Medical Subject Classification with BERT on MedMCQA

John Bang
john_bang@berkeley.edu

Mohammad Paracha
m.paracha@email.com

Abstract

We classify medical exam questions from the MedMCQA dataset, which contains over 193,000 expert-authored questions from AIIMS and NEET PG exams across 21 medical subjects. We evaluate five models: Naive Bayes, logistic regression, BERT, BioBERT, and Gemini 1.5 Flash. Unlike previous work that focuses on answer prediction or open-domain medical reasoning, our study reframes the problem as subject-level classification—an underexplored yet critical task for downstream applications. Our results show that BioBERT achieves high precision scores, highlighting the effectiveness of domain-specific pretraining for understanding medical language.

1 Introduction

Question Answering (QA) systems are extremely important in natural language processing and large-scale, diverse medical QA datasets such as MedMCQA incorporate difficult exam questions mimicking real-world multiple-choice questions for medical entrance exams. Previous literature has confirmed bidirectional encoder representations from transformers (BERT) models fine-tuned on biomedical corpora are effective at medical question answering (Pal, Umapathi, 2022).

Consequently, this paper attempts to classify questions into clinical specialty subjects using five models: Naive Bayes, logistic regression, BERT, and BioBERT, and Gemini 1.5 Flash. Successful applications of these are automated call routing to correct medical specialty departments in hospitals and/or identifying missed diagnoses from a patient's medical encounter/visit notes.

2 Background and Models

Recent advances in natural language processing (NLP), particularly in large language models (LLMs), have led to significant gains in biomedical QA. Foundational work such as BioBERT (Lee et al., 2020) demonstrated that domain-specific pretraining on biomedical corpora like PubMed and PMC abstracts yields substantial improvements over general-purpose models like BERT across downstream tasks, including named entity recognition, relation extraction, and QA. These results underscore the importance of domain adaptation in achieving high accuracy in specialized fields like medicine.

In the multiple-choice QA setting, MedMCQA (Pal et al., 2022) introduced a comprehensive benchmark composed of over 190,000 expert-authored medical questions spanning 21 clinical subjects drawn from Indian medical entrance exams (NEET PG and AIIMS). Each question is accompanied by four answer choices and an explanation. Beyond assessing factual recall, MedMCQA enables evaluation of models' reasoning, subject classification, and generalization abilities, particularly given the subtle distinctions across overlapping medical domains.

Building on this, recent studies have explored how general-purpose LLMs perform in the clinical QA domain. Liévin et al. (2023) benchmarked models such as GPT-3.5, GPT-4, PaLM, and LLaMA on MedMCQA and related datasets, finding that while these models excel at memorization, their reasoning remains brittle without structured prompting or external knowledge retrieval. Brown et al. (2020) addressed this by introducing *few-shot prompting*, a strategy that provides example inputs to the model, causing significant performance gains in generalizing on tasks. Similarly, Lewis et al. (2024) proposed a variational retriever-reader model that captures uncertainty in open-domain QA, an approach that could help disambiguate under-specified medical questions.

Motivated by these developments, our work investigates the task of medical subject classification—predicting the medical discipline (e.g., Surgery, Pediatrics, Ophthalmology) associated

with a given MedMCQA question. We evaluate a spectrum of models of increasing complexity:

- **Baseline classifiers**: We apply traditional machine learning models such as multinomial Naive Bayes and logistic regression using TF-IDF features over the concatenated question and answer options.
- **Transformer models**: We fine-tune BERT and BioBERT on the subject classification task. BioBERT, with its biomedical vocabulary and pretraining, is expected to outperform general-purpose models due to its domain alignment.
- **Large language model**: We evaluate Gemini 1.5 Flash (2024), a multimodal LLM from Google, using few-shot prompting to classify subject areas, assessing how a commercially available model generalizes without fine-tuning.

3 Data

We are using the MedMCQA dataset to predict the medical subject (`subject_name`) of each multiple-choice question. As referenced earlier, MedMCQA consists of over 193,000 expert-written medical exam questions drawn from AIIMS and NEET PG, spanning 21 subjects in two Indian medical school exams.

To constrain the scope of the task, we omitted the correct option from the models and only used the question and options A–D as our input data. The other columns, while informative, provided a depth of information that introduced noise. The explanations contained complex rationale that was more targeted toward the answer (`cop`) column as opposed to the `subject_name`. Since `subject_name` was more complete with few missing values, we used it as the label.

