MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset

Jing Li, Shangping Zhong and **Kaizhi Chen**
College of Computer and Data Science, Fuzhou University, Fuzhou, China
`{N190320027, spzhong, ckz}@fzu.edu.cn`

Abstract

Question **A**nswering (**QA**) has been successfully applied in scenarios of human-computer interaction such as chatbots and search engines. However, for the specific biomedical domain, QA systems are still immature due to expert-annotated datasets being limited by category and scale. In this paper, we present MLEC-QA, the largest-scale Chinese multi-choice biomedical QA dataset, collected from the National Medical Licensing Examination in China. The dataset is composed of five subsets with 136,236 biomedical multi-choice questions with extra materials (images or tables) annotated by human experts, and first covers the following biomedical sub-fields: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Traditional Chinese Medicine Combined with Western Medicine. We implement eight representative control methods and open-domain QA methods as baselines. Experimental results demonstrate that even the current best model can only achieve accuracies between 40% to 55% on five subsets, especially performing poorly on questions that require sophisticated reasoning ability. We hope the release of the MLEC-QA dataset can serve as a valuable resource for research and evaluation in open-domain QA, and also make advances for biomedical QA systems.[1]

1 Introduction

As a branch of the QA task, **B**iomedical **Q**uestion **A**nswering (**BQA**) enables effectively perceiving, accessing, and understanding complex biomedical knowledge by innovative applications, which makes BQA an important QA application in the biomedical domain (Jin et al., 2021). Such a task has recently attracted considerable attention from the NLP community (Zweigenbaum, 2003; He et al., 2020b; Jin et al., 2020), but is still confronted with the following three key challenges: (1) Most work attempt to build BQA systems with deep learning and neural network techniques (Ben Abacha et al., 2017, 2019b; Pampari et al., 2018) and are thus data-hungry. However, annotating large-scale biomedical question-answer pairs with high quality is prohibitively expensive. As a result, current expert-annotated BQA datasets are small in size. (2) Multi-choice QA is a typical format type of BQA dataset. Most previous work focus on such format type of datasets in which contents are in the field of clinical medicine (Zhang et al., 2018b; Jin et al., 2020) and consumer health (Zhang et al., 2017, 2018a; He et al., 2019; Tian et al., 2019). However, there are many other specialized sub-fields in biomedicine that have not been studied before (e.g., Stomatology). (3) Ideal BQA systems should not only focus on raw text data, but also fully utilize various types of biomedical resources, such as images and tables. Unfortunately, most BQA datasets are either texts (Tsatsaronis et al., 2015; Pampari et al., 2018; Jin et al., 2019) or images (Lau et al., 2018; Ben Abacha et al., 2019a; He et al., 2020a); as a result, BQA datasets that are composed by fusing different biomedical resources are relatively limited.

To push forward the variety of BQA datasets, we present MLEC-QA, the largest-scale Chinese multi-choice BQA dataset. Questions in MLEC-QA are collected from the **N**ational **M**edical **L**icensing **E**xamination in **C**hina (**NMLEC**)[2], which are carefully designed by human experts to evaluate professional knowledge and skills for those who want to be medical practitioners in China. The NMLEC has a total number of 24 categories of exams, but only five of them have the written exams in Chinese. Every year, only around 18-22% of applicants can pass one of these exams, showing the complexity and difficulty of passing

[1] `https://github.com/Judenpech/MLEC-QA`

[2] http://www.nmec.org.cn/Pages/ArticleList-12-0-0-1.html

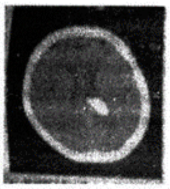
Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8862–8874
November 7–11, 2021. ©2021 Association for Computational Linguistics

them even for skilled humans.

There are three main properties of MLEC-QA: (1) MLEC-QA is the largest-scale Chinese multi-choice BQA dataset, containing 136,236 questions with extra materials (images or tables), Table 1 shows an example. (2) MLEC-QA first covers the following biomedical sub-fields: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Traditional Chinese Medicine Combined with Western Medicine (denoted as Chinese Western Medicine). Only one (Clinic) of them has been studied in previous research. (3) MLEC-QA provides extra labels of five question types (A1, A2, A3/A4 and B1) for each question, and an in-depth analysis of the most frequent reasoning types of the questions in MLEC-QA, such as lexical matching, multi-sentence reading and concept summary, etc. Detailed analysis can be found in Section 3.2. Examples of sub-fields and question types are summarized in Table 2. We set each example of five question types corresponding to one of the sub-fields due to page limits.

[Question]

男，63 岁。3 小时前长跑后头痛伴呕吐。查体：查体不合作，嗜睡，双瞳孔对光反射存在、颈项强直，四肢活动自如，肌张力略高。双侧 Babinski 征明显，头颅 CT 如图，该患者的诊断是（　）



Male, 63 years old. Had headache with vomiting after long-distance running 3 hours ago. Physical Examination: not cooperative, somnolence, double pupillary light reflex exists, neck rigidity, free movement of limbs, muscle tension slightly high. Bilateral Babinski sign is evident, CT of the head as shown in the figure. Which option is the correct diagnosis of the patient:

[Options]

A：蛛网膜下腔出血 / Subarachnoid hemorrhage
B：脑室肿瘤 / Tumor of ventricle
C：脑室囊肿 / Ventricular cyst
D：脑室出血 / Ventricular hemorrhage
E：脉络膜钙化 / Choroidal calcification

[Answer]

D

Table 1: An example of questions with additional images in MLEC-QA dataset.

As an attempt to solve MLEC-QA and provide strong baselines, we implement eight representative control methods and open-domain QA methods by a two-stage retriever-reader framework: (1) A retriever finding documents that (might) contain an answer from a large collection of documents. We adopt Chinese Wikipedia dumps[3] as our information sources, and use a distributed search and analytics engine, ElasticSearch[4], as the document store and document retriever. (2) A reader finding the answer in given documents retrieved by the retriever. We fine-tune five pre-trained language models for machine reading comprehension as the reader. Experimental results show that even the current best model can only achieve accuracies of 53%, 44%, 40%, 55%, and 50% on the five categories of subsets: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Chinese Western Medicine, respectively. The models especially perform poorly on questions that require understanding comprehensive biomedical concepts and handling complex reasoning.

Question Type	Example (∗ represents the correct answer)
A1	(Public Health) **What should be used to compare the results of two samples?** A. T test B. x^2 test C. μ test D. F test ∗E. Rank sum test
B1	(Chinese Western Medicine) A. Heart and liver B. Spleen and lungs C. Liver and kidneys D. Heart and kidneys E. Spleen and kidneys **1. What is the viscera that "Yikui" in Yikui homologous refers to? (E)** **2. What is the viscera that "water and fire" in harmonization of water and fire refers to? (D)**
A2	(Clinic) **Primipara, 29 years old, at 37 weeks of gestation. Had jet vomiting once this morning, suddenly convulsed an hour ago and then went to hospital in a coma. Physical examination: BP 180/120mmhg, urine protein (+++). The most likely diagnosis of this patient is:** ∗A. Eclampsia B. Hematencephalon C. Hysteria D. Epilepsy E. Cerebral thrombosis
A3	(Traditional Chinese Medicine) **Female, 28 years old. In the recent month, has oral ulcer attacks repeatedly, upset, difficulties to sleep at night, dry stool, defecates every 1-2 days, has dry mouth, does not like drinking water, and has yellow urine, red tongue, greasy fur, and rapid pulse.** **1. The drug of choice for the treatment of the pattern is:** A. Mirabilite B. Arctium lappa ∗C. Bamboo leaf D. Baikal skullcap E. Gypsum **2. The appropriate compatible drug for the treatment of the disease is:** A. Angelica dahurica B. Cassia twig C. Rhizoma Zingiberis ∗D. Rheum officinale E. Ash bark **3. Which of the following drugs should be used with caution during menstruation?** A. Semen sojae praeparatum ∗B. Rheum officinale C. Coptis chinensis D. Lophatherum gracile E. Rhizoma phragmitis
A4	(Stomatology) **A 50-year-old patient comes to the outpatient clinic 2 years after the end of radiotherapy for nasopharyngeal carcinoma outside hospital. Examination: full mouth multiple teeth dental surfaces with different degrees of caries, some of the affected teeth have become residual crowns, residual roots, less intraoral saliva, more soft scaling of the dental surfaces and sulci.** **1. The diagnosis of this patient is:** A. Acute caries ∗B. Rampant caries C. Chronic caries D. Secondary caries E. Smooth surface caries **2. There are several treatment designs as follows, except:** A. Design treatment of full mouth caries B. Endodontic treatment of teeth with hypodontia ∗C. Filling metal material D. Remineralization adjunctive therapy E. Regular review

Table 2: Examples of sub-fields and question types in MLEC-QA. The Chinese version is in Appendix D.

In summary, the major contributions of this paper are threefold:

- We present MLEC-QA, the largest-scale Chinese multi-choice BQA dataset with extra materials, and it first covers five biomedical sub-fields, only one of which has been studied in previous research.

[3] https://dumps.wikimedia.org/

[4] https://www.elastic.co/

- We conduct an in-depth analysis on MLEC-QA, revealing that both comprehensive biomedical knowledge and sophisticated reasoning ability are required to answer questions.
- We implement eight representative methods as baselines and show the performance of existing methods on MLEC-QA, and provide an outlook for future research directions.

2 Related Work

Open-Domain BQA The **T**ext **RE**trieval **C**onference (**TREC**) (Voorhees and Tice, 2000) has triggered the open-domain BQA research. At the time, most traditional BQA systems were employing complex pipelines with question processing, document/passage retrieval, and answer processing modules. Examples of such systems include EPoCare (Niu et al., 2003), MedQA (Yu et al., 2007; Terol et al., 2007; Wang et al., 2007) and AskHERMES (Cao et al., 2011). With the introduction of various BQA datasets that are focused on specific biomedical topics, such as BioASQ (Tsatsaronis et al., 2015), emrQA (Pampari et al., 2018) and PubMedQA (Jin et al., 2019), pioneered by Chen et al. (2017), the modern open-domain BQA systems largely simplified the traditional BQA pipeline to a two-stage retriever-reader framework by combining information retrieval and machine reading comprehension models (Ben Abacha et al., 2017, 2019b).

Moreover, the extensive use of medical images (e.g., CT) and tables (e.g., laboratory examination) has improved results in real-world clinical scenarios, making the BQA a task lying at the intersection of **C**omputer **V**ision (**CV**) and NLP. However, most BQA models focus on either texts or images (Lau et al., 2018; Ben Abacha et al., 2019a; He et al., 2020a); as a result, BQA datasets that are composed by fusing different biomedical resources are relatively limited.

Open-Domain Multi-Choice BQA Datasets With rapidly increasing numbers of consumers asking health-related questions on online medical consultation websites, cMedQA (Zhang et al., 2017, 2018a), webMedQA (He et al., 2019) and ChiMed (Tian et al., 2019) exploit patient-doctor QA data to build consumer health QA datasets. However, the quality problems in such datasets are that the answers are written by online-doctors and the data itself has intrinsic noise. By contrast, medical licensing examinations, which are designed by human medical experts, often take the form of multi-choice questions, and contain a significant number of questions that require comprehensive biomedical knowledge and multiple reasoning ability. Such exams are the perfect data source to push the development of BQA systems. Several datasets have been released that exploit such naturally existing BQA data, which are summarized in Table 3. Collecting from the Spain public healthcare specialization examination, HEAD-QA (Vilares and Gómez-Rodríguez, 2019) contains multi-choice questions from six biomedical categories, including Medicine, Pharmacology, Psychology, Nursing, Biology and Chemistry. NLPEC (Li et al., 2020) collects 21.7k multi-choice questions with human-annotated answers from the National Licensed Pharmacist Examination in China, but only a small number of sample data is available for public use.

Last but not least, clinical medicine, as one of the 24 categories in NMLEC, has been previously studied by MedQA (Zhang et al., 2018b) and MEDQA (Jin et al., 2020). However, the former did not release any data or code, and the latter only focused on clinical medicine with 34k questions in their cross-lingual studies, questions with images or tables were not included, and none of the remaining categories in MLEC-QA were studied.

Dataset	Size	Content	Metric	Available	Language	Extra
cMedQA v1.0	54k	Consumer Health	P@1	Yes	Chinese	No
cMedQA v2.0	108k	Consumer Health	P@1	Yes	Chinese	No
webMedQA	63k	Consumer Health	P@1 & MAP	Yes	Chinese	No
ChiMed	24.9k	Consumer Health	Acc	Yes	Chinese	No
NLPEC	21k	Pharmacology	Acc	No*	Chinese	No
MedQA	235k	Clinical Medicine	Acc	No	Chinese	No
MEDQA	61k	Clinical Medicine	Acc	Yes	Chinese & English	No
HEAD-QA	6.8k	Multi-Category	Acc	Yes	Spanish & English	Yes
MLEC-QA	136k	Multi-Category	Acc	Yes	Chinese	Yes

Table 3: Comparison of MLEC-QA with existing open-domain multi-choice BQA datasets. No* indicates a small number of sample data is available. Extra indicates if the dataset provides extra material to answer questions.

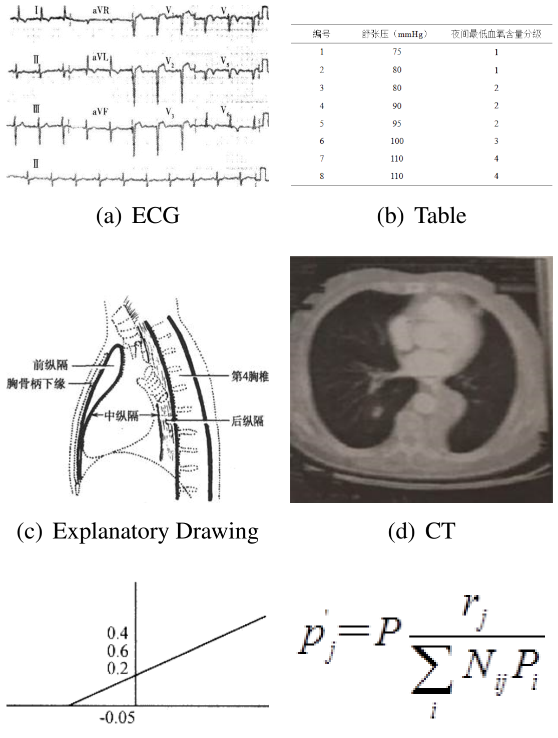
3 MLEC-QA Dataset

3.1 Data Collection

We collect 155,429 multi-choice questions from the 2006 to 2020 NMLEC and practice exercises from the Internet. Except for the categories that do not use Chinese in examinations, all categories are included in MLEC-QA: Clinic (Cli), Stomatology (Sto), Public Health (PH), Traditional Chinese

Medicine (TCM), and Chinese Western Medicine (CWM). After removing duplicated or incomplete questions (e.g., some options missing), there are 136,236 questions in MLEC-QA, and each question contains five candidate options with one correct/best option and four incorrect or partially correct options. We describe in detail the JSON data structure of MLEC-QA in Appendix B.

MLEC-QA contains 1,286 questions with extra materials that provide additional information to answer correctly. As shown in Figure 1, the extra materials are all in a graphical format with various types, such as ECG, table of a patient's condition record, formula, CT, line graph, explanatory drawing, etc. We include these questions with extra materials in MLEC-QA to facilitate future BQA explorations on the crossover studies of CV and NLP, although we will not exploit them in this work due to the various specifics involved in extra materials.



编号	舒张压（mmHg）	夜间最低血氧含量分级
1	75	1
2	80	1
3	80	2
4	90	2
5	95	2
6	100	3
7	110	4
8	110	4

(a) ECG (b) Table

(c) Explanatory Drawing (d) CT

$$p_j^{'} = P \frac{r_j}{\sum_i N_{ij} P_i}$$

(e) Line Graph (f) Formula

Figure 1: Examples of extra materials.

Basically, as shown in Table 2 and Table 4, the questions in MLEC-QA are divided into five types including:

- A1: single statement question;
- B1: similar to A1, with a group of options shared in multiple questions;
- A2: questions accompanied by a clinical scenario;
- A3: similar to A2, with information shared among multiple independent questions;
- A4: similar to A3, with information shared among multiple questions, new information can be gradually added.

We further classify these questions into Knowledge Questions (KQ) and Case Questions (CQ), where KQ (A1+B1) focus on the definition and comprehension of biomedical knowledge, while CQ (A2+A3/A4) require analysis and practical application for real-world medical scenarios. Both types of questions require multiple reasoning ability to answer.

Subset	Knowledge Questions		Case Questions		Total
	A1 (Extra)	**B1 (Extra)**	**A2 (Extra)**	**A3/A4 (Extra)**	
Cli	16,996 (19)	5,327 (21)	6,823 (7)	4,495 (26)	33,641
Sto	14,796 (269)	5,084 (84)	3,528 (325)	3,041 (311)	26,449
PH	10,413 (14)	3,949 (2)	2,469 (25)	1,693 (55)	18,524
TCM	15,235 (10)	8,045 (14)	6,044 (48)	1,541 (9)	30,865
CWM	14,051 (9)	7,336 (22)	5,370 (16)	0	26,757
Total	71,491	29,741	24,234	10,770	136,236

Table 4: Statistics of question types in MLEC-QA, where "Extra" indicates number of questions with extra materials. Only A1, A2, and B1 are used in the examination of Chinese Western Medicine.

For the Train/Dev/Test split, randomly splitting may cause data imbalance because the number of the five question types are various from each other (e.g., A1 is far more than others). To ensure that the subsets have the same distribution of the question types, we split the data based on the question types, with 80% training, 10% development, and 10% test. The overall statistics of the MLEC-QA dataset are summarized in Table 5. We can see that the length of the questions and the vocabulary size in Clinic are larger than the rest of the subsets, explaining that clinical medicine may involve more medical subjects than other specialties.

Metric	Cli	Sto	PH	TCM	CWM
# of options per question	5	5	5	5	5
Avg./Max. question len	46.51 / 332	37.12 / 341	36.76 / 352	32.24 / 340	30.63 / 280
Avg./Max. option len	9.07 / 100	9.71 / 101	9.72 / 125	7.25 / 200	7.77 / 130
Vocabulary size	46,175	41,178	35,790	38,904	38,187
# of extra materials	73	989	96	81	47
# of questions					
Train	26,913	21,159	14,818	24,692	21,406
Dev	3,365	2,645	1,852	3,086	2,676
Test	3,363	2,645	1,854	3,087	2,675
Total	33,641	26,449	18,524	30,865	26,757

Table 5: The overall statistics of MLEC-QA. Question/option length is calculated in characters. Vocabulary size is measured by Pkuseg (Luo et al., 2019) in words.

3.2 Reasoning Types of the Questions

Since the annual examination papers are designed by a team of healthcare experts who try to follow the similar reasoning types distribution. To better understand our dataset, we manually inspected 10 sets of examination papers (2 sets for each subfield), and summarize the most frequent reasoning types of the questions from MLEC-QA and previous works (Lai et al., 2017; Zhong et al., 2020). The examples are shown in Table 6. Notably, the "Evidence" is well-organized by us to show how models need to handle these reasoning issues to achieve promising performance in MLEC-QA. The definition of reasoning types of the questions are as follows:

Lexical Matching This type of question is common and the simplest. The retrieved documents are highly matched with the question, the correct answer exactly matches a span in the document. As shown in the example, the model only needs to check which option is matched with.

Multi-Sentence Reading Unlike lexical matching, where questions and correct answers can be found within a single sentence, multi-sentence reading requires models reading multiple sentences to gather enough information to generate answers.

Concept Summary The correct options for this type of question do not appear directly in the documents. It requires the model to understand and summarize the question relevant concepts after reading the documents. As shown in the example, the model needs to understand and summarize the relevant mechanism of "Thermoregulation", and infer that when an obstacle arises in thermoregulation, the body temperature will not be able to maintain a relatively constant level, that is, it will rise with the increase of ambient temperature.

Numerical Calculation This type of question involves logical reasoning and arithmetic operations related to mathematics. As shown in the example, the model first needs to judge the approximate age of month according to the height of the infant, and then reverse calculate the age of months according to the height formula of infants 7~12 months old to obtain the age in months: (68 - 65) / 1.5 + 6 = 8.

Multi-Hop Reasoning This type of question requires several steps of logical reasoning over multiple documents to answer. As shown in the example, the patient's hemoglobin (HB) value is low, indicating that the patient has anemia, and the supply of iron should be increased in their diet. The model needs to compare the iron content of each option: the iron content of C, D and E is low and that of A, B is high, but B is not easily absorbed, so the best answer is A.

Reasoning Type	Example (∗ represents the correct answer)
Lexical Matching	**The main hallmark of peritonitis is:** A. Significant abdominal distension B. Abdominal mobility dullness C. Bowel sounds were reduced or absent D. Severe abdominal cramping ∗E. Peritoneal irritation signs Evidence: The hallmark signs of peritonitis are peritoneal irritation signs, i.e., tenderness, muscle tension, and rebound tenderness.
Multi-Sentence Reading	**Which is wrong in the following narrative relating to the appendix:** A. The appendiceal artery is the terminal artery B. Appendiceal tissues contain abundant lymphoid follicles C. Periumbilical pain at appendicitis onset visceral pain ∗D. Resection of the appendix in adults will impair the body's immune function E. There are argyrophilic cells in the deep part of the appendiceal mucosa, which are associated with carcinoid tumorigenesis Evidence: (1) The appendiceal artery is a branch of the ileocolic artery and is a terminal artery without collaterals; (2) The appendix is a lymphoid organ[...]Therefore, resection of the adult appendix does not compromise the body's immune function; (3) The nerves of the appendix are supplied by sympathetic fibers[...]belonging to visceral pain; (4) Argyrophilic cells are found in the appendiceal mucosa and are the histological basis for the development of appendiceal carcinoids.
Concept Summary	**The main hallmark of thermoregulatory disorders in hyperthermic environments is:** A. Developed syncope B. Developed shock C. Dry heat of skin ∗D. Increased body temperature E. Decreased body temperature Evidence: The purpose of thermoregulation is to maintain body temperature in the normal range. In hyperthermic environments, the thermoregulatory center is dysfunctional and cannot maintain the body's balance of heat production and heat dissipation, so the body temperature is increased by the influence of ambient temperature.
Numerical Calculation	**A normal infant, weighing 7.5kg and measuring 68cm in length. Bregma 1.0cm, head circumference 44cm. Teething 4. Can sit alone and can pick up pellets with a hallux and forefinger. The most likely age of the infant is:** ∗A. 8 months B. 24 months C. 18 months D. 12 months E. 5 months Evidence: A normal infant measured 65cm at 6 months and 75cm at 1 year of age. The infant's 7 to 12 month length is calculated as: length = 65 + (months of age - 6) x 1.5.
Multi-Hop Reasoning	**6-month-old female infant, artificial feeding mainly, physical examination revealed a low hemoglobin (HB) value, the dietary supplement that should be mainly added is:** ∗A. Liver paste B. Egg yolk paste C. Tomato paste D. Rice paste E. Apple puree Evidence: (1) Low HB value indicates anemia tendency. Iron deficiency anemia is the most important and common type of anemia in China. (2) Iron supply should be increased in diet. (3) Liver paste is rich in iron. (4) The iron content of egg yolk paste is lower than that of liver paste, and it is not easy to be absorbed. (5) The iron content of tomato paste, rice paste and apple puree is lower than that of liver paste.

Table 6: Examples of reasoning types of the questions in MLEC-QA. The Chinese version is in Appendix E.

4 Methods

Notation We represent MLEC-QA task as: (D, Q, O, A), where Q_i represents the i^{th} **Q**uestion, D_i represents the collection of retrieved question relevant **D**ocuments, $O_i = \{O_{iA}, O_{iB}, O_{iC}, O_{iD}, O_{iE}\}$ are the candidate **O**ptions, A_i represents **A**nswer, and we use A_i' to denote the **P**redicted **A**nswer.

4.1 Document Retriever

Both examination counseling books and Wikipedia have been used as the source of supporting materials in previous research (Zhong et al., 2020; Jin et al., 2020; Vilares and Gómez-Rodríguez, 2019). However, because examination counseling books are designed to help examinees pass the examination, knowledge is highly simplified and summarized; even the easily confused knowledge points are compared. Using examination counseling books as information sources may make the retriever-reader more likely to exploit shallow text matching, and complex reasoning is seldom involved.

Therefore, to help better understand the improvement coming from future models, we adopt Chinese Wikipedia dumps as our information sources, which contain a wealth of information (over 1 million articles) of real-world facts. Building upon the whole Chinese Wikipedia data, we use a distributed search and analytics engine, ElasticSearch, as the document store and document retriever, which supports very fast full-text searches. The similarity scoring function used in Elasticsearch is the BM25 algorithm (Robertson and Zaragoza, 2009), which measures the relevance of documents to a given search query. As defined in Appendix C, the larger this BM25 score, the stronger the relevance between document and query.

Specifically, for each question Q_i and each candidate option O_{ij} where $j \in \{A, B, C, D, E\}$, we define $Q_iO_{ij} = Q_i + O_{ij}$ as a search query to Elasticsearch and is repeated for all options. The document with the highest BM25 score returned by each query is selected as supporting materials for the next stage machine reading comprehension task.

4.2 Control Methods

In general, each option should have the same correct rate for multi-choice questions, but in fact, the order in which the correct options appear is not completely random, and the more the number of options, the lower the degree of randomization (Poundstone, 2014). Given the complex nature of multi-choice tasks, we employ three control methods to ensure a fair comparison among various open-domain QA models.

Random $A' = Random(O)$. For each question, an option is randomly chosen as the answer from five candidate options. We perform this experiment five times and average the results as the baseline of the Random method.

Constant $A' = Constant_j(O)$, where $j \in \{A, B, C, D, E\}$. For each question, the j^{th} option is always chosen as the answer to obtain the accuracy distribution of five candidate options.

Mixed $A' = Mixed(O)$. Incorporating the previous experiences of NMLEC and multi-choice task work (Vilares and Gómez-Rodríguez, 2019), the Mixed method simulates how humans solving uncertain questions, and consists of the following three strategies: (1) the correct rate of choosing "All of the options above is correct/incorrect" is much higher than the other options. (2) Supposing the length of options is roughly equal, only one option is obviously longer with more detailed and specific descriptions, or is obviously shorter than the other options, then choose this option. (3) The correct option tends to appear in the middle of candidate options. The three strategies are applied in turn. If any strategy matches, then the option that matches the strategy is chosen as the answer.

4.3 Fine-Tuning Pre-Trained Language Models

We apply an unified framework UER-py (Zhao et al., 2019) to fine-tuning pre-trained language models on the machine reading comprehension task as our reader. We consider the following five pre-trained language models: Chinese BERT-Base (denoted as BERT-Base) and Multilingual Uncased BERT-Base (denoted as BERT-Base-Multilingual) (Devlin et al., 2019), Chinese BERT-Base with whole word masking and pre-trained over larger corpora (denoted as BERT-wwm-ext) (Cui et al., 2019), and the robustly optimized BERTs: Chinese RoBERTa-wwm-ext and Chinese RoBERTa-wwm-ext-large (Cui et al., 2019).

Specifically, given the i^{th} question Q_i, retrieved question relevant documents D_i, and a candidate option O_{ij}, where $j \in \{A, B, C, D, E\}$. The input sequence for the framework is constructed by concatenating [CLS], tokens in D_i, [SEP], tokens in Q_i, [SEP], tokens in an option O_{ij}, and [SEP], where [CLS] is the classifier token, and [SEP] is the sentence separator in pre-trained language models. We pass each of the five options in turn, and the model outputs the hidden state representation $S_{ij} \in R^{1 \times H}$ of the input sequence, then performs the classification and output an unnormalized log probability $P_{ij} \in R$ of each option O_{ij} being correct by $P_{ij} = S_{ij}W^T$, where $W \in R^{1\times H}$ is the weight matrix. Finally, we pass the unnormalized log probabilities of each option through a softmax layer and obtain the option with the highest probability as the predicted answer A_i'.

5 Experiments

5.1 Experimental Settings

We conduct detailed experiments and analyses to investigate the performance of control methods and open-domain QA methods on MLEC-QA. As shown in Figure 2, we implement a two-stage retriever-reader framework: (1) a retriever first retrieves question relevant documents from Chinese Wikipedia using ElasticSearch, (2) and then a reader employs machine reading comprehension models to generate answers in given documents retrieved by the retriever. For the reader, all machine reading comprehension models are trained with 12 epochs, an initial learning rate of 2e-6, a maximum sequence length of 512, a batch size of 5. The parameters are selected based on the best performance on the development set, and we keep the default values for the other hyper-parameters (Devlin et al., 2019). We use accuracy as the metric to evaluate different methods, and provide baseline results, as well as human pass mark (60%) instead of human performance due to the wide variations exist in human performance, from almost full marks to cannot even pass the exam.

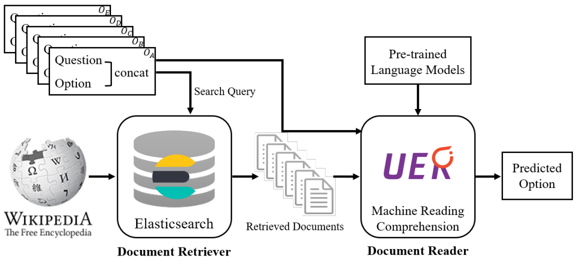


Figure 2: Overview of the two-stage retriever-reader framework on MLEC-QA.

5.2 Retrieval Performance

The main drawbacks of the Chinese Wikipedia database in biomedicine are that it is not comprehensive and thorough, that is, it may not provide complete coverage of all subjects. To evaluate whether retrieved documents can cover enough evidence to answer questions, we sampled 5% (681) questions from the development sets of five categories using stratified random sampling, and manually annotate each question by five medical experts with 3 labels: (1) Exactly Match (EM): the retrieved documents exactly match the question. (2) Partial Match (PM): the retrieved documents partially match the question, can be confused with the correct options or are incomplete. (3) Mismatch (MM): the retrieved documents do not match the question at all. Table 7 lists the performance of the retrieval strategy as well as the results of the annotation for KQ and CQ questions on five subsets.

Subset	EM (KQ / CQ)		PM (KQ / CQ)		MM (KQ / CQ)	
Cli	15.63	(18.75 / 12.5)	75	(68.75 / 81.25)	9.38	(12.5 / 6.25)
Sto	20.83	(16.67 / 25)	54.17	(50 / 58.33)	25	(33.33 / 16.67)
PH	6.25	(0 / 12.5)	43.75	(50 / 37.5)	50	(50 / 50)
TCM	10.71	(14.29 / 7.14)	50	(42.86 / 57.14)	39.29	(42.86 / 35.71)
CWM	20.83	(8.33 / 33.33)	54.17	(58.33 / 50)	25	(33.33 / 16.67)

Table 7: Matching rate (%) of retrieved documents that exactly match, partial match or mismatch with the questions in the MLEC-QA dataset.

From the table, we make the following observations. First, most retrieved documents indicate PM with the questions, while the matching rates of EM and MM achieve maximums of 20.83% (CWM) and 50% (PH), respectively. Second, the matching rate of CQ is higher than KQ in most subsets as CQ are usually related to simpler concepts, and use more words to describe questions, which leads to easier retrieval. By contrast, KQ usually involve more complex concepts that may not be included in the Chinese Wikipedia database. Therefore, the mismatching rate of KQ is significantly higher than that of CQ. Third, among different subsets, the performance in the subset Cli achieves the best as clinical medicine is more "general" to retrieve compare with other specialties. Whereas the performance in the subset PH achieves the worst as the Public Health is usually related to "confusing concepts", which leads to poor retrieval performance.

5.3 Baseline Results

Tables 8 and Figure 3 show the performance of baselines as well as the performance on KQ and CQ questions. As we can see, among control methods, the correct option has a slight tendency to appear in the middle (C and D) of candidate options, but the margins are small. The performance of the Mixed method is slightly better than a random guess, which indicates that the flexible use of

Method		Cli		Sto		PH		TCM		CWM	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Random		19.73	19.61	19.41	19.43	19.70	20.13	20.17	20.11	19.69	20.16
Constant	Option [A]	17.12	18.05	16.82	17.01	16.09	17.80	19.05	20.60	18.68	20.56
	Option [B]	20.30	20.93	20.95	20.53	21.49	21.36	**22.52**	20.70	20.70	21.20
	Option [C]	21.66	21.23	**22.31**	20.95	21.65	20.23	22.29	**22.35**	**23.39**	**21.27**
	Option [D]	**22.53**	**21.38**	20.64	**22.76**	**22.68**	**21.84**	19.80	20.67	21.00	19.51
	Option [E]	18.40	18.41	19.28	18.75	18.09	18.77	16.33	15.68	16.22	17.46
Mixed		24.40	24.68	24.46	24.42	23.97	23.68	22.07	22.38	23.62	23.25
BERT-Base		47.26	48.30	40.53	40.08	38.99	37.40	48.51	49.14	44.32	45.14
BERT-wwm-ext		50.27	50.89	43.26	42.05	**41.75**	**40.04**	**54.57**	**54.94**	49.89	50.04
BERT-Base-Multilingual		46.61	47.68	39.85	38.76	36.61	36.70	45.50	46.61	42.26	42.86
RoBERTa-wwm-ext		49.94	51.97	41.14	40.88	38.40	38.91	50.45	49.82	47.38	46.00
RoBERTa-wwm-ext-large		**53.25**	**53.22**	**44.92**	**43.75**	39.10	38.75	47.99	48.65	**50.49**	**50.11**
Human Pass Mark		60									

Table 8: Performance of baselines in accuracy (%) on the MLEC-QA dataset.

guessing skills may add wings to the tiger as humans can exclude some certain wrong options, but if the cart before the horse is reversed, it is impossible to pass the exam only through opportunistic guessing. RoBERTa-wwm-ext-large and BERT-wwm-ext perform better than other models on five subsets. However, even the best-performing model can only achieve accuracies between 40% to 55% on five subsets, so there is still a gap to pass the exams.

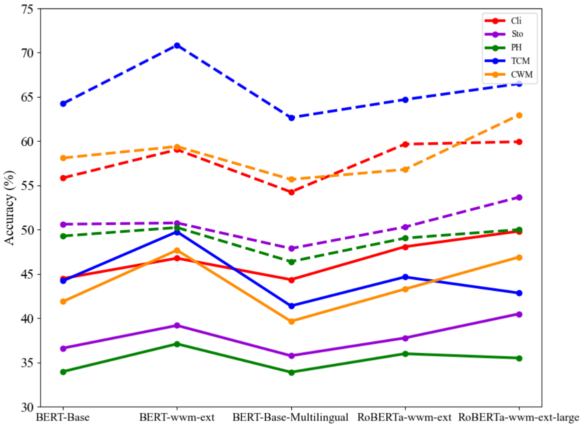


Figure 3: Performance in accuracy (%) on KQ (solid lines) and CQ (dashed lines) questions.

Comparing the performance between KQ and CQ questions, most models achieve better performance on CQ, which is positively correlated with CQ's better retrieval performance. Among different subsets, the subset TCM is the easiest (54.95%) one to answer across the board, while the subset PH is the hardest (40.04%), which does not totally correspond to their retrieval performance as shown in Table 7. The possible reason is that the diagnosis and treatment of diseases in traditional Chinese medicine are characterized by "Homotherapy for Heteropathy", that is, treating different diseases with the same method, which may result in some patterns or mechanisms that can be used by the models to reach such results.

5.4 Comparative Analysis

Given that we use the Chinese Wikipedia database as our information sources and apply a two-stage retriever-reader framework, the reason for such poor baseline performance could come from both our information sources and the retriever-reader framework.

Information Sources Both books and Wikipedia have been used as the information sources in previous research. One of our subsets, Clinic, has been studied by MEDQA (Jin et al., 2020) as a subset (MCMLE) for cross-lingual research. MEDQA uses 33 medical textbooks as their information sources and the evaluation result shows that their collected text materials can provide enough information to answer all the questions in MCMLE. We compare the best model (RoBERTa-wwm-ext-large) performance on both datasets as shown in Table 9. Notably, questions in MCMLE have four candidate options due to one of the wrong options being deleted. Therefore, the random accuracy on MCMLE is higher than ours.

From the results we can see that even with 100% covered materials, the best model can only

Dataset	Accuracy (%)	
	Dev	Test
MEDQA (MCMLE)	69.30	70.10
MLEC-QA (Clinic)	53.25	53.22

Table 9: Comparison of best model (RoBERTa-wwm-ext-large) performance on MEDQA and our MLEC-QA dataset.

achieve 16.88% higher accuracy on the test set than ours, which indicates that using Wikipedia as information sources is not that terrible compared with medical books, and the main reason for baseline performance may come from machine reading comprehension models that lack sophisticated reasoning ability.

Retriever-Reader We also perform an experiment that sampled 5% (92) questions from the development set of Public Health, and manually annotate each question by a medical expert to determine whether that can exactly or partially match with the top K retrieved documents, as shown in Table 10. Notably, the actual number of retrieved documents is $5 \times K$ as we define $Q_iO_{ij} = Q_i + O_{ij}$ as a search query and is repeated for all options. From the results, we can see that more documents even bring more noise instead, as the best match documents have already been fetched in the top 1 documents. It indicates that the poor performance of machine reading comprehension models is coming from the insufficiency of reasoning ability rather than the number of retrieved documents.

Top K	1	2	3	4	5
Match	52.08	42.19	36.25	32.29	29.46

Table 10: Matching rate (%) of Top K retrieved documents that exactly or partial match with the questions in the Public Health subset.

6 Conclusion

We present the largest-scale Chinese multi-choice BQA dataset, MLEC-QA, which contains five biomedical subfields with extra materials (images or tables) annotated by human experts: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Chinese Western Medicine. Such questions correspond to examinations (NMLEC) to access the qualifications of medical practitioners in the Chinese healthcare system, and require specialized domain knowledge and multiple reasoning abilities to be answered. We implement eight representative control methods and open-domain QA methods by a two-stage retriever-reader framework as baselines. The experimental results demonstrate that even the current best approaches cannot achieve good performance on MLEC-QA. We hope MLEC-QA can benefit researchers on improving the open-domain QA models, and also make advances for BQA systems.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This work is supported by the National Natural Science Foundation of China (NSFC No. 61972187).

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 LiveQA. In *Proceedings of the Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019a. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019b. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019*, volume 264 of *Studies in Health Technology and Informatics*, pages 25–29. IOS Press.

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont D. Antieau, Andrew S. Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *J. Biomed. Informatics*, 44(2):277–288.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Yiming Cui, W. Che, T. Liu, B. Qin, Ziqing Yang, S. Wang, and G. Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *ArXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to Chinese medical question answering: A study and a dataset. *BMC Medical Informatics and Decision Making*, 19(2):52.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020a. PathVQA: 30000+ questions for medical visual question answering. *CoRR*, abs/2003.10286.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020b. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv:2009.13081 [cs]*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2021. Biomedical Question Answering: A Comprehensive Review. *arXiv:2102.05281 [cs]*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438, Online. Association for Computational Linguistics.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019. PKUSEG: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 73–80, Sapporo, Japan. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

William Poundstone. 2014. *Rock Breaks Scissors*. Little Brown &amp; Co.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Rafael M. Terol, Patricio Martínez-Barco, and Manuel Palomar. 2007. A knowledge based method for the medical question answering problem. *Comput. Biol. Medicine*, 37(10):1511–1521.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015.

An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Weiming Wang, Dawei Hu, Min Feng, and Liu Wenyin. 2007. Automatic clinical question answering based on UMLS relations. In *Third International Conference on Semantics, Knowledge and Grid, Xian, Shan Xi, China, October 29-31, 2007*, pages 495–498. IEEE Computer Society.

Hong Yu, Minsuk Lee, David R. Kaufman, John W. Ely, Jerome A. Osheroff, George Hripcsak, and James J. Cimino. 2007. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J. Biomed. Informatics*, 40(3):236–251.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. 2018a. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Applied Sciences*, 7(8):767.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018b. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5706–5713. AAAI Press.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. UER: An open-source toolkit for pre-training models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 241–246, Hong Kong, China. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A Legal-Domain Question Answering Dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.*, pages 9701–9708.

Pierre Zweigenbaum. 2003. Question answering in biomedicine. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*.

A Source of Data Collection

For five subsets in MLEC-QA, we collect 2006 to 2020 Sprint Paper for the National Medical Licensing Examination - Tianjin Science and Technology Press in PDF format, and then converted them into digital format via Optical Character Recognition (OCR). We manually checked and corrected the OCR results with confidence less than 0.99 to ensure the quality of our dataset. We also scraped practice exercises from offcn (http://www.offcn.com/yixue/yszg/), which are freely accessible online for public usage.

B Data Structure

The data structure below describe the JSON file representation in MLEC-QA.

```
{
   "qid":The question ID,
   "qtype":["A1 型题", "B1 型题", "A2 型题", "A3/A4 型题"],
   "qtext":Description of the question,
   "qimage":Image or table path (if any),
   "options":{
      "A":Description of the option A,
      "B":Description of the option B,
      "C":Description of the option C,
      "D":Description of the option D,
      "E":Description of the option E
},
   "answer":["A", "B", "C", "D", "E"]
}
```

C BM25 Score Function

The BM25 algorithm is defined as:

$$s(D,Q)=\sum_{i=1}^{n} IDF\left(q_i\right)\cdot\frac{f\left(q_i,D\right)\cdot\left(k_1+1\right)}{f\left(q_i,D\right)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{avgdl}\right)}, \tag{1}$$

where q_i is the i^{th} query term of a query Q, $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and $avgdl$ is the average document length in the text collection from which documents are drawn. b determines the effects of the length of the document on the average length. $k1$ is a variable which helps determine term frequency saturation characteristics. By default, $b, k1$ has a value of 0.75, 1.2 in Elasticsearch, respectively. $IDF(q_i)$ is the **I**nverse **D**ocument **F**requency (**IDF**) weight of the query term q_i. It is usually computed as:

$$IDF(q_i)=\ln\left(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}+1\right), \tag{2}$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

D Chinese version of examples of sub-fields and question types

Question Type	Example (* represents the correct answer)
A1	(公共卫生) 在比较两样本等级资料结果有无差别时，宜用（　） A. t 检验　B. x^2 检验　C. μ 检验　D. F 检验　∗E. 秩和检验
B1	(中西医结合) A. 心、脾　B. 肝、肺　C. 脾、肾　D. 心、肾　E. 肝、肾 1. 乙癸同源的乙癸所指的脏是（E） 2. 水火既济的水火所指的脏是（D）
A2	(临床) 初孕妇，29 岁。妊娠 37 周，今晨喷射性呕吐 1 次，1 小时前突然抽搐并随即昏迷入院。查体:BP 180/120mmHg，尿蛋白 (+++)。该患者最可能的诊断是（　） ∗A. 子痫　B. 脑出血　C. 癔症　D. 癫痫　E. 脑血栓形成
A3	(中医) 女性，28 岁。近一月以来，口腔溃疡反复发作，心烦，夜晚难以入睡，大便干，1 ~ 2 日一行，口干不喜饮，小便黄，舌质红，苔腻，脉数。 1. 治疗此证宜首选的药物是（　） A. 芒硝　B. 牛蒡子　∗C. 竹叶　D. 黄芩　E. 石膏 2. 治疗此证可以适当配伍的药物是（　） A. 白芷　B. 桂枝　C. 干姜　∗D. 大黄　E. 秦皮 3. 患者月经期间，下列何药宜慎用（　） A. 淡豆豉　∗B. 大黄　C. 黄连　D. 淡竹叶　E. 芦根
A4	(口腔) 患者 50 岁，因鼻咽癌外院放疗结束后 2 年，来门诊就诊。检查：全口多个牙齿牙面不同程度龋，部分患牙已成残冠，残根，口内唾液较少，牙面及龈沟软垢多。 1. 该患者的诊断为（　） A. 急性龋　∗B. 猛性龋　C. 慢性龋　D. 继发龋　E. 平滑面龋 2. 治疗设计有以下几项，除了（　） A. 设计治疗全口龋齿　B. 牙髓病变的牙齿牙髓治疗 ∗C. 充填金属材料　D. 再矿化辅助治疗　E. 定期复查

Table 11: Chinese version of Table 2.

E Chinese version of examples of reasoning types

Reasoning Type	Example (∗ represents the correct answer)
Lexical Matching	腹膜炎的主要标志是（　） A. 明显的腹胀　B. 腹部移动性浊音　C. 肠鸣音消失或减弱 D. 剧烈的腹部绞痛　∗E. 腹膜刺激征 **Evidence:** 腹膜炎的标志性体征是腹膜刺激征，即压痛、肌紧张和反跳痛。
Multi-Sentence Reading	下列与阑尾相关的叙述，错误的是（　） A. 阑尾动脉是终末动脉　B. 阑尾组织中含有丰富的淋巴滤泡　C. 阑尾炎发病时的脐周痛属内脏性疼痛　∗D. 成人切除阑尾将损害机体的免疫功能　E. 阑尾黏膜深部有嗜银细胞，与类癌发生有关 **Evidence:** (1) 阑尾动脉是回结肠动脉的分支，是无侧支的终末动脉；(2) 阑尾是一个淋巴器官 [...] 故切除成人阑尾无损于机体的免疫功能；(3) 阑尾的神经由交感神经纤维 [...] 属内脏性疼痛；(4) 阑尾黏膜部有嗜银细胞，是发生阑尾类癌的组织学基础。
Concept Summary	高温环境中体温调节障碍的主要标志是（　） A. 出现晕厥　B. 出现休克　C. 皮肤干热　∗D. 体温升高　E. 体温降低 **Evidence:** 体温调节的目的在于维持体温在正常范围。高温环境中，体温调节中枢功能障碍，不能维持机体的产热和散热平衡，体温受环境温度影响而升高。
Numerical Calculation	正常婴儿，体重 7.5kg，身长 68cm。前囟 1.0cm，头围 44cm。出牙 4 个。能独坐并能以拇、食指拿取小球。该儿最可能的月龄是（　） ∗A. 8 个月　B. 24 个月　C. 18 个月　D. 12 个月　E. 5 个月 **Evidence:** 正常婴儿 6 个月时身长 65cm，1 岁时身长 75cm。婴儿 7~12 个月身长计算公式为：身长 = 65 + (月龄 - 6) x 1.5。
Multi-Hop Reasoning	6 个月女婴，人工喂养为主，体格检查发现血红蛋白（Hb）值偏低，应添加的辅食主要是（　） ∗A. 肝泥　B. 蛋黄泥　C. 西红柿泥　D. 米粉　E. 苹果泥 **Evidence:** (1)Hb 值偏低表明有贫血倾向，缺铁性贫血是我国最主要、最常见的贫血类型。(2) 饮食应增加铁的供给。(3) 肝泥含有丰富的铁。(4) 蛋黄泥铁含量比肝泥铁含量低，且不易吸收。(5) 西红柿泥、米粉和苹果泥的铁含量均不如肝泥高。

Table 12: Chinese version of Table 6.

F Data Statement

F.1 CURATION RATIONALE

In order to benefit researchers on improving the open-domain QA models, and also make advances for **B**iomedical **Q**uestion **A**nswering (BQA) systems, we present MLEC-QA, the largest-scale Chinese multi-choice BQA dataset to date.

F.2 LANGUAGE VARIETY

The data is represented in simplified Chinese (zh-Hans-CN), and collect from the 2006 to 2020 NMLEC, as well as practice exercises from the Internet.

F.3 SPEAKER DEMOGRAPHIC

Since the data is designed by a team of anonymous human healthcare experts, we are not able to directly reach them for inclusion in this dataset and thus could not be asked for demographic information. It is expected that most of the speakers come from China with professionals working in the area of biomedicine, and speak Chinese as a native language. No direct information is available about age and gender distribution.

F.4 ANNOTATOR DEMOGRAPHIC

The experiments involve annotations from 5 medical experts with at least have a master's degree and have passed the NMLEC. They ranged in age from 28–45 years, included 3 men and 2 women, all come from China and speak Chinese as a native language.

F.5 SPEECH SITUATION

All questions in MLEC-QA are collected from the **N**ational **M**edical **L**icensing **E**xamination in **C**hina (**NMLEC**), which are carefully designed by human experts to evaluate professional knowledge and skills for those who want to be medical practitioners in China.

F.6 TEXT CHARACTERISTICS

The topics include in MLEC-QA are in 5 biomedical sub-fields: Clinic, Stomatology, Public Health, Traditional Chinese Medicine, and Traditional Chinese Medicine Combined with Western Medicine.

F.7 RECORDING QUALITY

N/A.

F.8 OTHER

N/A.

F.9 PROVENANCE APPENDIX

N/A.