What Does *Infect* Mean to *Cardio*?
Investigating the Role of Clinical Specialty Instructions in Medical LLMs

Anonymous ACL submission

Abstract

In this paper, we introduce S-MedQA, a medical question-answering (QA) dataset for benchmarking large language models (LLMs) in fine-grained clinical specialties. Using S-MedQA, we gauge the role of instructions for knowledge-intense scenarios by checking the applicability of two popular hypotheses related to knowledge injection and style/format learning. We show that in the medical domain, more instructions result in better performance. However, the improvement in performance derives neither from the extra knowledge contained in the instructions nor the style/format learned from them. Thus, we suggest rethinking the role of instruction data in the medical domain. We release S-MedQA for the community.[1]

1 Introduction

Multiple-choice question-answering (QA) datasets are often used as benchmarks to evaluate large language models (LLMs) in the medical domain (Labrak et al., 2024; Singhal et al., 2023) and are crucial to guide the development of medical LLMs (e.g., PubMedQA, Jin et al., 2019; MedQA, Jin et al., 2021; MedMCQA, Pal et al., 2022). However, specialized hospitals that may be interested in deploying LLMs to address specific clinical problems are typically only interested in the performance of LLMs *in a few clinical specialties* (e.g., obstetrics or oncology). Moreover, to the best of our knowledge, there are no open-source medical QA datasets with annotations of medical specialties. Thus researchers could not investigate how well knowledge transfers across clinical specialties due to this lack of fine-grained benchmarks.

To address the gap, we develop **S-MedQA**, the first medical QA dataset with clinical specialty annotations. We build S-MedQA based on the widely used MedQA dataset (Jin et al., 2021) and use `gpt-3.5-turbo-0125`—henceforth GPT-3.5—and medical experts to map samples onto clinical specialties. We first prompt the model with carefully designed prompts and retain only annotations agreed upon by a majority. Further experts' examination guarantees the quality of the annotations (more details in §2.3), showing that our dataset maintains a high accuracy of categorization (97.8%). S-MedQA contains 15 specialties, each with hundreds of samples (see §2 for details).

We use our dataset to investigate the applicability of two popular hypotheses in the medical domain regarding the role of instruction data in the context of LLMs: **H1)** little-to-no knowledge injection occurs during instruction tuning as a few instructions yield results comparable to large data (Zhou et al., 2024); and **H2)** the role of instruction data is solely to learn downstream language styles or formats for the AI assistants (Lin et al., 2023). We carefully design control experiments to test these two claims.

Hypothesis 1 (H1): there is little-to-no knowledge injection due to instruction tuning. We fine-tune LLMs on one specialty and test them on all the others. In most cases, the best results are **not** achieved by fine-tuning with data from the same specialty. E.g., in the *cardio* domain, we get the best results when fine-tuning with the *infect* specialty, whereas the clinical knowledge contained in *infect* is almost completely irrelevant to *cardio*. This raises questions about the extent to which knowledge injection can be explained as the source of score improvements, partly supporting the hypothesis that instructions bring little knowledge.

Hypothesis 2 (H2): task improvements result from learning language styles or formats. Here, we change the answer of each training sample in S-MedQA to a random wrong option while keeping the format the same. Our results show that 1) training models using the same format with wrongly answered instruction data results in nearly random

[1] `https://anonymous.4open.science/r/S-MedQA-85FD/`

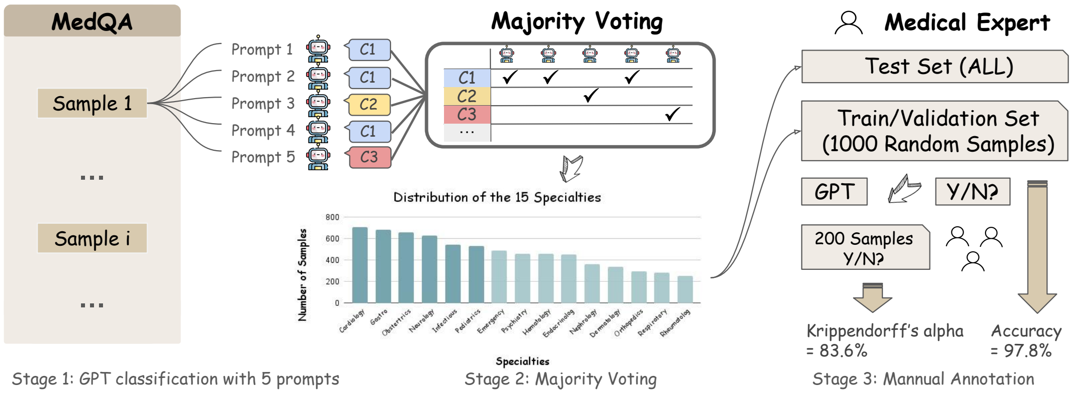

Figure 1: Overview of S-MedQA's construction process. For each sample, we generate predictions using 5 different prompts and only keep those where predictions agree (3+, 4+, or 5 times). We keep the top 15 predicted specialties in the dataset, each containing 250–705 samples when using 3+ majority (or 98-454 samples when requiring agreement among all 5 prompts; more details in Table 3). To annotate, we randomly sample $1,000$ questions from our train set and ask a medical expert to evaluate GPT-3.5's predictions, achieving an accuracy ranging from 97.8% (coverage of 49.2%) and 90.8% (coverage of 89.1%). The expert also manually annotates S-MedQA's whole validation and test set. Three medical students additionally annotate the same 200 samples out of the original $1,000$ samples annotated by the medical expert (for computing inter-annotator agreements; see §2.3 for details).

guessing ($\sim 25\%$), moreover, 2) this negative impact transfers to all other specialties, even though the model is only fine-tuned on "bad data" from a certain specialty. Our findings challenge H2 as we clearly show that correct instructions are critical to model performance, and its impact on the model is far greater than changing the language style alone.

We argue that in the medical domain—and possibly in other knowledge-intensive scenarios—large amounts of high-quality instruction data are still necessary for better performance. However, the improvements cannot be completely explained as the extra knowledge injection from the instructions.

2 A Benchmark of Clinical Specialties

We now describe how we create S-MedQA, a high-quality benchmark for medical QA with clinical specialty annotations. We release multiple 'versions' of S-MedQA with different accuracy/coverage trade-offs. We create these versions using different thresholds for majority voting to include an example (with its predicted clinical specialty[2]) in the final dataset. Dataset users can thus choose to have a *cleaner version of data with fewer examples*, or a *more noisy version with more examples* (more details in §2.3).

