms of trainable parameters, allowing for a fair comparison of their performance in the generative question-answering task. By analyzing the outcomes of this comparison, we aim to identify the most suitable model for addressing the challenges posed by lengthy input texts in HR chatbot applications, with T5 serving as our baseline model.

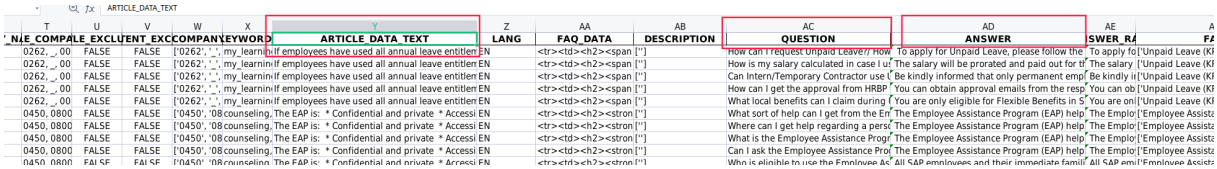
3.3 Dataset Preparation

3.3.1 Data Collection

SAP SE internally prepared two datasets for this study. The first dataset, referred to as Dataset A, consists of highly structured employee questions, context, and answers. The second dataset, referred to as Dataset B, comprises real user questions collected during actual user interaction. In this dataset, a retrieval model was employed to match real user questions with the most similar employee questions from Dataset A. The associated contexts and answers of the retrieved employee questions were then extracted and used as the paired contexts and answers for the real user questions.

[1] https://huggingface.co/t5-base

[2] https://huggingface.co/google/long-t5-local-base



T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	A
N|E_COMPA	LE_EXCLU	ENT_EXC	COMPANY|EYWORD		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 (KI
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 u	The salary will be prorated and paid out for th	The salary	['Unpaid Leave (KI
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 U	Be kindly informed that only permanent emp	Be kindly in	['Unpaid Leave (KI
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 (KI
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 onl	['Unpaid Leave (KI
0450, 0800	FALSE	FALSE	['0450', '08	counseling,	The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron	['']	What sort of help can I get from the En	The Employee Assistance Program (EAP) help	The Employ	['Employee Assista
0450, 0800	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 Employ	['Employee Assista
0450, 0800	FALSE	FALSE	['0450', '08	counseling,	The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron	['']	What is the Employee Assistance Prog	The Employee Assistance Program (EAP) help	The Employ	['Employee Assista
0450, 0800	FALSE	FALSE	['0450', '08	counseling,	The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron	['']	Can I ask the Employee Assistance Pro	The Employee Assistance Program (EAP) help	The Employ	['Employee Assista
0450, 0800	FALSE	FALSE	['0450', '08	counseling,	The EAP is: * Confidential and private * Accessi	EN	<tr><td><h2><stron	['']	Who is eligible to use the Employee As	All SAP employees and their immediate famili	All SAP em	['Employee Assista

(a) Snippet of Dataset A.

	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	CORRECTY	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, wit			
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 l			
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 Aetnaâ€™s sys			
11	0063, 0265	0.67305	0.501949	0.701227	0.601588	FALSE	FALSE	1.58E+09	internal	Logging into Benefitf	17:00.0	1.86E+09	How can I access my Aetna	Once your account has been created in the Aetnaâ€™s sys			

(b) Snippet of Dataset B.

Figure 1: Screenshots of Dataset A and B.

Both datasets also include additional information that is not relevant to this experiment. Furthermore, Dataset B contains manual annotations indicating the correctness of the retrieved employee questions. If the retrieved employee question was found to be incorrect, domain experts provided additional annotations with the correct employee questions. A screenshot of the raw Datasets A and B can be found in Figure 1.

3.3.2 Preprocessing

The preprocessing stage involved several steps to ensure the quality and relevance of the data used for training and evaluation. Initially, general preprocessing was applied to both datasets, which included the removal of samples containing invalid values, such as NaN or completely numeric values.

