Comparative Analysis of 47 Context-Based Question Answer Models Across 8 Diverse Datasets

Muhammad Muneeb[1,2], David B. Ascher[1,2,*], and Ahsan Baidar Bakht[3]

[1]School of Chemistry and Molecular Biology, The University of Queensland, Brisbane, 4067, Australia
[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, 3004, Australia
[3]Mechanical Engineering Department, Khalifa University, Abu Dhabi, UAE
[*]d.ascher@uq.edu.au

ABSTRACT

Context-based question answering (CBQA) models provide more accurate and relevant answers by considering the contextual information. They effectively extract specific information given a context, making them functional in various applications involving user support, information retrieval, and educational platforms. In this manuscript, we benchmarked the performance of 47 CBQA models from Hugging Face on eight different datasets. This study aims to identify the best-performing model across diverse datasets without additional fine-tuning. It is valuable for practical applications where the need to retrain models for specific datasets is minimized, streamlining the implementation of these models in various contexts. The best-performing models were trained on the SQuAD v2 or SQuAD v1 datasets. The best-performing model was ahotrod/electra_large_discriminator_squad2_512, which yielded 43% accuracy across all datasets. We observed that the computation time of all models depends on the context length and the model size. The model's performance usually decreases with an increase in the answer length. Moreover, the model's performance depends on the context complexity. We also used the Genetic algorithm to improve the overall accuracy by integrating responses from other models. ahotrod/electra_large_discriminator_squad2_512 generated the best results for bioasq10b-factoid (65.92%), biomedical_cpgQA (96.45%), QuAC (11.13%), and Question Answer Dataset (41.6%). Bert-large-uncased-whole-word-masking-finetuned-squad achieved an accuracy of 82% on the IELTS dataset. Palak/microsoft_deberta-large_squad accuracy was 31% on the JournalQA dataset. Twmkn9/albert-base-v2-squad2 was the best-performing model for the ScienceQA dataset with an accuracy of 24.6%. This study contributes to the broader goal of optimizing the use of CBQA models across diverse datasets by providing insights into the impact of context complexity, answer length, and question types on model performance.

Introduction

In the era of information abundance, the need for efficient extraction and retrieval of relevant information from vast textual repositories has become paramount[1,2]. Question-answering (QA) models are pivotal in facilitating human-machine interaction by interpreting complex linguistic structures, discerning contextual nuances, and generating accurate responses[3]. QA tasks are categorized into three classes: Answer generation, where the answer is generated based on the question; answer selection, where models choose answers from multiple options; and answer extraction, where models extract answers from the context[4]. We considered models that use a question and context to extract an answer.

A context-based question answering (CBQA) model is a type of natural language processing (NLP) model designed to understand and respond to questions in the context of a given passage or document[5–9]. The exigency for QA models is underscored by their applications across diverse domains, including information retrieval[10], customer support, education[11], and medical literature[12–14], QA models also aid in extracting information from research articles, enhancing the efficiency of literature reviews and evidence-based decision-making[15].

CBQA models initiate their process with tokenization[16] that combines the question with the context, and the data is then segmented into smaller units known as tokens. These tokens are represented as vectors in a high-dimensional space through embedding, encompassing token, position, and segment embedding (capturing the semantic meaning and relationships between words)[17]. The embeddings are passed to the transformer, which uses a self-attention mechanism[18] to capture dependencies between words, irrespective of their position in the sequence. Subsequently, QA models undergo pre-training on extensive datasets with a specific objective[19]. For instance, in the Masked Language Model (MLM) objective[20], a portion of tokens is masked (replaced with a unique token called MASK), and the model predicts the correct token in place of MASK to acquire contextualized language representations[21]. This pretraining phase aids the model in developing a broad understanding of language. Following pretraining[22], the model undergoes fine-tuning[23] on a task-specific dataset for QA. The question-answering head is added at this stage for fine-tuning the QA datasets, which predicts the answer's start and end token.

The evaluation of CBQA models typically involves metrics such as Exact Match (EM) and F1 score, assessing the model's ability to generate the exact answer span or a close approximation[24,25]. Additionally, researchers are exploring more nuanced evaluation methods that consider aspects such as answer relevance, coherence, and informativeness[26].

Benchmarking question-answering (QA) models is critical to assess and compare their performance systematically, providing a basis for understanding their capabilities and limitations across diverse tasks and datasets[27,28]. Benchmarking also serves to identify areas of improvement and challenges in QA models. The diverse nature of benchmarks allows for a comprehensive evaluation of models' performance in different domains, contextual complexities, and linguistic nuances. Such evaluations aid researchers in identifying strengths and weaknesses, refining existing models, and inspiring the development of novel methodologies.

Hugging Face is a popular platform for natural language processing models[29]. It hosts an extensive repository of question-answering models, including transformer-based models like BERT[21], GPT, and RoBERTa[30], Dilbert[31], ELECTRA[32] and T5[33]. Pre-trained on large corpora, these models have demonstrated state-of-the-art performance across a spectrum of NLP tasks, including question-answering. We considered 47 context-based question-answer models from the Hugging Face.

This study explored several crucial aspects of model performance and efficiency across diverse datasets. Our critical investigations encompassed (1) Identifying the top-performing models based on accuracy and computation time across eight distinct datasets, (2) Analyzing the impact of model size and context length on execution time, (3) Investigating the relationship between context complexity and model performance, (4) Assessing the impact of answer length on the accuracy of the model, providing insights into how different answer lengths influence the overall effectiveness of the model, (5) Examining the influence of answer types on model performance, offering valuable perspectives on how diversity of answers challenge the model's predictive capabilities, and (6) Implementing an ensemble-based approach that leverages genetic algorithms to discover optimal combinations of models, resulting in superior overall performance.

