CLIFT : Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain

Ankit Pal
Saama AI Research
Chennai, India
ankit.pal@saama.com

Abstract

This paper introduces a new testbed CLIFT (**Cli**nical Sh**ift**) for the clinical domain Question Answering task. The testbed includes 7.5k high-quality question-answering samples to provide a diverse and reliable benchmark. We performed a comprehensive experimental study and evaluated several QA deep-learning models under the proposed testbed. Despite impressive results on the original test set, the performance degrades when applied to new test sets, which shows the distribution shift. Our findings emphasize the need for and the potential for increasing the robustness of clinical domain models under distributional shifts. The testbed offers one way to track progress in that direction. It also highlights the necessity of adopting evaluation metrics that consider robustness to natural distribution shifts. We plan to expand the corpus by adding more samples and model results. The full paper and the updated benchmark are available at openlifescience-ai.github.io/clift

1 Introduction

Digital healthcare data has significantly expanded during the last several years. Question Answering (QA) in the clinical domain is a particular task involved with clinical texts in Electronic Health Records (EHR). Clinical notes & documents in EHRs are frequently used by medical practitioners to find information that can aid in decision-making. [22] Recent studies using neural QA models [20, 3, 4] have shown promising outcomes. As a result, significant effort has been invested in developing question-and-answer algorithms & datasets in this field. The objective of developing clinical domain question-answering systems is not to achieve the highest possible accuracy. Instead, the goal is to generalize the model on new clinical data from various regions, hospitals, and organizations.

In statistical learning theory, it is often assumed that test data points originate from the same distributional underpinnings as training data for a machine learning model. However, in reality, data may be sampled from various distributions with considerable domain changes [7, 21, 9]; therefore, this assumption is valid sometimes. Despite impressive results of QA models trained in the clinical domain, when data distribution unexpectedly changes, models that seem to be doing well in test set accuracy fail when deployed in real-time. [27] research indicates that when predicting questions that have been paraphrased and completely new, the F1 score drops by around 40% to 60%. We are evaluating the Models trained on emrQA to investigate natural distribution shifts for different test sets, for example, changing from an existing test set to a cancer, heart, and other test sets. Our Study focuses on a question:

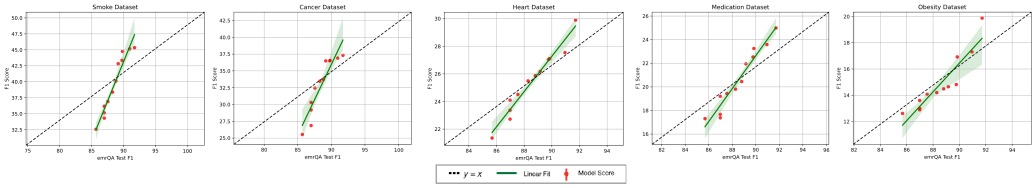
Are Clinical Models fine-tuned on emrQA robust to natural distribution shifts?

We perform a comprehensive experimental study on the emrQA dataset [14] in this paper. We introduce a new testbed derived from the MIMIC-III database. Our results show that emrQA models

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

exhibit robustness to some natural distribution shifts. Despite impressive results on the training dataset, when applied to new test sets, the performance degrades, and models in our testbed experiences a drop in F1 metric. Our experiments underline the need and the potential for improving the robustness of clinical domain models under distributional shifts. The results also highlight the necessity of adopting evaluation metrics that consider robustness to natural distribution shifts.

In brief, the contributions of this study are as follows. (i) We are proposing five new test datasets. The testbed covers different clinical domain diseases and multiple sub-topics. These questions are from MIMIC-III [8], and human experts validated all the questions, answers, and evidence triples. (ii) This dataset offers several advantages over existing datasets. They evaluate the diverse reasoning abilities of clinical models over a vast array of disease topics and subtopics. The testbed is a comprehensive evaluation of clinical domain data. (iii) Detailed statistics and fine-grained evaluation of natural distribution shift are provided in this study. We examine and quantify the shortcomings of the clinical deep learning models trained on emrQA when evaluated with new test data. Moreover, We also analyzed the dataset on difficulty metrics such as answer diversity and Syntactic divergence. (iv) We conduct our experiments with the help of the HuggingFace Transformers library [26]. Furthermore, we also open-source the code and model checkpoints to reproduce the results and test with new models in the future. (v) We provide a new QA tool to validate the Question Answering Dataset. We hope this will accelerate the tagging process and robust ML research in this domain.



