MedQA-SWE

A Clinical Question & Answer Dataset for Swedish

Niclas Hertzberg[1,2], **Anna Lokrantz**[1,2]
[1]AI Sweden
[2]Xsilico Ai
niclas.hertzberg@ai.se, anna.lokrantz@ai.se

Abstract

Considering the rapid improvement of large generative language models, it is important to measure their ability to encode clinical domain knowledge in order to help determine their potential utility in a clinical setting. To this end we present MedQA-SWE – a novel multiple choice, clinical question & answering (Q&A) dataset in Swedish consisting of 3,180 questions. The dataset was created from a series of exams aimed at evaluating doctors' clinical understanding and decision making and is the first open-source clinical Q&A dataset in Swedish. The exams – originally in PDF format – were parsed and each question manually checked and curated in order to limit errors in the dataset. We provide dataset statistics along with benchmark accuracy scores of seven large generative language models on a representative sample of questions in a zero-shot setting, with some models showing impressive performance given the difficulty of the exam the dataset is based on.

Keywords: Datasets, Clinical text, Evaluation

1. Introduction

Clinicians today face a substantial administrative workload (Gaffney et al., 2022; Toscano et al., 2020) and the continuously improving capabilities of large transformer-based language models (Vaswani et al., 2017) has led to questions regarding their potential usefulness in the healthcare sector. Recently, models like GPT-4, Med-PaLM, and Med-PaLM-2 have received passing scores on tests based on the *United States Medical Licensing Examination* (USMLE) (OpenAI, 2023; Singhal et al., 2022, 2023). Medical devices based on large language models (LLMs) could potentially benefit the healthcare sector by alleviating some of the administrative burdens faced by clinicians today by aiding in tasks such as information retrieval, summarizing text, radiologic decision making and clinical note generation (Dave et al., 2023; Moons and Bulck, 2023; Patel and Lam, 2023; Rao et al., 2023).

However, there are considerable risks associated with deploying LLMs in the clinical domain (Tian et al., 2023). Besides ethical and legal issues (Harrer, 2023) LLMs have a tendency to hallucinate (Ji et al., 2023) and prior to deployment they need to be thoroughly evaluated – in a multitude of ways – in order to better understand the potential and limitation of each particular model.

One such evaluation method involves testing the degree to which the model has parameterized knowledge of the clinical domain. Language models have the potential to encode knowledge (Petroni et al., 2019a), and the degree to which a model has parameterized knowledge in some domain have previously been evaluated using multiple choice question and answer datasets (MCQA) (Hendrycks et al., 2021).

Solving MCQA-tasks requires the model to correctly answer a question by selecting the correct alternative(s) from a number of candidate alternatives (Rogers et al., 2023).

Open-sourced MCQA datasets have frequently been used in order to evaluate language models on particular tasks, including those related to the clinical field (Singhal et al., 2022, 2023; Nori et al., 2023). However, high scores on clinical MCQA exams in English by certain LLMs – trained primarily on English text – does not necessarily imply a similar proficiency on non-English exams (Petrov et al., 2023; Zhang et al., 2023).

One advantage of testing a model on clinical MCQA tasks is that it is a relatively straightforward evaluation – it requires no input from clinicians nor any access to patient data. These two issues could otherwise hinder investigations into the capabilities of medical devices based on LLMs by limiting the number of participants to those with access to patient data and clinicians.

Several clinical MCQA datasets now exist, primarily in English (Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2021)[1], English and Chinese (Jin et al., 2020) and Spanish (Vilares and Gómez-Rodríguez, 2019).

Swedish on the other hand, is an under-resourced language (Holmström et al., 2023) and as such lacks the comparatively large number of datasets available for more high-resource languages. To the best of our knowledge there are currently no clinical MCQA datasets available in

[1]A subset of the MMLU dataset called clinical knowledge

LREC-COLING 2024, pages 11178–11186
20-25 May, 2024.

Swedish.