In §2.1, we explain how we use GPT-3.5 to categorize the dataset into distinct specialties. In §2.2, we show how we split the dataset and select the specialties. Finally, in §2.3, we describe how we manually validate our data to ensure its quality. We show an overview of our benchmark in Figure 1.

2.1 Medical Specialty Categorization

We source examples from MedQA (Jin et al., 2021), a commonly used dataset in evaluating medical LLMs. We use GPT-3.5 to annotate samples with clinical specialties. However, we observed low accuracy in our preliminary evaluation when using a single prompt ($\sim 75\%$). To improve this accuracy and counter the non-deterministic characteristics of the outputs generated by GPT-3.5, and to minimize the effort of manual annotation, we follow Ding et al. (2023) and Goel et al. (2023) and design five prompts, generate predictions with GPT-3.5 for each sample, and apply majority voting.

2.2 Dataset Splits

We show the resulting distribution of all specialties in §A.2. We exclude $1,324$ (13%) samples where there is no majority vote[3] and the 308 (3%) samples categorized into *Others*, as they contain clinically irrelevant information. We provide more examples and discuss reasons for exclusions in §A.3.

After the abovementioned steps, 15 out of 55 specialties contain more than 200 samples. We only include these 15 specialties in S-MedQA to ensure their statistical reliability. The final dataset comprises $7,125$ / 899 / 893 samples in train/validation/test sets (after applying the proce-

[2]We restrict the response of GPT-3.5 to the 55 medical specialties recognized in the European Union (see Appendix A.1).

[3]After manual inspection of some of these examples, we hypothesize this is due to their ambiguity in terms of specialty.

dure we describe next in §2.3). Specialties along with their number of samples are described in Table 3 in §A.4. Note that in the following experiments, we use the top 6 representative specialties with the largest number of samples (*Cardiology*, *Gastroenterology*, *Infectious diseases*, *Neurology*, *Obstetrics and Gynecology*, and *Pediatrics*).

2.3 Manual Validation

We ask a medical expert to label each example in the validation and test sets with the correct clinical specialty. This expert also validates 1000 random samples from the train set whether the specialties predicted by GPT-3.5 are correct. The performance after using multiple prompts and voting greatly improves compared to using single prompts (e.g., from 72.8%–80.2% to 90.8%–97.8%; see Appendix A.5 for details).

	# votes (out of 5)		
	3+	4+	5
Accuracy (%)	90.8	94.8	97.8
Coverage (%)	89.1	69.0	49.2

Table 1: Accuracy vs. coverage for majority voting under different minimum number of votes.

However, as we show in Table 1, we must decide on the trade-off between accuracy and coverage when determining the minimal number of votes in agreement for the example to be included in S-MedQA. A higher quorum results in higher accuracy (90.8 $\rightarrow$ 97.8) but greatly decreases coverage (89.1 $\rightarrow$ 49.2). We release individual categorizations and votes of all examples for users of the dataset to decide their preference between accuracy and coverage — more data but possibly more noise or less noise but less data — based on their specific use cases. We select 3 as the quorum in this study for adequate fine-tuning data. To assess the trustworthiness of the expert, we then randomly sample 200 from the $1,000$ examples and further ask three medical master students to validate the same way as the expert. We use Krippendorff's alpha (Hayes and Krippendorff, 2007) to measure the inter-annotator agreement among the four annotators (medical students and expert) on the 200 examples and achieve 83.6% (95% CI [69.0%, 93.9%]).

3 Experimental Setup

3.1 Cross-Specialty Evaluation

We experiment on four variants of two open-source LLMs: Llama2-Chat-7b and 13b (Touvron et al., 2023)—henceforth **Llama2-7b** and **Llama2-13b**—and Mistral-Instruct-v0.1 and v0.2 (Jiang et al., 2023), henceforth **Mistral-v0.1** and **Mistral-v0.2**. We fine-tune each LLM on the six per-specialty training datasets with prompts shown in Appendix A.6 and measure each resulting model's performance on all six per-specialty test sets. In each experiment, we train for 10 epochs and select the checkpoint based on the best per-specialty validation scores. We also train with a combined set containing all six specialties' training data to evaluate whether exposure to a more diverse and extensive set of instructions could affect the model's performance. More details and hyperparameters are found in Appendix A.7.

3.2 Bias Mitigation

In all test sets, we shuffle the answers 5 times for each sample and add all these 5 entries to the final test set in case the model prefers an option due to its position (Zheng et al., 2023). To further improve the reliability of results, we follow Wang et al. (2024) to generate the entire answer with the model and train a classifier to match model outputs to the options in a post-hoc step, instead of using the maximum probability of options {A, B, C, D} with a single next-token prediction step. More concretely, we randomly select 150 training samples and generate answers for these with all four LLMs (**Llama2-7b**, **Llama2-13b**, **Mistral-v0.1**, and **Mistral-v0.2**), resulting in 600 responses. We manually annotate all the responses with the right options and use these annotations to train a Mistral-Instruct-v0.2 model as the classifier, with 400 (200) train (test) samples. Our classifier achieves 96.5% accuracy and we use it in all experiments. The illustration of our approach to evaluate LLMs performance on S-MedQA can be found in Appendix A.8.

4 Results

Does Instruction Tuning Data Inject Knowledge? Table 2 'Right answers' shows the performance of Mistral-v0.2 fine-tuned independently on each specialty training set and tested on all six specialty test sets. We also report the performance after fine-tuning on the combined training set. We observe that the models fine-tuned on the combined dataset, as well as each single specialty, consistently outperform the base model, demonstrating the effectiveness of instruction fine-tuning.

However, when looking at the results of mod-

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
	Mistral-v0.2†	52.0	45.9	**48.2**	37.0	52.9	43.5	46.9
Right answers								
Train Sets	Cardio	55.8	54.7	**47.4**	44.0	54.8	47.2	**50.6**
	Gastro	54.0	**58.0**	41.7	46.5	**55.4**	45.2	49.8
	Infect	**57.0**	52.6	44.3	**47.3**	49.6	48.0	49.6
	Neuro	52.5	52.4	40.6	43.5	51.5	**50.3**	48.3
	Obstetrics	52.8	51.5	41.9	44.3	54.4	46.6	48.4
	Pediatrics	53.0	45.9	43.2	40.5	49.6	44.3	46.0
	Combined‡	61.5	63.1	45.3	51.1	57.9	49.1	54.3
Wrong answers								
Train Sets	Cardio	25.5	24.4	28.1	24.5	22.3	**26.7**	25.1
	Gastro	25.8	25.2	**29.4**	24.7	23.1	26.1	**25.6**
	Infect	25.5	26.1	24.7	23.4	22.3	25.3	24.5
	Neuro	**28.7**	**27.2**	24.7	23.9	22.5	20.5	24.7
	Obstetrics	20.3	20.0	22.9	22.0	21.9	23.0	21.6
	Pediatrics	24.3	25.4	29.2	**25.5**	**23.8**	23.3	25.2
	Combined‡	29.5	26.5	26.8	23.1	25.8	23.0	25.9