Subsequently, the datasets were processed as follows:

- For Dataset A: Only the Question, Context, and Answer columns were retained, discarding any irrelevant information.
- For Dataset B: For correct retrievals, the user question, context, and answer of the retrieved employee question were preserved. For incorrect retrievals, since the domain experts only annotated the correct employee questions without providing the corresponding context and answer, we additionally searched Dataset A to find the matching context and answer.

3.3.3 Creation of Datasets

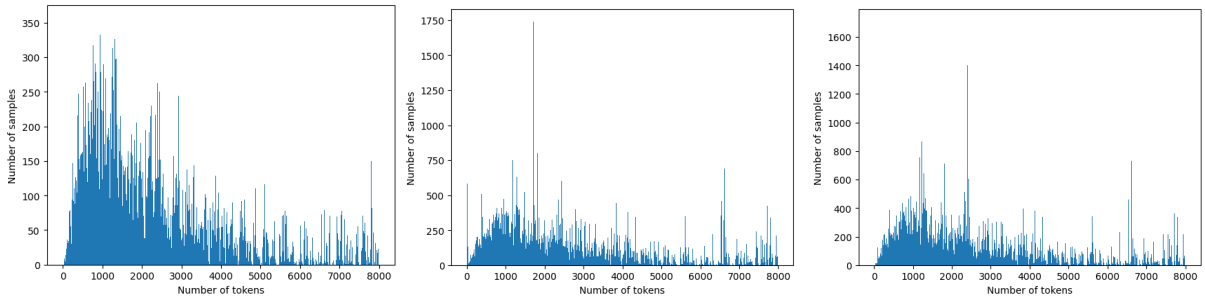
Using the preprocessed data, we created three distinct datasets for our experiments:

1. Highly structured dataset (N≈48k): This dataset is derived entirely from the preprocessed Dataset A, containing highly structured question-context-answer triples. This dataset provides a controlled environment for model training and evaluation.

2. Real user question dataset with inaccuracies (N≈89k): This dataset is constructed by merging the preprocessed Datasets A and B while excluding the manual corrections made by domain experts. Consequently, some samples in this dataset may have mismatched contexts and answers due to incorrect retrievals.

3. Real user question dataset with manual corrections (N≈89k): This dataset is also constructed by merging the preprocessed Datasets A and B, and it incorporates the manual corrections made by domain experts. For samples with incorrect retrievals, the erroneous context and answer are replaced with the corrected context and answer provided by the domain experts, resulting in a more accurate dataset.

These datasets offer a range of scenarios for training the chatbot models, from highly structured and controlled data to more realistic and challenging cases involving incorrect retrievals and manual corrections. After constructing these datasets, we divided them into training, validation, and test sets, with the validation and test sets each comprising 10% of the data.



(a) Highly structured dataset. (b) Real user question dataset with inaccuracies. (c) Real user question dataset with manual corrections.

Figure 2: The distribution of token counts across the three constructed datasets.

3.4 Model Training

Our experiments were conducted using a GPU with 16GB of memory: Tesla P100-PCIE.

For the T5 model, which has a maximum input length constraint of 512 tokens, we post-processed the datasets by filtering samples containing 512 tokens or fewer. This ensures that the contexts contain correct answers to the questions. During the training process, we utilized a learning rate of 1e-4 and a batch size of 8, and the T5 model was trained for a total of 15 epochs.

For the LongT5 model, considering the constraints imposed by our GPU's memory capacity, we filtered samples comprising 5120 tokens or fewer. This selection covers over 90% of the data in each of the three datasets. The distribution of token counts of the three datasets can be found in Figure 2. LongT5 was trained on both the short-input (512 tokens) and the longer (5120 tokens) versions of the datasets using the same learning rate of 1e-4 and batch size of 8 as with T5, completing 15 epochs on the short-input dataset and only 5 epochs on the long-input dataset due to limited training resources and the project's deadline. By training LongT5 on datasets with different length constraints, we aim to better understand its performance in handling longer input sequences compared to T5.

3.5 Evaluation Methods

In this subsection, we outline the evaluation methods used to evaluate the performance of the T5 and LongT5 models trained on the three distinct datasets. To gauge the question-answering capabilities in real-world situations, we construct an additional "Real User Question Only" dataset. We discuss the dataset, evaluation metrics, and model configurations in further detail below.

