BPQA Dataset: Evaluating How Well Language Models Leverage Blood Pressures to Answer Biomedical Questions

Anonymous ACL submission

Abstract

Clinical measurements such as blood pressures and respiration rates are critical in diagnosing and monitoring patient outcomes. It is an important component of biomedical data, which can be used to train transformer-based language models (LMs) for improving healthcare delivery. It is, however, unclear whether LMs can effectively interpret and use clinical measurements. We investigate two questions: First, can LMs effectively leverage clinical measurements to answer related medical questions? Second, how to enhance an LM's performance on medical question-answering (QA) tasks that involve measurements? We performed a case study on blood pressure readings (BPs), a vital sign routinely monitored by medical professionals. We evaluated the performance of four LMs: BERT, BioBERT, MedAlpaca, and GPT-3.5, on our newly developed dataset, BPQA (Blood Pressure Question Answering). BPQA contains 100 medical QA pairs that were verified by medical students and designed to rely on BPs . We found that GPT-3.5 and MedAlpaca (larger and medium sized LMs) benefit more from the inclusion of BPs than BERT and BioBERT (small sized LMs). Further, augmenting measurements with labels improves the performance of BioBERT and Medalpaca (domain specific LMs), suggesting that retrieval may be useful for improving domain-specific LMs. [1]

1 Introduction

Clinical measurements, such as blood pressure and respiration rate, are crucial in healthcare for accurate diagnosis and disease monitoring. Misinterpreting these measurements can be life-threatening. For example, blood pressure readings (BPs), which indicate the force of blood against the walls of arteries, is a vital sign that is measured for all patients. Abnormal BPs are associated with various diseases, such as cardiovascular disorders and kidney disease. Accurate interpretation of such measurements is crucial for patient health.

Clinical measurements are ubiquitous in biomedical datasets which can be used to train and evaluate transformer-based language models (LMs). Many LMs (Singhal et al., 2022; Saab et al., 2024; Yang et al., 2022; Krishna et al., 2021; Jiang et al., 2023) were trained and evaluated (Luo et al., 2022; Jin et al., 2019; Pampari et al., 2018; Jin et al., 2020) on such datasets for healthcare applications. Appendix B shows that these data are rich in clinical measurements.

It is unclear, however, whether LMs can effectively interpret and use clinical measurements. Although existing medical benchmarks evaluate LMs' capability on answering medical questions, they contain substantial additional information. For example, Appendix B shows that MedQA, PubMedQA, emrQA contain both clinical measurements and additional contexts such as patient medical history, symptoms, and primary diagnoses (examples shown in Appendix C). This makes it challenging to isolate and assess LMs' performance on using clinical measurements alone, as LMs may rely on the other available information to answer questions. Previous studies on numerical reasoning over text have investigated the capability of LMs to understand and work with numbers (Dua et al., 2019; Wallace et al., 2019; Thawani et al., 2021; Wu et al., 2021) and all kinds of measurements Park et al., 2022. However, none focus specifically on clinical measurements, which directly relate to health outcomes and requires domain knowledge to interpret. Focusing on clinical measurements allows a more targeted evaluation of LMs' ability to use numerical data in a medical context.

Here we assessed how LMs (GPT-3.5 (Brown

[1] Our code and data are available at the anonymous GitHub Repo: https://anonymous.4open.science/r/BPQA-evaluating-LM-for-biomedical-QA-616C

Question Type	Example
Abnormality Detection under Special Context	. . . an 80-year-old individual has a blood pressure reading of 155/65 mmHg. Is this considered hypertension?
Intervention Opinion	. . . a patient with a blood pressure of 150/100 mmHg significantly reduces sodium intake, can this dietary change alone normalize their blood pressure?
Symptoms and Illness	. . . a patient reports snoring heavily and feeling fatigued during the day. They have a blood pressure of 150/95 mmHg. Could sleep apnea be the cause?
Medical Research	. . . an individual with cerebrovascular disease, had an initial blood pressure of 160/100 mmHg and a high amino-terminal-pro-B-type natriuretic peptide (NT-proBNP) level. Will perindopril-based blood pressure-lowering therapy help with this individual's situation?

Table 1: Examples of four types of questions in BPQA.

et al., 2020), MedAlpaca (Han et al., 2023), BERT (Devlin et al., 2018), and BioBERT (Lee et al., 2020)) can interpret and use BPs in medical question-answering (QA) tasks. We focused on BPs, a routinely monitored vital sign, as a case study because it is the most common clinical measurements. We selected the four models to compare the effect of different size, type (encoder or decoder), and pretrain corpus. We designed a new dataset called BPQA (Blood Pressure Question Answering) with QA pairs that are verified by medical students and designed to rely on BP. Using BPQA, we investigate the impact of BPs and their text label (low, normal, high) on the performance of the selected LMs in medical QA tasks. The results indicate: 1) GPT-3.5 (large sized LM) and MedAlpaca (medium sized LM) benefit more from the inclusion of BP measurements than BERT and BioBERT (small sized LMs), 2) augmenting labels to BPs improves performance for BioBERT and MedAlpaca (domain specific LMs), suggesting that retrieval may be useful for improving domain-specific LMs.

2 Method

We investigated the effect of adding BPs and labels (low, normal, high) on LMs' performances using BPQA, our newly designed medical QA dataset whose answers specifically depend on BPs. We compared the performances of four LMs: GPT-3.5, MedAlpaca, BERT, and BioBERT. Their sizes, types (encoder or decoder), pretrain corpus, and versions can be found in Appendix D. The models were evaluated across four variants using accuracy.

2.1 Data

We created a new dataset, BPQA which contains 100 medical QA pairs verified by medical students and designed to rely on BPs.[2] All questions are formatted to be binary (answer "yes" or "no"). The dataset is balanced, with 50 questions having 'yes" answers and 50 questions having "no" answers.

The dataset contains four categories of questions (25 QA pairs in each category) to evaluate LMs' ability of answering BPs-related questions in both clinical and research settings. The first three categories (Abnormality Detection under Special Context, Symptoms and Illness, and Intervention Opinion) are designed to assess LMs' ability to use BPs in clinical environments, similar to how doctors use BPs in patient care. These categories focus on identifying normal and abnormal BP ranges for specific populations, recognizing possible symptoms associated with specific BPs, and providing appropriate recommendations based on BPs, respectively. The fourth category, Medical Research, assesses LMs' ability to use BPs in research contexts, with manually selected and adapted BP-related questions from PubMedQA (Jin et al., 2019). Table 1 provides examples of the four question types.

We also created four variants of BPQA to assess the effect of adding BPs and adding BP labels (Examples shown in Table 2).

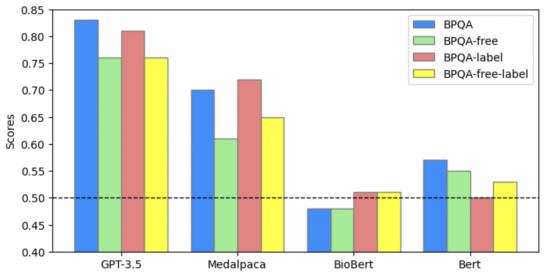
1. BPQA-free: Replaced the BPs with s/d (systolic and diastolic).
2. BPQA-label: Augmented categorical labels (low, normal, high) according to Centers for Disease Control and Prevention (CDC)[3]. Specific BP threshold is in Appendix A.2.
3. BPQA-free-label: Replaced the BPs with s/d and augmented categofical label.

[2]Two senior medical students reviewed the dataset in May 2024.

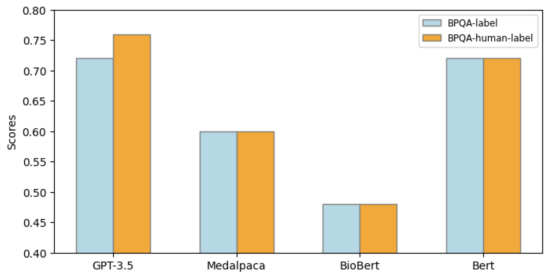
[3]https://www.cdc.gov/high-blood-pressure/about/index.html

Dataset	Example
BPQA	. . . blood pressure is **130/90** mmHg. . . Answer Yes or No.
BPQA-free	. . . blood pressure is s/d mmHg . . . Answer Yes or No.
BPQA-label	. . . blood pressure is 130/90 mmHg (high). . . Answer Yes or No.
BPQA-free-label	. . . blood pressure is s/d mmHg (high) . . . Answer Yes or No.
BPQA-human-label	. . . blood pressure for a pregnant woman is 130/90 mmHg (normal) . . . Answer Yes or No.

Table 2: Examples of BPQA dataset and its variants we created to assess the effect of adding BPs and augmenting labels. s/d is short for systolic/diastolic, whose definitions are in Appendix A.1.



(a) Zero-shot performance on BPQA shows that larger LMs benefit more from seeing BPs, label augmentation helps domain specific LMs, and GPT3.5 performs better with raw BPs.



(b) LMs Performance on the Special Context questions shows that GPT-3.5 benefits from context specific labels.

Figure 1: Comparison of model performance on different BPQA variants.

4. BPQA-human-label: Augmented context specific categorical labels from our review.

In BPQA-human-label, the human assigned labels are used only for the Special Context questions. This category requires human labels because minority groups, such as pregnant women and infants, have BP evaluation different from CDC guidelines, while questions in other categories can be adequately labeled using CDC guidelines. For example, a blood pressure of 130/85 mmHg would be labeled as "high" in BPQA-label (according to CDC guidelines). However, this reading is considered "normal" for a pregnant woman. In the BPQA-human-label dataset, such a reading would be labeled as "normal" to account for the specific context of pregnancy. 13 out of 25 Abnormality Detection under Special Context questions was changed compare to BPQA-label.

2.2 Experiments

We performed zero-shot QA evaluation for selected LMs on all variants of BPQA. We chose zero-shot evaluation to measure LMs' inherent capabilities in using BPs, avoiding potential biases introduced by fine-tuning or few-shot learning. For BERT-based LMs, we used fill-mask classification with single token (details in Appendix E); for MedAlpaca and GPT-3.5, we used text-generation with [Yes/No] choices in the prompts.

3 Results

GPT-3.5 and MedAlpaca (a large and a medium sized LM) benefit more than BERT and BioBERT (small sized LM) from seeing BPs. Figure 1a shows that GPT-3.5, MedAlpaca, and BERT achieve higher accuracy (9%, 15%, 3% gain respectively) when tested on the BPQA (blue bar) compared to the BPQA-free (green bar). However, BioBERT (third column) shows no change between the blue and green bars. This suggests that larger LMs may leverage BPs more effectively when answering medical questions, while smaller LMs may benefit less or do not utilize BPs at all.

Although all the questions are designed to rely solely on BPs, **some LMs achieve above random accuracy for BPs-redacted questions**, possibly because they used contextual information in the question as a spurious correlation. The green bars in Figure 1a show that BERT, MedAlpaca, and GPT-3.5 all reach accuracy above 0.5 (dashed line) on BPQA-free (green bar, BPs removed). We speculate that BP-related tokens (such as "old" and "pregnant" because they are risk factors for hypertension) in the questions may contribute to this.

LMs' performances vary on the BPQA-label dataset: **Seeing BPs with label augmenta-**

tion helps BioBERT and MedAlpaca (domain-specific LMs), but does not help or hurts BERT and GPT-3.5 (small and large sized general LMs). Figure 1a shows that adding BP labels (red bar) improves BioBERT and MedAlpaca's performances from original (blue bar) by 6% and 3% respectively. However, label augmentation hurts BERT's performance by 12% (the red bar is lower than the blue bar). We speculate this drop might be related to the increased sentence complexity from the inserted label. GPT-3.5's performance remains largely unchanged with a 1% decrease from label augmentation (blue versus red bar).

GPT-3.5 peforms better with raw BPs and without CDC labels. Focusing on the first column of Figure 1a, we compare GPT-3.5's performance on datasets with (blue and red bars) and without (green and yellow bars) raw BPs. The best performance (0.83) comes from original dataset with raw BPs (blue bar). The second best (red bar) has a 1% drop with augmented CDC label. And the worst peformance (0.76) are variants without BPs (green and yellow bar). This suggests GPT-3.5's performance hurt from including CDC labels and benefits from seeing raw measurements.

Context-specific labels helps GPT-3.5 achieve higher accuracy. Focusing on the Special Context questions (questions related to minority groups like pregnant women, whose BP evaluation might be different from CDC guidelines) in Figure 1b, GPT-3.5 improves its performance by 6% on BPQA-human-label compared to BPQA-label (orange versus blue bar) after human-labeling to conform to the specific contexts. This shows the incompatibility between BPs under special contexts and the general CDC labels might have hurt GPT3.5's performance on BPQA-label. For example, 130/85 mmHg is labeled "high" by CDC but is actually "normal" for a pregnant woman. With context-specific labeling, GPT-3.5 benefits from the label augmentation. Please refer to Appendix G for detailed explanation on BPQA-human-label result.

4 Discussion

The positive effect of context-specific labels suggests that patient-context-aware **retrieval augmentation may improve clinical language model's generalization to minority patients**, such as pregnant women with unique needs and vitals. While we manually modified CDC labels to special contexts, retrieval models (Zakka et al., 2024) can incorporate relevant contextual information from specialized database during the training and inference, helping LMs better understand and respond to the variability of different demographics. The negative effect of using general CDC labels also indicates the importance of accounting for the specific needs of different patient demographics.

Our work highlights the **need for more specialized benchmarks for clinical LMs**. In our case, the skill of understanding clinical measurements is crucial for practical use, but our early experiments showed that existing benchmarks such as MedQA involve too much additional information to target clinical measurements (see Appendix H and Appendix C for details). Our BP-targeted BPQA dataset showed the disparity in LMs' abilities to leverage BPs and the room for improvements. While BPQA serves as an initial step, we need more specialized benchmarks for targeted evaluation of LMs' skills in interpreting and reasoning over quantitative medical data.

A future direction for improving LM's skills with clinical measurements is modifying tokenizers. The unique numeric format in medical contexts, like special characters and units (e.g. 130/85 mmHg), may benefit from specialized tokenizers. Most existing tokenizers (Kudo and Richardson, 2018; Sennrich et al., 2016) are designed for natural text and may struggle with such formatted numerical representations while preserving quantitative semantics.

5 Limitations

Our study used BPs as a representative for clinical measurement, which does not fully capture the diversity of clinical scenarios. We also did a very preliminary study with a synthetic dataset (only 100 QA pairs) on 4 LMs. Future works include improving the dataset's breadth by including more types of measurements, more rigorously testing our finding with more LMs, testing the effect of retrieval model on tasks involving minority groups, and designing tokenizers that work well with clinical measurements.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Lavender Jiang, Xujin Liu, Nima Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, and Eric Oermann. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619:1–6.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Preprint*, arXiv:1808.06226.

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

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Sungjin Park, Seungwoo Ryu, and Edward Choi. 2022. Do language models understand measurements? *arXiv preprint arXiv:2210.12694*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *Preprint*, arXiv:2404.18416.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Preprint*, arXiv:1508.07909.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Qinzhuo Wu, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2021. Math word problem solving with explicit numerical values. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5859–5869, Online. Association for Computational Linguistics.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *Preprint*, arXiv:2203.03540.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Stacey Tullis, Melissa Ann Vogelsong, John Patrick Cunningham, and William Hiesinger. 2024. Almanac - Retrieval-Augmented language models for clinical medicine. *NEJM AI*, 1(2).

A Synthetic BPQA datasets

A.1 BP Introduction

BP is recorded as two numbers, in the form of systolic/diastolic mm Hg (millimeters of mercury): Systolic blood pressure indicates how much pressure blood is exerting against your artery walls when the heart contracts. Diastolic blood pressure indicates how much pressure blood is exerting against your artery walls while the heart muscle is resting between contractions.

A.2 BP Threshold

The threshold we used for labeling BP refers to Centers for Disease Control and Prevention (CDC).

Systolic BP (s) (mmHg)	Diastolic BP (d) (mmHg)	Label
$s < 90$		Low
$90 \leq s < 120$	$d < 60$	Low
	$60 \leq d < 80$	Normal
	$d \geq 80$	High
$s \geq 120$		High

Table 3: BP Threshold

B Examples of Datasets Involving Clinical Measurements

Table 4 shows that biomedical data used to train and evaluate LMs are rich in clinical measurements.

C Substantial Information Contained in Example Questions of MedQA, PubMedQA, and emrQA

1. MedQA: A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He denies gross hematuria or pelvic pain. He is sexually active with his girlfriend, and they consistently use condoms. . . . His mother has rheumatoid arthritis. **The patients temperature is 99 F (37.2 C), blood pressure is 112/74 mmHg, and pulse is 81/min.** On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is positive for leukocyte esterase. A urine microscopic evaluation shows a moderate number of white blood cells but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms? A: Chlamydia trachomatis, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum

2. PubMedQA: . . . Associations were assessed by logistic regression with respect to systolic, diastolic and pulse pressure, with adjustment for education, work status, physical activity, smoking, body mass and lipid levels. . . . In the prospective study of disease-free women, baseline pulse pressure and systolic pressure were inversely associated with risk of low back pain **[odds ratio (OR) 0.93 per 10 mmHg increase in pulse pressure, 95% confidence interval (CI) 0.89-0.98, p=0.007; OR 0.95 per 10 mm Hg increase in systolic pressure, 95% CI 0.92-0.99, p==0.005.]** . . . Does high blood pressure reduce the risk of chronic low back pain?

3. emrQA: 08/31/96 ascending aortic root replacement with homograft with omentopexy. The patient continued to be hemodynamically stable making good progress. Physical examination: **BMI: 33.4** Obese, high risk. **Pulse: 60. resp. rate: 18**. Has the patient ever had an abnormal BMI?

Here are example questions from MedQA, PubMedQA, and emrQA. While these questions contain clinical measurements (highlighted in bold), they often provide additional context (shown in red) that allow LMs to infer the answer without directly interpreting or using the measurements.

Dataset	Example
PubMed abstracts	. . . estimation error of **-2.06 ± 6.89 mmHg** for systolic BP, and **0.89** and **-4.66 ± 4.91 mmHg** for diastolic BP. . .
MIMIC-III (Johnson et al., 2016)	includes vital signs, medications, **laboratory measurements**, and more.
MedQA	. . . vital signs are: blood pressure, **148/90 mm Hg**, heart rate, **88/min** . . .
PubMedQA	. . . mortality rates doubled at **< 100 mm Hg**, tripled at **< 90 mm Hg** and were 5- to 6-fold at **< 70 mm Hg**, irrespective of age . . .
emrQA	. . . physical examination: BMI:**33.4**. Pulse: **60**. resp. rate: **18**. . .

Table 4: Examples of datasets involving clinical measurements

Model	Size	type
Bert	110M	encoder
BioBERT	110M	encoder
MedAlpaca	6.74B	decoder
GPT-3.5-turbo-instruct	NA	decoder

Table 5: We selected encoders and decoders models of different sizes and pretrain corpus

D Model Selection

Our model selection allows us to observe the performance of models with differing size, type, and pre-training data. See table 5 for sizes and type of selcted models. BERT is trained on Wikipedia articles and Book Corpus. BioBERT, on top on Bert, is trained on biomedical articles from PubMed abstracts. MedAlpaca is trained on a variety of medical texts, encompassing resources such as medical flashcards, wikis, and dialogue datasets. GPT-3.5 is trained on a large and diverse of data like webtexts, books, and Wikipedia. Here, we used **gpt-3.5-turbo-instruct** whose training data was updated up to Sep 2021.

E Fill-mask Classification for BERT Based Model

For BERT and BioBERT, we used fill-mask classification with single token. The BPQA datasets in 2 are modified by removing "Answer Yes or No" and adding "The answer is [MASK]". The candidates for [MASK] are constrained to "yes" and "no".

F Full Evaluation Results

Model	BPQA	BPQA -free	BPQA -label	BPQA -free-label
GPT-3.5	0.83	0.76	0.82	0.76
MedAlpaca	0.70	0.61	0.72	0.65
BioBERT	0.48	0.48	0.51	0.51
Bert	0.57	0.55	0.50	0.53

Table 6: Evaluation results of different LMs on various BPQA datasets.

G BPQA-human-label Result Explanation

13 out of 25 questions in the Special Context group were modified based on the need to align with specific question contexts. Focusing on the 13 modified questions, the performances of BERT, BioBERT and MedAlpaca showed no change before and after the modification. GPT-3.5 increased its performance from 0.61 to 0.69, indicating that GPT-3.5 is able to detect the trivial change and benefit from context-specific labeling.

H Experiments using MedQA-USMLE

On the early stage of our study, we extracted 3500 questions that contain BPs in the MedQA-USMLE benchmark as the evaluation dataset (**MedQA-BP**), and created a variant with BPs removed (**MedQA-BP-free**). We tested BERT, BioBERT and GPT-3.5 on the two datasets. Looking at the results in Table 7, the largest performance gap for all the models between MedQA-BP and MedQA-BP-Free is 0.01, which we believe is not significant enough to support any claims. This might be because questions in MedQA contain abundant additional information besides BPs, so adding or removing single BPs can only cause trivial differences on model performances. In this case, our experiments cannot be optimally employed on MedQA. Thus, we developed a synthetic BP-targeted dataset to display more significant results.

Model	**MedQA-BP**	**MedQA-BP-Free**
GPT-3.5	0.521	0.531
BioBERT	0.262	0.252
Bert	0.242	0.249

Table 7: Result on MedQA-BP