Table 2: Accuracy matrix for Mistral-v0.2 as the base model. †Model is applied without finetuning. ‡Model is trained on the combination of all 6 specialty train sets. **Right answers:** For each specialty, we highlight the best performance when fine-tuning on different specialty datasets in **bold**. We underline scores for models fine-tuned on a training set of the same specialty. Surprisingly, 5 out of 6 best performances are not achieved by the model tuned on the corresponding training set. **Wrong answers:** Models show near-random performance.

els fine-tuned on individual specialties, 5 out of the 6 best performances on each test set were not achieved by the model trained on the corresponding specialty's data. E.g., the best performance on the *Cardio* test set (57.0%) was achieved by the model trained on *Infect*. This raises the question: are these score improvements truly indicative of knowledge acquisition or injection? If the improvements were due to knowledge injection, we would not expect the model trained on *Infect* to perform the best on the *Cardio* test set, as there is almost no knowledge existing in *Infect* that is relevant and useful to *Cardio*. This inconsistency suggests that the model's enhanced performance may not solely reflect an increase in knowledge that is injected from the instruction data. The results of the other models are seen in Appendix A.9.

Is Instruction Tuning Only Superficial? Lin et al. (2023) explains the score improvements after fine-tuning as learning the style or formats from instruction data rather than acquiring knowledge. To explore this hypothesis, we randomly changed the answers of the training set into one of the wrong options under the question while keeping the rest (formats/style) the same. Then we train LLMs with wrong-answer instructions under the same setting and test on the original shuffled test sets.

Table 2 'Wrong answers' shows the results. It is easy to see that most performances drop to around 25%, meaning that incorrect information severely harms the performance of LLMs. Moreover, the negative impact is transferred to all other specialties, even though the models have never been fine-tuned on the corresponding "corrupted" data of those specialties. We also note that the accuracy is close to random guessing ($\sim 25\%$) instead of constantly producing wrong answers ($\sim 0\%$), implying that the models almost completely lose the ability to tackle medical tasks accurately. Therefore, we answer hypothesis H2 arguing that instruction data that contains correct clinical knowledge is critical to model performance, and its impact is far greater than changing language styles alone.

5 Conclusions

In this paper, we present S-MedQA, the first medical instruction dataset annotated across 15 distinct specialties. We use S-MedQA to investigate two popular hypotheses but now in the medical domain. Our findings show that 1) fine-tuning with medical instruction data can improve LLMs' performance, but the improvements can not be solely explained as extra knowledge injection; and 2) LLMs acquire content information that is far more than stylistic adaptations from instruction data. However, the actual effect of instruction data is still unclear. We suggest future research to further explore the role of instruction tuning.

Limitations

We limited our experiments to the medical domain. However, the findings' generalizability to other knowledge-intense domains is unknown. Also, we only allow GPT-3.5 to assign a single specialty to each example to obtain a unique 'ground truth'. This might have led to suboptimal performance since some examples could be relevant to multiple specialties and might not reflect the multifaceted nature of our scenario along with other real-world cases. Further research is needed to investigate the role of instruction data in different domains and to explore the possibility of multi-view annotation.

References

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In Machine Learning for Health (ML4H), pages 82–100. PMLR.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1):77–89.

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. arXiv preprint arXiv:2106.09685.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14):6421.

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.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. arXiv preprint arXiv:2312.01552.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Proceedings of the Conference on Health, Inference, and Learning, volume 174 of Proceedings of Machine Learning Research, pages 248–260. PMLR.

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. 2023. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. arXiv preprint arXiv:2402.14499.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. arXiv preprint arXiv:2309.03882.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36.

A Appendix

A.1 Prompts used for specialty classification

In Figures 4–8 we show the 5 prompts we use with GPT-3.5 for specialty classification. Prompt 1 is zero-shot, while we add 6 examples to the other prompts (one example from each top-6 specialty) to leverage the in-context ability of LLMs. We moved the list of specialties to the end of the user prompt in prompt 4 and changed the format of the user prompt to follow the examples by adding *"Question:"* and *"Answer:"* in prompt 5.

A.2 Distribution of predicted specialties

In Figure 2 we show the distribution of samples across specialties. We show the 15 specialties we include in S-MedQA in dark blue, comprising in total 70.0% / 70.7% / 70.1% of the entire train / validation / test sets. We do not include the rest of the specialties due to too few samples.

A.3 Examples and reasons for excluding samples

We carefully look into the samples that did not reach a vote of three together with the medical expert and noticed that most of these examples are ambiguous in terms of medical specialties. They are therefore difficult to be classified into one single specialty. For instance, many disagreements occur with *Neurology* and *Emergency Medicine* in an emergent neurological issue, such as the following question:

> *A 78-year-old man is brought to the emergency department by ambulance 30 minutes after the sudden onset of speech difficulties and right-sided arm and leg weakness. Examination shows paralysis and hypoesthesia on the right side, positive Babinski sign on the right, and slurred speech. A CT scan of the head shows a hyperdensity in the left middle cerebral artery and no evidence of intracranial bleeding. The patient's symptoms improve rapidly after pharmacotherapy is initiated and his weakness completely resolves. Which of the following drugs was most likely administered?*

According to the expert, both *Neurology* and *Emergency Medicine* apply to this situation, as they contain clinical knowledge from both specialties and require collaboration of these two specialties in clinical practices. Also, classifying it exclusively into one of the specialties requires extra expertise that could be beyond the capabilities of GPT-3.5, e.g. classify as *Emergency Medicine* if the question itself mainly focuses on maintaining vital signs, and *Neurology* when it comes to subsequent treatment phases. It is hard and unclear whether we should classify this type of questions into either specialty, thus we do not include these examples.

Another kind of sample we exclude are those classified as *"Others"*, i.e., not belonging to any specialty in the given list of 55 specialties recognized by the EU. Here is an example:

> *A resident in the department of obstetrics and gynecology is reading about a randomized clinical trial from the late 1990s that was conducted to compare breast cancer mortality risk, disease localization, and tumor size in women who were randomized to groups receiving either annual mammograms starting at age 40 or annual mammograms starting at age 50. One of the tables in the study compares the two experimental groups with regard to socioeconomic demographics (e.g., age, income), medical conditions at the time of recruitment, and family history of breast cancer. The purpose of this table is most likely to evaluate which of the following?*

This question belongs to *Clinical Trial Design* instead of any listed clinical specialties and does not contain knowledge required for daily clinical practices. Similar cases also include *Toxicology*, *Epidemiology*, and *Medical Ethics*. We thus also exclude such samples from S-MedQA.

