Efficient Medical Question Answering with Knowledge-Augmented Question Generation

Julien Khlaut [*1], **Corentin Dancette**[*1], **Elodie Ferreres**[*1], **Alaedine Bennani**[*2], **Paul Hérent**[1], **Pierre Manceron**[1]
[1]Raidium
[2] Service de médecine vasculaire, Hôpital européen Georges Pompidou (HEGP), AP-HP, Université Paris-Cité, Paris, France
[1]first.last@raidium.fr [2]alaedine.benani@aphp.fr

Abstract

In the expanding field of language model applications, medical knowledge representation remains a significant challenge due to the specialized nature of the domain. Large language models, such as GPT-4 (OpenAI, 2023), obtain reasonable scores on medical question answering tasks, but smaller models are far behind. In this work, we introduce a method to improve the proficiency of a small language model in the medical domain by employing a two-fold approach. We first fine-tune the model on a corpus of medical textbooks. Then, we use GPT-4 to generate questions similar to the downstream task, prompted with textbook knowledge, and use them to fine-tune the model. Additionally, we introduce ECN-QA, a novel medical question answering dataset containing "progressive questions" composed of related sequential questions. We show the benefits of our training strategy on this dataset. The study's findings highlight the potential of small language models in the medical domain when appropriately fine-tuned.

1 Introduction

Deep Learning led to a breakthrough in natural language processing, reaching human performances on many tasks like question answering or translation. However, their performances are still subpar in complex domains, such as medicine. This domain presents unique challenges, mainly due to its specialized vocabulary, complex concepts, and fast-changing medical literature. Language-based medical tasks, such as medical question answering, require vast knowledge and reasoning abilities to make correct diagnoses. Traditional language models (LMs), while effective in general language processing, struggle when faced with medical knowledge learning mainly because sufficient data for medical knowledge is not necessarily readily available for training. Moreover, in the context of language models, their number of parameters often plays a pivotal role in performances. Large models, although powerful, come with high computational costs and resource requirements, both for training and inference, making them less accessible and practical for widespread use. On the other hand, small models, which are more economical, face challenges in generalization and adapting to specialized domains like medicine. These models require careful fine-tuning to grasp the depth and breadth of medical knowledge effectively. The diversity of general, non-medical datasets on which LMs are trained poses another challenge. These datasets, encompassing a wide array of topics and styles, do not specifically cater to the medical domain. As a result, small models trained on such datasets might fail to develop the necessary understanding for answering more specialized medical questions.

Therefore, we tackle these issues for medical question answering tasks. First, we design a new dataset, ECN-QA. Existing medical question answering (QA) datasets such as MedQA (Jin et al., 2020) and others (Jin et al., 2019), (Pal et al., 2022) are usually single-question multiple answers, which do not encompass the complexity of making a medical diagnosis, which requires multiple turns of questions. Our dataset is based on the French medical residency examination and contains multiple related questions that require models to remember previous questions and reasoning over multiple steps. We then propose a method to train small to mid-size language models for medical question answering. We leverage a corpus of medical textbooks for pre-training. The pre-training set is enriched with specialized questions generated by large language models prompted with medical data from books. This helps to specialize the model on the target task with a small amount of original data.

[*]Equal Contribution

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 10–20

Our code will be made available online.

2 Datasets

2.1 ECN-QA Dataset

We design ECN-QA, a medical question answering dataset. The questions are collected from FreeCN[1], a website established by French medical students to facilitate ECN (*Examen Classant National*, the national ranking exam before medical residency), with their authorization. This website includes questions from past exams and additional questions ("custom" questions) to simulate exam conditions and aid in studying.