3.5.1 Evaluation Datasets

We build a "Real User Question Only" dataset (N≈4k): This dataset comprises only real user questions with correct context and answer pairs. It was constructed by filtering the test set of the "Real user question dataset with manual corrections" dataset to include only the real user questions. Notice that this dataset is different from the original test set since the original one also includes very standardized questions. We also post-process the dataset by filtering samples with less than or equal to 512 tokens for T5.

3.5.2 Evaluation Metrics

To evaluate the quality of the generated responses, we use the Rouge score (Lin, 2004) and BERTScore (Zhang et al., 2019). The Rouge score measures the similarity between the generated response and the ground truth answer by comparing the overlap of n-grams, with higher scores indicating better performance. BERTScore, on the other hand, computes token-wise cosine similarities between the contextual embeddings of the generated and ground truth sentences, using the F1 score for aggregation. By combining these metrics, we can assess the models' performance in terms of both lexical and semantic similarities to the ground truth answers.

3.5.3 Model Configurations for Evaluation

We have a total of 9 configurations in our experiments, which are as follows:

1. T5 model trained on the highly structured dataset, filtered to a maximum input length of 512 tokens.

2. T5 model trained on the real user question dataset with inaccuracies, filtered to a maximum input length of 512 tokens.

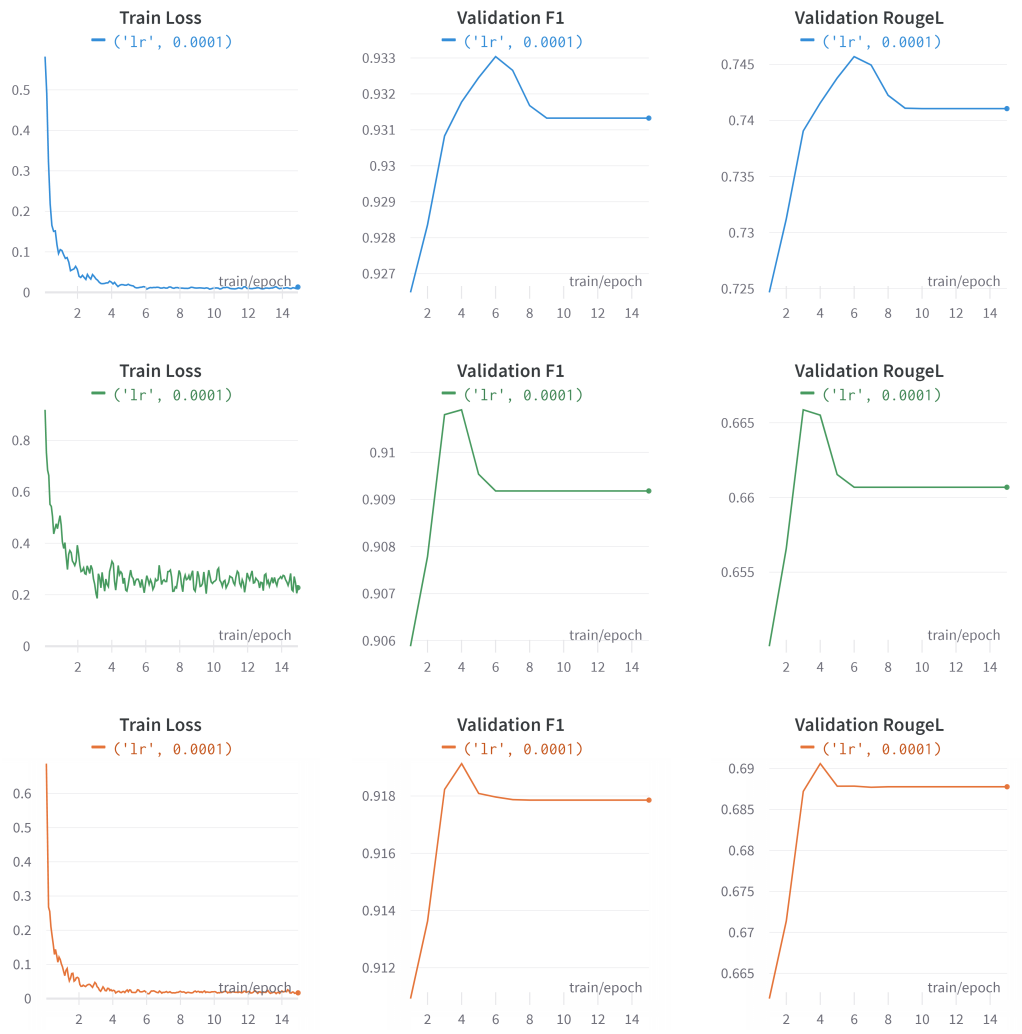


Figure 3: Training and validation results for T5 on the three distinct datasets with running average smoothing applied for better visualization. The blue, green, and orange curves represent the highly structured dataset, the real user questions dataset with inaccuracies, and the real user questions dataset with manual corrections respectively.

3. T5 model trained on the real user question dataset with manual corrections, filtered to a maximum input length of 512 tokens.

4. LongT5 model trained on the highly structured dataset, filtered to a maximum input length of 512 tokens.

5. LongT5 model trained on the real user question dataset with inaccuracies, filtered to a maximum input length of 512 tokens.

6. LongT5 model trained on the real user question dataset with manual corrections, filtered to a maximum input length of 512 tokens.

7. LongT5 model trained on the highly structured dataset, filtered to an maximum input length of 5120 tokens.

8. LongT5 model trained on the real user question dataset with inaccuracies, filtered to an maximum input length of 5120 tokens.

9. LongT5 model trained on the real user question dataset with manual corrections, filtered to an maximum input length of 5120 tokens.

These configurations enable us to compare the performance of T5 and LongT5 models trained on the three different datasets and assess the influence of varying input lengths on the quality of the generated answers.

For every configuration, we choose the model checkpoint with the highest F1 score on the validation set, obtained at a specific training epoch. This ensures that we evaluate the performance using the most optimized version of each configuration.

4 Results

In this section, we present the training and evaluation results.

4.1 Training Result

In this subsection, we discuss the training results of the T5 model on the three constructed datasets as shown in Figure 3.

The training loss curves for the first and third datasets appear relatively smooth in comparison to the second dataset which exhibits significant fluctuations. This suggests that the second dataset, containing erroneous samples, introduces a noticeable amount of noise into the training process. Additionally, the final training loss on the second dataset (around 0.2) is considerably higher than that on the first and third datasets, which are close to 0.

In terms of validation scores, the T5 model's BERTScore (F1) and Rouge scores (RougeL) on all three datasets exhibit a consistent trend. The T5 model achieves its highest scores (both BERTScore and Rouge) on the first dataset, which can be attributed to the highly structured nature of the dataset. As a result, the distribution of the training set is closely aligned with that of the validation set, leading to more impressive validation scores. On the other hand, the scores on the third dataset, which includes real user questions, are lower than those on the first dataset. We believe this is due to a higher degree of variance in the data distribution, which can be attributed to the presence of real user questions. This higher variance results in a less close alignment between the training and validation set distributions, which in turn leads to a modest reduction in scores. Similarly, the second dataset exhibits an even larger distribution variance due to real user questions and inaccurate contexts and answers. This leads to a greater dissimilarity between the training and validation sets. We hypothesize that this contributes to the noticeable fluctuations during the training phase and the lowest scores observed on the validation set.

The LongT5 (512) and LongT5 (5120) models also exhibit similar curve patterns, as detailed in Appendix A. The observations made for the T5 model also apply to these models, indicating that they show comparable behavior when trained on the different datasets.

4.2 Evaluation Result

In this subsection, we report the evaluation results for the nine model configurations on the "Real User Question Only" dataset. The performance of each model, measured using RougeL and F1 scores, is presented in Tables 1 and 2.