A.4 Clinical specialty benchmark description

In Table 3, we show the 15 specialties we include in S-MedQA, as well as their respective numbers of samples in the train set. The 3 numbers in each block represent the number of samples obtained via majority voting of 3+, 4+, and 5.

A.5 Accuracy vs. coverage trade-off of GPT-3.5 predictions

	Prompts				
	#1	#2	#3	#4	#5
Accuracy(%)	76.0	72.8	73.0	73.8	80.2

Table 4: Accuracy of each prompt. Prompt #i refers to Figures 4–8 in Appendix A.1.

In Table 4, we list the results of our manual validation. The accuracy when using only a single prompt ranges from 73% to 80%. We also report

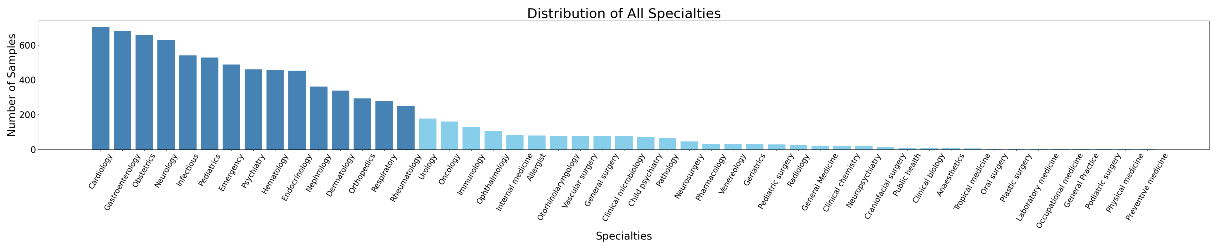

Figure 2: The distribution of all specialties classified by GPT-3.5. The dark blue specialties are the 15 we finally included in our benchmark.

Number of Votes (out of 5)	3+	4+	5
Cardiology	705	576	454
Gastroenterology	681	562	418
Obstetrics and gynecology	658	577	444
Neurology	630	507	325
Infectious diseases	541	346	151
Pediatrics	529	328	144
Emergency medicine	488	340	207
Psychiatry	460	405	331
Hematology	457	387	300
Endocrinology	452	368	275
Nephrology	362	313	236
Dermatology	339	298	237
Orthopedics	293	244	194
Respiratory medicine	280	202	98
Rheumatology	250	216	168
Total	7125	5669	3982

Table 3: Train sets description. Number of samples of the 15 specialties using different minimal numbers of votes (3+, 4+, 5) in the train sets included in S-MedQA.

the coverage and accuracy after applying different majority voting strategies, (i.e. at least 3, 4, 5 responses reach an agreement). Only the questions that obtain at least this number of votes are kept. There is an inherent trade-off between accuracy and coverage when deciding the threshold to use for majority voting (i.e., requiring a minimum of 3, 4, or 5 votes to agree in order to include an example).

In practice, we release all individual prompt predictions, as well as three versions of the dataset for majority voting with a minimum of 3+, 4+, and 5 votes. A coverage of 89.1% of the data leads to clinical specialties that are 90.8% accurate, whereas in the other side of the spectrum, we can obtain an accuracy of 97.8% while the coverage decreases to 49.2%. By sharing multiple versions of S-MedQA, we cater to different users' needs. Users can then use more data (coverage of 89.1%) if their use-case can cope with mistakes in the order of 10% (majority voting 3+); if the use-case requires data akin to gold-standard, i.e., error-free, users can use majority voting 5 (which basically requires all 5 prompts to agree for an example to be included), which provides an accuracy of 97.8%.

A.6 Prompt for LLM tuning and inferring

An example of the prompt we use for LLM tuning and inferring in all our experiments is as follows:

[INST] Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.
Question: A 62-year-old woman presents for a regular check-up. She complains of lightheadedness and palpitations which occur episodically. Past medical history is significant for a myocardial infarction 6 months ago and NYHA class II chronic heart failure. She also was diagnosed with grade I arterial hypertension 4 years ago. Current medications are aspirin 81 mg, atorvastatin 10 mg, enalapril 10 mg, and metoprolol 200 mg daily. Her vital signs are a blood pressure of 135/90 mm Hg, a heart rate of 125/min, a respiratory rate of 14/min, and a temperature of 36.5ŏ0b0C (97.7ŏ0b0F). Cardiopulmonary examination is significant for irregular heart rhythm and decreased S1 intensity. ECG is obtained and is shown in the picture (see image). Echocardiography shows a left ventricular ejection fraction of 39%. Which of the following drugs is the best choice for rate control in this patient?
A. Atenolol
B. Diltiazem
C. Propafenone
D. Digoxin
Answer: [/INST] D. Digoxin

A.7 Training settings and hyperparameters

We use LoRA (Hu et al., 2021) on all projection layers for the fine-tuning process in all experiments. The hyperparameters are as follows: learning rate=2e-5, rank=32, alpha=16, dropout