Consequently, we present MedQA-SWE[2], a dataset consisting of 3,180 multi-choice questions in Swedish that aims to test doctors' clinical knowledge and decision making. The MedQA-SWE dataset was created from exam questions posed in the theoretical exam given to assess the knowledge of foreign doctors wanting to obtain a Swedish medical license. The dataset consists of clinical context-based questions with 5 answer alternatives per question and one correct answer. The dataset aims to test clinical knowledge and decision making in a variety of ways. For example, a short synthetic patient report may be given and the task could range from determining the most likely disease, or what care a patient should receive and what necessary actions need to be taken, to more general questions, such as "The epineurium is the outermost layer that surrounds a peripheral nerve. What is the peripheral nerve epineurium made of?". We refer to section 2 for an in-depth description of the dataset.

Our primary contributions in this work are as follows,

- We present the MedQA-SWE dataset, the first clinical Q&A datset in Swedish.
- We furthermore evaluate seven LLMs on a sample of 300 questions and provide benchmark accuracy scores for each model.
- We show that while the test is difficult for most open-sourced models - Falcon-180B, GPT-3.5 and especially GPT-4 performs very well.

2. Dataset

2.1. Description

Medical doctors who received their license outside the EU/EES are not automatically granted license to practice medicine in Sweden. On behalf of the National Board of Health and Welfare, Umeå University is since 2016 responsible for the Swedish medical licensing examination. Prospective candidates need to demonstrate both sufficient theoretical knowledge and practical skills in order gain their Swedish medical license. The theoretical part is a standardized exam, while the practical part involves assessment of routine clinical tasks. The exam, known as the *Kunskapsprov för läkare* ("knowledge exam for doctors"), is given several times a year, usually four, and exam papers from previous years are made available on Umeå University's website (University, 2023a).

[2] https://huggingface.co/datasets/nicher92/medqa-swe

MedQA-SWE is based on the theoretical part of this exam, which is divided into three parts:

- *"Pre-clinical"*, which usually includes general clinical Q&A, e.g. *"Anemia due to iron deficiency is common. What is the typical blood imaging for iron deficiency?"*. The number of questions in this part of the exam is usually around 140.
- *"Clinical cases"*, which usually include a description of a synthetic patient followed by a question. The description varies in length from a few sentences to a more substantial piece of text which might include parts of a patient's medical record, results from a blood test, symptoms etc. This part usually comprises around 30 questions. One distinct feature of this subset of the dataset are reoccurring clinical cases that build on a specific patient's previous clinical case. Furthermore, the answer to the previous question is sometimes part of the input to the next question. For example – question $n-1$ might ask about where to send a patient given some background information, question n could then contain all of the the previous information and in addition where the patient was sent, ie the answer to question $n-1$ and pose a new question about what to do next. We refer to appendix A for examples of exam questions.
- *"Scientific article"*, consist of around 15 questions about a scientific article. Some of the scientific articles used might not be open source, we therefore omit these questions from MedQA-SWE.

For each context and question $i \in \{1, 2, \ldots, n\}$, there is one correct answer among a set of candidate answers $a_i \in \{1, 2, \ldots, 5\}$.

Passing the exam requires an ability to understand, reason and draw conclusions from the information provided in order to select the correct answer from the set of available alternatives. Some questions in the pre-clinical and clinical part also include images as part of the information given. To receive a passing grade, at least 50% of the questions in the clinical cases part of the exam need to be correctly answered and a minimum total score of 60% needs to be achieved (University, 2023b).

2.2. Dataset Collection

The data was originally in a PDF format with each PDF corresponding to part of an exam taken that year, there were 20 exams in total. The PDFs were generally structured in one of two ways which required two different parsing approaches. The exam papers prior to, and including, the one for the exam given on 2020-09-10 were possible to parse using text extraction libraries whereas more recent exams required parsing by Optical Character Recognition (OCR).

For the text extraction algorithm we relied heavily on regular expressions to split each PDF into sections corresponding to each question, related text and options. We then further structured each section into the background text and question in one part, the five alternatives in another, and the correct answer(s) in its separate key-value pair. The OCR-based algorithm was similar, but required prior processing of the PDF using OCR.

The automatic identification of the correct choice among the optional answers to each question involved identifying the choice denoted by tick marks in the PDF. Some questions, usually two to three per exam, had more than one correct answer – we removed those questions and kept only questions with one correct answer. We furthermore checked the exams for duplicate questions, of which 5 was found and subsequently removed.

Extensive curation was needed once the data had been collected as we wanted to discard as little of the data as possible. See 2.3 for a thorough discussion of quality issues and how we resolved them.