Table 1 presents the results of T5 and LongT5 models on the evaluation dataset filtered to a maximum of 512 tokens. For both T5 and LongT5 models, the highest RougeL and F1 scores are achieved when trained on the "Real user question dataset with manual corrections." We believe this is due to the distribution of the third dataset being closer to real user data. In contrast, the first dataset, "Highly Structured dataset," contains very standardized questions which often differ significantly from real-world data. This could explain why the models trained on this dataset perform the worst.

Although the "Real user question dataset with inaccuracies" includes real user data, the presence of inaccurate samples creates a gap between this dataset and real-world data, which may account for the intermediate scores of the models trained on it.

On the other hand, when trained on the same dataset, the T5 models consistently outperform the LongT5 models. We conjecture that this is because the T5 models are specifically pre-trained on data with a maximum length of 512 tokens, while LongT5 models are designed to handle much longer input texts (up to 16,384 tokens). The T5 models seem to be better suited for the datasets with a 512-token limit, which might explain their superior performance.

Table 2 shows the evaluation results of LongT5 on the evaluation dataset filtered to a maximum of 5120 tokens. It demonstrate a similar trend to the short input (512) evaluation: The models trained on the "Real user question dataset with manual corrections" achieve the highest scores, followed by those trained on the "Real user question dataset with inaccuracies," and finally, the models trained on the "Highly Structured dataset" perform the worst.

However, even after only five training epochs, with the metric scores for the three configurations not yet fully converging (as shown in Figure 5), the LongT5 models exhibit promising potential on handling long-input inputs. Specifically, the evaluation scores for LongT5 (5120) show noticeable improvements across all three configurations compared to their performance on the short-input

Model	Training Dataset	RougeL	F1
T5	(Highly Structured dataset, 512)	0.568	0.879
T5	(Real user question dataset with inaccuracies, 512)	0.581	0.883
T5	(Real user question dataset with manual corrections , 512)	**0.677**	**0.913**
LongT5	(Highly Structured dataset , 512)	0.331	0.798
LongT5	(Real user question dataset with inaccuracies , 512)	0.409	0.827
LongT5	(Real user question dataset with manual corrections, 512)	0.506	0.859

Table 1: Evaluation results of the first six configurations on the "Real User Question Only" dataset, filtered to a maximal 512 tokens. The best score for each metric is made bold.

Model	Training Dataset	RougeL	F1
LongT5	(Highly Structured dataset, 5120)	0.410	0.838
LongT5	(Real user question dataset with inaccuracies, 5120)	0.432	0.849
LongT5	(Real user question dataset with manual corrections, 5120)	**0.601**	**0.906**

Table 2: Evaluation results of the last three configurations on the "Real User Question Only" dataset, filtered to a maximal 5120 tokens. The best score for each metric is made bold.

dataset. We believe this can be attributed to the fact that LongT5 is inherently pre-trained on long texts.

Furthermore, the LongT5 scores on the long-input dataset are already approaching the T5 scores on the short-input dataset. Given that the LongT5 models have been trained for only five epochs and have not yet fully converged, it is possible that their performance on the long-input dataset may surpass that of the T5 models once the training is complete and fully converged.

An additional observation for all the configurations is that the F1 scores are substantially higher than the RougeL scores. We believe this can be attributed to the fact that large-scale generative models like T5 and LongT5 are capable of generating answers that may not have appeared in the training set. As a result, the generated responses might not have a close token-level match with the gold answers (Rouge), but they still convey similar meanings (BERTScore).

5 Discussion

In this section, we discuss the implications of our findings and the potential limitations of the current study.

5.1 Implications

The results of our experiments have several important implications for the development and application of language models in real-world question-answering tasks:

1. The choice of training dataset has a significant impact on the performance of the models. Highly structured datasets ease the learning process for models due to their standardized nature. Nevertheless, our results indicate that models trained on real user questions with manual corrections perform the best in real-world scenarios, as they closely resemble real-world data. This suggests that creating high-quality datasets containing realistic user-generated content is crucial for achieving better performance in practical applications.

2. The T5 models consistently outperformed the LongT5 models on the short-input dataset, while LongT5 showed promising potential in handling longer inputs. This indicates that selecting an appropriate model architecture based on the input length is an important factor to consider when developing question-answering systems.