We comprehensively explored critical factors influencing model behavior, offering a nuanced understanding of their collective impact on computation time and overall performance.

Results

Among the top ten models, overall performance exceeded 30 percent 1. While many models exhibited similar overall performance, models' performance on individual datasets varied. Top-performing models were trained on the SQuAD v2 or SQuAD v1 datasets 1.

	Model Name	**Model Size (MB)**	**Dataset**	**Overall Performance**
1	ahotrod/electra_large_discriminator_squad2_512	319	SQuAD v2	0.432683
2	bert-large-uncased-whole-word-masking-finetuned-squad[34]	320	BookCorpus and English Wikipedia	0.422568
3	Palak/microsoft_deberta-large_squad	387	SQuAD v1	0.422074
4	bhadresh-savani/electra-base-squad2	104	SQuAD v2	0.407044
5	twmkn9/albert-base-v2-squad2	12	SQuAD v2	0.384236
6	csarron/mobilebert-uncased-squad-v2	24	SQuAD v2	0.383692
7	valhalla/bart-large-finetuned-squadv1	388	SQuAD v1	0.383498
8	deepset/bert-base-cased-squad2	104	SQuAD v2	0.37514
9	deepset/roberta-base-squad2	119	SQuAD v2	0.370862

Table 1. This table shows information about the top 9 models, including the model name, size in megabytes, the dataset used for training, and the overall performance.

The second-best performer, Bert-large-uncased-whole-word-masking-finetuned-squad, was trained on BookCorpus and English Wikipedia. Other algorithms trained on the same dataset did not yield better results, potentially due to differences in model sizes. For instance, this model had a size of 320 MB, while other models trained on the same dataset had a size below 100 MB, resulting in low accuracy.

Table 1 highlights the top-performing models. ahotrod/electra_large_discriminator_squad2_512 achieved an overall accuracy of 43%, with a size of 319 MB. In contrast, twmkn9/albert-base-v2-squad2 achieved an accuracy of 38% with a significantly smaller size of 12 MB. Model size indeed impacted performance; however, the base model used for training also played a crucial role. Models trained on various base models (Electra, Albert, MobileBERT, BERT, and Roberta) exhibited performance in descending order, as shown in Table 1. We conducted a t-test (p-value < 0.1) to evaluate the difference in the performance of the models. It involved comparing the performance of each model against every other model across eight datasets. The results indicate no statistically significant difference among the top 11 best-performing models 2.

We observed that the context length and the model size have a noticeable impact on the model execution time. The correlation between the overall execution time of the models and their size across all datasets was 51%, indicating a significant

Figure 1. This heatmap illustrates the performance of each model across eight datasets. We computed the overall performance by summing the model performance for a specific dataset and dividing it by the number of datasets. We sorted the entries in the heatmap based on overall performance.

increase in execution time with larger models. However, the correlation between the overall execution time and overall performance across all datasets was 13%, suggesting that longer execution time or larger models may not necessarily improve the performance. It aligns with previous inferences, emphasizing the importance of selecting datasets and the base model for training to achieve better results. There was a correlation of -8% between overall performance and model size. The negative correlation implies that an increase in model size does not necessarily lead to an increase in overall performance. Examining specific models, the first (ahotrod/electra_large_discriminator_squad2_512) and fourth (bhadresh-savani/electra-base-squad2) models 3 demonstrated comparable performance (approximately 0.43 and 0.41, respectively). However, a substantial difference in execution time was observed due to the varying sizes of the models, which were 319 MB and 104 MB, respectively. In scenarios where rapid results are required and a compromise on performance is acceptable, opting for the smaller-sized model is advisable.

We conducted a t-test (p-value < 0.1) to evaluate the difference in the computation time of the models. It involved comparing the computation time of each model against every other model across eight datasets. The results indicate no statistically significant difference among the top 6 best-performing models 4.

Figure 2. This heatmap depicts the t-test comparing the performance of each model across datasets. We sorted the values based on performance and calculated the t-test (p-value < 0.1). If the p-value for a t-test was less than 0.1, we assigned a value of 1 to that particular entry; otherwise, it is marked as 0. This representation indicates whether a specific pair of two models is statistically significant.

We calculated the correlation between the model execution time on eight datasets and overall performance (Figure: 3 and Table 2). There was an increase in correlation between performance and model execution time, suggesting more execution time can bring better performance; however, this correlation might be coincidental. To further explore, we computed the correlation between the model execution time for each dataset and the model size. Initially, there was an increase, followed by a decrease as the context length increased. It implies that the model size alone does not explain the execution time; the context length also plays a significant role. If this correlation had increased linearly, it would have suggested that context length does not

Figure 3. This heat map illustrates the average execution time in minutes for each model across all datasets, overall execution time, overall performance, and the respective model sizes. The average execution time for a specific model on a given dataset was calculated by dividing the model execution time on a specific dataset by the number of questions in that dataset. This calculated average provides valuable insight into visualizing the impact of context length on execution time. Model sizes were normalized between 0 and 1 using min-max normalization.

impact model performance and that only the model size influences execution time. Additionally, we examined the correlation between the model execution time for each dataset and overall execution time, revealing that as the context length increases, the execution time also increases. All these observations are shown in Table 2.