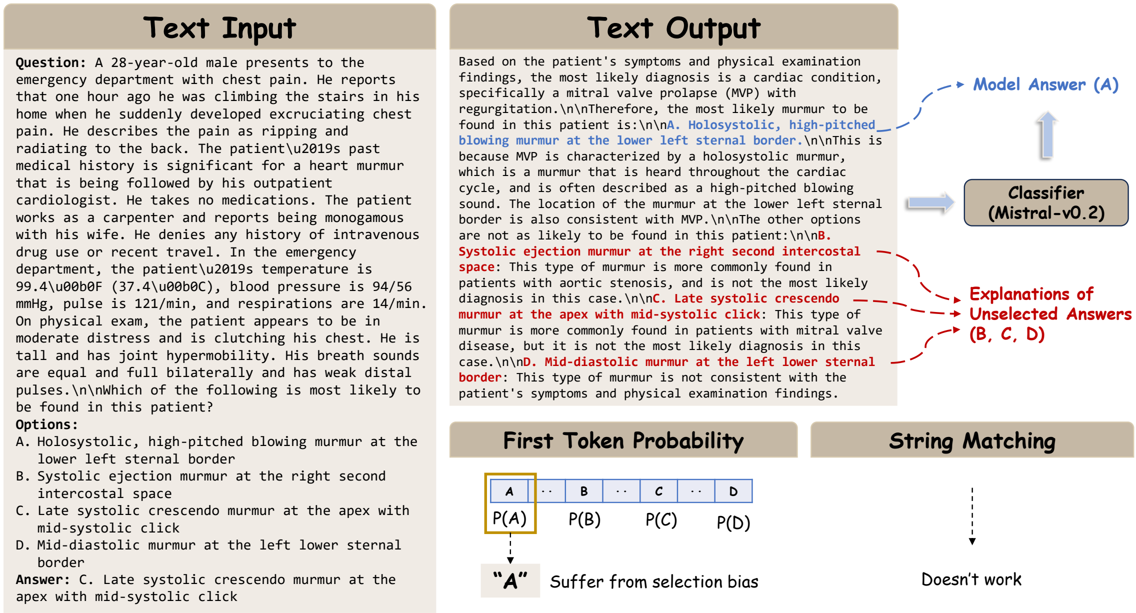

Figure 3: The illustration of first token probability, string matching, and our approach (classifier) to evaluating LLMs performance on S-MedQA. We use text output instead of first token probability for evaluation because first token probability suffers heavily from selection bias in multiple-choice question answering (Wang et al., 2024). However, string matching does not work in some cases. Our classifier trained on Mistral-v0.2 works successfully with an accuracy of 96.5%.

rate=0.1, batch size=8.

A.8 Illustration of Evaluation Process

Figure 3 illustrates the approach (classifier) we use to evaluate the performance of the models and counter the shortage of first token probability (Wang et al., 2024) and simple string matching. The classifier is trained based on Mistral-v0.2 and applied in all experiments.

A.9 Cross-specialty evaluation results for Llama 13B, Mistral 7B v0.1 and v0.2

In Tables 5, 6, and 7, we show cross-specialty evaluation matrices for Llama2-7b-chat, Llama2-13b-chat, and Mistral-7b-instruct-v0.1 in addition to our main results (in §4). Here we also observe that the best performance on each per-specialty test set is not achieved by the model that is tuned on training data from the same specialty for Llama models. However, Mistral-7b-instruct-v0.1 shows an opposing trend: the best performance is almost always obtained with the model trained on the same specialty.

When comparing these results with those obtained with Mistral-7b-instruct-v0.2 (in Table2 in our main paper), we note that results with Mistral-7b-instruct-v0.2 (in our main paper) are overall better than those obtained with Mistral-7b-instruct-v0.1 (in Table 7). We believe that answering why some models better transfer within-specialty than across-specialties (or vice-versa) warrants further research on this topic.

	Train Sets	Test Sets						
		Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
Train Sets	Cardio	<u>34.3</u>	29.1	32.6	28.3	31.5	29.5	31.0
	Gastro	31.3	<u>**32.5**</u>	30.7	28.8	**38.3**	30.1	32.0
	Infect	**35.0**	31.3	<u>32.8</u>	31.0	33.3	29.0	32.1
	Neuro	32.8	26.3	**35.4**	<u>**31.3**</u>	33.5	**36.1**	32.7
	Obstetrics	34.5	30.0	34.6	30.7	<u>37.9</u>	33.2	33.6
	Pediatrics	30.5	27.8	34.1	25.8	33.5	<u>31.0</u>	30.7
	Combined	42.0	40.1	38.5	37.2	42.5	36.1	39.4
	Llama2-7b	36.0	36.3	36.7	34.6	40.6	41.4	37.7

Table 5: Cross-specialty Accuracy Matrix of Llama2-7b.

	Train Sets	Test Sets						
		Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
Train Sets	Cardio	<u>38.3</u>	34.7	**41.1**	28.5	**38.8**	31.8	35.8
	Gastro	**38.5**	<u>**35.8**</u>	31.8	31.3	36.0	32.7	34.3
	Infect	36.0	31.3	<u>36.5</u>	32.9	39.2	32.7	34.9
	Neuroy	31.3	29.3	36.2	<u>28.8</u>	35.4	33.5	32.7
	Obstetrics	33.5	29.5	36.5	30.4	<u>36.0</u>	32.7	33.3
	Pediatrics	37.5	34.9	37.0	**33.2**	38.5	<u>**36.1**</u>	36.3
	Combined	44.0	45.9	42.4	40.2	45.8	42.9	43.6
	Llama2-13b	43.3	34.9	40.6	36.1	45.4	39.2	40.0

Table 6: Cross-specialty Accuracy Matrix of Llama2-13b.

	Train Sets	Test Sets						
		Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
Train Sets	Cardio	<u>**52.8**</u>	46.1	47.7	39.9	47.9	44.3	46.6
	Gastro	48.3	<u>**50.4**</u>	41.1	40.2	47.5	44.3	45.2
	Infect	51.5	42.7	<u>**49.0**</u>	**44.3**	47.5	43.5	46.5
	Neuro	50.7	46.3	47.1	<u>44.0</u>	49.6	**48.9**	47.8
	Obstetrics	45.8	44.2	44.3	41.3	<u>**53.8**</u>	46.3	46.1
	Pediatrics	48.8	45.5	42.7	35.1	51.0	<u>**48.9**</u>	45.5
	Combined	52.8	47.6	46.4	49.5	56.3	47.2	49.9
	Mistral-v0.1	41.0	39.0	40.0	30.7	42.7	37.5	38.9

Table 7: Cross-specialty Accuracy Matrix of Mistral-v0.1

Figure 4: Prompt-1

System: Please classify the medical multiple choice question into one of the following clinical specialties: *Emergency medicine*, *Allergist*, *Anaesthetics*, *Cardiology*, *Child psychiatry*, *Clinical biology*, *Clinical chemistry*, *Clinical microbiology*, *Clinical neurophysiology*, *Craniofacial surgery*, *Dermatology*, *Endocrinology*, *Family and General Medicine*, *Gastroenterologic surgery*, *Gastroenterology*, *General Practice*, *General surgery*, *Geriatrics*, *Hematology*, *Immunology*, *Infectious diseases*, *Internal medicine*, *Laboratory medicine*, *Nephrology*, *Neuropsychiatry*, *Neurology*, *Neurosurgery*, *Nuclear medicine*, *Obstetrics and gynecology*, *Occupational medicine*, *Oncology*, *Ophthalmology*, *Oral and maxillofacial surgery*, *Orthopedics*, *Otorhinolaryngology*, *Pediatric surgery*, *Pediatrics*, *Pathology*, *Pharmacology*, *Physical medicine and rehabilitation*, *Plastic surgery*, *Podiatric surgery*, *Preventive medicine*, *Psychiatry*, *Public health*, *Radiation Oncology*, *Radiology*, *Respiratory medicine*, *Rheumatology*, *Stomatology*, *Thoracic surgery*, *Tropical medicine*, *Urology*, *Vascular surgery*, *Venereology*, *Others*

User: A 39-year-old woman comes to the physician because of an 8-month history of progressive fatigue, shortness of breath, and palpitations. She has a history of recurrent episodes of joint pain and fever during childhood. She emigrated from India with her parents when she was 10 years old. Cardiac examination shows an opening snap followed by a late diastolic rumble, which is best heard at the fifth intercostal space in the left midclavicular line. This patient is at greatest risk for compression of which of the following structures?