The ECN exams themselves consist of two parts. The first part, known as *Individual Questions* (IQ), features general medicine questions with 5 possible answers. Among these answers, one or multiple may be correct, and candidates must identify the true ones. We display an example in Table 1. The dataset contains 4481 IQ, 721 of which come from the historical data of previous exams. The rest, the "custom" subset, contains 3760 additional IQ-like questions created by the FreeCN team to help students prepare for the exam. The second part is known as *Progressive Questions* (PQ), which features clinical cases. Each PQ consists of an introduction followed by a series of successive questions. Similar to the IQ section, these questions also offer 5 possible answers, with 1 to 5 correct answers. A single PQ can contain numerous successive sub-questions, sometimes more than 20. We have 1050 sub-questions in all PQ. We show an example in Table 5 of Appendix E. We also show a whole progressive question in Appendix E.1. We use the accuracy as our evaluation metric. Each proposition in the question is answered separately and gets a score of 0 or 1. The accuracy is then averaged over the five propositions, i.e., for one question, the possible score can be 0, 0.2, 0.4, 0.6, 0.8, or 1.0. For example, in Table 1, if the model answers a, b, c, e as wrong and d as right, it would have one error since c is right. The accuracy would, therefore, be 0.8. If the model answers a and e as wrong and b, c, and d as right, it would also have an accuracy of 0.8.

All the original data is in French, but all models are pre-trained using mostly English data. Therefore, we translate all the questions and answers into English using the Azure AI Translation API[2].

Question: A woman of Martinican origin has just given birth. The child's father is also of Martinican origin. The child has a cleft lip and palate. With regard to regulatory newborn screening of this child, what is the exact proposal(s)? **Propositions**: (a) Phenylketonuria is the only disease of amino acid and organic acid metabolism currently being screened for newborn in France (b) General screening test can detect hypothyroidism of pituitary origin (c) **This couple can refuse the screening after information** (d) **Completion before 48 h of life decreases the sensitivity and/or specificity of the screening test** (e) Targeted screening for sickle cell disease is not indicated in this child

Table 1: Example of Individual Question (IQ) in the ECN, translated to English. Correct answers are in bold.

2.2 Medical Textbooks

Additionally, we use classical French medical textbooks designed for medical students, containing comprehensive medical knowledge and established protocols for managing various medical conditions. We detail in Section F how we extract sections from medical textbooks in PDF format.

In total, we worked with 17,509 PDF files. We grouped text in sections rather than pages, recognizing that a single topic might span multiple pages and should not be truncated. The sections are defined by the book titles and correspond to chapters or important parts. This approach resulted in a total of 234,495 sections. The full dataset is composed of 174,242,531 tokens (with the GPT-3 tokenizer). We detail how we extract sections from PDF files in Appendix F.

We use them for pre-training, and to generate additional questions, as explained in Section 3.

3 Method

We detail our training strategy in this section. The strategy is depicted in Figure 1. We detail related works in the Appendix A.

3.1 Baseline Model

For our baseline, we use the BioMedLM model (Bolton et al., 2022). This 2.7-billion-parameter model is built upon the GPT-2 architecture (Radford et al., 2019) and has been trained on a substantial corpus of medical and biological

[1] https://www.freecn.io

[2] https://learn.microsoft.com/en-us/azure/ai-services/translator/

data. BioMedLM's specialized biomedical tokenizer sets it apart, enhancing its comprehension of specialized terminology. BioMedLM's training data contains all PubMed abstracts and full documents from The Pile (Gao et al., 2020), ensuring a rich knowledge base. Notably, BioMedLM reported state-of-the-art scores on the MedQA (Jin et al., 2020) dataset.

However, this model does not possess the scale needed to achieve impressive zero-shot generalization on new tasks, and medical question answering datasets are limited in scale. Therefore, we aim to train it on specific high-quality data that resembles our benchmark. As our training dataset is small (4967 questions), we propose a method to augment it with question generation using a large language model prompted by some medical knowledge extracted from textbooks.

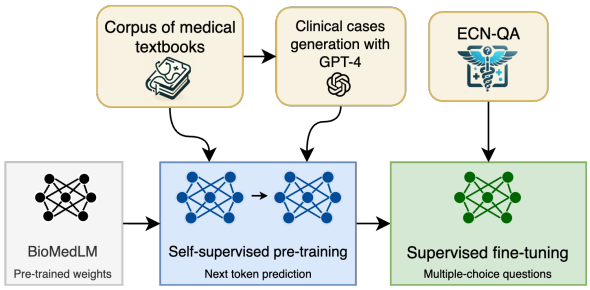

Figure 1: Our training strategy. Starting from an existing language model such as BioMedLM, we continue the pre-training on our corpus of medical textbooks. Then, we use GPT-4, prompted with knowledge from the textbooks, to generate clinical cases that are used to fine-tune the model.

