MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering

Ankit Pal   ANKIT.PAL@SAAMA.COM
Saama AI Research Chennai, India

Logesh Kumar Umapathi   LOGESH.UMAPATHI@SAAMA.COM
Saama AI Research Chennai, India

Malaikannan Sankarasubbu   MALAIKANNAN.SANKARASUBBU@SAAMA.COM
Saama AI Research Chennai, India

Abstract

This paper introduces MedMCQA, a new large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions. More than 194k high-quality AIIMS & NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects are collected with an average token length of 12.77 and high topical diversity. Each sample contains a question, correct answer(s), and other options which requires a deeper language understanding as it tests the 10+ reasoning abilities of a model across a wide range of medical subjects & topics. A detailed explanation of the solution, along with the above information, is provided in this study.

Data and Code Availability The dataset to reproduce these experiments and the leaderboard to track the progress of MedMCQA is available at medmcqa.github.io

1. Introduction

Question Answering (QA) is an important and challenging research area in Natural Language Processing (NLP). QA systems enable efficient access to the vast amount of information available that exists in text format.

In recent times, a significant amount of work has been done on constructing a question-answer dataset (Rajpurkar et al., 2016, 2018; Reddy et al., 2019; Kwiatkowski et al., 2019; Yang et al., 2015) reading comprehension datasets (Yang et al., 2018; Lai et al., 2017; Zellers et al., 2018; Yagcioglu et al., 2018; Dua et al., 2019; Bajaj et al., 2018; Huang et al.,

Figure 1: Samples from the MedMCQA dataset, along with the answer's explanation. (✓ : the correct answer)

2019), extractive question answering (Hermann et al., 2015; Trischler et al., 2017), healthcare domain QA (Jin et al., 2019; Šuster and Daelemans, 2018; Möller et al., 2020) and the organization of workshops & competitions such as the Question Answering in the medical domain & BioASQ Challenge (Abacha et al., 2019; Nentidis et al., 2020)

However, despite these successful efforts, automatic questions answering for real medical examination is still a challenge that is less explored. This type of real-world examination dataset on complex medical subjects like pharmacology, medicine, surgery, etc., is scarce. Apart from their scarcity, the requirement of a comprehensive understanding of the domain, matching human experts, makes them appealing for research pursuits. Before this attempt, very few works have been done to construct biomedical MCQA datasets (Vilares and Gomez-Rodr, 2019), and they are (1) mostly small, containing up to few thousand questions, and (2) cover a limited number of Medical topics and Subjects.

Thus, a large-scale, diverse medical QA dataset is needed to accelerate research and facilitate more consistent and effective open-domain QA models in Medical-QA. This paper addresses the aforementioned limitations by introducing MedMCQA, a new large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions. The dataset consists of 194k high-quality medical domain MCQs covering 2.4k healthcare topics and 21 medical subjects to provide a reliable and diverse benchmark. Apart from the question, the correct answer(s), and other options., it also consists of various ancillary data, the primary being a detailed explanation of the solution.

Questions are taken from AIIMS & NEET PG entrance exam MCQs, where graduate medical students are evaluated on their professional knowledge. Questions in these exams are challenging and generally require deeper domain and language understanding as it tests the 10+ reasoning abilities across a wide range of medical subjects & topics. Hence a model must be trained to find relevant information from the open domain knowledge base, reason over them, and choose the correct answer.

Fig.1 shows two example questions, their corresponding explanation, and answers from the study dataset.

An in-depth analysis & a thorough evaluation of the dataset are conducted. The baseline experiments on this dataset with the current state-of-the-

art methods can only answer 47% of the question correctly, which is far behind the performance of human candidates (merit candidates of these exams score an average of 90% marks). Error analysis and results indicate possibilities for improvement in the current methods' reasoning and medical domain question answering. It is believed that this dataset would be an appropriate testbed for future research in this direction.

In brief, the contributions of this study are as follows.

