SleepQA: A Health Coaching Dataset on Sleep for Extractive Question Answering

Iva Bojic [1] IVA.BOJIC@NTU.EDU.SG
Qi Chwen Ong [1] QICHWEN.ONG@NTU.EDU.SG
Megh Thakkar [1] MEGH.1211@GMAIL.COM
Esha Kamran [2] ESHA.KAMRAN20@IMPERIAL.AC.UK
Irving Yu Le Shua [1] SHUA0002@E.NTU.EDU.SG
Jaime Rei Ern Pang [1] PA0001ME@E.NTU.EDU.SG
Jessica Chen [2] JESSICA.CHEN20@IMPERIAL.AC.UK
Vaaruni Nayak [2] VAARUNI.NAYAK20@IMPERIAL.AC.UK
Shafiq Joty [1,3] SRJOTY@NTU.EDU.SG
Josip Car [1,2] JOSIP.CAR@NTU.EDU.SG
[1] *Nanyang Technological University, Singapore*
[2] *Imperial College London, United Kingdom*
[3] *Salesforce Research, USA*

Abstract

Question Answering (QA) systems can support health coaches in facilitating clients' lifestyle behavior changes (e.g., in adopting healthy sleep habits). In this paper, we design a domain-specific QA pipeline for sleep coaching. To this end, we release SleepQA, a dataset created from 7,005 passages comprising 4,250 training examples with single annotations and 750 examples with 5-way annotations[1]. We fine-tuned different domain-specific BERT models on our dataset and perform extensive automatic and human evaluation of the resulting end-to-end QA pipeline. Comparisons of our pipeline with baseline show improvements in domain-specific natural language processing on real-world questions. We hope that this dataset will lead to wider research interest in this important health domain.

Keywords: Factual Question Answering, Dense Passage Retrieval, Evidence-based Knowledge, Domain-specific Natural Language Processing

1. Introduction

Chronic diseases account for more than two-thirds of all deaths globally[2], causing a huge burden on the healthcare system. Healthcare focuses primarily on the detection and management of chronic diseases. In recent years, there has been greater emphasis on primary prevention which aims at preventing the occurrence of diseases by controlling the causes and risk factors, such as altering risky behaviors (Kisling and Das, 2021). Despite the presence of robust evidence showing association between sleep duration and the risk of chronic diseases (Lu et al., 2020), other lifestyle factors such as smoking and physical activity are more commonly recognized by the public as modifiable risk factors for chronic diseases (Von Ruesten et al., 2012).

1. Code and dataset are available at: https://github.com/IvaBojic/SleepQA

2. https://who.int/news-room/fact-sheets/detail/noncommunicable-diseases

In recent years, health coaching has gained recognition for its effectiveness in the prevention of chronic diseases (Yang et al., 2020) by supporting lifestyle behavior change. Question Answering (QA) systems can empower health coaches with evidence-based knowledge and support them in facilitating clients' lifestyle behavior changes, so as to improve overall sleep quality and quantity. The use of QA systems to develop a novel health coaching model could potentially revolutionize health coaching industry and current practice of preventive medicine.

Typically, extractive QA pipeline work in two phases, where a passage *reader* follows a passage *retriever* (Chen et al., 2017). For a given question, the retriever first retrieves a set of relevant passages from a knowledge base (i.e., a text corpus), and then the reader selects an answer (e.g., a text span) from one of the passages (Zhu et al., 2021). QA systems can be classified into two groups: *Open-domain* QA systems that answer a question based on large-scale unstructured documents (Zhu et al., 2021), and *domain-specific* systems that tend to cover questions restricted to a single field of study, such as medicine or architecture (Kia et al., 2022).

Open-domain QA systems are typically trained on large-scale general domain corpus like Wikipedia (Chen et al., 2017). Despite its broad coverage, information is expressed through various text patterns and thus models trained on such corpus generally cannot generalize well beyond the training domains (Sun et al., 2018). It is however difficult to generate sufficiently large domain-specific datasets as the creation process is costly and time-consuming due to the need for domain experts to interpret field-specific jargons, especially for niche fields like the biomedical sector. As it is not feasible to train a large-scale base QA model on a large, close-domain dataset, various transfer learning techniques have been used to adapt QA systems trained on large open-domain dataset to a specific domain (Wiese et al., 2017).

Our key contributions are summarized as:

- we present SleepQA, a dataset composed of 5,000 human-generated passage-question-answer triples (i.e., labels) annotated from 7,005 passages;
- we fine-tune different domain-specific BERTs, build the best performing QA pipeline and compare it with Lucene BM25 (Lin et al., 2021) + *BERT SQuAD2* (Rajpurkar et al., 2018) pipeline. We performed both automatic and human evaluation of the models and pipelines, as well as error analysis to better understand the drawbacks of the models and to suggest future directions.

Intrinsic evaluation showed that Lucene BM25 strongly outperformed all fine-tuned retrieval models on corpus-specific questions (i.e., test labels). However, extrinsic evaluation on real-world questions showed that our QA pipeline outperformed Lucene BM25 + *BERT SQuAD2* pipeline, showing promise in using domain-specific fine-tuning settings for QA pipelines.

2. Related Work

Pretraining neural language models on abundant unlabeled text is among the successful strategies used for transfer learning in Natural Language Processing (NLP). Previous work on domain-specific tasks such as biomedical QA mostly adopted a mixed-domain pretraining approach, where a general-domain pretrained model is continually pretrained on biomedical text. Some of the notable examples include *BioBERT*, which is pretrained on PubMed abstracts and full texts (Lee et al., 2020), and *ClinicalBERT*, which is initialized with *BioBERT* and continually pretrained on MIMIC-III clinical text (Alsentzer et al., 2019).