3. The discrepancy between RougeL and F1 scores in our evaluation suggests that large-scale generative models 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. This implies that solely relying on token-level matching metrics for QA system evaluation may not provide a comprehensive assessment of the model's performance, and incorporating additional evaluation metrics that capture semantic similarity would be valuable.

Despite the insights gained from this study, there are several limitations to consider:

1. The evaluation metrics used in this study, namely Rouge and BERTScores, may not fully capture the semantic richness and diversity of the generated answers. Future studies could incorporate additional evaluation metrics, such as human evaluation or more sophisticated automatic metrics, to better assess the quality of the generated responses.

2. Our experiments focused on a limited set of model configurations and training datasets. Further research could explore the performance of other language models, such as BERT or GPT-4, on the same datasets, or investigate the impact of additional training data sources on model performance.

3. The results presented in this study are based on a single evaluation dataset, the "Real User Question Only" dataset. It is possible that the performance of the models could vary depending on the evaluation dataset used. Future studies could utilize multiple evaluation datasets to validate the findings of this study.

4. Due to limited training resources, LongT5 has not been completely trained on long-input datasets. Consequently, the potential performance of the LongT5 model for handling longer inputs might not be entirely realized. Further training and experiments are necessary for future work.

6 Conclusion

In this study, we aimed to develop a generative question-answering (GQA) chatbot tailored for HR domain. We investigated the performance of T5 and LongT5 language models on generative question-answering tasks, particularly focusing on the impact of different training data distributions and input lengths. To achieve this, we constructed three datasets that range from standardized questions to real-user inquiries with human intervention and further divided them based on the input length constraints. Our findings revealed that the choice of training dataset plays a crucial role in model performance, with models trained on real user questions with human intervention achieving the best results. Moreover, the T5 models performed better on short-input datasets, while LongT5 demonstrated promising potential for handling longer inputs.

Due to limited training resources, this study is subject to certain limitations, such as the incomplete training of LongT5 on the long-input dataset, reliance on a single evaluation dataset, and the use of only two evaluation metrics. Future research should aim to complete the training process of LongT5, employ more models for comparison, collect more realistic datasets for both training and evaluation, and employ a broader range of metrics for assessment.

Despite the limitations of this study, our experiments offer valuable insights that can guide future research and development in the field of HR chatbots: (1) High-quality and realistic training data are crucial for achieving practical chatbot performance. As a possible direction for future work, a semi-supervised learning approach could be employed: initially training the chatbot on standardized datasets and later incorporating real user questions collected during deployment for continuous refinement. (2) Model architecture selection should be based on input length. For example, if the input length is less than 512 tokens, traditional transformer models like T5 may be suitable, while efficient transformers such as LongT5 should be considered for longer input sequences.

In conclusion, our research represents a significant step towards creating effective and adaptable GQA chatbots for HR departments. As organizations increasingly seek to leverage AI-powered chatbots to enhance employee experience and streamline HR processes, the insights gained from this study will be instrumental in guiding the development of high-performing chatbots that can effectively address diverse real-world scenarios.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Anum Afzal, for her continuous guidance, support, and encouragement throughout this project. Her knowledge and insights have been invaluable in shaping my research and its outcomes. I am also deeply thankful to Prof. Dr. Florian Matthes, who has provided me with invaluable advice and guidance during the course of this project. His expertise and dedication to fostering a positive research environment have been truly inspiring. I would like to extend my appreciation to SAP SE for their cooperation and support, which have been instrumental in the success of this research. Their valuable input and collaboration have greatly con-

tributed to the development of the work presented in this paper. Finally, I am grateful to my friends and family for their unwavering support and encouragement during my academic journey. Their belief in me has motivated me to persevere and achieve my goals.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mandy Guo, Yizhong Chen, Xiaodong Li, Xiaodan Liang, Fei Liu, Jianfeng Sun, and Ming Zhou. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

Zihang Liu, Paul Michel, and Xiaodong Liu. 2020. Etc: Encoding long and structured inputs in transformers.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, and Yanqi Zhou. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Sinong Wang and Belinda Z Li. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2021. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:2102.01569*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Training Results of LongT5

Figure 4: Training results of LongT5 (512) on the three datasets with running average smoothing applied for better visualization. The blue, green, and orange curves represent the highly structured dataset, the real user questions dataset with inaccuracies, and the real user questions dataset with manual corrections respectively.