We analyzed the impact of answer length on model performance 5. Our findings revealed that the performance of the top-performing model could exhibit variations with changes in answer length. Specifically, we identified a consistent decline in model performance as the answer length increased. For instance, the model ahotrod/electra_large_discriminator_squad2_512 emerged as the best performer overall. However, when predicting answers consisting of only two words, the model bhadresh-savani/electra-base-squad2 demonstrated a noteworthy accuracy of 55%, outperforming the best model, which achieved an accuracy of 0.47%.

Table 3 presents the models that yielded the best results across various answer lengths for all datasets. The values are displayed in pairs, following the format [Answer length, Best Accuracy]. The optimal performing model can vary based on the

Figure 4. This heatmap depicts the t-test comparing the execution time of each model across datasets. We sorted the values based on performance and calculated the t-test (p-value < 0.1). If the p-value for a t-test was less than 0.1, we assigned a value of 1 to that particular entry; otherwise, it is marked as 0. This representation indicates whether a specific pair of two models is statistically significant.

length of the answer. When predicting an answer of a specific length, it becomes imperative to identify which model is likely to produce the best results for the answer of a specific length. This variation is influenced by factors such as model architecture and other design considerations. Some models are tailored to pinpoint the precise answer, while others are designed to return more contextual information around the actual answer, resulting in a more detailed response.

We conducted a thorough analysis to investigate context complexity's influence on the performance of models. Our findings reveal a notable correlation between varying context complexities and the deviations in model performance 6. However,

	Model Name	Average Context Length	Overall Performance	Model Size (MB)	Overall Time in Minutes
1	atlas-math-sets	3	0.04	0.72	0.78
2	QuAC	84	0.05	0.77	0.78
3	ScienceQA	147	0.07	0.71	0.68
4	biomedical_cpgQA	161	0.04	0.78	0.8
5	bioasq10b-factoid	351	0.05	0.78	0.81
6	IELTS	907	0.04	0.77	0.82
7	Question Answer Dataset	5227	0.13	0.42	0.99
8	JournalQA	5568	0.16	0.43	0.99

Table 2. This table depicts the correlation between the execution time of all models for a specific dataset and the overall performance, model size, and total execution time for each dataset. The average context length is presented in the second column.

	models	atlas-math-sets	bioasq10b-factoid	biomedical_cpgQA	ielts	JournalQA	QuAC	ScienceQA	Question Answer Dataset
1	twmkn9/albert-base-v2-squad2	-	-	-	-	-	-	[[1, 0.36], [3, 0.08]]	-
2	valhalla/bart-large-finetuned-squadv1	[[4, 1.0]]	[[1, 0.74]]	[[5, 1.0]]	-	-	-	[[2, 0.21]]	-
3	deepset/roberta-base-squad2	-	[[3, 0.58]]	-	[[2, 0.94]]	-	-	-	-
4	Palak/microsoft_deberta-large_squad	-	-	[[3, 1.0]]	-	[[3, 0.18], [4, 0.43]]	-	-	-
5	bhadresh-savani/electra-base-squad2	[[1, 0.04], [2, 1.0]]	-	-	-	[[2, 0.43]]	-	-	[[1, 0.31], [5, 0.57]]
6	gpt2	-	-	-	[[4, 1.0]]	-	-	-	-
7	allenai/longformer-large-4096-finetuned-triviaqa	-	[[2, 0.65]]	-	-	-	[[3, 0.1], [4, 0.05]]	-	-
8	csarron/mobilebert-uncased-squad-v2	-	-	[[1, 0.98]]	-	-	-	-	[[3, 0.63]]
9	SRDdev/QABERT-small	-	-	-	-	-	-	[[4, 0.18]]	-
10	bert-large-uncased-whole-word-masking-finetuned-squad	-	-	-	[[3, 0.92]]	[[1, 0.27]]	-	-	-
11	ahotrod/electra_large_discriminator_squad2_512	[[3, 0], [5, 0]]	[[4, 0.47], [5, 0]]	[[2, 0.95], [4, 0.95]]	[[1, 0.89], [5, 0]]	[[5, 0.6]]	[[1, 0.23], [2, 0.09], [5, 0]]	[[5, 0]]	[[2, 0.74], [4, 0.61]]

Table 3. This table illustrates the comparative performance of different models for each answer length. The values are presented in the format [Answer Length, Best Accuracy]. "-" denotes that a particular model did not emerge as the best performer for any specific answer length.

intriguingly, we did not find any discernible pattern regarding the specific impact of context complexity on model performance. Further exploration is needed to unravel the nuanced dynamics that contribute to the observed variations in performance.

We employed a genetic algorithm to discover combinations of models that could enhance overall performance. We aimed to generate a more robust and powerful model by leveraging the weaker models. While a combination of models can obtain optimal performance, the improvement is not statistically significant. In many instances, the performance gains achieved by combining models were either on par with the best individual model or marginally less, at the expense of increased computation time.

To conduct our analysis, we employed a 5-fold cross-validation approach to generate training and test sets. The genetic algorithm was then utilized to optimize the number of models included in the optimal configuration, ranging from 2 to 47 models. The genetic algorithm selected the model with the highest answer score, optimizing the accuracy of the training data. Subsequently, we evaluated the performance of the combined model on the test set and reported improvements compared to the baseline 7.