The finished dataset was saved in a CSV-format, with separate columns for questions, answer options, correct answers, dates of exam papers and parts of exam.

2.3. Quality Checks and Possible Data Issues

In order to reduce the risk of errors in the dataset and ensure proper formatting we manually inspected all of the parsed questions and compared them to the corresponding PDF the question was extracted from.

There were four main issues found in the parsed dataset:

1. Incorrect formatting of the answer, question or the alternatives
2. Images as part of the information required to answer the question
3. Correct answer not found
4. OCR errors, often related to the Swedish letters Å,Ä and Ö

The first three issues were relatively simple to detect and resolve manually. Incorrect formatting was remedied by manually restructuring the question into the desired format. Questions containing images or graphs as part of the information required to answer were removed and whenever the algorithm failed to automatically detect the correct answer we manually added it.

Problems with the OCR occasionally occurred when detecting the Swedish characters “Å",“Ä" and “Ö", which caused the parsed text to contain additional spaces next to some of those characters. We alleviated this issue with the use of regular expressions and some manual curation. Nevertheless, with a total of over 3000 questions – some being over 1000 words long – it's challenging to guarantee their quality when reviewing them manually.

Furthermore, some of the PDFs contained mathematical notation that might not transfer well to our dataset format. Considering that the mathematical notation issue was rare and its overall effect on the data unknown, we left it as is.

2.4. Statistics

In this section, we provide some statistics for the dataset. The correct answers were nearly uniformly distributed, with each option among A, B, C, D, E being close to equally probable, as can be seen in Table 1 below.

Correct answer	Number of occurences
A	667
B	659
C	618
D	598
E	638

Table 1: Correct answer alternatives distribution

In Table 2, we present the total number of questions, as well as the number of questions from each part of the exam together with maximum and average lengths of questions and answers respectively.

	Pre-clinical	Cases	All
Questions	2,656	524	3,180
Avg Q words	42.1	204.5	68.8
Max Q words	267	1667	1667
Avg A words	4.1	4.7	4.2
Max A words	34	22	34

Table 2: Dataset statistics, averages rounded to the nearest tenth, words are counted by splitting on whitespace. Q = Question + Background, A = Individual answer alternatives

3. Benchmarking

We provide benchmarks for MedQA-SWE by evaluating the zero-shot results, i.e. when no prior examples or other information is given, of seven

LLMs. All tested models had been trained on some amount of Swedish data. The models used were the open sourced GPT-SW3 (Ekgren et al., 2023), Llama2-70B-chat and Llama2-13B-chat (Touvron et al., 2023), Falcon-180B-chat and Falcon-40B-instruct (Phillip Schmid, Omar Sanseviero, Pedro Cuenca, Leandro von Werra, Julien Launay, 2023) and finally OpenAI's GPT-3.5-turbo and GPT-4 (via API), both successors to GPT-3 (Brown et al., 2020).

The open source models were loaded in float16 because of computational limitations. Additionally, due to its size the Falcon-180B-chat model was quantized and loaded with Int-4 weights (Wu et al., 2023).

The evaluation dataset was created by randomly sampling 10 pre-clinical questions and 5 clinical cases from each exam for a total of 300 questions. We took each context and question, added "Choose an alternative" and two newlines before the 5 alternatives and used this as input to each model.

For each question, we prompted the models to generate the correct answer given the answer options. The decoding algorithm used was greedy search, i.e choosing the most probable next token for the entire generation. We report the accuracy of each model in Table 3.

The same prompt – in Swedish – was used for all seven models, with slight modifications to adapt to the syntax of each model. For example, the Llama2 models were prompted with special tokens "[INST]" and "[/INST]" before and after each input, as per recommendations (Huggingface, 2023). See appendix B for the prompt and its English translation.

Outputs from the models varied slightly and were post-processed for ease of comparison to the correct answer. Parentheses, spaces and newlines were removed from each output and the first letter was counted as the answer given by the model.

The larger models performed better overall and the two models accessed via API (GPT-3.5 and GPT-4) performed the best. However, the current regulations regarding patient data in Sweden prevents the API models from being used in a clinical settings that involve patient data. Therefore, there is a clear distinction between models that could potentially be used by clinicians and those that could not and from a clinical utility perspective the results of the API-models are not very relevant.