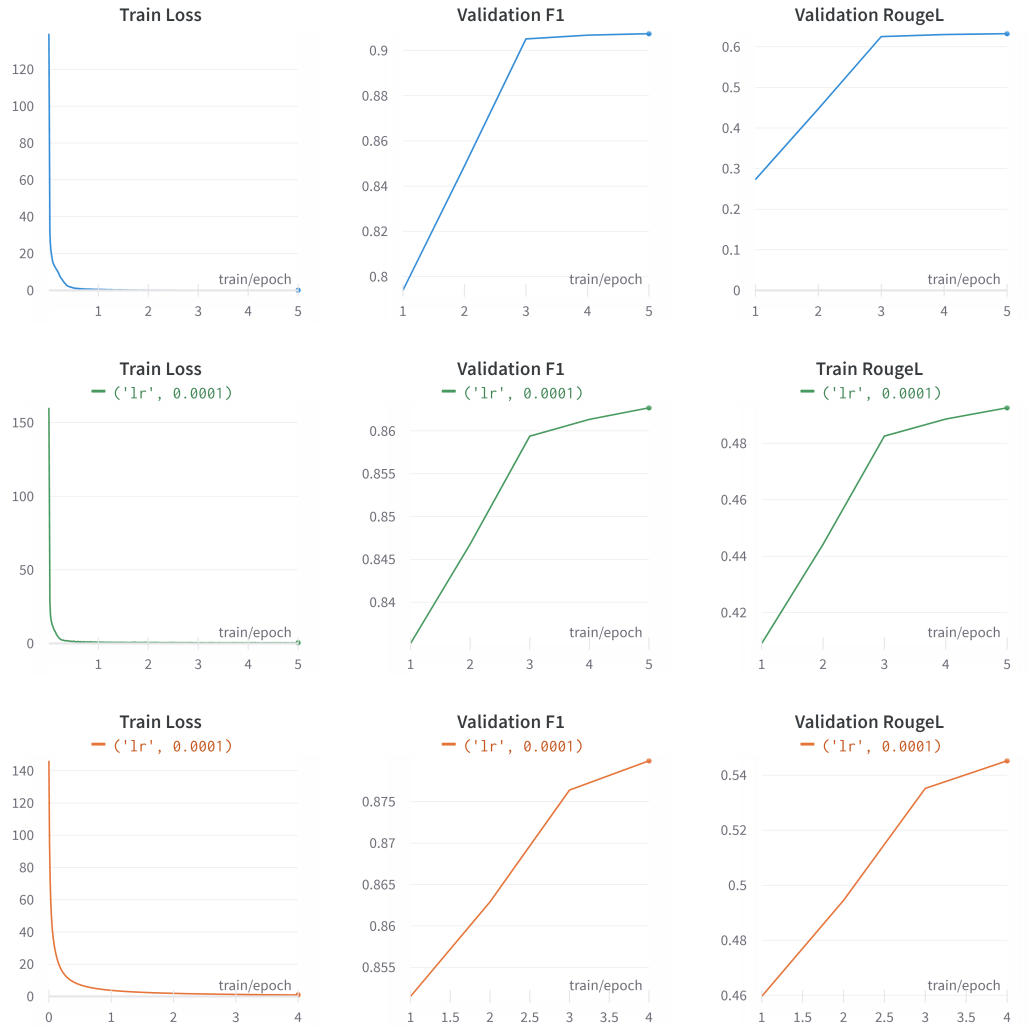


Figure 5: Training results of LongT5 (5120) on the three datasets with running average smoothing applied for better visualization. The blue, green, and orange curves represent the highly structured dataset, the real user questions dataset with inaccuracies, and the real user questions dataset with manual corrections respectively.