The JournalQA dataset comprised questions derived from two papers, one centered on solar power prediction and the other on the polygenic risk associated with diabetes. Our observations revealed that the effectiveness of the question-answer model varied depending on the type of information extracted from the papers. To simplify this diversity, we categorized information into ten groups: Type, Performance, Location, Time, Date, Features, Output, Evaluation metric, Best Model, Number, and Dataset. The performance category means the optimal achievements of specific algorithms, while location details the specific geographical information. Time means date ranges. Output means the evaluation metrics or output of the system discussed in the paper, and evaluation metrics detail the specific metrics. The best model identified the highest-performing model, the number conveyed numerical information, and the data specified the dataset used in the specific study. Lastly, the type category encompassed anything related to categorization.

We visualized the performance of all models through a heatmap 8, offering insights into the efficacy of models for specific question types using research articles as plain text as context. Notably, one model can only extract some types of questions. Question types impact the model's performance, requiring various models to extract a diverse range of question types.

Table 4 shows which model is best for which type of question time regarding journal-related question answering.

We analyzed models that exhibited poor performance and observed that such models were designed for masked language, multi-language (cross-language) tasks, Chinese language, or other specific language. Additionally, for a significant portion of underperforming models, the corresponding documentation lacked information about the datasets used. This limitation constrained our analysis, as we could only glean insights from the limited performance of these models.

Figure 5. This heatmap illustrates the impact of answer length on model performance. As the question length increases, the models' performance may either decrease or increase. The answer length spans from 1 to 5 words, and we calculated the percentage of correctly and incorrectly classified answers by each model for each answer length across all datasets. This information is then utilized to generate the heatmap.

Index	Model	Journal question type
1	twmkn9/albert-base-v2-squad2	['Dataset', 'Type']
2	deepset/roberta-base-squad2	['Date', 'Performance']
3	Palak/microsoft_deberta-large_squad	['Evaluation metric', 'Features', 'Time']
4	bhadresh-savani/electra-base-squad2	['Best Model', 'Output']
5	bert-large-uncased-whole-word-masking-finetuned-squad	['Number']
6	ahotrod/electra_large_discriminator_squad2_512	['Location']

Table 4. Table showing the best model for specific question types for the JournalQA dataset.

The best-performing model for each dataset and the best accuracy are shown in table 5.

Figure 6. This heatmap illustrates the impact of complexity on model performance. As the complexity increases, the models' performance may either decrease or increase. The complexity measure spans from 5 to 17, with an interval of 2. No specific pattern was observed for this analysis.

Discussion

This comprehensive study utilized 47 context-based question-answering models across eight diverse datasets. We conducted a thorough analysis to explain the impact of dataset diversity, model size, context length, context complexity, answer length, and question type on model performance. The insights from this research provide valuable considerations for optimizing question-answering tasks.

The performance analysis identified the top 11 models as statistically significant compared to others. This finding suggests that any of these models, particularly those with less computation time, can be effectively deployed for question-answer prediction tasks. The ahotrod/electra_large_discriminator_squad2_512 model appeared as the best-performing algorithm, making it a recommended option, especially when dealing with custom datasets that differ significantly from those considered in this study. Furthermore, models trained on SQuAD v2 or SQuAD v1 datasets showed the potential for outperforming models trained on

Figure 7. The heatmap represents the performance improvements achieved by the genetic algorithm when optimizing performance using a combination of models. Each cell displays the averaged five-fold training and test accuracy for each dataset across combinations of 2 to 47 models shown on the y-axis. The heatmap illustrates that, on the whole, there was no substantial improvement in performance.

other datasets despite having similar architectures. Examining the impact of answer length on performance revealed a decrease in model performance as the answer length increased. For tasks (medical treatment and disease name extraction) requiring accurate results, selecting a model suitable for handling long answers, such as ahotrod/electra_large_discriminator_squad2_512, is crucial. While larger models may lead to increased computation time, for tasks where a slight decrease in accuracy is acceptable for faster results, opting for a model with a small size is advisable. We observed no specific variation in performance concerning context complexity. Contrary to expectations, combining models based on their scores yielded negligible improvement. We recommend simultaneous training and a combination of multiple architectures during the pretraining stage to enhance model performance. This strategy allows for the concurrent updating of weights in the combined model, thereby improving its learning capability.

Furthermore, our analysis of question types within the JournalQA dataset revealed that the best-performing model varied

Figure 8. The heatmap illustrates the performance of models across various question types within the JournalQA dataset. Notably, there is variation in model performance across different question types.

for each question type. Therefore, a tailored approach based on question type classification is recommended. We suggest incorporating a classification head during pretraining to categorize question types, enabling selecting the best-performing model based on the specific question type. The referenced models, datasets, and their corresponding performance implications contribute to the broader understanding of effective model selection for diverse applications, particularly in extracting answers from journal texts. The code used for the analysis conducted in this study is accessible on GitHub at the following link: https://github.com/MuhammadMuneeb007/Comparative-Analysis-of-47-Context-Based-Question-Answer-Models-Across-8-Diverse-Datasets. Additionally, the pre-and post-processed datasets can be downloaded from The Open Science Framework (OSF) using the following link: https://osf.io/rd3x2/.

Index	Dataset name	Best Model	Accuracy
1	atlas-math-sets	bhadresh-savani/electra-base-squad2	3.72%
2	bioasq10b-factoid	ahotrod/electra_large_discriminator_squad2_512	65.92%
3	biomedical_cpgQA	ahotrod/electra_large_discriminator_squad2_512	96.45%
4	IELTS	bert-large-uncased-whole-word-masking-finetuned-squad	82%
5	JournalQA	Palak/microsoft_deberta-large_squad	31%
6	QuAC	ahotrod/electra_large_discriminator_squad2_512	11.13%
7	ScienceQA	twmkn9/albert-base-v2-squad2	24.6%
8	Question Answer Dataset	ahotrod/electra_large_discriminator_squad2_512	41.6%

Table 5. This table shows the dataset, the best-performing model for each dataset, and the final accuracy achieved by the model.