3.2 Questions Generation

Our objective is to create cases that closely resemble genuine ECN cases, as this offers the most effective training for the model. The format we desire for these cases closely resembles that of the progressive questions: an introduction, a list of questions and their possible answers, with a label (true or false) for each answer.

To create our clinical cases, we concatenate several prompts using different approaches. The design of each prompt begins with adopting the prompt used by FreeCN, which primarily comprises an introduction to the task. We refer to this as the pre-prompt. Next, we compile a list of all the specific details we want the case to encompass. This list is informed by the insights of medical experts and the manner in which they typically structure questions for the ECN. We refer to this list as the "constitution." When we initially applied this approach, we encountered somewhat disappointing results. The clinical cases exhibited two major shortcomings. First, they often had very similar subjects, causing the model to struggle to generate diverse cases. Additionally, the main issue was that the questions posed were consistently identical, revolving around topics such as "What is the diagnosis?" or "What tests would you conduct for confirmation?" and "How would you manage the patient?". To solve this issue and to introduce diversity in disease scenarios, we also supplied it with specific knowledge that could be utilized to construct these cases. This section was named the "knowledge part," and it drew upon information extracted from sections of medical books. An example can be seen in Annex 4.

We introduce an additional "justification" field. This component explains why a particular answer to a question is deemed suitable or not.

We build our pipeline using the OpenAI API, employing the GPT-4 model (Achiam et al., 2023) to generate clinical cases. We use the GPT-4 function calling JSON mode, which allows us to specify the output structure.

Following this approach, we generated a dataset with GPT-4, containing about 10,237,240 tokens. In some instances, the dataset underwent meticulous filtering procedures to rectify issues such as missing or alternative fields. This approach is inspired by phi (Gunasekar et al., 2023; Li et al., 2023) and Orca (Mitra et al., 2023; Mukherjee et al., 2023).

We gathered feedback and validation from the FreeCN team, composed of medical doctor students, for assistance and insights to ensure the quality of the questions. The prompt given to GPT-4 is displayed in Appendix B, and an example of a generated progressive question in Appendix G.

3.3 Pre-training

The initial phase involves pre-training the model on a dataset, partly composed of medical books and the additional generated questions. We start from BioMedLM's weights and use a next-token prediction loss to pre-train for three epochs. After training on the books, the model is further trained on the generated cases. The 160,889 generated questions are composed of 10,237,240 tokens. The training

is performed on one case at a time and the final loss is computed only on the model's answer and justification. Since the context length of BioMedLM is 2048, we truncate more prolonged cases. The training parameters are detailed in Appendix C.

3.4 Fine-tuning

Following the pre-training phase, the next step is fine-tuning the model on the ECN-QA dataset. For fine-tuning, the dataset is split into 90 % for training and 10% for testing set. There are multiple ways of getting the model to output an answer, for example, generating tokens with a specific format. However, since generating consistent word-by-word answers proved challenging for the model, often resulting in gibberish rather than accurate responses, we opted for a more traditional approach during fine-tuning. Similarly to previous work (Bolton et al., 2022), a classification head was added to the model. It operates at the proposition level: the model takes as input the question and a single proposition among the five. It then has to predict if the proposition is right or wrong, as a binary classification task. One possible approach to this binary classification involves predicting a single scalar value for each answer, training it with binary cross-entropy, and selecting a threshold value for inference. Another approach consists of adding the words "true" or "false" to the end of the sentence, feeding both sentences to the model, and selecting the answer with the highest score. Empirically, the second approach provided the best results. This modification allowed us to obtain more reliable responses from the model during evaluation.

4 Results

4.1 Evaluation of GPT models

We first evaluate the GPT models on our dataset to obtain baseline scores. For both GPT-3.5 and GPT-4 models, the 2023-12-01 version of the API is used (available on Azure).

We encountered occasional issues during evaluation, where specific prompts may have been blocked, possibly due to sensitive subjects like pediatric medicine. In such cases, we considered the model's response incorrect. The prompts were designed to be straightforward, typically asking the model to provide a true or false answer. Moreover, questions were asked in English using the translated dataset.