As biomedicine is a domain with open-access large-scale corpora like PubMed, it has been proposed and proven that models with domain-specific pretraining from scratch like *PubMedBERT* substantially outperform models using mixed-domain pretraining, achieving new state-of-the-art results for various biomedical NLP tasks (Gu et al., 2021). While *SciBERT* is also pretrained from scratch, it is considered an exception as the corpus of scientific literature comes from both computer science and biomedical domain (Beltagy et al., 2019). Since computer science text is not within the biomedical domain, this approach still belongs to mixed-domain pretraining.

Biomedical QA can be further categorized into multiple sub-domains, namely scientific, clinical, consumer health and examination (Jin et al., 2022). Within the wider healthcare ecosystem, several QA systems have aimed to provide clinicians with succinct comprehensible answers through the sifting of large knowledge bases to output the most relevant information. Notable examples as documented by (Mutabazi et al., 2021) include AskHermes, a system that relies on a binary approach for question classification and an adjusted BM25 model for information retrieval, producing a summary output allowing easy comprehension by the clinician (Cao et al., 2011).

Artificial Intelligence (AI) in sleep medicine has demonstrated immense potential in evaluation of sleep disorders, with several successful examples leveraging machine learning to improve the clinical efficacy of automated polysomnography (PSG) sleep staging (Sun et al., 2017; Biswal et al., 2018). Improved identification of distinct subtyping of sleep disorders through machine learning analysis of large datasets could potentially reveal the role of sleep in health and disease and allow for more specific population-based interventions (Goldstein et al., 2020). Currently there is paucity of studies that employ NLP in health coaching, let alone sleep coaching. To the best of our knowledge, SleepQA is the first within the space of health coaching and preventive medicine to adopt QA system for sleep coaching and among the first to explore AI-aided health coaching model in the field of healthcare.

3. Dataset Collection

We collect our dataset in three phases: passage curation, crowdsourcing passage-question-answer triplets, and obtaining additional annotations for inter-annotator agreement evaluation. Additionally, we collect real-world questions, which we use for extrinsic evaluation of our QA system.

3.1. Passage Curation

We download more than 1,000 articles from two web pages[3] to obtain a high-quality, evidence-based, and medically reviewed sleep health information. The final text corpus consists of 527 articles after removal of irrelevant articles such as interviews and reviews of products. Duplicated sentences across different articles (usually advertisement lines) and short sentences less than five words (e.g., subtitles) are retrieved for manual screening and removed if deemed irrelevant. We remove unnecessary punctuation, superscripts (e.g., footnotes or cited references), symbols and other irrelevant information such as authors bibliography and references. Subsequently, we reorganize all passages and divide the content into passages with lengths of 100 to 150 words. Our pre-processing work produced 7,005 clean passages covering a wide range of topics related to sleep health.

3. https://www.sleepfoundation.org and https://thesleepdoctor.com

3.2. QA Collection

Crowdsourcing We use small-scale crowdsourcing efforts to create labels, which are passage-question-answer triplets. The annotation process consists of three conceptual stages: (1) passage reading, (2) question formation, and (3) answer identification. We recruit five annotators who are medical students and whose first or native language is English. A web interface is created to support the label collection process (see Appendix A).

Passage-question-answer Triplets Annotators are tasked to read and comprehend the content of the passage, following which they form a question based on their understanding and identify an answer embedded within the passage. Data collection training with pre-defined guidelines is delivered to the annotators to reduce heterogeneity and ensure good quality of collected labels. Label collection starts with comprehending the passage and ensuring it is 'clean', i.e., without unnecessary punctuation, double space etc. The web interface allows the annotators to flag uncleaned passages, after which they are manually checked and cleaned.

For each pair of labels, the question formed must start with one of the following words: "who", "what", "where", "when", "why" or "how" (i.e., factoid question). The question has to be entered into the text field and end with a question mark. A text span from the passage is selected as answer and entered into the text field. Annotators are not allowed to change the structure, grammar, vocabulary, or punctuation of the selected answer as extractive question answering is a task that finds an answer span from a passage. Unfilled text field or alteration in answer's text is disabled through the web interface. Annotators are given the option to skip the passage if they are unable to provide a question and answer. All three stages of this process were completed sequentially and independently by each annotator.

3.3. Inter-annotator Agreement

Exact Match and F1 Score We calculate inter-annotator agreement by evaluating a subset of our dataset with two metrics: Exact Match (EM) and (Macro-averaged) F1 score. EM measures the percentage of predictions that match the ground truth answers exactly. F1 score measures the average overlap between ground truth answer and predictions, where both are treated as bag of tokens.

Additional Annotation We randomly sample 150 labels from each annotator and assign them to four other annotators to collect 5-way annotation needed to calculate inter-annotator agreement. With a passage and question shown, they repeat the annotation process independently without knowing each other's input of answers. Each of them identifies answers for 600 questions formed by their counterparts, yielding 750 passages with one question and five answers. In this way, each of the questions in the set of 750 labels has a 5-way annotation. An example of label with a 5-way annotation is shown in Appendix A.

To calculate inter-annotator agreement, we used same methodology as proposed in Karpukhin et al. (2020). Namely, we treat the answers that are provided by the initial annotator as ground truth answers and keep the answers from four other annotators as human predictions. We take the maximum for both EM and F1 scores over all predictions (i.e., for four comparisons) and average them over all questions. The resulting score in this subset is **0.85** for the EM and **0.91** for F1 score. These numbers are above 0.8 ensuring reasonable quality of annotations (Artstein and Poesio, 2008).

3.4. Real-world Questions

In addition to collecting labels, we also collect 650 real-world questions related to sleep from more than 70 students who attended our health coaching course. Students are instructed to pose random sleep-related questions in natural language. They are neither presented with the text corpus nor given any specific instruction to form a certain type of questions, such as factoid questions. By collecting real-world questions, in addition to performing intrinsic evaluation of our system using test labels, we are also able to do extrinsic evaluation (Belz and Reiter, 2006).

4. Dataset Analysis

Questions and respective answers were generated manually for a total of 5,000 randomly chosen passages. This data was then rigorously analyzed. We compared the general characteristics of our dataset (e.g., average number of words in passages, questions and answers) with six datasets (Fan et al., 2019) containing extractive and short abstractive answers (see Table 1).

The datasets include CoQA (Yatskar, 2018), a multi-domain dialogue dataset with 127,000 labels, and NarrativeQA (Kočiskỳ et al., 2018), a dataset of narrative works summaries totaling 47,000 labels. Both collect responses from crowdworkers and contains written answers are largely extractive and short. TriviaQA (Joshi et al., 2017), which has 110,000 labels, contains longer multi-document web input, collected using Bing and Wikipedia whereas HotpotQA with 113,000 labels (Yang et al., 2018) is solely Wikipedia-based. As the dataset is built from trivia, most questions can be answered with a short extractive span. MS MARCO (Nguyen et al., 2016), a dataset of 183,000 crowdsourced responses to Bing queries, has sentence-long written answers together with input passages that are short.

4.1. Average Number of Words

The distribution of average number of words in each of the 7,005 clean passages is shown in Figure 1. Average length of passages in SleepQA dataset, around 120 words per passage, is compared to the one in SQuAD (2.0) dataset. As described in Section 3.1, we reorganized all passages and divided the content into passages with lengths between 100 to 150 following the methodology proposed in (Karpukhin et al., 2020).

To better understand the questions formed, we computed the distribution of question length, which can be seen in Figure 2. The length of most questions ranges from five to ten words. The minimum number of words in each question is three, which likely comprises of an interrogative word, such as “what”, a linking verb, like “is” or “are”, and a noun, which is the object of the given question. One example of a such question is: “What is sleepwalking?”.

The length of the question likely mirrors the specificity of the question. A longer question has a greater tendency to include more specific or defining domains that limit the answer. Since a majority of the questions lie between five and ten words, it can be concluded that most questions have a moderate degree of specificity. One example of a question with five words is: “What side-effects can clonazepam cause?”

The average length of SleepQA answers is longer than that of the other datasets (see Table 1). As responses from other datasets are based on summaries of topics, their answer would be shorter. However, for SleepQA, the responses are based on passages of articles, hence, the answers found within these passages would not be concise and easily expressed in a few words. Nevertheless, in future work, we will check the answers that are longer than the average length to see if we can make them more concise.

Table 1: Comparing large-scale QA datasets.

Dataset	Avg. # of words			1st question word frequency (%)							
	P	Q	A	Why	How	What	When	Where	Who	Which	OTHER
SleepQA	**120.6**	**9.9**	**10.3**	**5.9**	**16.8**	**68.4**	**4.9**	**1.0**	**2.9**	**0**	**0**
MS MARCO v2	56	6.4	13.8	1.7	16.8	35.0	2.7	3.5	3.3	1.8	35.3
TriviaQA	2895	14	2.0	0.2	3.9	32.6	2.0	2.1	16.8	41.8	0.6
CoQA	271	5.5	2.7	2	5	27	2	5	15	1	43
HotpotQA	917	17.8	2.2	0.1	2.6	37.2	2.8	2.2	13.8	28.5	12.8
NarrativeQA	656	9.8	4.7	9.8	10.7	38.0	1.7	7.5	23.4	2.2	6.8
SQuAD (2.0)	116.6	9.9	3.2	1.4	8.9	45.3	6.0	3.6	9.6	4.4	17.6

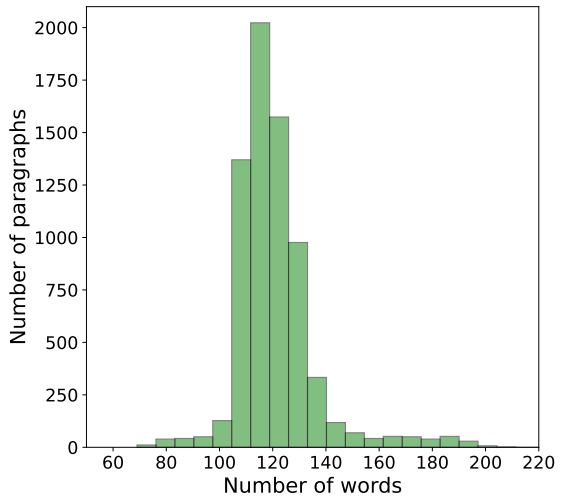

Figure 1: Distribution of average number of words in passages.

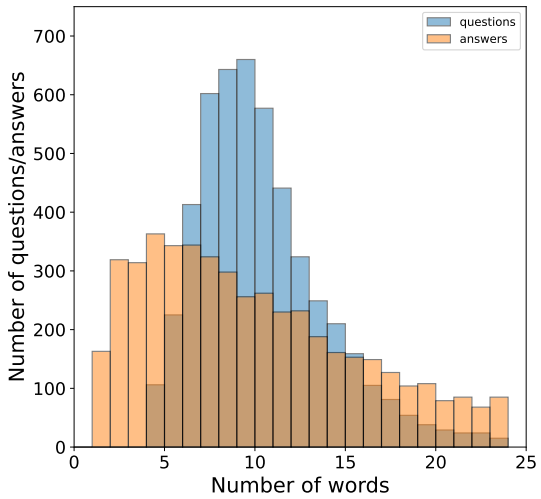

Figure 2: Distribution of average number of words in questions and answers.

4.2. Question Types

To better understand the type of questions in our dataset, we looked into the frequency of the first and second words in each question we collected (see Table 1 and Figure 3). Most of the collected questions were of the "what" type, which is similar to SQuAD (2.0) and MS MARCO v2 datasets. Other types include "how", "why", "when", "who", and "where" in descending order. This may be due to the versatility of the "what" type, which permits questioning of various topics, whereas the other interrogative types are more restrictive in terms of their questioning.

For instance, the "who" type asks about certain persons and the "when" type asks about the timing. Moreover, as we allowed only for factoid questions to be collected during the labeling process, we do not have category "OTHER", as some other datasets have (e.g., MS MARCO v2). Similarly, we also did not allow to start questions with a question word "which", so that category is also zero. Finally, with regards to the second word in the collected questions, the most common was of the "is/are" type, followed by "can", then "does/do", and lastly, category "others" (see Figure 3).

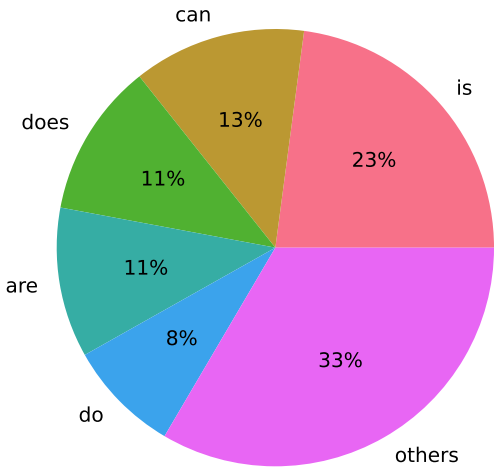

Figure 3: Pie charts of frequency of second words in all questions.

4.3. Question and Answer Entailment

Question A entails question B if every answer to question B is also exactly or partially correct answer to question A (Abacha and Demner-Fushman, 2016). Similarly, answer A and answer B can be considered to be entailed if both are able to answer the same question (Saikh et al., 2021). Examples of both can be seen in Appendix B. Understanding both, especially question entailment (Ben Abacha and Demner-Fushman, 2019), could improve QA.

We have indicated the occurrence of both cases in Table 2. Since different passages are used for each question, it is expected that both question and answer entailment would be rather low due to different choices of words and the unlikeliness of identical phrasing appearing in different passages. However, there is a greater than expected occurrence of identical answers. This can be attributed to two factors: the large proportion of questions that call for numerical answers such as "7", as well as the fact that answers are text spans as opposed to full sentences, leading to less variation. More detailed explanation about how both entailments were calculated is provided in Appendix B.

Table 2: Occurrence of entailment types.

Entailment type	**Occurrence**
Question	222
Answer	149

5. Model Fine-tuning

5,000 of the 7,005 passages were randomly selected and used for the labeling process. The final dataset consists of 5,000 triplets (p, q, a) without repetition, where p is a passage, q is a question, and a is an answer. We created a train-valid-test split using automated randomized partitioning, generating 80:10:10 split. We used k-fold (k=9) cross-validation.

We evaluated the quality of our dataset and performed retrieval/reader models fine-tuning on *BERT* model[4] and five domain-specific BERTs: 1) *BioBERT*[5], 2) *BioBERT BioASQ*[6], 3) *ClinicalBERT*[7], 4) *SciBERT*[8] and 5) *PubMedBERT*[9], using the code shared by authors[10]. We fine-tuned our models for 30 epochs, with a batch size equal to 16. We only set other negatives parameter to one, while hard negatives parameter was equal to zero. Other negatives were randomly chosen from the text corpus.

4. https://huggingface.co/bert-base-uncased
5. https://huggingface.co/dmis-lab/biobert-v1.1
6. https://huggingface.co/gdario/biobert_bioasq
7. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
8. https://huggingface.co/allenai/scibert_scivocab_uncased
9. https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
10. https://github.com/facebookresearch/DPR

6. Evaluation

We performed both *intrinsic* and *extrinsic* evaluation of fine-tuned models and our QA pipeline. Intrinsic evaluation evaluates properties of each models' output, while extrinsic evaluation evaluates the impact of the whole QA pipeline, by investigating to which degree it achieves the overarching task for which it was developed. In our case, the QA system was designed to provide health coaches with direct and accurate answers upon receiving factual sleep-related queries from their clients.

6.1. Intrinsic Evaluation

Intrinsic evaluation was done using automatic metrics on 500 test labels: recall@k for retrieval models and EM and F1 scores for reader models and QA pipelines. We evaluated our five fine-tuned domain-specific BERT retrieval models against Lucene BM25 model (Lin et al., 2021), while our fine-tuned domain-specific BERT reader models were compared against *BERT SQuAD2* (Rajpurkar et al., 2018). Finally, we compared the built QA pipeline (the best performing combination of fine-tuned retrieval and reader models) against Lucene BM25 + *BERT SQuAD2* QA pipeline.

Retriever Models Table 3 compares recall@1 on 500 corpus-specific questions from our test set using six retrieval models. Figure 4 shows recall@k for top three models: Lucene BM25, fine-tuned *PubMedBERT* and fine-tuned *SciBERT*. The results showed that Lucene BM25, a traditional sparse vector space model, outperformed both domain-specific BERT models fine-tuned on SleepQA dataset. This shows that there exists a significant margin of improvement for domain-specific dense retrieval models.

Lucene BM25 model measures the relative concentration of a term in a given piece of text (Pérez-Iglesias et al., 2009). In practice, it means that it will work better if the given question and the given answer have more significant overlap. The problem of similarities between questions and answers was previously already reported in the literature (Rajpurkar et al., 2016). In order to prevent it, annotators who participated in the labeling process in that paper were strongly encouraged to formulate a question in their own words, without copying word phrases from the passage. This reminder was prompted at the beginning of each passage and the function of copy-pasting on the passage was disabled. Similarity analysis for our dataset is given in Appendix B.

Table 3: Automatic evaluation of Lucene BM25, *BERT SQuAD2* and different domain-specific BERT models using recall@1, EM and F1 scores.

Name of the model	**recall@1**	**EM (oracle)**	**F1 (oracle)**
Lucene BM25 (retrieval)	**0.61**		
BERT SQuAD2 (reader)		**0.50**	0.64
Fine-tuned BERT (retrieval/reader)	0.35	0.56	0.68
Fine-tuned BioBERT (retrieval/reader)	0.35	0.58	0.70
Fine-tuned BioBERT BioASQ (reader)		**0.61**	0.73
Fine-tuned ClinicalBERT (retrieval/reader)	0.34	0.56	0.68
Fine-tuned SciBERT (retrieval/reader)	0.38	0.60	0.71
Fine-tuned PubMedBERT (retrieval/reader)	**0.42**	0.59	0.71

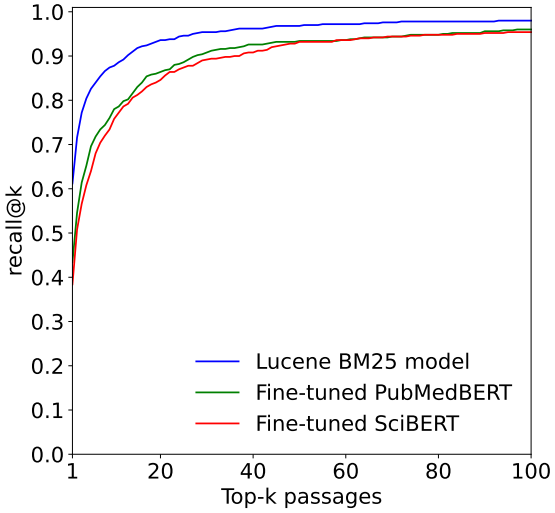

Figure 4: recall@k for top three retrieval models.

Reader Models The fine-tuned domain-specific BERT reader models were compared to *BERT SQuAD2* model. Reader models are evaluated independently from the retrieval models, meaning that the question and its exact passage ("oracle") are provided for each reader as its inputs (see Table 3). This allows us to find the best performing fine-tuned retrieval and reader models separately and consequently to build the best performing QA pipeline.

QA Pipeline To further perform evaluation of the best performing QA pipeline, we took the best fine-tuned retrieval model and the best fine-tuned reader model (based on results from Table 3) and compared them with Lucene BM25 + *BERT SQuAD2* QA pipeline. Table 4 shows automatic evaluation of two QA pipelines: *PubMedBERT + BioBERT BioASQ* (denoted as Pipeline 1) and Lucene BM25 + *BERT SQuAD2* (denoted as Pipeline 2) on 500 test labels. Pipeline 2 (with Lucene BM25 as a retrieval model) still performs better.

Table 4: Automatic evaluation of two QA pipelines on 500 test labels questions using EM and F1 scores.

Pipeline name	EM	F1
Pipeline 1	0.24	0.33
Pipeline 2	0.30	0.41

6.2. Extrinsic Evaluation

Outputs from each pipeline were presented in randomized order to avoid bias by hindering annotators from favoring our pipeline. Annotators were asked to give a score "1" if the answer 1 was better, "2" if the answer 2 was better, "3" if both answers were equally good, and "4" if both answers were equally bad. In addition to comparing two pipelines based on the text span answers (denoted as "w/o exp" in Table 5), we also asked annotators to repeat the same evaluation, but this time to give scores not only based on the text spans, but also on their corresponding passages (denoted as "w exp" in Table 5). By showing the retrieved passages in addition to the text spans, annotators were presented with an explanation about passages that the text spans were retrieved from.

Extrinsic evaluation was done by five annotators on 500 real-world questions. Each annotator evaluated answers for 100 questions. The answers for additional 150 questions were evaluated by all five annotators to allow for inter-annotator agreement calculation using Gwet's AC1 score (Gwet, 2001). Through this process we collected 500 answers with single evaluation and 150 answers with 5-way evaluations. The calculated Gwet's AC1 scores were **0.72** and **0.71** for span answers and span answers + explanations, respectively. Both scores imply a substantial agreement among annotators (Landis and Koch, 1977).

Results from Table 5 indicate that our pipeline performs better than Pipeline 2 on the task which it was designed for, i.e., on the task of providing health coaches with the correct answers on sleep-related queries from their clients. Moreover, chi-squared test showed that there was no statistically significant difference (p=0.29) between scores given to each pipeline when they look only at answers and when they are also provided by the explanations. However, in a small number of cases, after reading provided explanation, annotators change their score to "4", which means that explanation helped them to realize that answer was not as good as it seemed on its own. This might come from the fact that answers are short and in a small number of cases it is hard to judge their quality only on their own.

Table 5: Human extrinsic evaluation on 500 real-world questions.

	w/o exp	w exp
Pipeline 1 wins	37.2%	34.4%
Pipeline 2 wins	11.2%	8.6%
Equally good	12.2%	14.2%
Equally bad	39.4%	42.8%

6.3. Error Analysis and Challenges

Analysis of Different Question Types Our models were only fine-tuned on factoid questions as explained in Section 3.2. However, on other types of questions (e.g., "Yes/No" questions), our pipeline performed 28% better than Pipeline 2 (e.g., see **Q1** in Table 6), compared to 13% when Pipeline 2 performed better (e.g., see **Q2** in Table 6). This analysis indicates that our pipeline performed better at dealing with questions that do not start with wh-words. However, the majority of questions were incorrectly answered by both pipelines (in 45% of cases) indicating that more work is still needed in this area (e.g., see **Q3** in Table 6).

Challenges In addition to *hallucinations* (e.g., see **Q4** in Table 6), another detected problem was when both pipelines provided different, yet plausible answers for the given question (e.g., see **Q5** in Table 6). After inspection of the dataset, we noticed that the answer for this question was *not consistent* across different passages. Finally, last three questions from Table 6 show that answers can change even with a slight change in the question. For example, adding "my" to **Q6** changed answers for **Q7** for both pipelines. On the other hand, changing "my" to "our" (**Q7** and **Q8**) did not affect the answer for our pipeline, while it did change the answer for Pipeline 2, indicating that our pipeline might be more resilient to slight changes in questions that do not affect their meaning.

7. Conclusion

We introduced SleepQA, a new question-answering dataset for the specific health domain of sleep. We also provided several state-of-the-art baselines and measures. Our evaluation showed that our fine-tuned pipeline outperformed Lucene BM25 on real-world questions. We hope our dataset will serve as a useful benchmark for the development of other health-related NLP models.

Acknowledgments

The authors would like to acknowledge the funding support from Nanyang Technological University, Singapore. Josip Car's post at Imperial College London is supported by the NIHR NW London Applied Research Collaboration. The authors would also like to thank Mathieu Ravaut and Duy Phung Van for their valuable feedback while writing the paper and performing experiments.

Table 6: Sample answers from Pipeline 1 and Pipeline 2 on a subset of real-world questions.

Question	Pipeline 1	Pipeline 2
Q1: Does coffee affect my sleep?	coffee has the ability to disrupt sleep even when ingested six hours before bed	my kid does this in his sleep
Q2: Will drinking coffee affect my sleep?	may reduce the risk of cardiovascular diseases and cancer	might backfire
Q3: Should I take melatonin every day before I sleep?	between 30 minutes and two hours before bedtime	easier
Q4: What really happens if I don't get the optimum number of sleep for a university student?	I went into advertising, and I was highly stressed	your child is not sleeping in a way that's predictable and efficient
Q5: What time should I stop drinking coffee to have a good sleep?	at least three hours before bed	four hours before bedtime
Q6: How does alcohol affect sleep	alcohol's impact on sleep largely depends on the individual	allowing you to fall asleep more quickly
Q7: How does alcohol affect my sleep?	alcohol also lowers sleep quality and duration, making sleep unrestorative and choppy	drinking to excess will probably have a more negative impact on sleep
Q8: How does alcohol influences our sleep?	alcohol also lowers sleep quality and duration, making sleep unrestorative and choppy	comfortable and supportive

References

Asma Ben Abacha and Dina Demner-Fushman. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310, 2016.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4): 555–596, 2008.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of NLG systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320, 2006.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23, 2019.

Siddharth Biswal, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Jimeng Sun, and Matt T Bianchi. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650, 2018.

YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2): 277–288, 2011.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Cathy A Goldstein, Richard B Berry, David T Kent, David A Kristo, Azizi A Seixas, Susan Redline, and M Brandon Westover. Artificial intelligence in sleep medicine: Background and implications for clinicians. *Journal of Clinical Sleep Medicine*, 16(4):609–618, 2020.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Kilem Gwet. Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD: STATAXIS Publishing Company*, 2001.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Mahsa Abazari Kia, Aygul Garifullina, Mathias Kern, Jon Chamberlain, and Shoaib Jameel. Adaptable closed-domain question answering using contextualized CNN-attention models and question expansion. *IEEE Access*, 10:45080–45092, 2022.

Lisa A Kisling and Joe M Das. Prevention strategies. In *StatPearls [Internet]*. StatPearls Publishing, 2021.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328, 2018.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.

Chuntian Lu, Bing Liao, Jing Nie, Wei Wang, and Yafeng Wang. The association between sleep duration and chronic diseases: A population-based cross-sectional study. *Sleep Medicine*, 73:217–222, 2020.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*, 2016.

Joaquín Pérez-Iglesias, José R Pérez-Agüera, Víctor Fresno, and Yuval Z Feinstein. Integrating the probabilistic models BM25/BM25F into Lucene. *arXiv preprint arXiv:0911.5046*, 2009.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*, 2018.

Tanik Saikh, Sovan Kumar Sahoo, Asif Ekbal, and Pushpak Bhattacharyya. COVIDRead: A large-scale question answering dataset on COVID-19. *arXiv preprint arXiv:2110.09321*, 2021.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen. Open domain question answering using early fusion of knowledge bases and text. *arXiv preprint arXiv:1809.00782*, 2018.

Haoqi Sun, Jian Jia, Balaji Goparaju, Guang-Bin Huang, Olga Sourina, Matt Travis Bianchi, and M Brandon Westover. Large-scale automated sleep staging. *Sleep*, 40(10), 2017.

Anne Von Ruesten, Cornelia Weikert, Ingo Fietze, and Heiner Boeing. Association of sleep duration with chronic diseases in the European Prospective Investigation into Cancer and Nutrition (EPIC)-

Potsdam study. *PloS one*, 7(1):e30972, 2012.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*, 2017.

Juan Yang, Brent A Bauer, Stephanie A Lindeen, Adam I Perlman, Kasey R Boehmer, Manisha Salinas, Susanne M Cutshall, et al. Current trends in health coaching for chronic conditions: A systematic review and meta-analysis of randomized controlled trials. *Medicine*, 99(30), 2020.

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. *arXiv preprint arXiv:1809.09600*, 2018.

Mark Yatskar. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. *arXiv preprint arXiv:1809.10735*, 2018.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.

Appendix A. Interface for Labeling

A.1. Annotator Interface

Figures below show user interfaces for login, labels collection, additional annotation task, and administrator interfaces. The login interface (see Figure 5) requires annotators to enter their credentials, which are their username and password. The features displayed on the interface for labels collection (see Figure 6) include one passage, two text fields (for question and answer), two check boxes for marking unclean passages or skipping the passages, and one button for submission. The interface for additional annotation task (see Figure 7) is similar, except that it shows the question which was formed by another annotator. Annotators are only asked to find an answer span for this task.

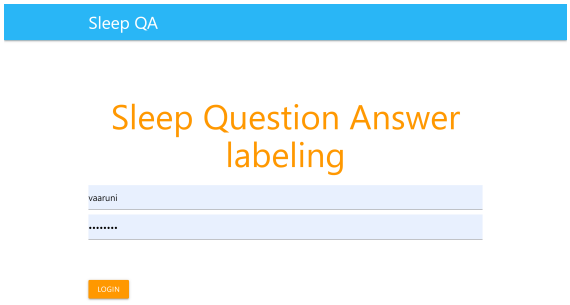

Figure 5: User interface for login.

A.2. Administrator Interface

Administrator interface (see Figure 8 and Figure 9) is designed to help the administrators (I.B. and Q.C.O.) monitor annotators' progress and make changes where necessary. The main features are passage number, question formed, selected answer span, and name of annotator who performed the task.

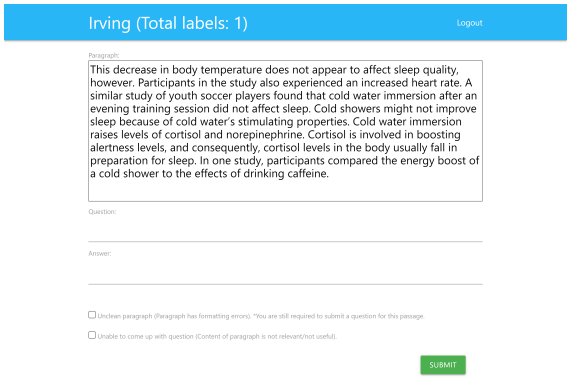

Figure 6: User interface for labels collection.

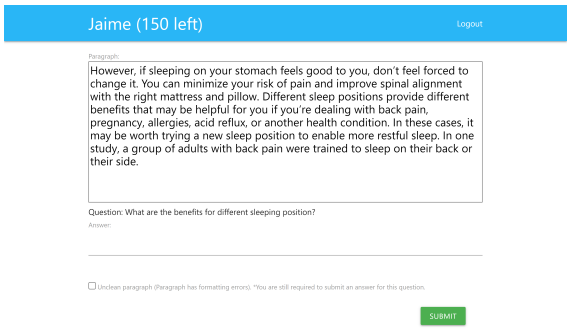

Figure 7: User interface for additional annotation task (evaluation of consistency).

A.3. 5-way Annotation

An example of label which have 5-way annotation is shown in Table 7. The passage comes from a set of 7005 passages that compose our text corpus. Question was posed by one annotator and consequently answered by other four annotators.

Table 7: A sample of label with 5-way annotation.

Passage	**Question**	**Answer 1**	**Answer 2**	**Answer 3**	**Answer 4**	**Answer 5**
Sleep issues arise for a variety of reasons. Pain seems to be a leading factor in poor quality of sleep. Spinal pain affects over 54% of adults, while migraines and chronic headaches affect over 38 million Americans and counting. In addition to pain, insomnia affects over 30% of adults and is often correlated with health issues. Other sleep disorders such as sleep apnea and narcolepsy, although not as prevalent in the general population as chronic pain, also create barriers to restorative sleep. Often, people seek out new mattresses or pillows to help alleviate their sleep issues.	How many adults are affected by spinal pain?	over 54% of adults	over 54%	over 54% of adults	over 54% of adults	Spinal pain affects over 54% of adults

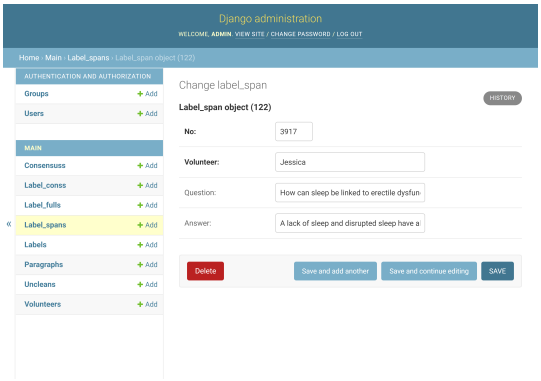

Figure 8: Administrator interface for inspecting all labels.

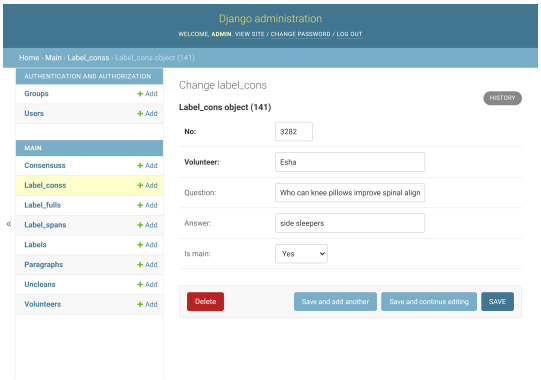

Figure 9: Administrator interface for consensus labels.

Appendix B. QA Analysis

B.1. Question and Answer Entailment

Entailment was calculated by annotators comparing the Questions and Answers for different passages (see Table 8). Firstly, if the Answers are noted to be identical, their respective Questions are then compared. If the pair of Questions are different, yet yield the same Answer, this pair of Question and Answers will be considered as a case of Question Entailment. This is then repeated with Questions being first compared. If the Questions are identical and their respective Answers are different, it is then considered a case of Answer Entailment.

B.2. Question Answer Similarities

In order to compare similarities between a given question and a given answer in a pair in our dataset against the other similar datasets, we downloaded five datasets using download script provided by authors[11]. From each training set, we then randomly selected 1,000 question-answer pairs and for each question we detected the full sentence where the answer came from. In that sense, for each dataset separately we built a subset of labels where answers were full sentences, rather than text spans. Finally, for each pair in a particular dataset, we calculated F1 score separately and then averaged them over all 1,000 pairs. The final calculated F1 scores are shown in Table 9.

Detected similarities between a question and an answer in a question-answer pair in our dataset were higher than those from other datasets. This could potentially be a result of labeling process during which annotators were encouraged to first find a potential answer from the passage and then formulate a question based on the chosen answer. This resulted in using similar phrases in the posed questions from the corresponding passages. Although, we do note that perhaps some of the overlap could be reduced by giving reminders to rephrase questions, this problem cannot be completely solved using just annotators' efforts. In future work, we will investigate whether using back translation for data augmentation could solve this problem. The main idea behind using back translation for data augmentation is that the training examples are machine-translated from a source to a pivot language and back, thus obtaining paraphrases.

11. https://github.com/facebookresearch/DPR/blob/main/dpr/data/download_data.py

Table 8: Example of question and answer entailment.

Entailment	Question	Answer
Question Entailment	What is a known cause of insulin resistance?	obesity
	What condition can increase the likelihood of experiencing obstructive sleep apnea?	
	What is a risk factor for sleep-related breathing disorders?	
Answer Entailment	What is the ideal bedroom temperature?	66 and 70 degrees fahrenheit
		somewhere around 65 degrees

Table 9: Similarities between a question and an answer in a question-answer pair.

Dataset name	F1 score
SleepQA	**0.17**
SQUAD 1.1	0.09
TriviaQA	0.07
CuratedTrec	0.05
NQ	0.04
WebQuestions	0.02

Appendix C. Model Hyperparameters

In order to compare the performance of our fine-tuned domain-specific BERT models/pipeline with the Lucene BM25 model, we used Pyserini Python toolkit[12]. Hyperparameters of retrieval model fine-tuning for domain-specific BERTs are shown in Table 10, and for reader models in Table 11.

Table 10: Hyperparameters of retrieval model fine-tuning for domain-specific BERTs.

Hyperparameter	Value
batch size	16
dev batch size	16
adam eps	$1e-8$
adam betas	(0.9, 0.999)
max grad norm	1.0
log batch step	100
train rolling loss step	100
weight decay	0.0
learning rate	$1e-5$
warmup steps	100
gradient accumulation steps	1
num train epochs	30
eval per epoch	1
hard negatives	0
other negatives	1
val av rank hard neg	0
val av rank other neg	10
val av rank bsz	128
val av rank max qs	10000

12. https://github.com/castorini/pyserini

Table 11: Hyperparameters of reader model fine-tuning for domain-specific BERTs.

Hyperparameter	**Value**
eval step	500
batch size	16
dev batch size	16
adam eps	$1e-8$
adam betas	(0.9, 0.999)
max grad norm	1.0
log batch step	100
train rolling loss step	100
weight decay	0.0
learning rate	$1e-5$
warmup steps	0
gradient accumulation steps	1
num train epochs	30