Materials and Methods

We considered eight different datasets from various sources to assess the performance of 47 question-answering models 6. **Atlas-math-sets** centers around mathematical concepts related explicitly to sets and mathematical operations (addition, subtraction, multiplication, and division). **Bioasq10b-factoid** contains questions from the BioASQ challenge, which contains biomedical question answers. **Bbiomedical_cpgQA** is tailored for question-answering tasks within the biomedical domain, explicitly focusing on Clinical Practice Guidelines (CPG). It encompasses a set of questions from medical practices and guidelines, providing a comprehensive resource for training and evaluating models in the healthcare domain. **QuAC** (Question Answering in Context) is a dataset for advancing question-answering research. Questions are posed within a conversational context, demanding the model to consider the complexities of previous dialogue. **ScienceQA** focuses on questions and answers related to scientific topics. The **Question Answer Dataset** consists of articles from Wikipedia containing context, questions, and answers. The **IELTS** (International English Language Testing System) dataset is designed to assess English language proficiency across listening, reading, writing, and speaking skills. The IELTS dataset is centered on the English language, encompassing reading comprehension with corresponding answers. We compiled the dataset of 50 questions from various online IELTS sample tests. The actual URL to the test, the passage or context, the question, and the answer to that question are available on GitHub. We considered only those IELTS questions that ended with a question mark sign. The **JournalQA** dataset is crafted from scholarly articles. Research articles contain information like the model's performance, the best-performing model, population, specific methodology, and other information that can be extracted using question-answer models. To generate such questions, we used two systematic literature reviews. The first review focused on articles related to solar power prediction[35], while the second centered on using polygenic risk scores for predicting type 2 diabetes[36].

	Dataset	Link	Preprocessing	Set	Answer Word Threshold	Type	Number of Questions
1	**atlas-math-sets**	https://huggingface.co/datasets/AtlasUnified/atlas-math-sets	Yes	Test - First 10000	5	Maths	10000
2	**bioasq10b-factoid**[37]	https://huggingface.co/datasets/legacy107/bioasq10b-factoid	Yes	Train	5	Biology	983
3	**biomedical_cpgQA**	https://huggingface.co/datasets/chloecchng/biomedical_cpgQA	Yes	Train	5	Biomedical	535
4	**IELTS**	Collected from online resources	No	-	-	IELTS	50
5	**JournalQA**	Collected from research articles	Yes	-	-	Science	273
6	**QuAC**[38]	https://quac.ai/	Yes	Validation	5	General	618
7	**ScienceQA**[39]	https://scienceqa.github.io/	Yes	Test	5	Science	1378
8	**Question-Answer Dataset**	https://www.kaggle.com/datasets/rtatman/questionanswer-dataset	Yes	S08,S09,S09	5	General	2037

Table 6. Overview of Question Answer Datasets. The first column denotes the dataset name, while the second column contains the download link. The third column indicates whether preprocessing was performed on the dataset. The information in each dataset is presented in different formats, and we transformed the information to generate context, questions, and answers for each dataset. Some datasets offer three sets (Training, Testing, and Validation), and for a specific dataset, questions were selected from a particular set, as indicated in the fourth column. Due to lengthy answers in all datasets, we considered answers having a length of less than or equal to 5 words. The sixth column classifies the dataset type (general questions, mathematical queries, or biology-related questions). The seventh column specifies the number of questions after preprocessing.

Table 7 displays the average context length, question length, response length, and context complexity for all datasets. These metrics provide insights into the parameters influencing the performance of CBQA models. Context complexity is calculated using the Flesch-Kincaid grade complexity formula.

Table 8 shows the model name, size, and the dataset used to fine-tune the models.

	Dataset Name	Average Context Length	Average Question Length	Average Response Length	Average Context Complexity
1	**atlas-math-sets**	3	8	1	-3
2	**bioasq10b-factoid**	351	9	2	17
3	**biomedical_cpgQA**	161	13	5	13
4	**IELTS**	907	12	2	10
5	**JournalQA**	5568	9	2	11
6	**QuAC**	84	6	2	11
7	**ScienceQA**	147	16	2	6
8	**Question-Answer Dataset**	5227	9	2	11

Table 7. Overview of Average Context, Answer, and Question length across all datasets.

Methodology

This section elaborates on the preprocessing and transforming datasets, the training and evaluation of models, and an ensemble-based approach to finding the optimal combination of models 9.

Figure 9. The flowchart illustrates the benchmarking process for context-based question-answer models. We processed datasets and generated a context, question, and answer for each dataset. Each dataset was passed through a context-based question-answer model, and the actual and predicted answers were compared using the evaluation procedure to calculate accuracy. We performed different kinds of analyses, including finding the best model, assessing the effect of context length on execution time, and determining the impact of question type on performance. Finally, we used a genetic algorithm to determine an optimal number and combination of models to optimize performance.

Index	Model Name	Model Size (MB)	Dataset
1	twmkn9/albert-base-v2-squad2	12	SQuAD v2
2	valhalla/bart-large-finetuned-squadv1	388	SQuAD v1
3	deepset/bert-base-cased-squad2	104	SQuAD v2
4	google/bigbird-roberta-base[40]	122	Books, CC-News, Stories and Wikipedia.
5	google/bigbird-pegasus-large-arxiv[40]	551	Arxiv dataset
6	dmis-lab/biobert-v1.1	104	NA
7	deepset/roberta-base-squad2	119	SQuAD v2
8	Splend1dchan/canine-c-squad	126	NA
9	YituTech/conv-bert-base	101	NA
10	Palak/microsoft_deberta-large_squad	387	SQuAD v1
11	microsoft/deberta-v2-xlarge	844	NA
12	distilbert-base-uncased[41]	64	BookCorpus and English Wikipedia
13	bhadresh-savani/electra-base-squad2	104	SQuAD v2
14	nghuyong/ernie-1.0-base-zh[42]	96	Chinese
15	xlm-mlm-en-2048[43]	637	Masked language modeling
16	google/fnet-base[44]	80	Colossal Clean Crawled Corpus (C4)
17	funnel-transformer/small[45]	125	BookCorpus, English Wikipedia, Clue Web, GigaWord, and Common Crawl
18	EleutherAI/gpt-neo-1.3B[46]	1255	Pile
19	hf-internal-testing/tiny-random-gptj	1	NA
20	gpt2	119	WebText
21	kssteven/ibert-roberta-base[47]	119	NA
22	allenai/led-base-16384	155	NA
23	allenai/longformer-large-4096-finetuned-triviaqa	415	NA
24	facebook/mbart-large-cc25	583	Multilingual mbart model
25	mnaylor/mega-base-wikitext	8	wikitext-103
26	csarron/mobilebert-uncased-squad-v2	24	SQuAD v2
27	microsoft/mpnet-base	105	NA
28	google/mt5-small	165	101 languages
29	RUCAIBox/mvp[48]	388	NA
30	sijunhe/nezha-cn-base	98	Chinese
31	uw-madison/nystromformer-512	104	BookCorpus and English Wikipedia
32	facebook/opt-350m[49]	316	BookCorpus, CC-Stories, The Pile, Pushshift.io, and CCNewsV2
33	bert-base-uncased[34]	105	BookCorpus and English Wikipedia
34	google/rembert[50]	550	Multilingual Wikipedia data over 110 languages.
35	roberta-base[51]	119	BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories
36	andreasmadsen/efficient_mlm_m0.40	339	NA
37	ArthurZ/dummy-rocbert-qa	112	NA
38	tau/splinter-base[52]	103	BookCorpus and English Wikipedia
39	squeezebert/squeezebert-uncased[53]	49	BookCorpus and English Wikipedia
40	t5-small[54]	58	Colossal Clean Crawled Corpus (C4)
41	xlm-mlm-en-2048[43]	637	Masked language modeling
42	xlnet-base-cased[55]	112	XLNet model pre-trained on English language
43	uw-madison/yoso-4096	121	Masked language modeling
44	SRDdev/QABERT-small[56]	64	30k samples from the Stanford Question Answering Dataset
45	bert-large-uncased-whole-word-masking-finetuned-squad[34]	320	BookCorpus and English Wikipedia
46	facebook/bart-large-cnn[57]	388	CNN Daily Mail
47	ahotrod/electra_large_discriminator_squad2_512	319	SQuAD v2

Table 8. The table shows language models from Hugging Face included in this study. For access to the documentation of each model, utilize the following pattern: `https://huggingface.co/+ModelName` (e.g., `https://huggingface.co/twmkn9/albert-base-v2-squad2`). The Model Name column presents the name of each model, and the Model Size column denotes the size of the respective models. Additionally, the dataset column specifies the datasets utilized to fine-tune each model.

Processing

All downloaded datasets in various formats (Parquet or JSON) were transformed into a tabular format. Fields containing information for context-based question answering were mapped to three sets: question, context, and answer. Questions within all datasets where the answer length fell below or equal to five were retained. Text cleaning was performed on the question and context columns, replacing newline characters and stripping unnecessary whitespaces. Rows with duplicate questions, entries, or missing values were removed to ensure data accuracy and consistency.

Models Training

Pre-trained CBQA models from Hugging Face[29] were utilized for question answering on processed datasets without additional fine-tuning on each dataset. The context and question columns, containing pertinent information for question answering, were extracted from all datasets and fed into the models. The model outputs, comprising both answers and their corresponding scores, were preserved for subsequent analysis. The execution time for all models was recorded due to its critical importance in performance assessments.

Table 9 shows the most occurring datasets used to train the models included in this study. SQuAD v2, BookCorpus, and English Wikipedia were among the highly used datasets to train most of the models provided by Hugging Face. A description of each model included in this study, along with the corresponding research paper, model size, and the dataset used to train each model, is shown in Table 8.

	Dataset to Train Models	**Frequency**
1	SQuAD v2[58]	6
2	BookCorpus[59] and English Wikipedia	6
3	Masked language modeling	3
4	Chinese	2
5	Colossal Clean Crawled Corpus (C4)[60]	2
6	SQuAD v1[58]	2

Table 9. This table shows the most occurring datasets used to train the models. The first column displays the dataset, and the second column shows the number of models trained using that dataset.

Models Evaluation

For accuracy evaluation, we used a combination of Exact Match (EM) and text similarity to assess the performance of the CBQA models[5]. First, we performed text processing on the predicted and actual answers before evaluating the model's performance. The actual and predicted answers underwent tokenization, and we applied a series of cleaning steps, including removing leading and trailing whitespaces, translating special characters, converting text to lowercase, and eliminating English stop words. A similarity score was then calculated between the actual and predicted answers, incorporating a comprehensive set of conditions. Specifically, a similarity score of 0 was assigned when the predicted answer was null. In cases where the predicted and actual answers matched perfectly, a similarity score of 1 was assigned. We used the fuzzy-wuzzy library to compute the Levenshtein distance ratio to enhance the accuracy assessment. If the Levenshtein distance ratio exceeded 80%, we assigned a similarity score of 1. Conversely, a similarity score of 0 was assigned when none of these conditions were met. This evaluation criterion facilitated the computation of nuanced similarity scores, offering a detailed measure of the model's accuracy.

Analysis

After evaluating the performance of all models across various datasets, we conducted diverse analyses to understand better why specific models performed satisfactorily or inadequately. Our initial analysis focused on assessing the impact of context length and model size on the execution time of the models. We observed variations in processing speed, with some models delivering rapid results while others required a significant amount of time. We determined the statistically significant model in terms of performance and execution time across eight datasets, providing insights into the comparative effectiveness of different models. Further analyses delved into understanding the factors contributing to superior or inferior performance, including the model specifications and the datasets used for training. This exploration aimed to uncover the nuances behind model behavior. Another crucial analysis investigated how model performance relates to the complexity and types of questions posed to it. The nature of this analysis differed for each dataset, depending on the available information in the database. To enhance overall performance, we experimented with a combination of different models using a genetic algorithm[61].

Ensemble-Based Approach

After obtaining individual performance metrics from each model, we tried to enhance overall performance through an ensemble approach. We aimed to leverage individual model strength and optimize overall predicting performance. Employing a 5-fold cross-validation technique, we partitioned the dataset into training (80%) and test (20%) sets to evaluate the effectiveness of the ensemble strategy.

We evaluated 47 models, providing a diverse set for potential ensemble combinations ranging from 2 to 47. Consider a scenario where we aim to construct an ensemble comprising three models, capitalizing on the predictive capabilities of this trio. We used a genetic algorithm to select the top 3 optimal models to address this. This optimization process focused on the training set, ensuring the ensemble consisted of the best-performing models. Each model within our ensemble architecture produces an

answer and an associated score. In the ensemble context, the ultimate prediction is determined by the model with the highest prediction score. The goal is to optimize the ensemble of models by selecting the combination that yields the highest accuracy.

Author contributions statement

M.M. wrote the code and generated the results. M.M. and D.A. contributed to writing and reviewing the manuscript and arranging resources for the analysis. All authors reviewed the manuscript.

Competing interests

The author(s) declare no competing interests.

Data availability

The code used for the analysis conducted in this study is accessible on GitHub at the following link: https://github.com/MuhammadMuneeb007/Comparative-Analysis-of-47-Context-Based-Question-Answer-Models-Across-8-Diverse-Datasets. Additionally, the pre-and post-processed datasets can be downloaded from The Open Science Framework (OSF) using the following link: https://osf.io/rd3x2/.

References

1. Landhuis, E. Scientific literature: Information overload. *Nature* **535**, 457–458, DOI: 10.1038/nj7612-457a (2016).
2. Alanazi, S. S., Elfadil, N., Jarajreh, M. & Algarni, S. Question answering systems: A systematic literature review. *Int. J. Adv. Comput. Sci. Appl.* **12**, DOI: 10.14569/ijacsa.2021.0120359 (2021).
3. Sen, P. & Saffari, A. What do models learn from question answering datasets? *arXiv preprint arXiv:2004.03490* (2020).
4. Allam, A. M. N. & Haggag, M. H. The question answering systems: A survey. *Int. J. Res. Rev. Inf. Sci. (IJRRIS)* **2** (2012).
5. Liddy, E. D., Diekema, A. R. & Yilmazel, O. Context-based question-answering evaluation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR04, DOI: 10.1145/1008992.1009094 (ACM, 2004).
6. Choi, E. *et al.* Quac : Question answering in context, DOI: 10.48550/ARXIV.1808.07036 (2018).
7. Li, L., Zhang, M., Chao, Z. & Xiang, J. Using context information to enhance simple question answering. *World Wide Web* **24**, 249–277, DOI: 10.1007/s11280-020-00842-7 (2020).
8. Li, H., Min, M. R., Ge, Y. & Kadav, A. A context-aware attention network for interactive question answering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, DOI: 10.1145/3097983.3098115 (ACM, 2017).
9. Seonwoo, Y., Kim, J.-H., Ha, J.-W. & Oh, A. Context-aware answer extraction in question answering, DOI: 10.48550/ARXIV.2011.02687 (2020).
10. Anshul Kumar, Abhinav Panwar & Anurag Mukesh Rawat. Research paper on question answering system using bert, DOI: 10.13140/RG.2.2.32542.20800 (2022).
11. Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A. & Prather, J. The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian Computing Education Conference*, 10–19 (2022).
12. Lee, J. *et al.* Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240, DOI: 10.1093/bioinformatics/btz682 (2019).
13. Cao, Y. *et al.* Askhermes: An online question answering system for complex clinical questions. *J. Biomed. Informatics* **44**, 277–288, DOI: 10.1016/j.jbi.2011.01.004 (2011).
14. Cao, Y.-g., Cimino, J. J., Ely, J. & Yu, H. Automatically extracting information needs from complex clinical questions. *J. Biomed. Informatics* **43**, 962–971, DOI: 10.1016/j.jbi.2010.07.007 (2010).
15. Brown, T. B. *et al.* Language models are few-shot learners, DOI: 10.48550/ARXIV.2005.14165 (2020).
16. Song, X., Salcianu, A., Song, Y., Dopson, D. & Zhou, D. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524* (2020).
17. Bai, H. *et al.* Segatron: Segment-aware transformer for language modeling and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 12526–12534 (2021).
18. Vaswani, A. *et al.* Attention is all you need, DOI: 10.48550/ARXIV.1706.03762 (2017).

19. Wang, T. *et al.* What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, 22964–22984 (PMLR, 2022).

20. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Masked language model scoring. *arXiv preprint arXiv:1910.14659* (2019).

21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, DOI: 10.48550/ARXIV.1810.04805 (2018).

22. Gururangan, S. *et al.* Don't stop pretraining: Adapt language models to domains and tasks, DOI: 10.48550/ARXIV.2004.10964 (2020).

23. Dodge, J. *et al.* Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305* (2020).

24. Risch, J., Möller, T., Gutsch, J. & Pietsch, M. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130* (2021).

25. Chen, A., Stanovsky, G., Singh, S. & Gardner, M. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, 119–124 (2019).

26. Rodrigo, A. & Peñas, A. A study about the future evaluation of question-answering systems. *Knowledge-Based Syst.* **137**, 83–93, DOI: 10.1016/j.knosys.2017.09.015 (2017).

27. Wang, Z. Modern question answering datasets and benchmarks: A survey. *arXiv preprint arXiv:2206.15030* (2022).

28. Usbeck, R. *et al.* Benchmarking question answering systems. *Semantic Web* **10**, 293–304, DOI: 10.3233/sw-180312 (2019).

29. Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing, DOI: 10.48550/ARXIV.1910.03771 (2019).

30. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach, DOI: 10.48550/ARXIV.1907.11692 (2019).

31. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, DOI: 10.48550/ARXIV.1910.01108 (2019).

32. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. Electra: Pre-training text encoders as discriminators rather than generators, DOI: 10.48550/ARXIV.2003.10555 (2020).

33. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer, DOI: 10.48550/ARXIV.1910.10683 (2019).

34. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018). 1810.04805.

35. Başaran, K., Bozyiğit, F., Siano, P., Yıldırım Taşer, P. & Kılınç, D. Systematic literature review of photovoltaic output power forecasting. *IET Renew. Power Gener.* **14**, 3961–3973, DOI: 10.1049/iet-rpg.2020.0351 (2020).

36. Padilla-Martínez, F., Collin, F., Kwasniewski, M. & Kretowski, A. Systematic review of polygenic risk scores for type 1 and type 2 diabetes. *Int. J. Mol. Sci.* **21**, 1703, DOI: 10.3390/ijms21051703 (2020).

37. Tsatsaronis, G. *et al.* An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinforma.* **16**, DOI: 10.1186/s12859-015-0564-6 (2015).

38. Choi, E. *et al.* Quac : Question answering in context, DOI: 10.48550/ARXIV.1808.07036 (2018).

39. Lu, P. *et al.* Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)* (2022).

40. Zaheer, M. *et al.* Big bird: Transformers for longer sequences (2021). 2007.14062.

41. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv* **abs/1910.01108** (2019).

42. Sun, Y. *et al.* Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

43. Lample, G. & Conneau, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

44. Lee-Thorp, J., Ainslie, J., Eckstein, I. & Onta
n
'on, S. Fnet: Mixing tokens with fourier transforms. *CoRR* **abs/2105.03824** (2021). 2105.03824.

45. Dai, Z., Lai, G., Yang, Y. & Le, Q. V. Funnel-transformer: Filtering out sequential redundancy for efficient language processing (2020). 2006.03236.

46. Black, S., Leo, G., Wang, P., Leahy, C. & Biderman, S. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, DOI: 10.5281/zenodo.5297715 (2021). If you use this software, please cite it using these metadata.

47. Kim, S., Gholami, A., Yao, Z., Mahoney, M. W. & Keutzer, K. I-bert: Integer-only bert quantization. *arXiv preprint arXiv:2101.01321* (2021).

48. Tang, T., Li, J., Zhao, W. X. & Wen, J.-R. Mvp: Multi-task supervised pre-training for natural language generation. *arXiv preprint arXiv:2206.12131* (2022).

49. Zhang, S. *et al.* Opt: Open pre-trained transformer language models (2022). 2205.01068.

50. Chung, H. W., F
'evry, T., Tsai, H., Johnson, M. & Ruder, S. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (OpenReview.net, 2021).

51. Liu, Y. *et al.* Roberta: A robustly optimized BERT pretraining approach. *CoRR* **abs/1907.11692** (2019). 1907.11692.

52. Ram, O., Kirstain, Y., Berant, J., Globerson, A. & Levy, O. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3066–3079, DOI: 10.18653/v1/2021.acl-long.239 (Association for Computational Linguistics, Online, 2021).

53. Iandola, F. N., Shaw, A. E., Krishna, R. & Keutzer, K. W. SqueezeBERT: What can computer vision teach nlp about efficient neural networks? *arXiv:2006.11316* (2020).

54. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).

55. Yang, Z. *et al.* Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR* **abs/1906.08237** (2019). 1906.08237.

56. Dixit, S. (2023).

57. Lewis, M. *et al.* BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR* **abs/1910.13461** (2019). 1910.13461.

58. Rajpurkar, P., Jia, R. & Liang, P. Know what you don't know: Unanswerable questions for squad, DOI: 10.48550/ARXIV.1806.03822 (2018).

59. Zhu, Y. *et al.* Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)* (2015).

60. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints* (2019). 1910.10683.

61. Mirjalili, S. & Mirjalili, S. Genetic algorithm. *Evol. Algorithms Neural Networks: Theory Appl.* 43–55 (2019).