4 Model Evaluation Approach

For evaluation, we use the F1 score as the primary metric, as it balances precision and recall, both critical in medical subject classification. This ensures that the model not only captures relevant cases but also avoids incorrect predictions, making F1 a suitable choice for assessing overall performance.

5 Feature Analysis and Fine-tuning

As part of our iterative experimentation process, we explored ways to improve model performance through both label consolidation and input representation enhancements.

5.1 Evaluating the "Unknown" Category

Throughout multiple iterations of model training and evaluation, the `Unknown` subject category consistently yielded the lowest F1 scores. Given its poor performance and the ambiguity it introduced, we ultimately made the decision to remove this category from our final model. Upon closer inspection, we found evidence suggesting that several questions labeled as `Unknown` may have been mislabeled. For example, consider the following question originally tagged as `Unknown`:

> "An 8-year-old child has a history since early childhood of malabsorption, ataxia, acanthocytes in the peripheral blood, and very low cholesterol and triglyceride levels. In addition, the patient has been developing progressive, bilateral, concentric contraction of the visual fields and loss of central vision. The underlying pathogenesis of this patient's disease is:"

This question contains strong clinical indicators relevant to `Pediatrics`, yet was not categorized as such. Cases like this suggest labeling inconsistencies within the dataset.

To further investigate, we conducted a BERTScore analysis comparing `Unknown` questions with those in known subject categories. The resulting F1 score of 0.82 indicates a high degree of semantic similarity between the two groups. This reinforces our suspicion that many `Unknown` questions are not fundamentally different in content, but rather under-annotated or mislabeled.

5.2 Subject Grouping Based on Confusion Patterns

Analysis of our confusion matrices revealed consistent misclassifications between certain subject areas. For example in our baseline models Medicine was predicted as Pathology in 10.5% of cases, Orthopaedics as Surgery in 27.6%, Pathology as Medicine in 14.3%, and Skin as Dental in 28%. Upon reviewing sample questions from these categories, the confusions were justifiable. For instance, Medicine and Pathology often cover overlapping diagnostic scenarios, and Surgery and Orthopaedics share common procedural language. Additionally, categories like Skin and Orthopaedics

had lower sample counts, further contributing to weaker performance due to limited training data.

To address these issues and reduce noise in classification, we consolidated closely related subjects into grouped categories: *Medicine & Pathology*, *Surgery & Orthopaedics*, and *Skin & Dental*. This grouping strategy not only aligns with the semantic overlap in the curriculum but also provides more robust training signals by increasing the effective sample size per class.

5.3 Combining Question and Answer Options

In our experiments, we enhanced the input representation by including the answer options alongside the question text. Concatenating the question with all four options (opa–opd) provides the model with a more comprehensive semantic context, aligning with how humans interpret multiple-choice questions—by evaluating the options to infer the underlying domain (e.g., antibiotics suggesting *Infectious Disease*, surgical tools pointing to *Surgery*). This strategy is particularly beneficial for underspecified or ambiguous prompts, where the subject becomes evident only through the nature of the answer choices.

For example, the question "Model analysis indicated for 2 quadrants:" is contextually ambiguous on its own and could plausibly belong to any of the subject classes. However, the answer options ("Hixon-old father," "Peck and peck index," "Pont's index," "All of the above") list dentition analysis terms and help ground the question in *Dental*. In such cases, the subject-revealing information lies more in the choices than in the question stem.

To empirically validate this design choice, we computed semantic similarity metrics between the question stem and the concatenated options: BERTScore F1 = 0.3692 and SBERT Cosine Similarity = 0.2558. These low similarity scores confirm that the options are not paraphrases or simple restatements of the question, but instead contribute distinct semantic content. Including all four options in the input allows the model to capture subject-specific signals that are otherwise absent from the question alone, thereby enhancing subject classification performance.

6 Results and Discussion

Both BERT Models performed strongly on the classification task. BERT performs strongly on Forensic Medicine (F1 = 0.90), Opthalmology (F1 = 0.91), and Social and Preventative Medicine (F1 = 0.87). Its corpus of training data contributes to its performance across subject areas. In Figure 1, the first word embedding layer, has 22 million parameters, with positional encoding for each word. These embeddings gain context in the transformer layer, which has 85 million parameters, and finally the output layers. This architecture guarantees that the model can learn signal from medical words, phrases, sentences and associate them against one another.

BioBERT builds upon this complex architecture with additional pretraining data on PubMED, a repository of biomedical literature. It includes 18 million words from PubMED articles and abstracts, which allows the model to see clearer relationships between medical words. As a test, we sampled an input question to see the embedding layers' functionality:

> "After periodontal surgery, epithelial reattachment to root by hemidesmosomes occurs within?"

BERT associated 'periodontal' and 'surgery' with a 0.68 cosine similarity and 'epithelial' and 'reattachment' with a similarity of 0.75. BioBERT had similarities of 0.80 and 0.83, respectively. This indicates that the model pretrained on biomedical literature learned the relationship between these sets of words more effectively.

BioBERT performed well with F1 scores of 0.79 to 0.90 across the subject classes. It was particularly strong on Biochemistry, Ophthalmology, and Skin and Dental, exceeding 0.90 F1 for those subjects.

In 2023, Liévin et al used the GPT 3.5 series to obtain between 60% and 78.2% accuracy on multiple choice questions on the USMLE and the MEDMCQA. Our suggested LLM, Gemini 1.5 Flash, released in September 2024 (and deprecated in September 2025) is a multimodal model supporting audio, images, videos, and text with an input token limit of 1,048,576 and an output token limit of 8,192. Being a general-purpose model, Gemini 1.5 Flash must be fine-tuned for higher performance.

Without few-shot examples and prompt-tuning, the baseline accuracy during experimentation of Gemini for classification resulted in an F1 score of 0.60. With few-shot examples in the system prompt, we gave the model five examples (five few shots) of questions for each subject (20 subjects). We used a simple but very specific prompt:

"""You are a medical expert. Your task is to classify the subject of each multiple-choice question. Choose only from these labels: ', '.join(le.classes$_)$.

Respond with only the subject name. Do not say anything else. Do not give a list.

Examples: shots

Now classify: 'current question'

Subject:"""

There are several adjustments we made. "Respond with the subject name" was an instruction we created because early iterations of Gemini on the dataset would provide unnecessary language. The model began responding with a list of subjects instead of simply one, so we added 'choose one.' In the prompt, the shots object contains 100 example questions, ensuring good representation. In this guided approach, we essentially give the model a few representative examples of each class for better classification. The model is tuned using 'few-shot' prompting (Brown, et al 2020). Using these techniques, we fine-tuned Gemini to an F1 score of 0.72 on 3,000 test samples.

Due to inference cost, we did not run Gemini on the full test set of 6,150 samples, but still obtained high F1 scores on Gynecology (0.90) and Opthalmology (0.94). However, there is high variance within the F1 scores across the subject categories.

7 Conclusion and Future Work

Our findings highlight the difficulty of subject classification in medical QA, even with advanced models. BioBERT, pretrained on biomedical literature, consistently outperforms general-purpose LLMs like Gemini 1.5 Flash, demonstrating the value of domain-specific pretraining. Despite advances in prompting and model scaling, domain adaptation remains critical for complex clinical reasoning.

Gemini was less competitive than the BioBERT model (F1 score of 0.85) and the BERT model (0.82) suggesting BERT's utility in real-world systems.

We also experimented with addressing class imbalance through oversampling of underrepresented subjects and undersampling of dominant ones. However, these adjustments led to only modest improvements in F1 scores. Future work could explore more sophisticated resampling methods to mitigate these effects.

Additionally, while our experiments focused on MedMCQA, the subject classification task can be extended and validated on other large-scale medical QA datasets such as PubMedQA, USMLE QA, and ClinicalQA. Comparative evaluations across these datasets would offer a more generalizable understanding of subject inference and specialization challenges in medical question answering.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

[2] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[3] Lewis, M., Wu, Y., Karpukhin, V., Fan, A., Kiela, D. (2024). Variational open-domain question answering. Retrieved June 14, 2025, from https://paperswithcode.com/paper/variational-open-domain-question-answering

[4] Liévin, V., Hother, C. E., Motzfeldt, A. G., Winther, O. (2023). Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143v4*. https://doi.org/10.48550/arXiv.2207.08143

[5] Pal, A., Umapathi, L. K., Sankarasubbu, M. (2022). MedMCQA: A large-scale multi-subject multiple-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, T. Naumann (Eds.), *Proceedings of the Conference on Health, Inference, and Learning* (Vol. 174, pp. 248–260). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v174/pal22a.html

A Appendix

See attached figures.