—

Figure 1: Comparison of models' F1 scores on emrQA and our proposed test datasets. The result demonstrates that despite impressive results on the training dataset when applied to new test sets, the model's performance degrades, which shows the distribution shift

2 Background

2.1 Natural Distribution shift

For a prediction algorithm $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$ We assume a clinical domain dataset from a Hospital A where input space is X and label space is Y. To evaluate M's performance on a test dataset $Q = \{x_i | i = 1, \ldots, n\}$ drawn from a data distribution P, we will assume that algorithm M is fixed. The initial steps would be choosing a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and calculating the loss $\ell_Q(M) = \sum_{i=1}^{n} \ell\left(M\left(x_i\right), y_i\right)$ and population loss $\ell_{\mathcal{P}}(M) = \mathbb{E}\mathcal{P}[\ell(M(x), y)]$ under the test distribution.While the input X and label Y spaces are fixed in this case, there may be many distinct joint distributions. Furthermore, Besides the dataset from Hospital A, we collect another clinical domain dataset from Hospital B where the test set is Q' from a new distribution P'

In addition to the test loss $l_Q(M)$, we are interested to know the loss $l_{Q'}(M)$ under real-time applications, known as distribution shift. The observations of models' expected loss to changes or differences would be helpful to improve the model's robustness, eliminate the biases and understand the dataset better. This difference may be broken down into three parts.[13]

$$l_Q(M) - l_{Q'}(M) = \underbrace{(l_Q(M) - l_{\mathcal{P}}(M))}_{\text{Adaptivity gap}} + \underbrace{(l_{\mathcal{P}}(M) - l_{\mathcal{P}'}(M))}_{\text{Distribution gap}} + \underbrace{(l_{\mathcal{P}'}(M) - l_{Q'}(M))}_{\text{Generalization gap}} \tag{1}$$

The adaptivity gap quantifies how adjusting the model to the test set affects the estimate of population loss. While the distribution gap indicates how much a change in the sample's distribution impacts the model's performance, the difference between the sample and population losses due to Q's random sampling is captured by the generalization gap.

3 emrQA Dataset

emrQA [14] is a large-scale extractive question-answering dataset generated semi-automatically from question templates, and human experts annotated the templates. The proposed task was to extract a continuous text span from the clinical note while giving a clinical note and a question as input to the QA model. The corpus contains almost 400,000 triplets of a question, logical form, and evidence/answer.

We used Medication and Relation datasets for fine-tuning the models. The following two comprise 80% of the overall emrQA dataset, and their structure is congruent with the span prediction task, which is more significant for clinical decision assistance. Table 1 describes the statistics of the emrQA dataset.

4 Proposed Testbed

The MIMIC-III [8] is the most popular dataset for Electronic Health Record (EHR) related tasks. In the dataset, A clinical record is maintained for each patient, which contains information such as ECGs, physician notes, discharge summaries, nurse progress reports and more. Due to its comprehensive collection of unstructured clinical notes, MIMIC-III has been extensively used in clinical NLP tasks.[15]

In order to evaluate the ability of healthcare domain QA systems to generalize to new clinical data, we created five new test sets from MIMIC-III. First, we randomly sampled the patient's records from the MIMIC-III dataset, and later, we filtered the sampled data based on the Heart, Medication, Obesity, Smoking, and Cancer subset's keywords. Finally, we utilized the same question templates provided by emrQA to create the question and answer pairs, along with a few more templates annotated by experts.

Heart Disease Test set The Heart disease test subset contains 1.5k annotated questions and answers related to heart disease and issues such as heart attack, heart failure, coronary artery disease Ischemia, and more.

Medication Test set The medication subset contains around 1.5k samples. It includes the question and answers related to medical prescriptions, drugs name, drug dosages, history, and & status of medications; it also contains questions related to the number of doses, side effects of medications, reasons for medications and more.

Obesity Test set Similar to the above datasets, we extracted 1k questions and answers related to obesity, such as comorbidities associated with the current issue, hypertension, Sleep Apnea, Diabetes, and more. The test set's diversity helps to quantify the degree to which current systems overfit the original test set.

Smoking Test set For the smoking subset, 2k questions were selected related to the current status & history of smoking, quantity & type of smoking, issues and more.

Cancer Test set The subset includes 1.5k questions and answers related to different cancers, such as Breast Cancer, Colon cancer, Rectal Cancer, Thyroid Cancer, and more. Moreover, the set also contains samples mentioning the Cancer cause, side effects, medications for current cancer, stage of cancer and more.

4.1 Preprocessing & Quality Checks

The following procedures were used to sanitize the raw data and to ensure that all questions could be answered using textual input. We lowercase the text in the subset test sets, remove punctuation marks and tokenize the text using whitespace. We set the min length to 100 and split the long documents into sub-documents. We also converted a few abbreviations to complete forms such as y.o., y/o, to a year old and yr, yr's, yrs to years. Also, additional data cleaning procedures were performed to ensure that the information gathered from the question, answer, and evidence triples met the data quality objectives. We generated 25k question-answers pairs from selected medical records; 15k samples were discarded due to low data quality, noisy and wrong answers. After the data team re-validated, at the time of writing the paper, 7.5k final samples were finalized for the testbed. However, We plan

to expand the corpus by adding more samples and model results. The full paper and the updated benchmark are available at openlifescience-ai.github.io/clift

4.2 Dataset Statistics & Analysis

This dataset covers different diseases and medical conditions based on patient information.

Table 2 in the Appendix summarises the general statistics of preprocessed data. Moreover, Figure 7 in the Appendix depicts an additional useful statistic: the number of unique tokens in the dataset. The performance of the models is influenced by the amount of vocabulary, which is a good indicator of the linguistic and domain complexity associated with a text corpus. The Obesity & Heart subset was observed to include more extended questions and a broader vocabulary than the others. As a result, the former subsets' question are more complex than the rest.

In addition, We use the two difficulty metrics provided by [19] to compare the original emrQA test set to our five new test sets. Figure 2 shows the Syntactic divergence histograms for the emrQA & proposed datasets.

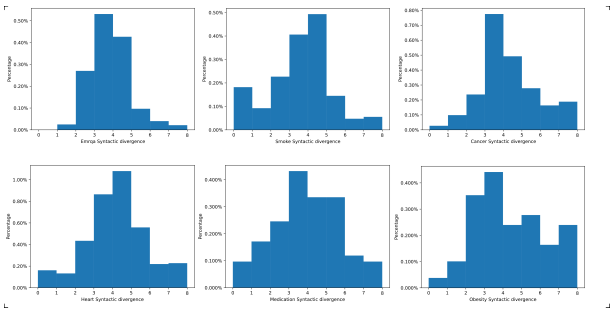


Figure 2: Syntactic divergence histograms for the emrQA & proposed datasets' question and response phrases.

5 Experimental Setup

We followed the [27] random seed and split the combined Medication and Relation datasets based on the clinical texts obtaining train, dev, and test with the ratio 7:1:2. Table 7 shows the descriptive statistics of the emrQA dataset. Similar to [19], we used the F1 score (F1) as our evaluation matrices. The hyperparameters are chosen based on [4]. In order to fine-tune it for four epochs, we set the batch size to 8, the learning rate to 2e5, and the maximum response length to 30. we optimize the model's parameters using the AdamW optimizer.

We implemented these models using Pytorch [16] and obtained pre-trained checkpoints from the HuggingFace library [25]. All models are fine-tuned on 4 GPUs with NVIDIA Quadro RTX 8000 and 256 GB of RAM.

5.1 Baseline Models

In our baseline studies, we consider 12 existing models as a baseline, which have shown strong performance in biomedical NLP tasks. These are developed on several pre-trained language models employing the Transformers architecture [24]. Notably, BioBERT [10], and variant BioClinicalBert and BioBert-squad-L pre-trained on large-scale biomedical corpora. DistilBert is a distilled version of the BERT base model. We used two different versions of DistilBert, including DistilBert-squad-B and BERT-ClinicalQA, fine-tuned on Squad and ClinicalQA, respectively.

To better utilize the knowledge of MIMIC and PubMed documents for QA tasks, we fine-tuned the BlueBert Models, including BlueBert-MIMIC and BlueBert-PubMed. Moreover, We also used the Roberta architecture models, including Biomed-Roberta, and Roberta-squad-B. In addition, we fine-tuned the PubMedBERT, Bert-Large, and Electra-Squad-L on the emrQA dataset. Section D in the Appendix describes the models in detail.

We finetuned the models on the training dataset of emrQA and evaluated on the proposed test sets.

6 Results & Discussion

The objective of the proposed testbed is to test the performance of a clinical model when evaluated on data relevant to the same Question-Answering task but from multiple data sources/hospitals and environments. We ran a battery of experiments and evaluated twelve models on proposed datasets. In addition to investing the state of art models in clinical domains for Question Answering, we examine the robustness of models on the natural distribution shift in the emrQA dataset. More information about baseline models is provided in Section 5.1 and Appendix D

6.1 Comparison of Transformer Based clinical Models on emrQA & Clift

Under the emrQA test set, PubMedBERT performed better than other models. Moreover, Biomed-Roberta gave the second-best results on the same. The Roberta-squad-Base model outperformed other models on the Medication, Obesity, and Heart test set. However, Bert-Large BlueBert-MIMIC gave the best accuracy on Cancer and Smoke datasets.

To examine the QA model's result closely, We also performed interpretability experiments and provided the results in Appendix B

6.2 Are Clinical Models fine-tuned on Emrqa robust to natural distribution shifts?

To test the hypothesis that the original test set performance is better than new test sets, we compare the F1 scores for all the models trained on the emrQA to the F1 scores on each of our additional test sets. Figure 1 shows that, when applied to proposed test datasets, All the models indicate a performance loss ranging from 43.82% to 73.07%. Table 3 in the Appendix shows the performance of different models on the medication test set. As we can see that BERT-ClinicalQA drops around 70% F1 points while evaluating the medication dataset, compared to a drop of about 61%, 55%, 75%, and 53 % F1 points when evaluating the model on heart, cancer, obesity, and smoke dataset respectively.

PubMedBERT fine-tuned on emrQA shows the worst performance on the Heart test set with a gap of 70.37 % F1. Moreover, Biomed-Roberta shows a large gap of 64.22 % on the Cancer dataset and DistilBert-squad-B, BioBert-B suffers from generalizing well on the Obesity and Smoke dataset and dropping by 73.07 and 54.98 respectively, in the performance. Roberta-squad-B outperformed all other models on 3 out of 5 datasets, including Medication, Heart, and Obesity, with an F1 score of 24.99%, 29.90%, and 19.88%, respectively.

These results suggest that the models are sensitive to distribution shifts. Real-world deployments of health machine learning systems may encounter these failure situations. Therefore, depending on the entity of interest, performance for downstream applications would vary greatly. Table 5, 6, 4 and 7 in the Appendix show the results of different models on cancer, obesity, heart, and smoke dataset respectively.

These results imply the potential for further study handling distribution shifts in clinical datasets. In addition, future research should look at specific model training parameters and the datasets' size, quality, and quantity that contribute to these variations.

7 Conclusion

Robust machine learning aims to create techniques that consistently work in different environments. To this end, we propose a CLIFT testbed to evaluate the clinical deep learning model's performance under the distributional shift of the different test sets. We evaluate this testbed on several algorithms trained on the emrQA dataset and find that the performance degrades when applied to new test sets despite impressive results on the original test set. Moreover, In this study, we performed an empirical analysis on emrQA and analyzed the proposed test sets on difficulty metrics such as answer diversity and syntactic divergence. Our extensive testbed will indicate progress towards trustworthy machine learning in healthcare.

Acknowledgments

We thank Samsudeen Mugaideen for building essential Front-end tools for data tagging and validation. We greatly appreciate the support of our lovely team members with medical backgrounds Srishti Awasthi, Amanpreet Kaur, Dipali Mishra, Chitvan Gautam and Nidhi, who played a crucial role in validating the tagged data, supported by Geetha Eswaran and Prashant Shukla. We also want to thank anonymous reviewers & Malaikannan Sankarsubbu for their valuable comments & feedback.

References

[1] Sultan Alrowili and Vijay K. Shanker. "BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA". In: *BIONLP*. 2021.

[2] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *ArXiv* abs/1904.03323 (2019).

[3] Danqi Chen et al. "Reading Wikipedia to Answer Open-Domain Questions". In: Association for Computational Linguistics, 2017, pp. 1870–1879. DOI: 10.18653/v1/P17-1171.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[5] Yuxian Gu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3 (2022), pp. 1–23.

[6] Suchin Gururangan et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *ArXiv* abs/2004.10964 (2020).

[7] Arthur Jochems et al. "Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept." In: *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* 121 3 (2016), pp. 459–467.

[8] Alistair E.W. Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3 (1 Dec. 2016), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35.

[9] Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. "Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges". In: *ACM Transactions on Computing for Healthcare* (2022).

[10] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* (Sept. 2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682.

[11] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36 (2020), pp. 1234–1240.

[12] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692 (2019).

[13] John Miller et al. "The Effect of Natural Distribution Shift on Question Answering Models". In: ed. by Hal Daumé III and Aarti Singh. Vol. 119. PMLR, Oct. 2020, pp. 6905–6916. URL: https://proceedings.mlr.press/v119/miller20a.html.

[14] Anusri Pampari et al. "emrQA: A Large Corpus for Question Answering on Electronic Medical Records". In: Association for Computational Linguistics, 2018, pp. 2357–2368. DOI: 10.18653/v1/D18-1258.

[15] Junwoo Park et al. "Knowledge Graph-based Question Answering with Electronic Health Records". In: *MLHC*. 2021.

[16] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *ArXiv* abs/1912.01703 (2019).

[17] Yifan Peng, Shankai Yan, and Zhiyong Lu. "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets". In: *BioNLP@ACL*. 2019.

[18] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *ACL*. 2018.

[19] Pranav Rajpurkar et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

[20] Bhanu Pratap Singh Rawat et al. "Entity-Enriched Neural Models for Clinical Question Answering". In: Association for Computational Linguistics, 2020, pp. 112–122. DOI: 10.18653/v1/2020.bionlp-1.12.

[21] Nicola Rieke et al. "The future of digital health with federated learning". In: *NPJ Digital Medicine* 3 (2020).

[22] Samuel T. Rosenbloom et al. "Data from clinical notes: a perspective on the tension between structure and flexible documentation". In: *Journal of the American Medical Informatics Association : JAMIA* 18 2 (2011), pp. 181–6.

[23] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *ArXiv* abs/1910.01108 (2019).

[24] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[25] Thomas Wolf et al. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv* abs/1910.03771 (2019).

[26] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

[27] Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. "Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset". In: *ArXiv* abs/2005.00574 (2020).

A Evaluation Metric

We use the following metric to measure the performance of the clinical question-answering models. Assuming the clinical text passage is t, a question is denoted by q, answers are denoted by a, and the dataset sample and the model are represented by X and f respectively.

A.0.1 F1-Score

F1 score is based on individual word level. It is a standard metric for classification tasks and is widely utilised in Question Answering. The F1 score measures the average overlapping between a model's output and ground-truth span. It is a weighted average of precision and recall. The ratio of shared words to the total number of words in the model's output is called precision, while the ratio of shared words to the total number of words in the ground truth is called recall.

$$\mathrm{F1}Score(f) = \frac{1}{|X|} \sum_{(t,q,(a^1,\ldots,a^n))\in X} \max_{i=1,\ldots,n} k\left(a^i, f(t,q)\right) \tag{2}$$

Where k is the harmonic mean of precision and recall between the two sets.

B Interpretability Analysis

In order to understand the attention layer and prediction on the dataset, we investigated the output from a model trained on the emrqa dataset. In addition, we utilized the interpretation hooks to examine and better understand embeddings, sub-embeddings and attention layers.

B.1 Probability Mass Function

To understand and compare the distributional attributions patterns across multiple layers, we visualized each layer and presented them in Figure 3.

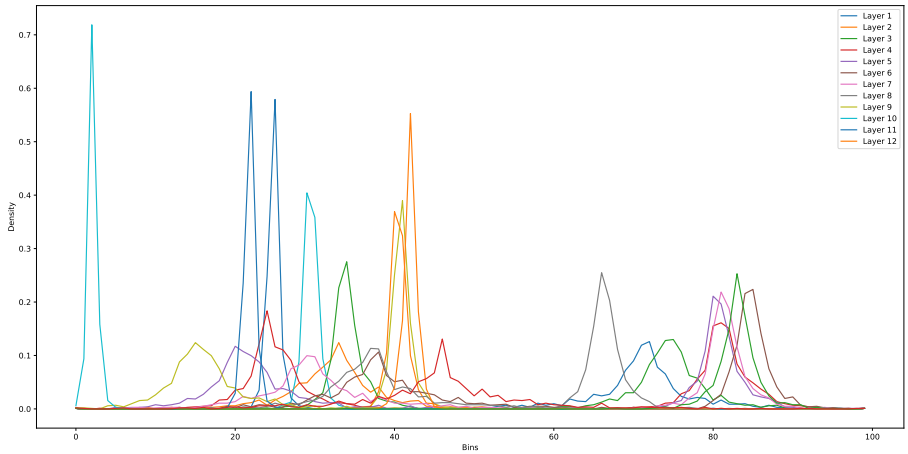


Figure 3: Visualization of the probability mass function of attributions for end position tokens.

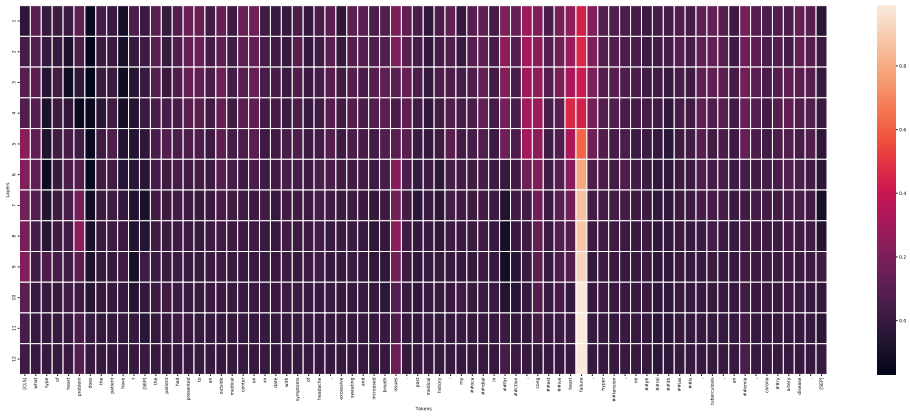


Figure 4: Heat map of all layers' tokens & attributions for the start position prediction.

In order to calculate the entropy of each layer's attributions, we represented the attributions as a probability density function (pdf) and estimated the entropy of that. The plot illustrates the Probability Mass Function (PMF) of each layer's attributions for the start and end position tokens. We can see from the figure that the distributions are taking on bell-shaped forms with various means and variances.

B.2 Attributions HeatMap

We used the layer conductance algorithm to visualize and understand the distribution of attribution scores for tokens across all layers.

Figure 4 shows the heat map of end position tokens across all layers. Interestingly, the word heart failure gains high attribution from layers seven to twelve and one to twelve respectively. Conversely, many other tokens have negative or close to zero attribution in the first six layers.

B.3 Distribution of attributions per layer

Figure 5 shows the end position's tokens box plot. From the diagram, one can see the presence of outliers from 4 to 9. We also observe that the interquartile range slowly decreases for start position prediction as we move deeper into the layers and finally diminish.

We also visualised the prediction of the start position's distribution box plot. Figure 6 shows the outliers from layer 1 to layer 7. The interquartile range stays mostly the same when moving deeper into the layers.

C In-depth Analysis

Based on the outcomes of our experiments and model evaluation, we can say that the proposed testbed is challenging & complex. In order to support this argument & take a closer look at the proposed test sets, We use the difficulty metric provided by [19] to compare the original emrQA test set to our four new test sets.

C.1 Syntactic divergence

Syntactic divergence assesses how similar the syntactic dependency tree structures of the question and response phrases are. Figure 2 shows the syntactic divergence histograms for our new test sets and the emrQA test set.

D Model Details

BioClinicalBert The MIMIC III database, which contains EHR from ICU patients at the Beth Israel Hospital in Boston, Massachusetts, was used to train the Bio ClinicalBERT model [2]. The NOTEEVENTS table's whole collection of notes (880M words) was included.

BioBert-B stands for Biomedical Bert, a domain-specific language representation model pre-trained on large-scale biomedical corpora. [11] shows that in many biomedical text mining tasks, BioBERT performs much better than BERT and earlier state-of-the-art models.

BioBert-squad-L The large variant of the BioBert model, was trained on the same PubMed abstracts with more parameters & domain-specific vocabulary based on WordPiece [11]. We utilised the Large BioBert model, which has been finetuned on the Squad dataset for the question-answering task.

BERT-ClinicalQA DistilBERT [23] is the distilled version of the BERT base model , and it is smaller and faster than BERT, which was pre-trained on the same corpus using the self-supervised mechanism. It uses the BERT base model as a teacher. We used the DistilBERT version, which has already been fined-tune on the clinical QA dataset.

DistilBert-squad-B is the base variant of DistilBERT [23] and it was fine-tuned using knowledge distillation on the SQuAD dataset.

BlueBert-MIMIC is a Transformer based Bert model [17] pre-trained on clinical notes from the MIMIC-III dataset. The result presented in the paper shows that Bluebert outperforms Bert and other general domain models on the BLUE benchmark, consisting of five tasks and ten datasets covering both biomedical and clinical literature with varying dataset sizes and difficulty.

BlueBert-PubMed Similar to the BlueBert-MIMIC model, BlueBert-PubMed [17] was trained on PubMed abstracts. The final training dataset contains around 4,000 million Words.

Biomed-Roberta is a variant of RoBERTa-base [12] architecture. The authors utilised the continued pretraining approach to train the model on 2.68 million scientific papers from the Semantic Scholar. [6] The training dataset consists of the full text of the papers with 7.55B tokens and 47GB of data in total.

Roberta-squad-B This is a Roberta Base architecture fine-tuned on the SQuAD2.0 dataset [18]. It has been trained for Question Answering Task on question-answer pairs dataset, including unanswerable questions.

PubMedBERT is based on Bert's architecture [5].The Model has been pre-trained on the full articles from PubMedCentral and abstracts from Pubmed from scratch. The results show that it outperforms different Transformer based models on Biomedical NLP tasks.

Bert-Large The Model is based on transformer-based architecture. First, it was trained with a technique called Word masking. The word masking technique masks all the tokens that correspond to a word while the overall masking rate remains constant. After the pre-training process, the Model was fine-tuned with the SQuAD dataset.

Electra-Squad-L The model [1] was pre-trained on abstract data from PubMed, fine-tuned on the SQuAD2.0 dataset, and later fine-tuned on the BioASQ8B-Factoid training dataset. The result shows that fine-tuning helped the model to achieve a better score on BioASQ challenge Benchmark.

	Medication	Relation	Total
Question #	222,957	904,592	1,127,549
Context #	261	423	684
Template #	80	139	219
Avg Q tokens	8.00	7.91	7.9
Avg A tokens	9.47	10.41	9.9
Avg C tokens	1062.66	889.23	975.9

Table 1: emrQA datasets' statistics, where Q, A, and C represent the Questions, Answers and Context respectively.

	Smoke	Heart	Medication	Obesity	Cancer
Question #	2k	1.5k	1.5k	1k	1.5k
Min Q tokens	2	3	3	3	3
Min A tokens	1	2	1	1	1
Min C tokens	100	200	100	100	100
Avg Q tokens	6.42	8.31	7.61	7.19	8.40
Avg A tokens	4.59	4.15	3.93	4.04	4.12
Avg C tokens	217.33	234.18	215.49	212.88	210.16

Table 2: Proposed datasets' statistics, where Q, A, and C represent the Question, Answer, and Context, respectively.

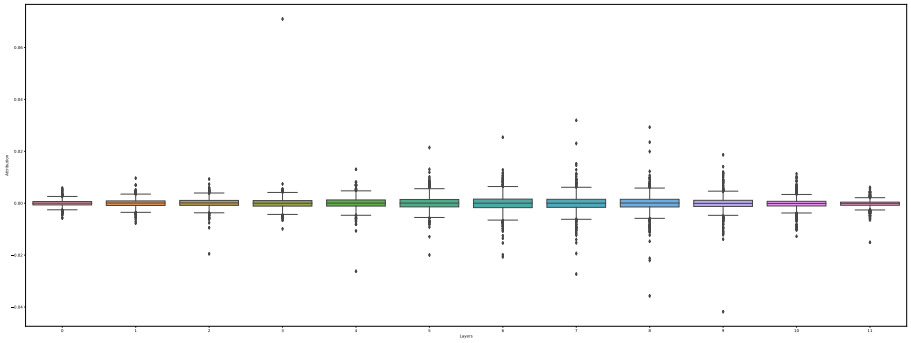


Figure 5: Distribution of attributions per layer in start position.

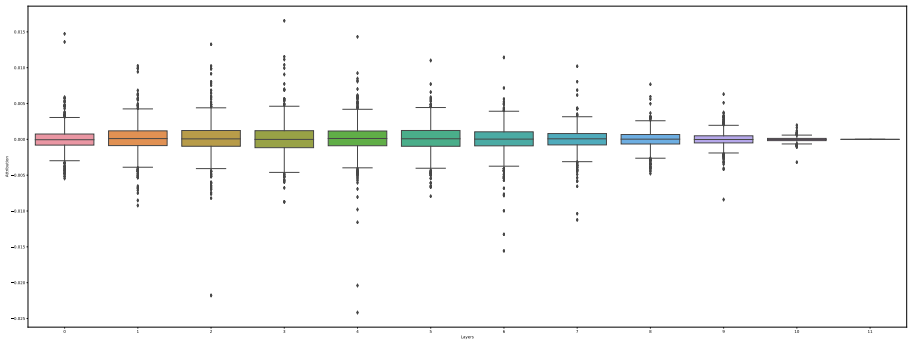


Figure 6: Distribution of attributions per layer in end position.

Model	emrQA F1	Clift F1	Gap (δ)
BERT-ClinicalQA	88.28	17.31	70.97
BioBert-B	87.55	17.38	70.17
BioClinicalBert	86.98	17.67	69.32
BlueBert-MIMIC	89.16	19.20	69.96
Electra-Squad-L	86.98	19.44	67.55
BlueBert-PubMed	86.98	19.81	67.18
DistilBert-squad-B	85.68	20.45	65.23
PubMedBERT	91.72	21.95	69.77
BioBert-squad-L	90.95	22.53	68.42
Bert-Large	88.81	23.25	65.56
Biomed-Roberta	89.76	23.59	66.18
Roberta-squad-B	89.84	24.99	64.85

Table 3: Performance of different models on the emrQA Testset and new Medication Testset

Model	emrQA F1	Clift F1	Gap (δ)
PubMedBERT	91.72	21.35	70.37
Biomed-Roberta	89.76	22.72	67.04
BioBert-B	87.55	23.37	64.18
Electra-Squad-L	86.98	24.09	62.89
BioClinicalBert	86.98	24.51	62.47
BlueBert-MIMIC	89.16	25.50	63.67
BioBert-squad-L	90.95	25.87	65.08
Bert-Large	88.81	26.17	62.63
BERT-ClinicalQA	88.28	27.04	61.24
DistilBert-squad-B	85.68	27.10	58.58
BlueBert-PubMed	86.98	27.54	59.44
Roberta-squad-B	89.84	29.90	59.94

Table 4: Performance of different models on the emrQA Testset and new Heart Testset

Model	emrQA F1	Clift F1	Gap (δ)
Biomed-Roberta	89.76	25.54	64.22
PubMedBERT	91.72	26.86	64.86
DistilBert-squad-B	85.68	29.19	56.49
BioBert-B	87.55	30.31	57.24
BERT-ClinicalQA	88.28	32.45	55.83
BioClinicalBert	86.98	33.46	53.52
Roberta-squad-B	89.84	33.73	56.11
BlueBert-MIMIC	89.16	36.49	52.67
BlueBert-PubMed	86.98	36.51	50.47
BioBert-squad-L	90.95	36.56	54.40
Electra-Squad-L	86.98	36.90	50.08
Bert-Large	88.81	37.31	51.49

Table 5: Performance of different models on the emrQA Testset and new Cancer Testset

Model	emrQA F1	Clift F1	Gap (δ)
DistilBert-squad-B	85.68	12.62	73.07
BioBert-B	87.55	12.89	74.66
BERT-ClinicalQA	88.28	13.00	75.28
BlueBert-MIMIC	89.16	13.60	75.56
BlueBert-PubMed	86.98	14.07	72.91
BioClinicalBert	86.98	14.20	72.78
BioBert-squad-L	90.95	14.49	76.46
PubMedBERT	91.72	14.65	77.08
Electra-Squad-L	86.98	14.81	72.17
Bert-Large	88.81	16.92	71.88
Biomed-Roberta	89.76	17.31	72.46
Roberta-squad-B	89.84	19.88	69.96

Table 6: Performance of different models on the emrQA Testset and new Obesity Testset

Model	emrQA F1	Clift F1	Gap (δ)
BioBert-B	87.55	32.57	54.98
BERT-ClinicalQA	88.28	34.31	53.97
Electra-Squad-L	86.98	35.13	51.85
DistilBert-squad-B	85.68	36.18	49.51
BioClinicalBert	86.98	36.90	50.08
Biomed-Roberta	89.76	38.38	51.39
BlueBert-PubMed	86.98	40.11	46.87
Bert-Large	88.81	42.84	45.97
BioBert-squad-L	90.95	43.36	47.60
PubMedBERT	91.72	44.75	46.97
Roberta-squad-B	89.84	45.16	44.68
BlueBert-MIMIC	89.16	45.34	43.82

Table 7: Performance of different models on the emrQA Testset and new Smoke Testset

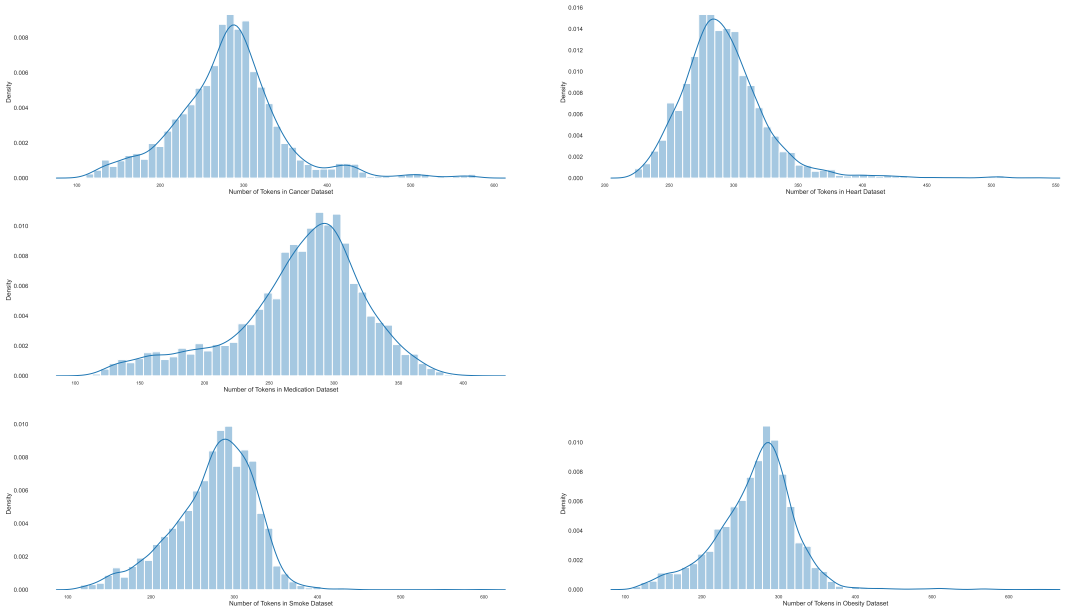


Figure 7: Subfigure (a), (b), (c), (d), and (e) shows the Distribution of the context length in the Cancer, Heart, Medication, Smoke and Obesity dataset, respectively