Model	Accuracy
GPT-3.5	69.36
GPT-4	79.04
GPT-4-32k	78.97
GPT-4-32k 5 few shot	81.42

Table 2: Results on the all evaluation dataset

The results are presented in Table 2. GPT-4's performances on our dataset are similar to those on MedQA and USMLE, reaching zero-shot performances of around 74% (Nori et al., 2023a). Overall, GPT-4 and its 32k-context variant is the strongest model. Additionally, we confirm (Nori et al., 2023b)'s findings that adding some questions in the prompt (*few shot*) increases the accuracy, in our case, by around 2.5 points.

4.2 Main Results

Model	Accuracy
BioMedLM	67.74
BioMedLM + Books	69.65
BioMedLM + MQG	68.62
BioMedLM + Books + MQG	**70.56**

Table 3: Final results for BioMedLM with various parameters. MQG stands for Medical Question Generation. The model is trained on books for three epochs and on MQG for two epochs. All models are then fine-tuned on ECN-QA.

The results of our experiments are shown in Table 3. We report the result of the original BioMedLM, as well as models pre-trained on the collection of books (*BioMedLM + Books*), pre-trained on the questions (*MQG* for Medical Question Generation) and our complete method (*BioMedLM + Books + MQG*). All models are fine-tuned on ECN-QA.

Including books as part of our training data improves the accuracy by approximately 2 points and the MQG method alone by 1 point. The best accuracy is achieved by combining the pre-training using books with the question-generation method. Overall, we significantly improve the baseline with our full method, getting +3 points in accuracy. We also surpass the GPT-3.5 model, as shown in Table 2.

In Figure 2, we display the number of questions for each score for our full method and GPT-4. We observe that our model still lags behind GPT-4.

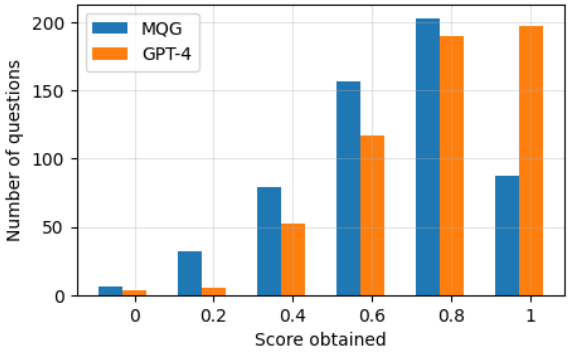

Figure 2: Accuracy distribution by question (number of correct propositions divided by number of total propositions) on the FreeCN dataset of GPT-4 and BioMedLM + Books + MQG

Since the model answers all propositions independently and has no knowledge of its answer to other propositions, the model can contradict itself, which makes it harder to obtain a score of 1 (i.e. having the right answer to all propositions). More detailed statistics per subject are available in Figure 3. Our method appears less effective on subjects it has not been trained on, such as pediatrics.

5 Conclusion

We introduced ECN-QA, a novel dataset for medical question answering that contains a novel type of exercise: progressive questions. We proposed a training strategy based on prompted question generation that improves results over our baseline model, enabling the model to surpass GPT-3.5 accuracy with a much lower parameter count.

Potential avenues for improving efficient medical question answering include increasing the size of the pre-training dataset and the number of generated questions and investigating retrieval-based answering (open-book exam). A model with significant capabilities in medical answering can aid in making informed decisions, especially in time-sensitive situations where rapid response is crucial. Such a model can offer up-to-date information, suggest potential diagnoses, and recommend treatment options based on the latest research and clinical guidelines.

6 Ethical Concerns

The model was trained on questions designed for students' examination, not for a real-world clinical setting. The generalization of this model to actual clinical settings is unknown. Indeed the model has potential biases and limitations in handling sensitive and complex medical cases and should not be used as so on real-world patients.

7 Acknowledgement

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013489R1 made by GENCI.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Biomedlm.

Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S. Glicksberg, Alexander W. Charney, Girish Nadkarni, and Eyal Klang. 2023. Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Scientific Reports*, 13.

Zeming Chen, Alejandro Hern'andez Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Kopf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *ArXiv*, abs/2311.16079.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

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. *ArXiv*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Yuan-Fang Li, Sébastien Bubeck, Ronen Eldan, Allison Del Giorno, Suriya Gunasekar, and Yin Tat Lee.

2023. Textbooks are all you need ii: phi-1.5 technical report. *ArXiv*, abs/2309.05463.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023b. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. GPT-4 Technical Report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu 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. *ArXiv*, abs/2307.09288.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *arXiv preprint arXiv:2305.10415*, 6.

A Related Work

A.1 Medical QA Datasets

Various datasets have been developed in medical question answering (QA). Among these, the MedQA dataset (Jin et al., 2020) stands out for its comprehensive coverage of multiple-choice questions derived from professional medical board exams. This dataset is particularly significant because it encompasses many questions, totaling 12,723 items. It aims to evaluate the depth of medical knowledge encoded in AI models.

Another dataset is PubMedQA Dataset for Biomedical Questions, (Jin et al., 2019). This dataset uniquely focuses on questions generated from article titles and abstracts within the biomedical literature, excluding conclusions, and provides answers in a format conducive to yes/no/maybe evaluations.

Further expanding the landscape, the MedMCQA dataset (Pal et al., 2022) is a large-scale, multi-subject repository of medical multiple-choice questions. This dataset has a large scope and relevance, covering many medical subjects.

A.2 Medical QA Models

Several strategies aim to construct a good model with high accuracy and reliability of responses on those medical tests. One method involves leveraging large language models (LLM) such as GPT-4. Through prompt engineering, (Brin et al., 2023) or (Nori et al., 2023a) have demonstrated excellent results on MedQA.

Further exploration into the efficacy of large-scale models has been conducted, with (Singhal et al., 2023a) and (Singhal et al., 2023b). These studies have assessed the performance of such models on MedQA and across a diverse array of medical datasets.

Moreover, the landscape of medical QA has been enriched by initiatives to fine-tune pre-existing LLMs. For instance, adaptations of Llama 2 (Touvron et al., 2023) have been proposed (Wu et al., 2023; Chen et al., 2023). These efforts signify a targeted move towards refining the capabilities of LLMs to meet the demands of the medical domain, illustrating a focus on customizing general models for specialized tasks.

In the context of smaller-scale models, (Bolton et al., 2022) has been recognized for its superior performance. This model stands out as a testament to the effectiveness of more compact models in handling medical QA tasks, offering an alternative to the larger, more complex systems.

B Question Generation

Table 4 shows the prompt we used to generate questions with GPT-4. The prompt is appended with a section coming from a medical textbook.

You are a French professor of medicine. You seek to test the level of medicine of your students. Your task is to generate 1 to 2 different clinical cases requiring the highest medical understanding. Each clinical case consists of an Introduction and 4-10 multiple-choice questions. They must be formatted as follows: Introduction, Propositions. Propositions contain several proposals with a justification and a field to know if they are correct. The clinical case needs to be very very hard and accurate. The level of difficulty is 10 out of 10. It should be very hard even for the best students. And you should have a very detailed justification. The case should be long with detailed questions and detailed justification. The criteria to be met are: 1. The introduction is common to all questions. 2. There must be 4-10 different questions. 3. A question can have 5-10 possible choices. 4. One or more proposals may be fair. 5. Justification must be specific, justified and sourced. It is very important to have a very good and long justification. It should be at least 3 lines long. 6. Uses the highest medical level possible. 7. Questions must be diversified to a minimum of 4. They must deal with the patient's disease but also with the examinations to be carried out, the follow-up and the possible developments of the case. They will make the case both nuanced and complex. 8. The case must be precise or even quantitative. It is a question of providing as much information as possible, and the solution to the questions may be found in detail. 9. Cases must be pedagogical and the questions must be linked to build a complete reasoning. 10. Responses should be directed to prioritize severe and frequent cases. 11. The student's expected behaviour is above all to avoid medical misconduct. 12. The student's method must be a probabilistic approach. 13. A language model must be able to answer questions. For example, do not ask the wizard to create images or audio. 14. The case must be written in English. 15. All fields must be completed. 16. The MA for the drug and the recommendations of the HAS and ANSM must be respected. In the absence of recommendations from HAS and ANSM, the current practices recommended by French speciality colleges and learned societies will be applied. ### To do that you can use the following information: [Extract of a medical book]

Table 4: Prompt used to generate progressive questions with GPT-4

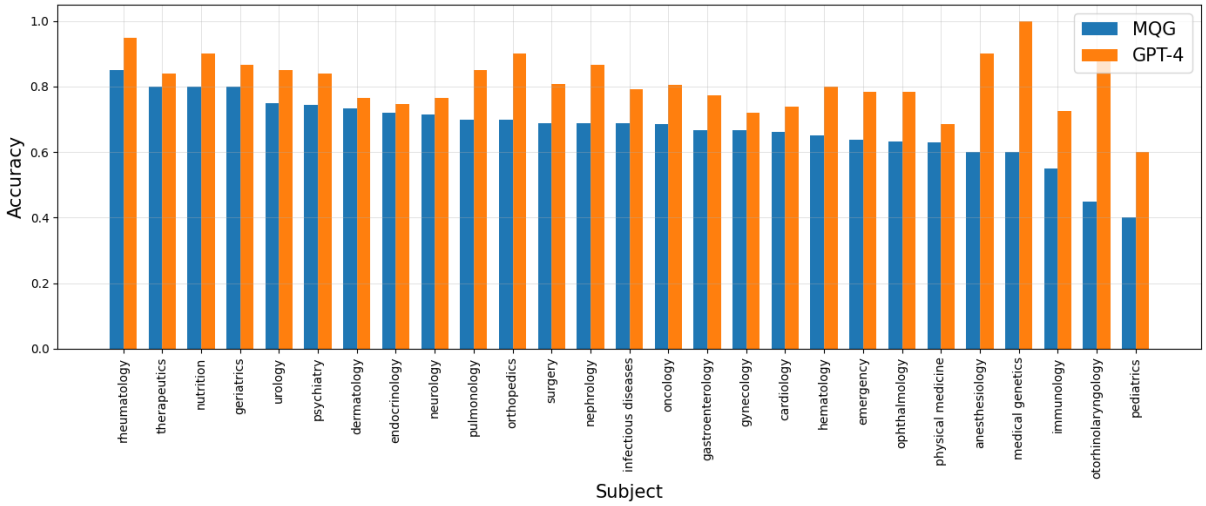

Figure 3: Accuracy per subject of BioMedLM and GPT-4

C Implementation and training details

For training, we use a node with 4 NVIDIA V100 gpus. The model is pre-trained on books for three epochs on generated cases for two epochs, and we fine-tune the final model for 20 epochs. We use a learning rate of 1.e-4 for pre-training and 2.e-6 for fine-tuning.

D Additional Results

We show in Figure 3 the accuracies for each topic of the ECN exam for GPT-4 and our model. Our model is close to GPT-4 for most subjects and performs worse on subjects that GPT-4 often refuses to generate questions, like pediatrics. These questions are from the test set but only come from the additional questions provided by the FreeCN team.

E Example of Cases

In Table 5, we display the first question of a progressive question from the ECN-QA dataset, with the propositions of answer.

Introduction: A 67-year-old man consults for right calf pain occurring after a walk that the patient estimates to be 350 meters away. He is a retired and sedentary taxi driver. This patient has been smoking a pack of cigarettes a day since the age of 30. You follow it for high blood pressure discovered by a systematic and balanced examination by perindopril. Blood sugar is normal as well as lipid balance.
Question: What is your main diagnostic hypothesis ?
Propositions:

- **Obliterating arterial disease of the lower limbs**
- Narrow lumbar canal
- Lombosciatica
- Hypokalemia
- Deep vein thrombosis

For the example the following questions are:

- You suspect arterial disease obliterating the lower limbs. Which of the following semiological elements will guide the diagnosis towards this hypothesis?
- The interrogation confirms the appearance of a pain when walking with a cramp localized in the right calf. The pain manifests itself early when the patient climbs a slope, thus supporting your diagnostic hypothesis of arterial obliterating disease of the lower limbs. [...] On the data of this clinical examination, which is(are) the arterial atheromatous lesion(s) that you should suspect?

Table 5: Example of PQ in the ECN-QA dataset. This particular PQ consists of 16 questions, and both this PQ and the previous IQ section are derived from the 2020 ECN. Correct answers are in bold.

E.1 Full Progressive Question

Here, we display a full progressive question with all possible answers. Correct answers are in bold font.

Full example of a progressive question

Introduction: A 54-year-old man, a long-term smoker who has been hypertensive for 12 years (calcium channel blocker treatment), consults his attending physician for an isolated episode of total gross hematuria, without a clot. His other history has been an appendectomy in childhood. The blood count is as follows: Hb 10.4 g/dL (MCV 78 µm3), GB 8 G/L, blisters 247 G/L. Creatinine is 110 µmol/L (estimated glomerular filtration rate of 65 ml/min/1.73 m2). A renal ultrasound showed a hyperechoic mass of 7 cm on the right kidney.
Questions: What are the elements (present or to be sought at the interrogation and clinical examination) that can evoke a malignant tumor of the kidney? (one or more correct answers)
- **Smoking**
- **Chronic high blood pressure**
- Long-term calcium channel blocker treatment
- A family history of multiple endocrine neoplasia
- **Low back pain**

Which exam(s) are you asking for as a first line?
- **Urinary cytology with pathological examination**
- **Cytobacteriological examination of urine**
- Serum erythropoietin assay
- **Abdominopelvic CT scan with and without contrast injection**
- Ultrasound-guided puncture of the mass

On the cut shown below, what are the True propositions? (one or more correct answers)
- **This is an abdominal CT scan with injection**
- This is a coronal cup
- Structure number 1 is the inferior vena cava
- **The cut passes through the third duodenum**
- The number 2 corresponds to the inferior mesenteric artery

What are the real propositions? (one or more correct answers)
- **The patient must receive red blood cells**
- The patient must receive platelet pellets
- In case of transfusion of red blood cells, you would prescribe O-negative pellets
- A search result for irregular agglutinins less than 48 h old must be available
- Since 2003, there has been no risk of transmission of infectious pathogens through red blood cell transfusion

What is the real proposal(s)?
- **This is acute renal failure**
- The glomerular filtration rate must be recalculated
- An obstacle on the contralateral kidney is likely
- **It may be functional renal failure**
- **An ionogram should be prescribed on a urine sample**

What are the exact proposals? (one or more correct answers)
- He has moderate chronic renal failure
- **His antihypertensive treatment must include an inhibitor of the renin-angiotensin system**
- The LDL cholesterol target to be achieved is 1.3 g/L
- He must follow a diet containing no more than 1.5 g/kg of protein weight
- It is necessary to advocate a diet low in fast sugars

What risk(s) does he run?
- **Gradual decrease in diuresis**
- **Increased cardiovascular risk**
- **Hyperphosphoremia**
- **Erectile dysfunction**
- **Contralateral kidney cancer**

What is the True answer(s)?
- The ALD file is completed by the patient and validated by the medical specialist
- **The attending physician must specify in the request the protocol of care envisaged including treatments, examinations and consultations**
- **The medical officer of the Health Insurance must validate the care protocol**
- In case of coverage in ALD, remains the responsibility of the patient only the co-payment
- The third-party payer is the part of the care paid by the insured whether or not he is registered in ALD

What is your interpretation of the electrocardiogram below?
- **Sinus rhythm**
- Sino-auricular block
- T-waves suggestive of hyperkalemia
- Expanded QRS Complexes
- **Left ventricular hypertrophy**

To reduce edematous syndrome, what do you recommend at this stage? (one or more correct answers)
- A low-salt diet (less than 6 g/d)
- Water restriction
- **A loop diuretic (furosemide)**
- A thiazide diuretic (hydrochlorothiazide)
- Blood ultrafiltration (start of hemodialysis)

What are the possible cause(s) in the context of the new biological abnormality observed?
- **Excessive calcium intake**
- Taking furosemide
- **Chronic renal failure**
- Secondary hyperparathyroidism
- **Bone metastases from kidney cancer**

What additional examination(s) do you recommend to explore this biological anomaly?
- **Ionized serum calcium**
- Test de PAK
- **PTH assay**
- PTHrp assay
- **Bone scintigraphy**

Which proposals are correct? (one or more correct answers)
- Metastatic cancer is a contraindication to dialysis
- Haemodialysis confers survival advantage over peritoneal dialysis
- The preparation of an arteriovenous fistula (AVF) is contraindicated given the prognosis
- **A tunneled central venous catheter may be placed to initiate hemodialysis**
- A transplant from a cadaveric donor must be discussed

In general, regarding living donors, what are the real proposals? (one or more correct answers)
- Only people with a genetic link to the recipient can be donors
- **Transplantation can be done in incompatible ABO condition**
- Rhesus compatibility must be respected

- HLA incompatibility between donor and recipient is a formal contraindication
- The donor is remunerated on a basis proportional to the recipient's waiting time

F Extraction of text from PDF files

We use the Azure AI Document Intelligence API[3] to extract text sections from PDF files. The API returns paragraphs of texts and titles, sorted in reading order, along with tables and figures. We remove all tables and figures and implement text filtering algorithms to remove useless paragraphs, like headers, footers, or paragraphs that are just a few characters long.