GPT-SW3 performed surprisingly poorly on the dataset, considering it has been trained on a substantial amount of Swedish text. Several of the models trained primarily on English performed relatively well on the exam. For example: LLama2 was trained on only 0.15% Swedish data while the main text-source the Falcon models were trained on was *the refined web* which contains roughly 1.35 % Swedish data (Penedo et al., 2023). Therefore

Name	**All**	**Pre-clinical**	**Cases**
Random	20.0%	20.0%	20.0%
GPTSW3-20b-inst	22.3%	24.0%	19.0%
Llama2-13b-chat	29.6%	26.5%	36.0%
Falcon-40b-instruct	41.6%	41.0%	43.0%
Llama2-70b-chat	45.3%	42.5%	51.0%
Falcon 180B-Chat	57.3%	59.5%	53.0%
Pass	**60.0%**	**NA**	**50.0%**
GPT-3.5-turbo	60.0%	62.0%	56.0%
GPT-4	84.3%	86.0 %	81.0%

Table 3: Benchmarking of seven LLMs on a sample of MedQA-SWE

the amount of Swedish data the models have been trained on seem to have little correlation with their performance on this task.

Furthermore, several of the models scored unevenly on the different parts of the exam, for example: Llama2-13B-chat performed similarly to GPT-SW3-20B-instruct on the pre-clinical part of the exam despite performing almost twice as well on the clinical part.

4. Conclusions and Future Work

Our work contributes the first clinical Q&A dataset in Swedish and our experimental results indicate that some models perform well on the task, in particular GPT-4. The impressive results achieved by the larger models on this difficult exam – even with minimal prompt engineering – suggest further exploration in the direction of LLMs to solve clinical MCQA tasks.

Future work might include further evaluation on the dataset along with prompt-engineering, prompt-tuning and fine-tuning approaches. The current method of evaluation requires the models to follow instructions well enough to format the output in a particular way. Although the models generally accomplished this task quite well it is worth exploring other methods, which might have caused the models to perform differently.

Furthermore, our dataset only represents one particular task of interest in clinical NLP, it is not fully representative of tasks that would be of interest for clinicians in their every day job. Therefore, future work might also include the creation of datasets that can be helpful in creating solutions that meet practicing doctors' specific needs, for example synthetic patient data.

5. Ethical Considerations

Since the dataset is related to the clinical domain we feel compelled to alleviate potential privacy concerns. It is therefore worth noting that Sweden has stringent patient data laws and regulations that permit actual patient data from being shared. Therefore none of the patients in MedQA-SWE are based off of any real patient, or else the original exam papers that MedQA-SWE was created from could not have been open sourced in the first place.

We hope that by making this dataset available to the community, we will encourage further research into applications of LLMs in the clinical domain and promote the development of ethically sound solutions that have the possibility to aid clinicians in their work.

6. Conflict of Interest

Nothing to report.

7. Acknowledgements

We would like to thank our colleagues at AI Sweden and Xsilico Ai for their support, Umeå University for keeping the exam papers open-source and Region Halland for allowing us to use their computing resources.

8. Bibliographical References

Johanna Berg, Carl Aasa, Bjorn Thorell, and Sonja Aits. 2023. Openchart-se: A corpus of artificial swedish electronic health records for imagined emergency care patients written by physicians in a crowd-sourcing project.

Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2023. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Peter Cangialosi, Brian Chung, Torin Thielhelm, Nicholas Camarda, and Dylan Eiger. 2020. Medical students' reflections on the recent changes to the usmle step exams. *Academic Medicine*.

Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. 2023. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6.

Kent Dezee, Anthony Artino, D Elnicki, Paul Hemmer, and Steven Durning. 2012. Medical education in the united states of america. *Medical teacher*, 34:521–5.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. Gpt-sw3: An autoregressive language model for the nordic languages.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2023. Medalign: A clinician-generated dataset for instruction following with electronic medical records.

Adam W. Gaffney, Stephanie Woolhandler, Christopher Cai, David H. Bor, Jessica Himmelstein, Danny McCormick, and David Himmelstein. 2022. Medical documentation burden among us office-based physicians in 2019: A national study. *JAMA internal medicine*.