- **Diversity and difficulty** This dataset offers several advantages over existing datasets: (i) Covers 2.4k healthcare topics and 21 medical subjects with an average token length of 12.77, the diversity of questions in MedMCQA demonstrate challenges unique to the dataset. (ii) It is larger than pre-existing Medical QA datasets, (iii) As these questions are from real-world and mock examinations, all the questions and candidate options are created by human experts. These questions are a comprehensive evaluation of a medical practitioner's professional skills, (iv) The questions are difficult & challenging. They test the 10+ reasoning abilities of a model across a wide range of medical subjects & topics.

- **Quality** Detailed statistics, analysis of the data, and fine-grained evaluation per medical subject are provided, yielding a more precise comparison between models. Each sample contains a question, correct answer(s), other options, and a detailed explanation of the solution.

- **Evaluation of quality** Extensive experiments are conducted using high-performance pretrained medical domain models. Error analysis is also provided to illustrate the major challenges of this task. The baseline experiments on this dataset with the most current state-of-the-art methods answer only 47% of the question correctly, which is far behind the human performance of 90%, indicating possibilities for improvement in models' reasoning ability & constitutes a challenging benchmark for future research.

- **Reproducible exam-based split** The dataset is split based on the exams instead of a question-based split (explained in section 2.4). This ensures that the evaluation is closer to the real-world examinations, model generalizability, and

reusability. Individual Examinations tend to have similar questions or pattern of questions repeated periodically. Exam based split avoid this leakage of similar questions into test set, hence helping in generalizability of the dataset. The dataset code to reproduce the experiments & the leaderboard to track the progress of MedMCQA are available at `medmcqa.github.io`

2. The MedMCQA Dataset

In this section, properties of the MedMCQA dataset are presented. Data collection, preparation, preprocessing, and train/test/development splits are discussed.

2.1. Task Definition

The MedMCQA task can be formulated as $\mathbf{X} = \{\mathbf{Q}, \mathbf{O}\}$ where $\mathbf{Q}$ represents the questions in the text, $\mathbf{O}$ represents the candidate options, multiple candidate answers are given for each question $\mathbf{O} = \{\mathbf{O_1}, \mathbf{O_2}, ..., \mathbf{O_n}\}$. The goal is to select the single or multiple answers from the option set.The ground truth label of a data point is $y \in \mathbb{R}^n$ where $y^i = \{\mathbf{0}, \mathbf{1}\}$ and n is the number of options, the objective is to learn a prediction function $f : X \rightarrow y$

2.2. Dataset collection

All India Institute of Medical Sciences (AIIMS PG) & National Eligibility cum Entrance Test (NEET PG) are the two medical entrance exams conducted by All India Institute for Medical Sciences (AIIMS) & National Board of Examinations (NBE), respectively, for providing admission to the postgraduate medical courses. The applicants must have obtained an Bachelor of Medicine and Bachelor of Surgery (MBBS) from a recognized institute to appear for the exams. The exams are used to evaluate the candidates in a structured format, namely, Diagnostic Reasoning and Treatment, Pharmacology, Psychology, Biology, Physical Examination, General Management Strategies, Medical Knowledge, and many other aspects of health and general attitude demeanor of the patient and the examiners. These exams are a comprehensive evaluation of the professional skills of a medical practitioner.

In this paper, the raw data is collected from open websites and books that put together several mock tests and online test series created by medical professionals. In addition to the collected data, AIIMS & NEET PG examination questions (1991- present) from the official websites are also used to create the MedMCQA.

The dataset contains MCQs with fine-grained human-labeled classes on various graduation level medical subjects. Each sample contains ID, question, correct answer, and options. Besides, an explanation of the solution is also provided.

2.3. Preprocessing & Quality Checks

To ensure that all the questions are answerable using textual input only, the following steps were taken to clean the raw data, considering questions from several data sources,

- Questions with an inconsistent format were excluded, e.g., a question where the number of options was not four(excluding punctuation marks).

- Questions with no best answer and missing or null candidates were also omitted.

- Questions whose validity relied on external information were filtered, i.e., the articles and questions containing images or tables.

- Questions containing the keywords "equation", "India", "graph", "map" etc., were removed using a manually curated list of words.

- Further, heuristic rules were also used. For example, in some cases, the question contained HTML tags, special symbols, URLs, extra whitespaces, and missing options. Different tools were used, e.g., a spell checker, an HTML parser, to identify and correct these cases.

- A proofreading tool, 'Grammarly' was used for all the questions, options, and explanations in the dataset to fix the grammar, punctuation, and spelling mistakes. Appropriate suggestions from the tool were applied to the content with human supervision to improve the dataset's quality. As a result, many errors could be corrected

- Lastly, all duplicated questions were removed.

Additional data cleansing steps were carried out to ensure that the question has provided information that matches the data quality goals. The final dataset contains 193,155 questions.

Dataset	# Question	# Subject	Publicly Available	Explanation	Split Type	Open Domain
MedQA	270,000	-	✗	✗	random	✓
HEAD-QA	13,530	6	✓	✗	yearwise	✓
MedMCQA	193,155	21	✓	✓	exam-based	✓

Table 1: Comparison of MedMCQA with several existing MCQA datasets(MedQA(Zhang et al., 2018), HEAD-QA(Vilares and Gomez-Rodr, 2019)) in the medical domain. ✓ represents the dataset that has the feature and ✗ represents it does not
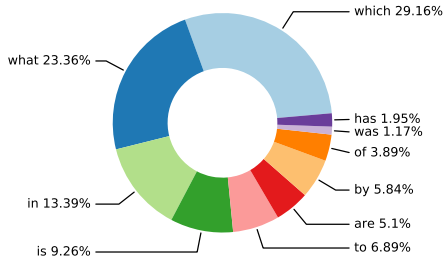

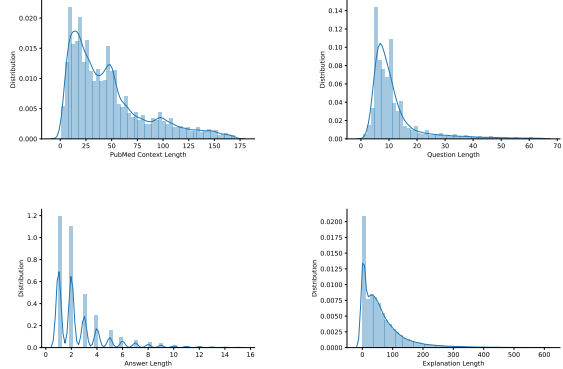Figure 2: Relative sizes of Question Types in MedMCQA

Figure 3: (a) distribution of Pubmed context length (b) Distribution of question length (c) Distribution of answer length (d) Distribution of explanation length

2.4. Split Criteria

The goal of MedMCQA is to emulate the rigor of real word medical exams. To enable that, a predefined split of the dataset is provided. The split is by exams instead of the given questions. This also ensures the reusability and generalization ability of the models.

The training set of MedMCQA consists of all the collected mock & online test series, whereas the test set consists of all AIIMS PG exam MCQs (years 1991-present). The development set consists of NEET PG exam MCQs (years 2001-present) to approximate real exam evaluation.

In the dataset, leakages of similar questions from the training data to test and dev could artificially inflate the models' performance. This is avoided by building the development and test set to include sufficiently different training data questions.

The Levenshtein distance between each pair of questions was computed in the entire dataset. If the similarity between the two documents was larger than 0.9, the question was excluded from the development and test set. The final dataset contains 183K train examples, 6K in the development set, and 4K in the test set.

3. Data statistics

This dataset covers many medical subjects based on the AIIMS & NEET PG entrance exams. The train, development, and test set consist of 182,822 , 4,183 & 6,150 questions with an average token length of 12.35, 13.91 & 9.68, respectively. The general statistics of preprocessed data are summarized in table 2

An additional informative statistic is the count of unique tokens in the dataset plotted in Fig. 4. Vocabulary size is a good measure of linguistic and domain complexity associated with a text corpus and influences the models' performance. It is observed that the length of questions and the vocabulary size in the AIIMS PG exams (test set) are larger than that of the NEET PG exams (dev. set). Hence, it can be inferred that questions from AIIMS are more complex than NEET.

4. Data Analysis

An analysis of the dataset is presented in the subsequent sections. The difficulty and diversity of questions and the answers were analyzed to understand

	Train	Test	Dev	Total
Question #	182,822	6,150	4,183	193,155
Vocab	94,231	11,218	10,800	97,694
Max Q tokens	220	135	88	220
Max A tokens	38	21	25	38
Max E tokens	3,155	651	695	3,155
Avg Q tokens	12.77	9.93	14.09	12.71
Avg A tokens	2.69	2.58	3.19	2.70
Avg E tokens	67.52	46.54	38.44	66.22

Table 2: MedMCQA dataset statistics, where Q, A, E represents the Question, Answer, and Explanation, respectively

the MedMCQA dataset's properties. The complexity of MedMCQA is demonstrated by considering the question and reasoning types covered in the dataset.

4.1. Difficulty and Diversity of Questions

In clinical medicine, a diverse number of questions are possible as it is spread over a range of topics. For example, given the description of a patient's condition, the question might be asked for the most probable diagnosis/the most appropriate treatment or examination required/mechanism of a certain condition, etc.

The majority of the dataset questions are non-factoid and open-ended in nature and seek detailed information about the health condition. Questions in MedMCQA are fairly long, with a mean length of 12.77 words, indicating the compositional nature of questions and different levels of complexity and details covered.

To understand the types of questions in MedMCQA, 25% of questions were sampled, and their properties were analyzed manually. It was observed that 68% of the questions started with an interrogative word, which generally tends to be open-ended. The dataset also contained many dichotomous questions, which often require explanations. The diversity of questions in the MedMCQA makes it a challenging dataset containing many aspects of medical knowledge. Another distinguishing factor of this dataset is that it has questions that were created for and by human domain experts.

4.2. Answer types

In the dataset, each question contains four options with an average length of 2.69 tokens. Out of which, 25% examples were sampled from the development set, and the answer types are presented

in Fig. 5. As shown, MedMCQA covers a broad range of answer types, which matches the analysis on questions' contribution. The answers were manually categorized, and it was observed that answers regarding drug/medicine's name accounted for 22.49%. Medical procedure/Treatment type aiming to determine, measure, or diagnose a condition or parameter accounted for 18.74% of answers. In comparison, 11.24% of answers were related to the quantity of dose(in unit). It was observed that side effects, causes & affected body parts accounted for 12.74%, 10.49% & 9.75% of the dataset. The rest of the answer groups contained fewer instances of the time period, adverse events & other types.

4.3. Subject & Topic Analysis

Fig. 8(A) in the Appendix presents the distribution of medical topics per subject for the datasets. Almost 95% of the subjects contain above 50 topics, while 70% of subjects exceed 100 topics exhibiting a plethora of medical content. Topics range from Medicine (Endocrinology, Infection, Haematology, Respiratory, etc.), Surgery (General Surgery, Endocrinology, breast, and Vascular surgery, etc.) to Radiology & Biochemistry. This wide range of topics increases the dataset's difficulty.

4.4. Reasoning Types

To provide a detailed & better understanding of the datasetś reasoning types, 25% of questions from MedMCQA were sampled randomly. The reasoning types required to answer were manually analyzed. The procedure was followed, and the annotation types presented in (Clark et al., 2018) were re-used to categorize them into the following reasoning types:

- **Question logic** In this, the reasoning is tested by excluding the distractor.

- **Factual** These are the questions that have facts as answers.

- **Explanation/definition** The questions that require selection of definition or explanation or a term/phenomenon.

- **MultiHop Reasoning** To answer these questions, the reasoning is required from multiple passages.
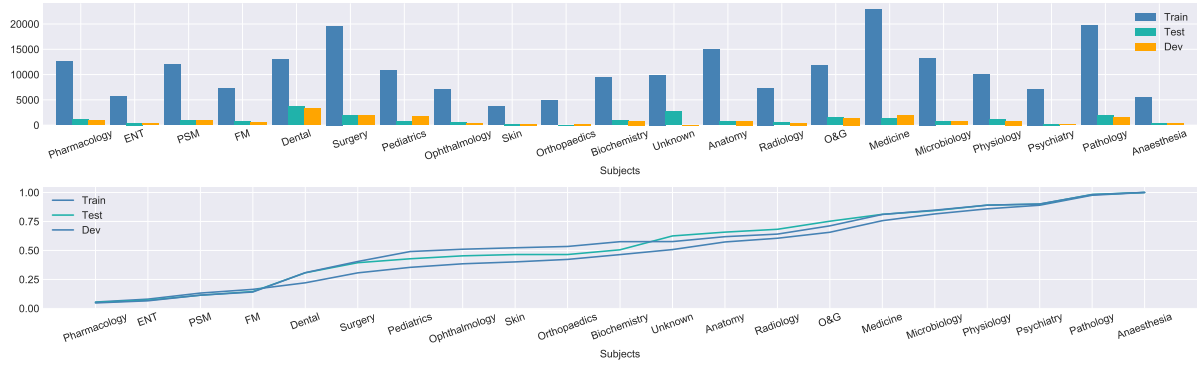
Figure 4: Distribution of unique tokens & Cumulative Frequency Graph in the union of Train, Test, and Development split in MedMCQA dataset. The vocabulary size in the AIIMS PG exams (Test Set) is larger than that of the NEET exams (Dev. Set). Thus indicating the correlation between vocabulary size and difficulty level of the exam.
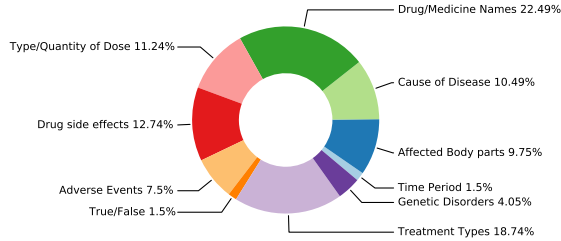

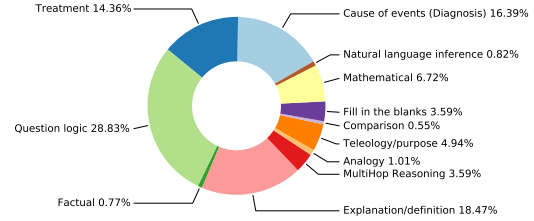Figure 5: Relative sizes of Answer Types in MedMCQA

Figure 6: Relative sizes of Reasoning Types in MedMCQA

- **Analogy** In these types of questions, the responder must select the most similar/analogous answer.

- **Teleology/purpose** Requires understanding of the purpose of a phenomenon/a thing.

- **Comparison** Questions that require reasoning by comparing multiple options.

- **Fill in the blanks** The responder selects the most appropriate answer suitable to fill the blanks.

- **Natural language inference** Determining whether a hypothesis is true, false (contradiction), or neutral given an assumption.

- **Mathematical** Questions that require mathematical critical thinking and logical reasoning.

- **Treatment** Questions that require selection of a correct treatment method for a given ailment / condition.

- **Diagnosis** Questions that require selection of a correct cause of a given ailment / condition.

Fig. 6 shows statistics & examples of major reasoning types in the dataset.

5. Baseline Models

The primary motivation of the baseline experiments is to understand the adequacy of the current models in answering multiple-choice questions meant for human domain experts (post-graduate medical students) and to understand the level of domain specificity required in the models. Therefore, models and knowledge sources with varying levels of specificity are selected. We consider four existing models in our baseline experiments.

They are based on different pre-trained language models using Transformers architecture (Vaswani et al., 2017) , including BERT (Devlin et al., 2019) , SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020) and PubmedBERT(Gu et al., 2022). We fine-tuned these models on our training dataset in a multiclass classification fashion. We consider models of

base size. BERT is evaluated for its out-domain pre-training, SciBERT and BioBERT for their mixed domain and in-domain continual training, and Pubmed-BERT for its in-domain pretraining. These models are explained in detail in the following section,

5.1. SciBERT

SciBERT (Beltagy et al., 2019) is a pretrained language model based on BERT. The model has been pre-trained from scratch on 1.14M papers on the semantic scholar. Even though SciBERT has been pre-trained from scratch, it has a mix of computer science (18%) and biomedical domain (82%), making it a mix-domain pretrained model. The uncased version of the model that uses a vocabulary called scivocab is used, which is a domain-specific vocabulary of size 30K

5.2. BioBERT

BioBERT (Lee et al., 2020) is the first biomedical domain-specific pretrained language model based on BERT. The model is initialized with standard BERT weights (pretrained from Wikipedia and BookCorpus), and continual pretraining is performed with PubMed abstracts and full texts. The model uses the same vocabulary as the standard BERT model. The base variant of the 1.1 version of the model is used in the experiments.

5.3. PubMedBERT

PubMedBERT (Gu et al., 2022) is a recent domain-specific pre-trained language model that is first to pretrain only on in-domain texts (PubMed abstracts and full texts). The base version of the model trained with both abstracts and full texts is used in the experiments. This model is used to evaluate the performance of a fully in-domain pre-trained model on the dataset.

5.4. Retriever models

With the recent success of neural retrievers, dense passage retrieval (Karpukhin et al., 2020), and Pub-MedBERT(Gu et al., 2022) were utilized to evaluate Wikipedia and PubMed as knowledge bases, respectively. Dense passage retriever follows a siamese/bi-encoder architecture; One encoder encodes the documents and another to encode the query, originally

trained with Maximum inner product search objective. The pretrained DPR model and Wikipedia index from Transformer's library (Wolf et al., 2020) were used in the experiments.

6. Experiments

To complement the motivation stated in section 5, The reader models were chosen with varying domain specificity levels. The contribution of external knowledge sources (Wikipedia and PubMed) was evaluated by providing these sources as contexts. Furthermore, an ablation study was also performed on context by training and evaluating all the models without context. This was done to understand the contribution of external context and the usefulness of the internal knowledge stored in these domain-specific models. The baseline experiments are broadly classified as follows,

- **Out-Domain**: Pre-trained models trained on out-domain corpora like Wikipedia and Book corpus were used in this experiment type.

- **Mix domain (continual)**: Pre-trained models trained on out-domain initially and later adapted to in-domain or trained from scratch on both out-domain and in-domain corpora were used in this experiment.

- **In-Domain**: Pre-trained models trained from scratch on in-domain corpora like PubMed abstracts and full texts were used in this experiment type.

All these experiments were repeated with and without external knowledge context.

6.1. Pubmed Data Preprocessing

Before encoding the passages, the passages were truncated to 250 token lengths to fit the memory.

6.2. Retriever

For the experiments that involve context, a retriever+reader pipeline approach was opted (as introduced in (Chen et al., 2017)). The out-of-the-box retriever models were used (explained in the section 6.2) from Huggingface's Transformers library (Wolf et al., 2020) to encode the passages and questions. The passage with the highest cosine similarity was retrieved and used as a context for training the reader models.

Figure 7: The Retriever+Reader Pipeline for Open-Domain Question Answering system used in our experiments.Dense passage retrieval (Karpukhin et al., 2020) and PubMedBERT (Gu et al., 2022) are used to evaluate Wikipedia and PubMed as knowledge bases respectively, while different transformer models (explained in section 5) as reader models.

6.3. Reader finetuning

The finetuning approach was followed as in (Devlin et al., 2019) to finetune the reader models. The highest scoring contexts for each question are retrieved from the retriever. These contexts are combined by [SEP] token with the concatenation of question and answer pair. This creates four input sequences per question.

[CLS] Context [SEP] Question [SEP] Option [SEP]

A linear layer with softmax is applied over the output of the [CLS] token of the encoder. This is to select the most appropriate option for a question and context pair.

For the experiments that do not use context, question and answer pair concatenation is encoded, and a linear layer with softmax is applied over the output of the [CLS] token of the encoder to select the most appropriate option for a question.

[CLS] Question [SEP] Option [SEP]

The models were finetuned on two Tesla T4 GPUs for 5 epochs with a learning rate of 2e-4 and a batch size of 16. The model checkpoint with the highest validation score in the 5 epochs was selected and used to evaluate the Test Set.

7. Error Analysis

The error analysis details on a sample set of mispredictions by the best baseline model (PubMedBERT) is given in this section. The analysis was done manually for about 100 mispredictions that were sampled.This could be used for further research to improve the models/methods on the dataset.

- **Multi-hop reasoning**: It was observed that the model often mispredicted the questions related to the cause of an event (diagnosis) and the right course of action (treatment) in a given medical situation. Such questions typically require information on multiple symptoms, ailments, and treatments to select the most appropriate choice. This multiplicity of information is not likely to be present in one passage, possibly the reason for the mispredictions.

- **Incorrect context passages**: It is observed that inadequate contexts from the retriever are also major contributors to the mispredictions.

- It is found that the models mispredicted the questions requiring arithmetic reasoning. This is in line with the observations in (Dua et al., 2019) on BERT-based models.

8. Result & Discussion

In this section, the results from the evaluation of the methods discussed in section 6 are presented.

- It is observed that PubMedBERT performs better than other models in all the categories. This aligns with the results from (Gu et al., 2022) where PubMedBERT surpasses all other biomedical models in the majority of BLURB tasks. Examples of correct and incorrect predictions of the model is presented in Table A

- PubMedBERT is followed by SciBERT (mix domain pretraining) and BioBERT (continual pretraining) in accuracy. From this result, it can be inferred that the model's performance decreases with a decrease in domain specificity of the models and external knowledge sources.

- It is observed that there is an insignificant improvement in the model's performance when

Wikipedia is used as context compared to without context results, and the model variants trained on PubMed, which have a 4-7% improvement in the performance. This can be attributed to the domain specificity of the external knowledge source required by the dataset. The majority of the reasoning types (Diagnosis, treatment, etc.) mentioned in 4.4 require domain expertise as these questions are intended for post-graduate medical students.

- The subject wise accuracies of the top PubMed-BERT model is presented in Table 3

Subject Name	Test	Dev
Anaesthesia	0.47	0.26
Anatomy	0.40	0.39
Biochemistry	0.48	0.49
Dental	0.43	0.36
ENT	0.47	0.52
FM	0.48	0.35
O&G	0.54	0.39
Medicine	0.49	0.47
Microbiology	0.50	0.44
Ophthalmology	0.60	0.51
Orthopaedics	-	0.33
Pathology	0.53	0.46
Pediatrics	0.39	0.45
Pharmacology	0.46	0.46
Physiology	0.47	0.47
Psychiatry	**0.67**	**0.56**
Radiology	0.42	0.31
Skin	0.50	0.29
PSM	0.44	0.35
Surgery	0.50	0.43
Unknown	0.44	1.0

Table 3: Fine-grained evaluation per medical subject in test and dev set

9. Conclusion

In this work, MedMCQA, a new large-scale, Multi-Choice Question Answering (MCQA) dataset, is presented, which requires a deeper domain and language understanding as it tests the 10+ reasoning abilities of a model across a wide range of medical subjects & topics. It is demonstrated that the dataset is challenging for the current state-of-the-art methods and domain-specific models, with the best baseline achieving only 47% accuracy. It is expected that this

Model	w/o Context		Wiki		PubMed	
	Test	Dev	Test	Dev	Test	Dev
Bert$_{Base}$	0.33	0.35	0.33	0.35	0.37	0.35
BioBert	0.37	0.38	0.39	0.37	0.42	0.39
SciBert	0.39	0.39	0.38	0.39	0.43	0.41
PubMedBERT	**0.4**	**0.40**	**0.4**	**0.41**	**0.47**	**0.43**

Table 4: Performance of all baseline models in accuracy (%) on MedMCQA test-dev set

dataset would facilitate future research in this direction.

Institutional Review Board (IRB)

This research does not require IRB approval.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *MEDIQA*, pages 370–379. Association for Computational Linguistics, 8 2019. doi: 10.18653/v1/W19-5039. URL https://www.aclweb.org/anthology/W19-5039.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset. *2016 Conference on Neural Information Processing Systems*, 2018.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *SciBERT: A Pretrained Language Model for Scientific Text*, pages 3615–3620. Association for Computational Linguistics, 11 2019. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and

Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *The Allen Institute for Artificial Intelligence*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, pages 4171–4186. Association for Computational Linguistics, 6 2019. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs*, 2019.

Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ArXiv*, abs/2007.15779, 2022.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 December 2015 Pages 1693–1701*, 2015.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. Covid-qa: A question answering dataset for covid-19. In *COVID-QA: A Question Answering Dataset for COVID-19*. Association for Computational Linguistics, 7 2020. URL https://www.aclweb.org/anthology/2020.nlpcovid19-acl.18.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. Results of the seventh edition of the bioasq challenge. *Communications in Computer and Information Science*, page 553–568, 2020. ISSN 1865-0937. doi: 10.1007/978-3-030-43887-6_51. URL http://dx.doi.org/10.1007/978-3-030-43887-6_51.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, pages 2383–2392. Association for Computational Linguistics, 11 2016. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

David Vilares and Carlos Gomez-Rodr. Head-qa: A healthcare dataset for complex reasoning. In *HEAD-QA: A Healthcare Dataset for Complex Reasoning*, pages 960–966. Association for Computational Linguistics, 7 2019. doi: 10.18653/v1/P19-1092. URL https://www.aclweb.org/anthology/P19-1092.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Transformers: State-of-the-Art Natural Language Processing*, pages 38–45. Association for Computational Linguistics, 10 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

Yi Yang, Wen tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *WikiQA: A Challenge Dataset for Open-Domain Question Answering*, pages 2013–2018. Association for Computational Linguistics, 9 2015. doi: 10.18653/v1/D15-1237. URL https://www.aclweb.org/anthology/D15-1237.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Simon Šuster and Walter Daelemans. Clicr: A dataset of clinical case reports for machine reading comprehension. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.
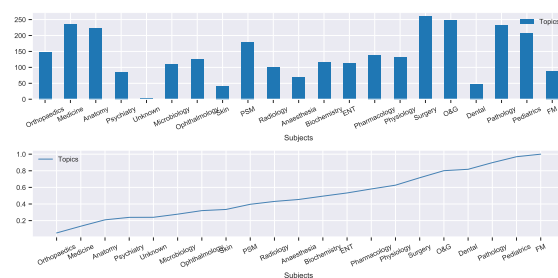
Appendix A. Topic Distribution

Figure 8: Distribution of topics per subject & Cumulative Frequency Graph for MedMCQA dataset.

Predictions from the best model

A.1. Correct Predictions

Question	Correct option	Options	Prediction
A 10-year-old boy is having sensory neural deafness. He showed no improvement with conventional hearing aids. Most appropriate management is:	D	A. Bone conduction hearing aids B. Fenestration C. Stapes fixation D. Cochlear implant	D
Meralgia paraesthetica is due to the involvement of:	A	A. Lateral cutaneous nerve of the thigh B. Sural nerve C. Medial cutaneous nerve of the thigh D. Femoral nerve	A
Xanthenuric acid is produced in metabolism of?	A	A.Tyrosine B.Glycine C.Methionine D.Tryptophan	A
A 10 years — old child is brought to the emergency room with seizures of the tonic — clonic type. His mother reports that these seizures have been occurring for the past 50 minutes. The treatment of choice is.	A	A.Diazepam B.Phenytoin C. Carbamazepine D.Valproate	A
Which drug is a selective COX 2 inhibitor?	A	A. Celecoxib B.Acetaminophen C.Ketorolac D.Aspirin	A

A.2. Incorrect Predictions

Question	Correct option	Options	Prediction
Drug of choice for American trypanosomiasis is?	D	A. Miltefosine B. Amphotericin C. Amphotericin D. Amphotericin	A
Which of the following drugs dosage interval should be maximum in a patient with creatinine clearance less than 10,	C	A. Amikacin B. Rifampicin C. Vancomycin D. Amphotericin	B
Filgrastrim is used for:	A	A.Neutropenia B.Anemia C.Polycythemia D.Neutrophilia	C
A 30 years old male is having prpductive cough with dysnea. Blood gas analysis shows low pa02. Chest x-ray is showing reticulonodular pattern. The causative agent is?	C	A.Staph aureus B.Pneumococcus P. jerovecii Pseudomonas	B
A population study showed a mean glucose of 86 mg/dL in a sample of 100 showing normal curve distribution, what percentage of people have glucose above 86 mg/dL?	B	A. 34 B.50 C.Nil D.68	A