We then regroup paragraphs in sections based on the titles and divide them into subsections, or regroup them, to have sections of similar lengths, between 500 and 1000 words.

G Generated Progressive Question

Below, we present an example of a progressive question generated by GPT-4. Correct answers are in bold font.

Introduction: A 45-year-old female is being admitted to the emergency department. She is complaining of a severe and sudden headache unlike any she has ever experienced before. The headache was followed by episodes of vomiting and photophobia. Her Glasgow Coma Scale (GCS) score on admission is 14 and her physical examination is unremarkable. Computed Tomography (CT) of the brain reveals subarachnoid hemorrhage (SAH).
Questions What is the most likely diagnosis in this patient?

- **Ruptured cerebral aneurysm** (The presence of SAH in the setting of a severe, sudden-onset headache (often described as a 'thunderclap headache'), especially when associated with vomiting and photophobia, is highly suggestive of a ruptured cerebral aneurysm. Other common symptoms of SAH due to a ruptured aneurysm include neck stiffness, altered mental status and cranial nerve palsies.)
- Migraine (While migraines can cause severe headaches, they do not typically lead to SAH. Moreover, the acute onset described in the vignette is uncharacteristic of migraines which usually have a more gradual onset.)
- Tension headache (A tension headache often presents as a dull, constant and bilateral headache. SAH or sudden severe headache is not typical of a tension headache.)
- Cluster headache (Cluster headaches are usually unilateral, severe, and around the eye, and are associated with autonomic symptoms such as tearing, nasal congestion and ptosis. They do not typically cause vomiting or lead to SAH.)
- Brain tumor (While brain tumors can cause headaches, they do not typically present acutely. Moreover, they generally result in neurological symptoms and signs that localize to the area of the brain where the tumor is located rather than leading to SAH.)

What is the next step in the management of this patient?

- **Cerebral angiography** (Cerebral angiography is the gold standard for diagnosing cerebral aneurysms. It can demonstrate the size, location and morphology of the aneurysm and help plan intervention. Cerebral angiography is also able to diagnose other causes of SAH such as arteriovenous malformation.)
- Magnetic resonance imaging (MRI) (While MRI may be used to evaluate for other causes of SAH such as arteriovenous malformations (AVM) if the angiogram is inconclusive cerebral angiography is the preferred primary imaging modality in this clinical scenario.)
- Lumbar puncture (Lumbar puncture may be performed when the clinical suspicion for SAH is high but initial CT scan is negative. Given the positive CT findings in this case lumbar puncture would not be the best next steps.)
- Blood cultures (Blood cultures are indicated in patient with suspected infective endocarditis that can lead to mycotic aneurysms. In this case there is no clinical suspicion of infective endocarditis.)
- Repeat CT scan (A repeat CT scan may be used in the surveillance of SAH treatment and complications but it is not the best choice for the acute management of suspected ruptured cerebral aneurysm which requires immediate intervention.)

If this patient's cerebral angiography reveals a vasospasm of the cerebral arteries, what is the most appropriate step that should be taken?

- Platelet antiaggregation (Platelet antiaggregation is not the standard management for cerebral vasospasm. Studies showed that antiplatelet medication would lead to a higher rate of rebleeding in cerebral aneurysm.)
- Anticoagulation (Anticoagulation is not the standard management for cerebral vasospasm as it could increase the risk of bleeding.)
- **Calcium channel blockers** (Nimodipine, a calcium channel blocker, is the only medication proven to improve neurologic outcomes in the case of SAH. It doesn't prevent or treat the vasospasm itself but reduces the severity of neurologic deficits and improves outcome.)
- Statins (Although statins have been shown to have some vasoprotective effects they are not the first-line treatment for cerebral vasospasm.)
- Analgesics (While managing pain is an important component of care in this patient population analgesics themselves do not treat vasospasm).

[3] https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence