Bridging the Knowledge Gap: Improving BERT models for answering MCQs by using Ontology-generated synthetic MCQA Dataset

Sahil Sahil
Department of Computer Science and Engineering
Indian Institute of Technology Madras
cs20s017@cse.iitm.ac.in

P Sreenivasa Kumar
Department of Computer Science and Engineering
Indian Institute of Technology Madras
psk@cse.iitm.ac.in

Abstract

BERT-based models possess impressive language understanding capabilities but often lack domain-specific knowledge, limiting their performance on specialised tasks such as medical multiple-choice question answering (MCQA). In this paper, we study how biomedical ontologies, rich repositories of medical knowledge, can be harnessed to enhance BERT-based models for medical MCQA task. Our contributions include OntoMCQA-Gen, a system which leverages different biomedical ontologies to construct BioOntoMCQA, a large synthetic MCQA dataset. OntoMCQA-Gen exploits the subclass-class relationships, definitions of concepts, and also synonym relationships from the ontologies to create this dataset of MCQs automatically. We then use this synthetic dataset to fine-tune various BERT-based models to answer medical MCQs. We evaluated these fine-tuned BERT models on the challenging MedMCQA and MedQA datasets of questions from admission examinations for medical degrees in India and USA, respectively. Our evaluation study on these datasets shows that fine-tuning the BERT-based models on BioOntoMCQA results in significantly improved accuracy scores. BioBERT and PubMedBERT, pretrained on the large medical corpus, have also shown significant improvements with our technique of fine-tuning ontology-generated synthetic data. This finding highlights the effectiveness of incorporating biomedical ontologies to enhance the BERT-based model in the medical domain. Moreover, our results underscore the importance of using ontology-generated data along with model adaptation for specialised domains, contributing to a novel advancement in natural language processing.

Introduction

Biomedical ontologies serve as structured knowledge representations in the realm of medical research. They encompass a systematic arrangement of medical concepts and their relationships, much like a comprehensive framework that captures the complex web of medical information. These ontologies play a pivotal role in categorising and organising various facets of medical language, including symptoms, diagnoses, treatments, and more. Biomedical ontologies form the backbone of medical knowledge management, underpinning advancements in healthcare, research, and understanding of human health and diseases.

In the context of the medical domain, multiple-choice question-answering (MCQA) is a particularly challenging task. Unlike in general domains, relevant medical knowledge isn't readily abundant in standard text corpora. The effectiveness of MCQA systems relies on achieving a good balance between understanding the language, using domain-specific logic, and incorporating a wide range of knowledgeable information.

Ontology-based QA systems like (Kwon et al. 2021) and (Midhunlal and Gopika 2016), have a lot of potential to capture medical knowledge and give accurate answers. By leveraging biomedical ontologies, these systems can effectively illustrate complex relationships among medical concepts, thereby furnishing more precise and contextually aware answers. However, employing such systems necessitates understanding ontology structure to formulate queries effectively.

Harnessing the capabilities of contextual word embedding models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018), advancements have been made in natural language processing (NLP). While initially designed for general domains, models like BioBERT (Lee et al. 2020), SciBERT (Beltagy, Lo, and Cohan 2019), and PubmedBERT (Gu et al. 2021) have been tailored to the biomedical field through pretraining on biomedical corpora. However, the integration of structured knowledge from biomedical ontologies into BERT-based models remains an area that hasn't been extensively explored.

To understand the significance of biomedical ontology knowledge, consider a scenario where a medical question pertains to a rare disease. A language model pretrained on extensive corpora may recognise related terms or phrases but may lack the domain-specific knowledge needed for accurate answers. In contrast, biomedical ontologies encompass structured, domain-specific knowledge, including relationships, hierarchies, and semantic insights about medical concepts. Integrating ontology knowledge into our models facilitates a comprehensive and precise representation of medical domain knowledge, leading to more accurate and contextually aware question-answering.

The research gap we have identified shows the need for novel approaches to connect ontology knowledge with pre-trained language models like BERT. This would make answering multiple-choice questions in the medical field more accurate and effective. In response to these challenges, we suggest a novel approach that combines the benefits of biomedical ontologies and BERT-based models. Our contributions can be summarised as follows:

1) We introduce the BioOntoMCQA dataset, a synthetic MCQA dataset generated using the proposed OntoMCQA-Gen system. By leveraging medical knowledge from ontologies, BioOntoMCQA provides a set of synthetic MCQAs specifically tailored to the medical domain. To the best of our knowledge, this is the first ontology-generated synthetic MCQA dataset for the medical field.

2) We demonstrate quantitative improvements by fine-tuning BERT-based models on the BioOntoMCQA dataset. Our results show substantial accuracy enhancements in answering the MedMCQA dataset, surpassing baseline models and underscoring the importance of domain-specific fine-tuning in natural language processing applications.

In the following sections, we explore the details of our methodology, experimental setup, and comprehensive analysis of the results. Furthermore, we discuss the implications of our findings and the broader significance of using domain-specific datasets and tailored fine-tuning strategies to optimise language models for specialised tasks and domains.

Related Work

Although ontology-based MCQ answering systems (Midhunlal and Gopika 2016) and (Kwon et al. 2021) exhibit promise in capturing domain-specific medical knowledge and providing accurate answers, these methods are not without limitations. Their reliance on template-based approaches can constrain the flexibility and adaptability of the system.

Complementing the ontology-based paradigm using pre-trained models has significantly propelled the advancement of MCQ answering systems. For instance, PubmedBERT (Gu et al. 2021), BioBERT (Lee et al. 2020), and SciBERT (Beltagy, Lo, and Cohan 2019) have showcased remarkable proficiency in capturing medical terminologies and comprehending complex medical questions. These models, pretrained on extensive corpora such as Pubmed abstracts and comprehensive medical datasets, have demonstrated enhanced performance in various medical question-answering and information retrieval tasks.

While incorporating external knowledge sources has been extensively explored in various domains (Sap et al. 2019) and (Lv et al. 2020), the biomedical MCQ answering domain has witnessed relatively fewer developments. BioOntoBERT (Sahil and P. Sreenivasa Kumar 2023) employs sentences derived from biomedical ontologies for the pre-training of BERT-based models. However, this approach did not result in a significant increase in accuracy.

Furthermore, (Zhou and Srikumar 2021), (Merchant et al. 2020), (Mosbach et al. 2020) and (Hao et al. 2020) contribute to the existing work by investigating the application of pre-trained contextualised representations and their fine-tuning effects. Focusing on the English BERT family, (Zhou and Srikumar 2021) employs two probing techniques to scrutinise the changes induced by fine-tuning. Their results indicate that fine-tuning improves performance by segregating label-associated points from others and adapting representations for downstream tasks.

In contrast to existing approaches, this paper introduces a novel approach wherein ontology knowledge is transformed into synthetic MCQs. These synthetic MCQs are then used to enrich the fine-tuning process, representing a unique and innovative strategy to enhance the MCQ answering task.

Biomedical Ontologies

Biomedical ontologies are crucial in medicine for organising disease, gene, anatomical, and medical knowledge. They create a standard framework integrating data, enabling sharing, interoperability, and knowledge discovery. Here, we highlight the key biomedical ontologies we are incorporating into our model. These are summarised in Table 1

1. **Foundational Model of Anatomy Ontology (FMAO)** (Rosse and Mejino Jr 2008) (v5.0.0): FMAO depicts human anatomy comprehensively, featuring a structured hierarchy of body structures. It captures spatial links and functional ties among body parts.
2. **Gene Ontology (GO)** (Ashburner et al. 2000) (v2023-04-01): A widely used ontology that focuses on representing the functional attributes of genes and gene products across different species. It comprises three domains: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). BP covers gene-involved processes, MF shows molecular functions, and CC defines cellular locations.
3. **Disease Ontology (DO)** (Schriml et al. 2012) (v1.2): The Disease Ontology provides standardised descriptions of human disease terms, phenotypes, and related medical concepts for the biomedical community's consistency and reusability.
4. **Precision Medicine Ontology** (Hou et al. 2020) (v4.0): It is a comprehensive ontology that standardises medical concepts and relationships. It encompasses domains like diseases, symptoms, treatments, diagnostics, and medical devices.
5. **Bioassay Ontology (BAO)** (Visser et al. 2011) (v1.1): BAO aims to create standard metadata terms and definitions for describing information about both low and high throughput drug and probe screening assays, including their results.
6. **Dental Ontology** (Duncan et al. 2020) (v2016-06-27): DO records dental concepts and relationships, offering a standardised vocabulary for dental conditions, procedures, materials, and anatomy. It aids dental data integration, benefiting research, education, and clinical dentistry.
7. **Pediatrics Ontology** (v2.0): It represents pediatric healthcare concepts and relationships, encompassing diseases, developmental milestones, treatments, and interventions.

Table 1: Different Biomedical Ontologies used

Ontology	Scope	Classes	# Annotations	# subClass
FMAO Ontology	Anatomy	104,721	51	262,548
Gene Ontology	Bioinformatics	84,108	60	192,606
Disease Ontology	Pathology	11,033	53	11,063
Precision Medicine Ontology	Medicine	76,155	23	122,760
Bioassay Ontology	Pharmacology	904	34	981
Dental Ontology	Dentistry	2,745	28	6,507
Paediatrics Ontology	Pediatrics	1,771	8	1,760
HPS Ontology	Physiology	2,920	34	3,143
Mental Disease Ontology	Psychiatry	879	102	940

8. **Human Physiology Simulation Ontology (HPSO)** (Gündel et al. 2013) (v1.1.1): HPSO captures and models human physiology concepts and relationships, offering a standard framework for physiological processes, organ interactions, and computational models.
9. **Mental Disease Ontology (MDO)** (Hastings et al. 2012) (v2020-04-26): MDO depicts mental disorders and concepts, providing a standardised vocabulary for categorising mental diseases, symptoms, treatments, and diagnostic criteria.

Datasets

MedMCQA Dataset (Pal, Umapathi, and Sankarasubbu 2022), contains an extensive collection of 194,000 multiple-choice questions covering approximately 2,400 healthcare topics and 21 medical subjects. These questions are drawn from one of India's most rigorous entrance exams for medical graduates, namely, AIIMS and NEET PG. The diversity of questions in MedMCQA poses a challenge as it covers various aspects of medical knowledge. Table 2 shows a sample question from the dataset.

An important characteristic that sets this dataset apart is that human experts authored and curated the questions, ensuring high expertise and accuracy. The dataset comprises three distinct parts: a training set of 1,82,822 questions, a validation set with 4,183 questions, and a test set containing 6,150 questions. The average token lengths for these sets are 12.35, 13.91, and 9.68, respectively.

In the dataset, the answer choices are provided under the 'labels' column and encoded as integers 0, 1, 2, and 3. It is important to note that the ground truth for the test set is not publicly available. As a result, our analysis will focus on the validation set, which offers a reliable basis for evaluating a model's performance.

MedQA Dataset (Jin et al. 2021) comprises a collection of multiple-choice questions, each offering four choices. The dataset is sourced from professional medical board exams and encompasses three languages: English, simplified Chinese, and traditional Chinese. Across these languages, there are 12,724, 34,251, and 14,123 questions, respectively. We exclusively utilise the English section derived from the United States Medical License Exams (USMLE) for our evaluation. Following the recommended data distribution, this portion is divided into 10,178 training questions, 1,273 validation questions, and 1,273 testing questions. Notably, the dataset lacks categorisation based on subjects or topics.

Table 2: Sample MCQA question from MedMCQA dataset with the correct answer as (C)

Question: Which one of the following is a muscle splitting incision?	
(A) Kocher's incision	(B) Rutherford-Morrison incision
(C) Pfannenstiel incision	(D) Lanz incision

Methodology

OntoMCQA-Gen

As already mentioned, Biomedical ontologies serve as comprehensive and well-structured representations of the knowledge of medical concepts. These ontologies have been found to be useful for pretraining language models, leading to improved performance on downstream tasks.

The richness of information in these biomedical ontologies allows for the generation of synthetic datasets for multiple-choice question answering in a controlled manner. This process employs utilising ontology triples, where the 'subject' is the entity, the 'predicate' is the relationship, and the 'object' is the concept.

Our proposed approach, OntoMCQAGen, involves designing various question templates based on ontology triples. Specific templates are utilised for 'subClass of', 'has synonym', and 'has definition' predicates such as "{Subject} is a type of ?", "Which of the following is a synonym for {Subject}?" and "Which of the following correctly defines {Subject}?" to ensure the generation of diverse and contextually relevant questions. This process is shown in Figure 1.

We employ RDFLib (Krech et al. 2023) and SPARQL (Harris, Seaborne, and Prud'hommeaux 2013) to extract relevant information from different biomedical ontologies. RDFLib is used for loading and effectively manipulating the biomedical ontology and helps us extract the ontology

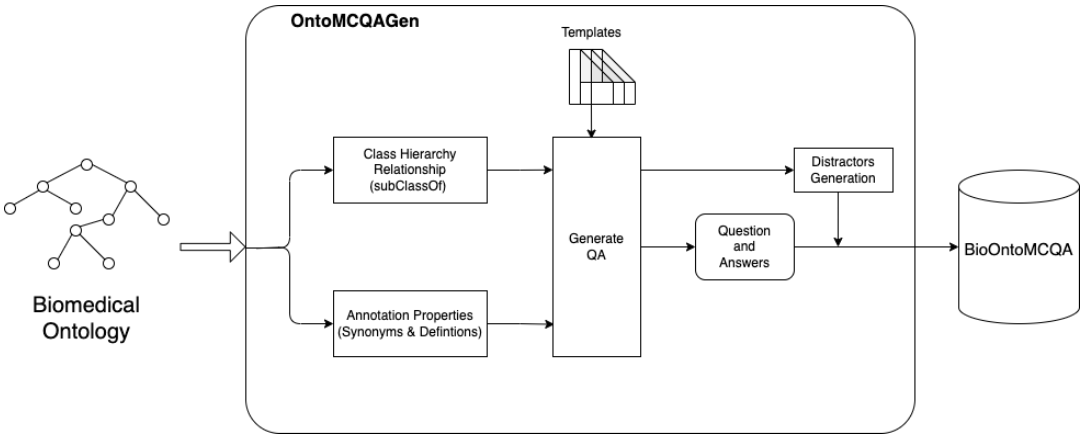

Figure 1: Generating Medical MCQA from Biomedical Ontology using OntoMCQAGen

triples, consisting of 'subject', 'predicate', and 'object' components.

Additionally, we extract the necessary information from the ontology triples using SPARQL. These SPARQL queries are designed to obtain information related to 'subClass of', 'has synonym', and 'has definition' predicates, which are then used to generate questions.

For each ontology triple, we generate a question by applying the appropriate template and substituting the 'subject' component while maintaining semantic alignment with the biomedical domain.

The 'object' component of the ontology triple serves as the correct answer choice. To create a challenging dataset, three incorrect answer choices are randomly selected from other 'object' values in the ontology. This ensures that the object chosen is not connected to the subject node, represented as distractor generation in Table 1. The answer choices, including the correct and incorrect options, are shuffled to avoid potential biases introduced by the position of the correct answer.

This approach results in the creation of BioOntoMCQA, a large synthetic dataset comprising 1,672,575 multiple-choice question answers (MCQA) from biomedical ontologies. The questions are generated using three types of predicates: 'subClass of', 'has synonym', and 'has definition', resulting in 815,920, 726,598, and 130,059 questions, respectively. The dataset provides a structured and controlled context for language understanding and reasoning, effectively leveraging the wealth of knowledge encapsulated in the ontologies.

The availability of such a comprehensive synthetic dataset derived from biomedical ontologies serves as an invaluable resource for training and evaluating language models in the biomedical domain. Even though other predicates from the ontologies can also be used for generating questions, in this paper, we confine to the predicates mentioned above as they are available commonly across all the ontologies used. Furthermore, the dataset generation process is scalable and adaptable, allowing for easy incorporation of new knowledge as biomedical ontologies continue to evolve.

Fine-tuning BERT-based models

Fine-tuning language models on domain-specific tasks is crucial for enhancing their performance in specialised fields like medical multiple-choice question answering. In the process of fine-tuning the BERT model for medical multiple-choice question answering, we followed two main steps: tokenisation and fine-tuning using two different datasets, BioOntoMCQA and evaluation datasets consisting of MedMCQA and MedQA datasets.

We employed the BERT tokeniser, which breaks down the input text into individual tokens and converts them into numerical representations. Special tokens like [CLS] and [SEP] are added to indicate the beginning and separation of sentences. Additionally, we formed four different combinations of each question with each answer choice (1, 2, 3, and 4) and fed them to the BERT model, as illustrated in Figure 2. This step allowed the model to learn the relationships between questions and answer choices effectively.

During the first phase of fine-tuning, our primary objective was to enhance the BERT model's accuracy by leveraging the synthetic BioOntoMCQA dataset, whose generation was described in the previous section. We selected a random subset of sample questions from the extensive BioOntoMCQA dataset containing approximately 1.6 million MCQs (to prevent overfitting). Initially, the size of this subset is chosen to be about 50% of the target training dataset (i.e., 91,411 for MedMCQA and 5089 for MedQA). A comparative study of different subset sizes is discussed in a later section of the paper. Through this strategic approach, we aimed to capitalise on the rich and organised knowledge representations of medical concepts within BioOntoMCQA.

In the second fine-tuning phase, we leveraged the real-world evaluation dataset, which contains medical questions and answers authored by human experts. This step further honed the model's ability to perform well in real medical multiple-choice question-answering scenarios, making it more reliable for practical use.

We carefully adjusted the model's parameters throughout both fine-tuning phases, optimised its performance through various settings, such as learning rate and training epochs,

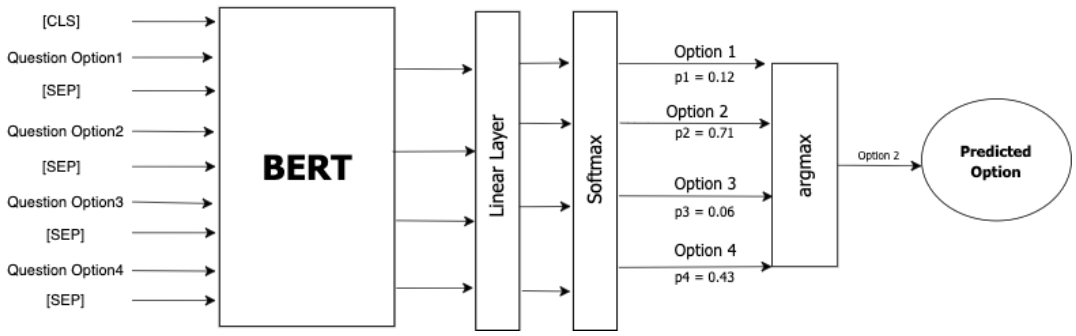

Figure 2: Fine-tuning BERT model on Multiple choice question

and represented the labels in a simple format. The labels were encoded in a one-hot format derived from integers to provide the necessary input for the multiple-choice question-answering task. The evaluation process involved generating probability distributions for each answer choice associated with a question (represented as $p1, p2, p3$, and $p4$) and selecting the index with the highest probability as the most likely answer choice. This comprehensive evaluation allowed us to measure the model's performance on new and unseen data, providing valuable insights into its ability to answer complex medical multiple-choice questions accurately.

By fine-tuning with domain-specific synthetic datasets, our approach significantly improves the BERT model's accuracy and performance in medical multiple-choice question answering. It is worth noting that during the fine-tuning process, we did not use any context for the questions, as each question was independently paired with the respective answer choices. We intend to release the BioOntoMCQA for other researchers as a valuable resource for training and evaluating language models in the medical domain.

Results

In this section, we present a detailed analysis of the results obtained from evaluating the performance of the BERT-based models on MedMCQA and MedQA datasets before and after fine-tuning on ontology-generated synthetic MCQA.

During fine-tuning, we retained the original architecture configuration of BERT, including the number of self-attention layers denoted as L, the number of heads denoted as A, and the hidden dimension of embedding vectors denoted as H, which were set to $L = 12, A = 12$, and $H = 768$ in this study. Fine-tuning was performed on a Tesla V100-PCIE-32GB GPU with a maximum sequence length of 128. A batch size of 16 and a learning rate of 5e-5 were selected for the fine-tuning process. It took approximately 42 hours to complete the fine-tuning process on BioOntoMCQA, and further fine-tuning on MedMCQA and MedQA took 28 hours and 5 hours, respectively. All accuracy results were derived from the average of 10 runs, each utilising the same experimental configurations.

After fine-tuning, we observed substantial improvements in the BERT model's performance on both evaluation datasets. Similarly, the finetuned SciBERT, BioBERT, BioOntoBERT and PubmedBERT models also exhibited impressive results. These results are shown in Table 3. The smaller size and lack of categorisation in MedQA influenced its moderate accuracy gains compared to the more substantial improvements in the structured MedMCQA dataset

Additionally, we conducted individual experiments with MCQs generated from each ontology property. Fine-tuning the BERT-base model solely on MCQs derived from subClass relationships resulted in higher accuracy compared to the accuracy achieved on MCQs based on the definition property and synonyms property, as shown in Table 4.

The variations in accuracy among the different experimental setups can be attributed to the nature of the generated MCQs. MCQs based on subClass relationships may offer a higher semantic similarity and context level, leading to slightly better performance. Conversely, MCQs generated from the definition and synonym properties may introduce a broader range of linguistic variations and nuances, presenting additional challenges for the models. While these results showcase the usefulness of various ontology properties in MCQ generation, they may not entirely represent the models' full capabilities. Factors such as the quantity, diversity of MCQs, and ontological relationship complexity significantly influence model performance.

We observed that using 50% of the BioOntoMCQA dataset for fine-tuning the BERT model resulted in the highest accuracy across both MedMCQA and MedQA datasets, as shown in Table 5. The finding suggests an optimal balance; smaller dataset proportions prevent overfitting by focusing on general patterns, while larger proportions may lead to memorisation of specific examples, reducing the model's generalisation ability. Thus, selecting the right amount of data for fine-tuning is crucial to maintaining a balance between leveraging the dataset's knowledge and ensuring the model's broad applicability in diverse medical question-answering scenarios.

Models like BioBERT, SciBERT, and PubmedBERT, pretrained on large medical corpora such as PubMed abstracts and articles, are being outperformed by models that undergo fine-tuning on synthetic datasets generated from ontologies. Intuitively, we can observe that the extensive corpus of abstracts and articles might not cover the foundational knowledge of the medical domain, as their primary purpose is to convey research results and not basic medical knowledge. In

Model	MedMCQA		MedQA			
	Validation	+ BioOntoMCQA	Validation	+ BioOntoMCQA	Test	+ BioOntoMCQA
BERT-base	35%	**43.2%**	31.1%	**35.8%**	30.9%	**36.9%**
BioBERT	38%	**46.9%**	32.4%	**36.6%**	33.1%	**33.8%**
SciBERT	39%	**46.5%**	32.8%	**36.8%**	33.5%	**34.8%**
PubmedBERT	40%	**49.7%**	33.7%	**37 .1%**	32.1%	**34.6%**
BioOntoBERT	42%	**47.1%**	34.2%	**36 .3%**	33.6%	**34.9%**

Table 3: Model Accuracy Evaluation with and without Fine-Tuning on BioOntoMCQA.

contrast, biomedical ontologies provide a fundamental understanding of the medical domain in a structured manner. This explains why incorporating ontological knowledge into the model leads to improvements.

Furthermore, our findings demonstrate that models fine-tuned on our synthetic dataset surpass the performance of BioOntoBERT (Sahil and P. Sreenivasa Kumar 2023), a model leveraging sentences generated from biomedical ontologies for pre-training. This outcome underscores the efficacy of task-specific fine-tuning. By aligning the model's learning process closely with the intricate demands of MCQA, task-specific fine-tuning ensures a more precise and effective application of the model's capabilities in practical scenarios.

The study highlights the effectiveness of fine-tuning BERT-based models on our ontology-generated synthetic dataset. The enhanced accuracy in answering medical multiple-choice questions underscores the importance of domain-specific fine-tuning in natural language processing applications.

Properties	MedMCQA	MedQA	
	Validation	Validation	Test
subClass	**43%**	**35.8%**	**36.2%**
Synonyms	42.2%	35.5%	35.4%
Definition	41%	34.2%	32.7%

Table 4: Comparison of Model Accuracy after Fine-Tuning using Synthetic MCQA Dataset using different Properties

Conclusions

This study introduces the novel BioOntoMCQA dataset, to the best of our knowledge, the first medical ontology-generated synthetic MCQA dataset. Fine-tuning BERT-based models on this dataset led to substantial improvement of accuracy figures for answering multiple-choice questions on the MedMCQA. Also, a moderate improvement was obtained on MedQA datasets. Incorporating ontology-generated MCQAs enabled the models to 'understand' the complexities of medical language better, resulting in enhanced performance.

Dataset	MedMCQA	MedQA	
Proportion (%)	Validation	Validation	Test
25	42%	34.3%	35.5%
50	**43.2%**	**35.8%**	**36.9%**
75	40%	34.9%	35.8%
100	38%	31.9%	33.4%

Table 5: Model Accuracy Comparison based on the Proportion of BioOntoMCQA dataset used for Fine-Tuning

The results underscore the significance of domain-specific fine-tuning in natural language processing applications. The BioOntoMCQA dataset proves to be a valuable resource for training language models in the medical field. This study also highlights the effectiveness of ontology-generated synthetic datasets in improving the accuracy and applicability of language models in specialised domains. As natural language processing advances, the adoption of fine-tuning strategies and domain-specific datasets will play a pivotal role in enhancing the capabilities of language models for specific tasks and datasets.

Future Work

Future work includes expanding the synthetic dataset with a broader range of medical questions, integrating more ontology properties, and exploring different fine-tuning strategies. Investigating the model's generalisation to related medical tasks and conducting human evaluations for practical insights are also important. Extending the approach to other specialised domains can unlock broader applications in natural language processing.

References

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25(1):25–29.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duncan, W. D.; Thyvalikakath, T.; Haendel, M.; Torniai, C.; Hernandez, P.; Song, M.; Acharya, A.; Caplan, D. J.; Schleyer, T.; and Ruttenberg, A. 2020. Structuring, reuse and analysis of electronic dental data using the oral health and disease ontology. *Journal of Biomedical Semantics* 11(1):1–19.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3(1):1–23.

Gündel, M.; Younesi, E.; Malhotra, A.; Wang, J.; Li, H.; Zhang, B.; de Bono, B.; Mevissen, H.-T.; and Hofmann-Apitius, M. 2013. Hupson: the human physiology simulation ontology. *Journal of biomedical semantics* 4(1):1–9.

Hao, Y.; Dong, L.; Wei, F.; and Xu, K. 2020. Investigating learning dynamics of bert fine-tuning. In *Proceedings of the 1st conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th international joint conference on natural language processing*, 87–92.

Harris, S.; Seaborne, A.; and Prud'hommeaux, E. 2013. Sparql 1.1 query language. *W3C recommendation* 21(10):778.

Hastings, J.; Ceusters, W.; Jensen, M.; Mulligan, K.; and Smith, B. 2012. Representing mental functioning: Ontologies for mental health and disease.

Hou, L.; Wu, M.; Kang, H. Y.; Zheng, S.; Shen, L.; Qian, Q.; and Li, J. 2020. Pmo: A knowledge representation model towards precision medicine. *Math. Biosci. Eng* 17:4098–4114.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11(14):6421.

Krech, D.; Grimnes, G. A.; Higgins, G.; Hees, J.; Aucamp, I.; Lindström, N.; Arndt, N.; Sommer, A.; Chuc, E.; Herman, I.; Nelson, A.; McCusker, J.; Gillespie, T.; Kluyver, T.; Ludwig, F.; Champin, P.-A.; Watts, M.; Holzer, U.; Summers, E.; Morriss, W.; Winston, D.; Perttula, D.; Kovacevic, F.; Chateauneu, R.; Solbrig, H.; Cogrel, B.; and Stuart, V. 2023. RDFLib.

Kwon, S.; Yu, J.; Park, S.; Jun, J.-A.; and Pyo, C.-S. 2021. Stroke medical ontology qa system for processing medical queries in natural language form. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 1649–1654. IEEE.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8449–8456.

Merchant, A.; Rahimtoroghi, E.; Pavlick, E.; and Tenney, I. 2020. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*.

Midhunlal, M., and Gopika, M. 2016. Xmqas-an ontology based medical question answering system. *International Journal of Advanced Research in Computer and Communication Engineering* 5(4):929–932.

Mosbach, M.; Khokhlova, A.; Hedderich, M. A.; and Klakow, D. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv preprint arXiv:2010.02616*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260. PMLR.

Rosse, C., and Mejino Jr, J. L. 2008. The foundational model of anatomy ontology. In *Anatomy ontologies for bioinformatics: principles and practice*. Springer. 59–117.

Sahil, and P. Sreenivasa Kumar. 2023. Leveraging biomedical ontologies to boost performance of bert-based models for answering medical mcqs. In *Proceedings of the 14th International Conference for Biomedical Ontologies (ICBO)*, volume 3603, 95–106. CEUR.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Schriml, L. M.; Arze, C.; Nadendla, S.; Chang, Y.-W. W.; Mazaitis, M.; Felix, V.; Feng, G.; and Kibbe, W. A. 2012. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research* 40(D1):D940–D946.

Visser, U.; Abeyruwan, S.; Vempati, U.; Smith, R. P.; Lemmon, V.; and Schürer, S. C. 2011. Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results. *BMC bioinformatics* 12(1):1–16.

Zhou, Y., and Srikumar, V. 2021. A closer look at how fine-tuning changes bert. *arXiv preprint arXiv:2106.14282*.