Steven Haist, Peter Katsufrakis, and Gerard Dillon. 2013. The evolution of the united states medical licensing examination (usmle). *JAMA : the journal of the American Medical Association*, 310:2245–6.

Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, 90.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Huggingface. 2023. Llama 2 is here - get it on Hugging Face. `https://huggingface.co/blog/llama2#how-to-prompt-llama-2/`. Accessed: 2023-10-18.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Philip Moons and Liesbet Van Bulck. 2023. Chatgpt: Can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. *European journal of cardiovascular nursing*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

OpenAI. 2023. Gpt-4 technical report.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet. Digital health*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019a. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019b. Language models as knowledge bases?

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages.

Phillip Schmid, Omar Sanseviero, Pedro Cuenca, Leandro von Werra, Julien Launay. 2023. Spread Your Wings: Falcon 180B is here. `https://huggingface.co/blog/falcon-180b`. Accessed: 2023-10-15.

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

Abdul Sohail Rao, J. N. Kim, Meghana Kamineni, Minxia Pang, Wh Lie, and Marc D. Succi. 2023. Evaluating chatgpt as an adjunct for radiologic decision-making. *medRxiv : the preprint server for health sciences*.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Malik Sallam and Affiliations. 2023. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. In *medRxiv*.

Nigam H. Shah, David N. Entwistle, and Michael A. Pfeffer. 2023. Creation and adoption of large language models in medicine. *JAMA*.

Prabin Sharma, Kisan Thapa, Dikshya Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. 2023. Performance of chatgpt on usmle: Unlocking the potential of large language models for ai-assisted medical education.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health.

Fabrizio Toscano, Eloise May O'Donnell, Joan E. Broderick, Marcella May, Pippa Tucker, Mark Aaron Unruh, Gabriele Messina, and Lawrence P. Casalino. 2020. How physicians spend their work time: an ecological momentary assessment. *Journal of General Internal Medicine*, 35:3166 – 3172.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Umeå University. 2023a. Kunskapsprov för läkare. `https://www.umu.se/utbildning/sok/kunskapsprov/kunskapsprov-for-lakare/`. Accessed: 2023-10-09.

Umeå University. 2023b. Test scores. `https://www.umu.se/nyheter/ytterligare-46-utlandska-lakare-godkanda-pa-teoridelen_10251419/`. Accessed: 2023-10-12.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Hadi Veisi and Hamed Fakour Shandi. 2020. A persian medical question answering system. *Int. J. Artif. Intell. Tools*, 29:2050019:1–2050019:29.

Abraham Verghese, Nigam Haresh Shah, and Robert A. Harrington. 2018. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA*, 319 1:19–20.

Xiaoxia Wu, Cheng Li, Reza Yazdani Aminabadi, Zhewei Yao, and Yuxiong He. 2023. Understanding int4 quantization for transformer models: Latency speedup, composability, and failure cases.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust gpt when your question is not in english.

9. Language Resource References

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning.

A. Examples from MedQA-SWE

A.1. Pre-Clinical Example

Swedish: *Question: Anna, 32 år, söker vård på grund av amenorré sedan cirka 8 månader tillbaka. Provsvar (referensvärden inom parentes): S-Prolactin 249 mIU/L (102- 496) S-TSH 1,19 mIU/L (0,27 -4,20) S-FSH 45 IU/L (follikelfas 3,5 -12; ovulation 4,7- 22; lutealfas 1,7- 7,7) S-LH 21 IU/L (follikelfas 2,4- 13; ovulation 14- 96; lutealfas 1,0- 11) På vilken nivå i hypothalamic -pituitary -gonadal -axeln finns den mest sannolika orsaken till Annas amenorré?*

Answer options:
A: *Uterus*
B: *Ovarier*
C: *Hypotalamus*
D: *Hypofys*
E: *Binjurebark*

Correct answer: B

English translation: *Question: Anna, 32 years old, seeks medical care due to amenorrhea for about 8 months. Test results (reference values in parentheses): S-Prolactin 249 mIU/L (102- 496) S-TSH 1.19 mIU/L (0.27 -4.20) S-FSH 45 IU/L (follicular phase 3.5 -12; ovulation 4.7- 22; luteal phase 1.7- 7.7) S-LH 21 IU/L (follicular phase 2.4- 13; ovulation 14- 96; luteal phase 1.0- 11) At which level in the hypothalamic-pituitary-gonadal axis is the most likely cause of Anna's amenorrhea?*

Answer options:
A: *Uterus*
B: *Ovaries*
C: *Hypothalamus*
D: *Pituitary gland*
E: *Adrenal cortex*

Correct answer: B

A.2. Clinical Cases Example

Swedish: *Question: Anna, 70 år, söker akut på grund av andfåddhet som debuterat relativt abrupt och därefter försämrats påtagligt de sista veckorna. De sista nätterna har hon vaknat på efternatten på grund av andnöd som släpper vid uppresning. Anna förnekar förekomst av bröstsmärtor och hon har inte noterat någon oregelbunden hjärtrytm. Anna har haft hypertoni sedan många år och är ordinerad ett tiazidpreparat för detta. Status: At: Påtagligt andfådd. De ytliga halsvenerna är synligt fyllda i sittande. Cor: Normala hjärttoner, inga säkra blåsljud. Oregelbunden hjärtrytm. Blodtryck: 180/90 mmHg. Pulm: Fuktiga rassel hörs över* *bägge lungornas basala delar. Buk: Leverkanten anas under revbensbågen. Du misstänker hjärtsvikt och ordinerar EKG och lungröntgen. EKG visar förmaksflimmer och vänsterkammarhypertrofi. Röntgen av lungorna visar att hjärtat är normalstort, men Kerley´s B -linjer ses i lungorna och det finns måttligt med pleuravätska bilateralt. Anna får nitroglycerin och furosemid intravenöst, och må r genast bättre. Hon skickas till en vårdavdelning. På grund av hennes förmaksflimmer ordineras Anna ett NOAK. Prover visar att hon är euthyroid Hjärtsvikten orsakas sannolikt av mångårig och otillräckligt behandlad hypertoni. Vilken av följande kombinationer av preparat ger bäst överlevnad vid hjärtsvikt?*

Answer options:
A: *Digoxin, furosemid och kalium*
B: *Atenolol och tiazid*
C: *Ramipril, metoprolol och spironolakton*
D: *Sotalol, kinidin och amilorid*
E: *Cordarone, tiazid och kalium*

Correct answer: C

English translation: *Question: Anna, 70 years old, seeks emergency care due to shortness of breath that started relatively abruptly and then significantly worsened over the last few weeks. The last few nights, she has woken up in the early hours due to difficulty breathing, which eases upon sitting up. Anna denies having chest pains and has not noticed any irregular heart rhythm. Anna has had hypertension for many years and has been prescribed a thiazide medication for this. Status: At: Noticeably short of breath. The superficial neck veins are visibly filled while sitting. Cor: Normal heart sounds, no definite murmurs. Irregular heart rhythm. Blood pressure: 180/90 mmHg. Pulm: Moist crackles heard over both lung bases. Abdomen: The liver edge is palpable under the rib cage. You suspect heart failure and prescribe ECG and chest X-ray. ECG shows atrial fibrillation and left ventricular hypertrophy. X-ray of the lungs shows that the heart is of normal size, but Kerley's B lines are seen in the lungs and there is a moderate amount of pleural fluid bilaterally. Anna is given nitroglycerin and furosemide intravenously, and immediately feels better. She is sent to a ward. Due to her atrial fibrillation, Anna is prescribed a NOAC. Tests show that she is euthyroid. The heart failure is likely caused by long-standing and inadequately treated hypertension. Which of the following combinations of medications provides the best survival in heart failure?*

Answer options:
A: *Digoxin, furosemide, and potassium*
B: *Atenolol and thiazide*
C: *Ramipril, metoprolol, and spironolactone*
D: *Sotalol, quinidine, and amiloride*
E: *Cordarone, thiazide, and potassium*

Correct answer: C

B. Prompt

The prompt used was as follows:

Instruktion: Du är en kompetent kliniker som svarar på frågor. Välj vilket av alternativet som bäst besvarar frågan {question}

Svar:

In English, this roughly translates to:

Instruction: You are a competent clinician who answers questions. From the provided options, choose the one that best answers the following question {question}

Answer: