saama

# MedMCQA

## Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering


Ankit Pal (Aaditya Ura)
Research Engineer, Saama AI Research Lab
ankit.pal@saama.com

Logesh Kumar Umapathi
Research Engineer, Saama AI Research Lab
logesh.umapathi@saama.com

Malaikannan Sankarasubbu
VP of AI, Saama AI Research Lab
Malaikannan.Sankarasubbu@saama.com

# About Authors

**Ankit Pal (Aaditya Ura)**

- Research Engineer @ Saama AI Research Lab
- Research interests involve
  Representation Learning on Graphs and Manifolds
- Interpretable Natural Language Processing,
  and their applications in Healthcare data
- Respiratory, Neurophysiological (EEG, ECG, EMG etc.)
  based Signal Processing

Email : ankit.pal@saama.com
Website : aadityaura.github.io
twitter : @aadityaura

**Logesh Umapathi**

- Research Engineer @ Saama AI Research Lab
- Research interests involve
  information retrieval
  and their applications in Healthcare , clinical trials.

Email : logesh.umapathi@saama.com
Website : http://logeshumapathi.com
twitter : @logesh_umapathi

# Outline

- **Introduction**
- **MedMCQA Dataset**
  - Sample data, Data Collection, Preprocessing & split criteria
- **Data statistics**
  - Total questions, vocab etc
- **Data Analysis**
  - Difficulty and Diversity of Questions
  - Answer types
  - Subject & Topic Analysis
  - Reasoning Types
- **Experiments**
  - Retriever
  - Reader finetuning
  - Results

# Introduction

| 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

# MedMCQA Dataset

The dataset is designed to address real-world medical entrance exam questions

- Question from AIIMS PG and NEET PG entrance exam
- 194k high-quality medical domain MCQs
- 2.4k healthcare topics
- 21 medical subjects
- Tests 10+ reasoning abilities

**Dataset :**

X = {Q,O} , Q - questions , O - candidate options

O = {O1,O2, ...,On}.

**Label :** $y \in R^n$ where $y^i = \{0,1\}$ , n is number of options

**Objective :** $f : X \rightarrow y$

# Sample Data

## Pharmacology

**Q** A 40-year-old man has megaloblastic anemia and early signs of neurological abnormality. The drug most probably required is

**A**
a) Folic acid
b) Iron sulphate
c) Erythropoietin
d) Vitamin B12 ✓

**E** Deficiency of vitamin B12 results in megaloblastic anemia and demyelination. It can cause subacute combined degeneration of the spinal cord and peripheral neuritis.

## Surgery

**Q** A five-year-old child presents with ballooning of prepuce after micturition. Examination reveals preputial adhesions. Which of the following is the best treatment?

**A**
a) Circumcision ✓
b) Dorsal slit
c) Adhesiolysis & dilatation
d) Conservative management

**E** The Treatment of phimosis in children is dependent on the parent's preference, however preputial if phimosis is causing ballooning of prepuce, circumcision is strongly considered.

# Exams

- Sources of the dataset are from
  - All India Institute of Medical Sciences Post Graduate Exam (AIIMS PG)
  - National Eligibility cum Entrance Test ( NEET PG)

- **Eligibility**
  - a Bachelor of Medicine and Bachelor of Surgery (MBBS) from a recognized institute
  - Completed a 12 months of mandatory rotating Internship

- Merit candidates of these exams score an average of 90% marks

# Data Collection & Preprocessing

Source:

- Historical Exam questions from official websites - AIIMS & NEET PG (1991- present)
- The raw data is collected from open websites and books

Quality Checks:

- Questions with an inconsistent format were excluded
- Questions with no best answer and missing or null candidates were also omitted.
- Questions containing images or tables
- Keywords: “equation”, “India”, “graph”, “map”

Preprocessing:

- Heuristics rules to clean - HTML tags, Special symbols , URLs , Missing options
- ‘Grammarly’ was used to fix the grammar, punctuation, and spelling mistakes.
- Duplicates were removed

# Split Criteria

**Rationale for exam based split:**

- To ensure the evaluation is closer to the real world examinations, model generalizability, and reusability
- Similar questions from train , test and dev set were removed based on similarity

**Exam based split :**

1. Training set : Mock & online test series - 183K examples
2. Test set : Real AIIMS exam MCQs (years 1991- present) - 6K examples
3. Dev set : NEET exam MCQs (years 2001- present) - 4K examples

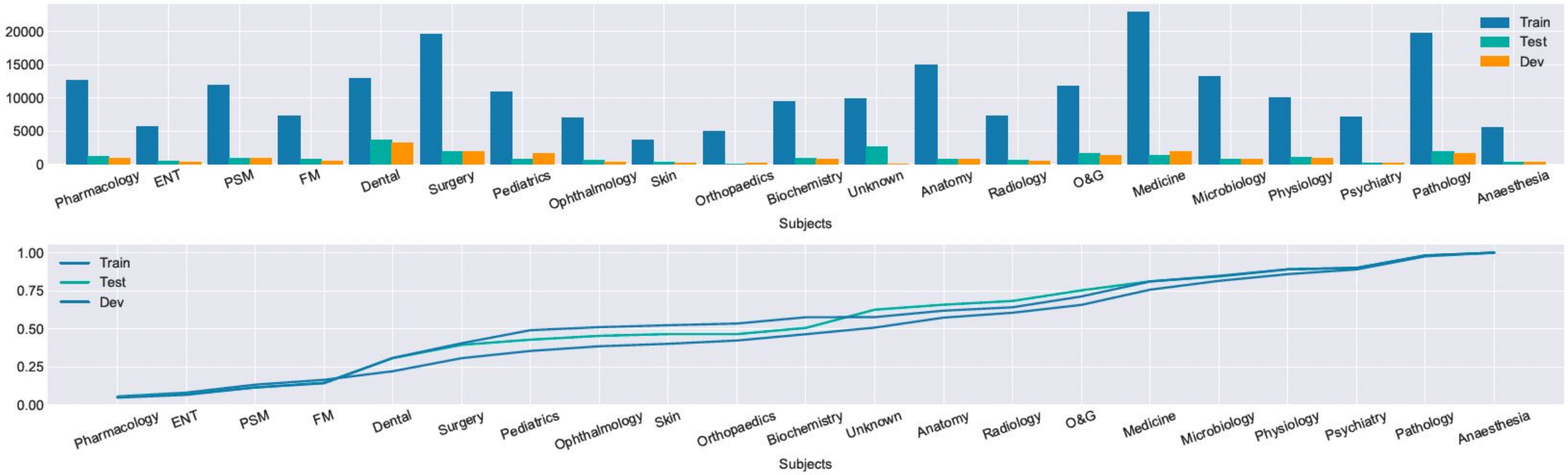
# Data Statistics

Difficulty and Diversity of Questions :

- majority of the dataset questions are non-factoid and open-ended in nature
- mean length of 12.77 words

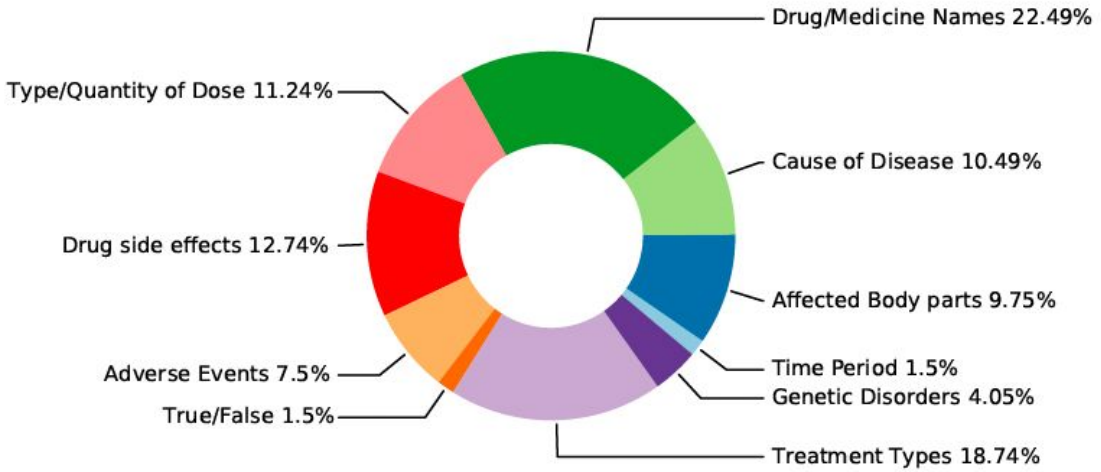
| | Train | Test | Dev | Total |
|---|---|---|---|---|
| Question # | 182,822 | 6,150 | 4,183 | 193,155 |
| Vocab | 94,231 | 11,218 | 10,800 | 97,694 |
| Max Q tokens | 220 | 135 | 88 | 220 |
| Max A tokens | 38 | 21 | 25 | 38 |
| Max E tokens | 3,155 | 651 | 695 | 3,155 |
| Avg Q tokens | 12.77 | 9.93 | 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 |

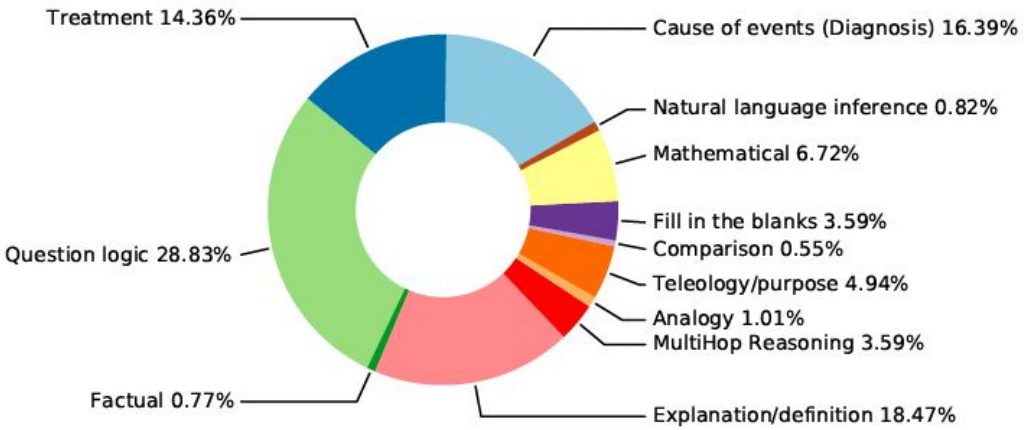
# Data Statistics




**Distribution of unique tokens & Cumulative Frequency Graph**

# Data Statistics




**Answering and Reason types**

# Baseline

**Experiment Design:**

- **Motivation** :
  - Adequacy of the current models in answering multiple-choice questions meant for human domain experts
  - understand the level of domain specificity required in the models

- To evaluate the need for external domain specific knowledge source:
  - No KB
  - Wikipedia
  - Pubmed

- To evaluate the effectiveness of pretraining source:
  - out of domain pretrained models (BERT (Devlin et al., 2019) )
  - Mixed domain pretrained models (SciBERT (Beltagy et al., 2019) & BioBERT (Lee et al.,2020)
  - In-domain pretrained models (PubmedBERT (Gu et al., 2022))

# Baseline

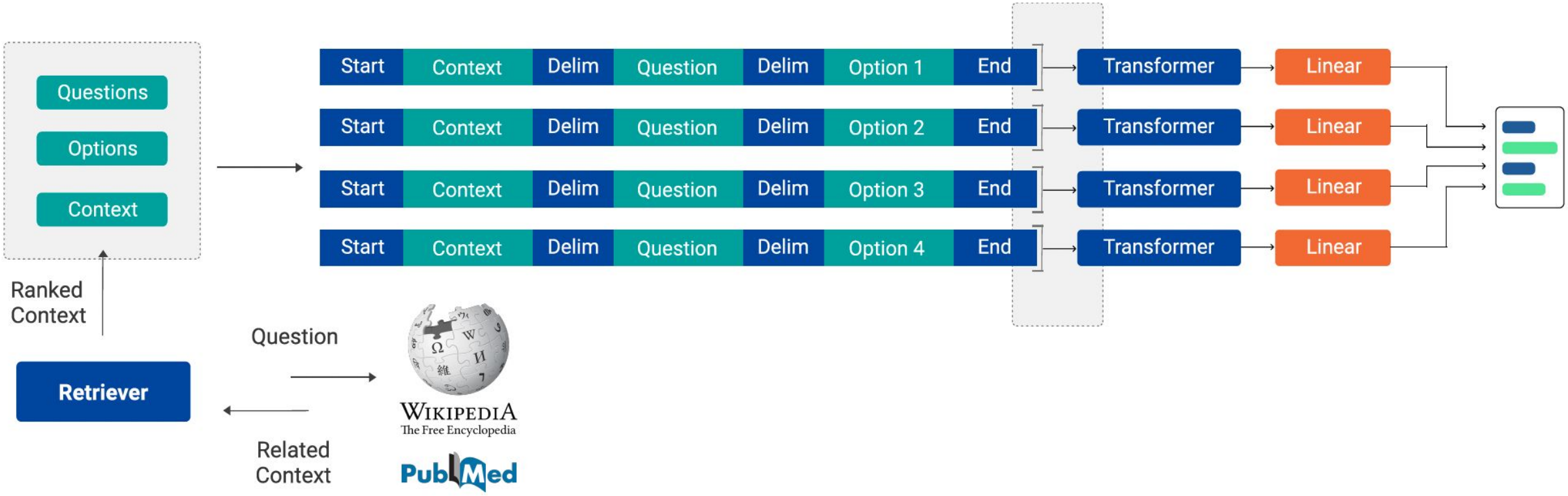
Retriever Models :

- Wikipedia : Dense passage retrieval ( Karpukhin et al., 2020 )
- Pubmed : PubmedBERT ( Gu et al., 2022)

Reader Models:

- BERT (Devlin et al., 2019) - Out of domain pretraining
- SciBERT (Beltagy et al., 2019) & BioBERT (Lee et al.,2020) - Mixed domain pretraining
- PubmedBERT ( Gu et al., 2022) - In domain pretraining

# Baseline


Questions
Options
Context
Ranked Context
Retriever
Question
Related Context
WIKIPEDIA
The Free Encyclopedia
PubMed
Start Context Delim Question Delim Option 1 End
Start Context Delim Question Delim Option 2 End
Start Context Delim Question Delim Option 3 End
Start Context Delim Question Delim Option 4 End
Transformer
Transformer
Transformer
Transformer
Linear
Linear
Linear
Linear

# Baseline Results

| | w/o Context | | Wiki | | PubMed | |
|---|---|---|---|---|---|---|
| Model | Test | Dev | Test | Dev | Test | Dev |
| $\text{Bert}_{Base}$ | 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** |

# Baseline Results - evaluation per medical subject

| 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 |

# Questions?

*Thanks !*

ankit.pal@saama.com, Logesh.umapathi@saama.com

twitter : @aadityaura, @logesh_umapathi