Table 1: F1 Score by Subject and Model

Subject	Bayes	Logistic	BERT	Fine-Tuned BERT	Fine-Tuned BioBERT	Gemini
Anaesthesia	0.11	0.61	0.60	0.81	**0.84**	0.79
Anatomy	0.61	0.64	0.65	0.82	**0.84**	0.72
Biochemistry	0.64	0.73	0.76	0.86	**0.90**	0.76
ENT	0.25	0.69	0.54	0.83	**0.86**	0.77
Forensic Med.	0.58	0.77	0.61	**0.90**	0.88	0.76
Gyn & Obs.	0.58	0.74	0.74	0.86	0.87	**0.90**
Medicine & Path.	0.49	0.63	0.55	0.79	**0.81**	0.57
Microbiology	0.57	0.68	0.56	0.82	**0.86**	0.61
Ophthalmology	0.48	0.81	0.75	0.91	0.90	**0.94**
Pediatrics	0.22	0.57	0.60	0.75	**0.80**	0.65
Pharmacology	0.64	0.69	0.64	0.85	**0.87**	0.79
Physiology	0.46	0.63	0.65	0.81	**0.84**	0.69
Psychiatry	0.23	0.70	0.11	0.86	**0.88**	0.33
Radiology	0.08	0.57	0.56	0.75	**0.80**	0.69
Skin & Dental	0.59	0.71	0.79	0.86	**0.90**	0.79
Social & Prev	0.70	0.73	0.55	0.87	**0.89**	0.52
Surgery & Ortho.	0.57	0.63	0.62	0.76	**0.79**	0.58
Overall	0.52	0.67	0.57	0.82	**0.85**	0.72

Note: BERT and BioBERT models were trained on 50% of the training set.

```
==========================================================================================
Layer (type:depth-idx)                                                  Param #
==========================================================================================
BertForSequenceClassification                                           --
├─BertModel: 1-1                                                        --
│    └─BertEmbeddings: 2-1                                              --
│    │    └─Embedding: 3-1                                              22,268,928
│    │    └─Embedding: 3-2                                              393,216
│    │    └─Embedding: 3-3                                              1,536
│    │    └─LayerNorm: 3-4                                              1,536
│    │    └─Dropout: 3-5                                                --
│    └─BertEncoder: 2-2                                                 --
│    │    └─ModuleList: 3-6                                             85,054,464
│    └─BertPooler: 2-3                                                  --
│    │    └─Linear: 3-7                                                 590,592
│    │    └─Tanh: 3-8                                                   --
├─Dropout: 1-2                                                          --
├─Linear: 1-3                                                           13,842
==========================================================================================
Total params: 108,324,114
Trainable params: 108,324,114
Non-trainable params: 0
==========================================================================================
```

Figure 1: BERT Layers

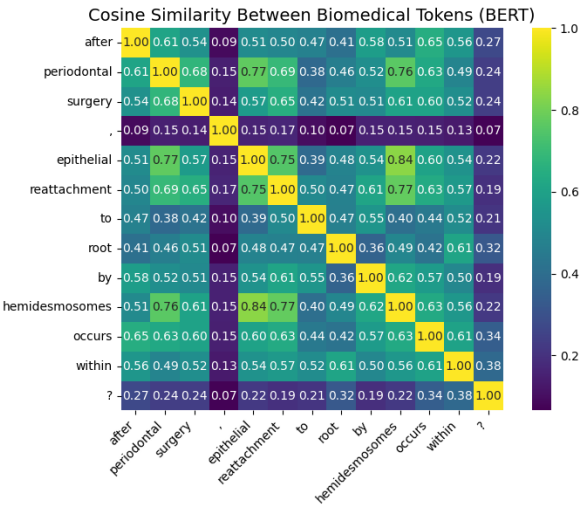

Figure 2: BERT Cosine Similarity

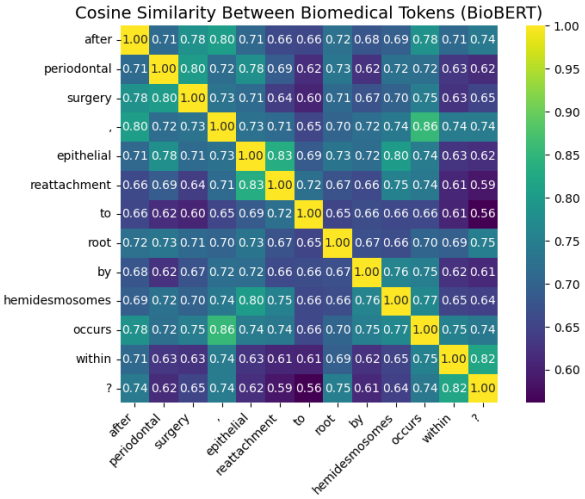

Figure 3: BioBERT Cosine Similarity