Figure 5: Prompt-2

System: You are medical student taking a multiple choice exam. The knowledge of which of the following clinical specialties is the most helpful to answering the question: *Emergency medicine*, *Allergist*, *Anaesthetics*, *Cardiology*, *Child psychiatry*, *Clinical biology*, *Clinical chemistry*, *Clinical microbiology*, *Clinical neurophysiology*, *Craniofacial surgery*, *Dermatology*, *Endocrinology*, *Family and General Medicine*, *Gastroenterologic surgery*, *Gastroenterology*, *General Practice*, *General surgery*, *Geriatrics*, *Hematology*, *Immunology*, *Infectious diseases*, *Internal medicine*, *Laboratory medicine*, *Nephrology*, *Neuropsychiatry*, *Neurology*, *Neurosurgery*, *Nuclear medicine*, *Obstetrics and gynecology*, *Occupational medicine*, *Oncology*, *Ophthalmology*, *Oral and maxillofacial surgery*, *Orthopedics*, *Otorhinolaryngology*, *Pediatric surgery*, *Pediatrics*, *Pathology*, *Pharmacology*, *Physical medicine and rehabilitation*, *Plastic surgery*, *Podiatric surgery*, *Preventive medicine*, *Psychiatry*, *Public health*, *Radiation Oncology*, *Radiology*, *Respiratory medicine*, *Rheumatology*, *Stomatology*, *Thoracic surgery*, *Tropical medicine*, *Urology*, *Vascular surgery*, *Venereology*, *Others*

Here are some examples:
Question: A 62-year-old woman presents for a regular check-up. She complains of lightheadedness and palpitations which occur episodically. Past medical history is significant for a myocardial infarction 6 months ago and NYHA class II chronic heart failure. She also was diagnosed with grade I arterial hypertension 4 years ago. Current medications are aspirin 81 mg, atorvastatin 10 mg, enalapril 10 mg, and metoprolol 200 mg daily. Her vital signs are a blood pressure of 135/90 mm Hg, a heart rate of 125/min, a respiratory rate of 14/min, and a temperature of 36.5°C (97.7°F). Cardiopulmonary examination is significant for irregular heart rhythm and decreased S1 intensity. ECG is obtained and is shown in the picture (see image). Echocardiography shows a left ventricular ejection fraction of 39%. Which of the following drugs is the best choice for rate control in this patient?
Answer: Cardiology
Question: A 68-year-old man comes to the physician because of recurrent episodes of nausea and abdominal discomfort for the past 4 months. The discomfort is located in the upper abdomen and sometimes occurs after eating, especially after a big meal. He has tried to go for a walk after dinner to help with digestion, but his complaints have only increased. For the past 3 weeks he has also had symptoms while climbing the stairs to his apartment. He has type 2 diabetes mellitus, hypertension, and stage 2 peripheral arterial disease. He has smoked one pack of cigarettes daily for the past 45 years. He drinks one to two beers daily and occasionally more on weekends. His current medications include metformin, enalapril, and aspirin. He is 168 cm (5 ft 6 in) tall and weighs 126 kg (278 lb); BMI is 45 kg/m2. His temperature is 36.4°C (97.5°F), pulse is 78/min, and blood pressure is 148/86 mm Hg. On physical examination, the abdomen is soft and nontender with no organomegaly. Foot pulses are absent bilaterally. An ECG shows no abnormalities. Which of the following is the most appropriate next step in diagnosis?
Answer: Gastroenterology
Question: A 6-year-old male who recently immigrated to the United States from Asia is admitted to the hospital with dyspnea. Physical exam reveals a gray pseudomembrane in the patient's oropharynx along with lymphadenopathy. The patient develops myocarditis and expires on hospital day 5. Which of the following would have prevented this patient's presentation and decline?
Answer: Infectious diseases
Question: A 35-year-old woman with a history of Crohn disease presents for a follow-up appointment. She says that lately, she has started to notice difficulty walking. She says that some of her friends have joked that she appears to be walking as if she was drunk. Past medical history is significant for Crohn disease diagnosed 2 years ago, managed with natalizumab for the past year because her intestinal symptoms have become severe and unresponsive to other therapies. On physical examination, there is gait and limb ataxia present. Strength is 4/5 in the right upper limb. A T1/T2 MRI of the brain is ordered and is shown. Which of the following is the most likely diagnosis?
Answer: Neurology
Question: A 23-year-old G1 at 10 weeks gestation based on her last menstrual period is brought to the emergency department by her husband due to sudden vaginal bleeding. She says that she has mild lower abdominal cramps and is feeling dizzy and weak. Her blood pressure is 100/60 mm Hg, the pulse is 100/min, and the respiration rate is 15/min. She says that she has had light spotting over the last 3 days, but today the bleeding increased markedly and she also noticed the passage of clots. She says that she has changed three pads since the morning. She has also noticed that the nausea she was experiencing over the past few days has subsided. The physician examines her and notes that the cervical os is open and blood is pooling in the vagina. Products of conception can be visualized in the os. The patient is prepared for a suction curettage. Which of the following is the most likely cause for the pregnancy loss?
Answer: Obstetrics and gynecology
Question: An 8-month-old boy is brought to a medical office by his mother. The mother states that the boy has been very fussy and has not been feeding recently. The mother thinks the baby has been gaining weight despite not feeding well. The boy was delivered vaginally at 39 weeks gestation without complications. On physical examination, the boy is noted to be crying in his mother's arms. There is no evidence of cyanosis, and the cardiac examination is within normal limits. The crying intensifies when the abdomen is palpated. The abdomen is distended with tympany in the left lower quadrant. You suspect a condition caused by the failure of specialized cells to migrate. What is the most likely diagnosis?
Answer: Pediatrics

User: A 39-year-old woman comes to the physician because of an 8-month history of progressive fatigue, shortness of breath, and palpitations. She has a history of recurrent episodes of joint pain and fever during childhood. She emigrated from India with her parents when she was 10 years old. Cardiac examination shows an opening snap followed by a late diastolic rumble, which is best heard at the fifth intercostal space in the left midclavicular line. This patient is at greatest risk for compression of which of the following structures?

Figure 6: Prompt-3

System: Please classify the medical multiple choice question into one of the following clinical specialties: *Emergency medicine*, *Allergist*, *Anaesthetics*, *Cardiology*, *Child psychiatry*, *Clinical biology*, *Clinical chemistry*, *Clinical microbiology*, *Clinical neurophysiology*, *Craniofacial surgery*, *Dermatology*, *Endocrinology*, *Family and General Medicine*, *Gastroenterologic surgery*, *Gastroenterology*, *General Practice*, *General surgery*, *Geriatrics*, *Hematology*, *Immunology*, *Infectious diseases*, *Internal medicine*, *Laboratory medicine*, *Nephrology*, *Neuropsychiatry*, *Neurology*, *Neurosurgery*, *Nuclear medicine*, *Obstetrics and gynecology*, *Occupational medicine*, *Oncology*, *Ophthalmology*, *Oral and maxillofacial surgery*, *Orthopedics*, *Otorhinolaryngology*, *Pediatric surgery*, *Pediatrics*, *Pathology*, *Pharmacology*, *Physical medicine and rehabilitation*, *Plastic surgery*, *Podiatric surgery*, *Preventive medicine*, *Psychiatry*, *Public health*, *Radiation Oncology*, *Radiology*, *Respiratory medicine*, *Rheumatology*, *Stomatology*, *Thoracic surgery*, *Tropical medicine*, *Urology*, *Vascular surgery*, *Venereology*, *Others*

Here are some examples:
Question: A 62-year-old woman presents for a regular check-up. She complains of lightheadedness and palpitations which occur episodically. Past medical history is significant for a myocardial infarction 6 months ago and NYHA class II chronic heart failure. She also was diagnosed with grade I arterial hypertension 4 years ago. Current medications are aspirin 81 mg, atorvastatin 10 mg, enalapril 10 mg, and metoprolol 200 mg daily. Her vital signs are a blood pressure of 135/90 mm Hg, a heart rate of 125/min, a respiratory rate of 14/min, and a temperature of 36.5°C (97.7°F). Cardiopulmonary examination is significant for irregular heart rhythm and decreased S1 intensity. ECG is obtained and is shown in the picture (see image). Echocardiography shows a left ventricular ejection fraction of 39%. Which of the following drugs is the best choice for rate control in this patient?
Answer: Cardiology
Question: A 68-year-old man comes to the physician because of recurrent episodes of nausea and abdominal discomfort for the past 4 months. The discomfort is located in the upper abdomen and sometimes occurs after eating, especially after a big meal. He has tried to go for a walk after dinner to help with digestion, but his complaints have only increased. For the past 3 weeks he has also had symptoms while climbing the stairs to his apartment. He has type 2 diabetes mellitus, hypertension, and stage 2 peripheral arterial disease. He has smoked one pack of cigarettes daily for the past 45 years. He drinks one to two beers daily and occasionally more on weekends. His current medications include metformin, enalapril, and aspirin. He is 168 cm (5 ft 6 in) tall and weighs 126 kg (278 lb); BMI is 45 kg/m2. His temperature is 36.4°C (97.5°F), pulse is 78/min, and blood pressure is 148/86 mm Hg. On physical examination, the abdomen is soft and nontender with no organomegaly. Foot pulses are absent bilaterally. An ECG shows no abnormalities. Which of the following is the most appropriate next step in diagnosis?
Answer: Gastroenterology
Question: A 6-year-old male who recently immigrated to the United States from Asia is admitted to the hospital with dyspnea. Physical exam reveals a gray pseudomembrane in the patient's oropharynx along with lymphadenopathy. The patient develops myocarditis and expires on hospital day 5. Which of the following would have prevented this patient's presentation and decline?
Answer: Infectious diseases
Question: A 35-year-old woman with a history of Crohn disease presents for a follow-up appointment. She says that lately, she has started to notice difficulty walking. She says that some of her friends have joked that she appears to be walking as if she was drunk. Past medical history is significant for Crohn disease diagnosed 2 years ago, managed with natalizumab for the past year because her intestinal symptoms have become severe and unresponsive to other therapies. On physical examination, there is gait and limb ataxia present. Strength is 4/5 in the right upper limb. A T1/T2 MRI of the brain is ordered and is shown. Which of the following is the most likely diagnosis?
Answer: Neurology
Question: A 23-year-old G1 at 10 weeks gestation based on her last menstrual period is brought to the emergency department by her husband due to sudden vaginal bleeding. She says that she has mild lower abdominal cramps and is feeling dizzy and weak. Her blood pressure is 100/60 mm Hg, the pulse is 100/min, and the respiration rate is 15/min. She says that she has had light spotting over the last 3 days, but today the bleeding increased markedly and she also noticed the passage of clots. She says that she has changed three pads since the morning. She has also noticed that the nausea she was experiencing over the past few days has subsided. The physician examines her and notes that the cervical os is open and blood is pooling in the vagina. Products of conception can be visualized in the os. The patient is prepared for a suction curettage. Which of the following is the most likely cause for the pregnancy loss?
Answer: Obstetrics and gynecology
Question: An 8-month-old boy is brought to a medical office by his mother. The mother states that the boy has been very fussy and has not been feeding recently. The mother thinks the baby has been gaining weight despite not feeding well. The boy was delivered vaginally at 39 weeks gestation without complications. On physical examination, the boy is noted to be crying in his mother's arms. There is no evidence of cyanosis, and the cardiac examination is within normal limits. The crying intensifies when the abdomen is palpated. The abdomen is distended with tympany in the left lower quadrant. You suspect a condition caused by the failure of specialized cells to migrate. What is the most likely diagnosis?
Answer: Pediatrics

User: A 39-year-old woman comes to the physician because of an 8-month history of progressive fatigue, shortness of breath, and palpitations. She has a history of recurrent episodes of joint pain and fever during childhood. She emigrated from India with her parents when she was 10 years old. Cardiac examination shows an opening snap followed by a late diastolic rumble, which is best heard at the fifth intercostal space in the left midclavicular line. This patient is at greatest risk for compression of which of the following structures?

Figure 7: Prompt-4

System: Please classify the medical multiple choice question into one of the clinical specialties.

Here are some examples:
Question: A 62-year-old woman presents for a regular check-up. She complains of lightheadedness and palpitations which occur episodically. Past medical history is significant for a myocardial infarction 6 months ago and NYHA class II chronic heart failure. She also was diagnosed with grade I arterial hypertension 4 years ago. Current medications are aspirin 81 mg, atorvastatin 10 mg, enalapril 10 mg, and metoprolol 200 mg daily. Her vital signs are a blood pressure of 135/90 mm Hg, a heart rate of 125/min, a respiratory rate of 14/min, and a temperature of 36.5°C (97.7°F). Cardiopulmonary examination is significant for irregular heart rhythm and decreased S1 intensity. ECG is obtained and is shown in the picture (see image). Echocardiography shows a left ventricular ejection fraction of 39%. Which of the following drugs is the best choice for rate control in this patient?
Answer: Cardiology
Question: A 68-year-old man comes to the physician because of recurrent episodes of nausea and abdominal discomfort for the past 4 months. The discomfort is located in the upper abdomen and sometimes occurs after eating, especially after a big meal. He has tried to go for a walk after dinner to help with digestion, but his complaints have only increased. For the past 3 weeks he has also had symptoms while climbing the stairs to his apartment. He has type 2 diabetes mellitus, hypertension, and stage 2 peripheral arterial disease. He has smoked one pack of cigarettes daily for the past 45 years. He drinks one to two beers daily and occasionally more on weekends. His current medications include metformin, enalapril, and aspirin. He is 168 cm (5 ft 6 in) tall and weighs 126 kg (278 lb); BMI is 45 kg/m2. His temperature is 36.4°C (97.5°F), pulse is 78/min, and blood pressure is 148/86 mm Hg. On physical examination, the abdomen is soft and nontender with no organomegaly. Foot pulses are absent bilaterally. An ECG shows no abnormalities. Which of the following is the most appropriate next step in diagnosis?
Answer: Gastroenterology
Question: A 6-year-old male who recently immigrated to the United States from Asia is admitted to the hospital with dyspnea. Physical exam reveals a gray pseudomembrane in the patient's oropharynx along with lymphadenopathy. The patient develops myocarditis and expires on hospital day 5. Which of the following would have prevented this patient's presentation and decline?
Answer: Infectious diseases
Question: A 35-year-old woman with a history of Crohn disease presents for a follow-up appointment. She says that lately, she has started to notice difficulty walking. She says that some of her friends have joked that she appears to be walking as if she was drunk. Past medical history is significant for Crohn disease diagnosed 2 years ago, managed with natalizumab for the past year because her intestinal symptoms have become severe and unresponsive to other therapies. On physical examination, there is gait and limb ataxia present. Strength is 4/5 in the right upper limb. A T1/T2 MRI of the brain is ordered and is shown. Which of the following is the most likely diagnosis?
Answer: Neurology
Question: A 23-year-old G1 at 10 weeks gestation based on her last menstrual period is brought to the emergency department by her husband due to sudden vaginal bleeding. She says that she has mild lower abdominal cramps and is feeling dizzy and weak. Her blood pressure is 100/60 mm Hg, the pulse is 100/min, and the respiration rate is 15/min. She says that she has had light spotting over the last 3 days, but today the bleeding increased markedly and she also noticed the passage of clots. She says that she has changed three pads since the morning. She has also noticed that the nausea she was experiencing over the past few days has subsided. The physician examines her and notes that the cervical os is open and blood is pooling in the vagina. Products of conception can be visualized in the os. The patient is prepared for a suction curettage. Which of the following is the most likely cause for the pregnancy loss?
Answer: Obstetrics and gynecology
Question: An 8-month-old boy is brought to a medical office by his mother. The mother states that the boy has been very fussy and has not been feeding recently. The mother thinks the baby has been gaining weight despite not feeding well. The boy was delivered vaginally at 39 weeks gestation without complications. On physical examination, the boy is noted to be crying in his mother's arms. There is no evidence of cyanosis, and the cardiac examination is within normal limits. The crying intensifies when the abdomen is palpated. The abdomen is distended with tympany in the left lower quadrant. You suspect a condition caused by the failure of specialized cells to migrate. What is the most likely diagnosis?
Answer: Pediatrics

User: A 39-year-old woman comes to the physician because of an 8-month history of progressive fatigue, shortness of breath, and palpitations. She has a history of recurrent episodes of joint pain and fever during childhood. She emigrated from India with her parents when she was 10 years old. Cardiac examination shows an opening snap followed by a late diastolic rumble, which is best heard at the fifth intercostal space in the left midclavicular line. This patient is at greatest risk for compression of which of the following structures?
Please classify the medical multiple choice question into one of the following clinical specialties: *Emergency medicine*, *Allergist*, *Anaesthetics*, *Cardiology*, *Child psychiatry*, *Clinical biology*, *Clinical chemistry*, *Clinical microbiology*, *Clinical neurophysiology*, *Craniofacial surgery*, *Dermatology*, *Endocrinology*, *Family and General Medicine*, *Gastroenterologic surgery*, *Gastroenterology*, *General Practice*, *General surgery*, *Geriatrics*, *Hematology*, *Immunology*, *Infectious diseases*, *Internal medicine*, *Laboratory medicine*, *Nephrology*, *Neuropsychiatry*, *Neurology*, *Neurosurgery*, *Nuclear medicine*, *Obstetrics and gynecology*, *Occupational medicine*, *Oncology*, *Ophthalmology*, *Oral and maxillofacial surgery*, *Orthopedics*, *Otorhinolaryngology*, *Pediatric surgery*, *Pediatrics*, *Pathology*, *Pharmacology*, *Physical medicine and rehabilitation*, *Plastic surgery*, *Podiatric surgery*, *Preventive medicine*, *Psychiatry*, *Public health*, *Radiation Oncology*, *Radiology*, *Respiratory medicine*, *Rheumatology*, *Stomatology*, *Thoracic surgery*, *Tropical medicine*, *Urology*, *Vascular surgery*, *Venereology*, *Others*

Figure 8: Prompt-5

System: Please classify the medical multiple choice question into one of the following clinical specialties: *Emergency medicine*, *Allergist*, *Anaesthetics*, *Cardiology*, *Child psychiatry*, *Clinical biology*, *Clinical chemistry*, *Clinical microbiology*, *Clinical neurophysiology*, *Craniofacial surgery*, *Dermatology*, *Endocrinology*, *Family and General Medicine*, *Gastroenterologic surgery*, *Gastroenterology*, *General Practice*, *General surgery*, *Geriatrics*, *Hematology*, *Immunology*, *Infectious diseases*, *Internal medicine*, *Laboratory medicine*, *Nephrology*, *Neuropsychiatry*, *Neurology*, *Neurosurgery*, *Nuclear medicine*, *Obstetrics and gynecology*, *Occupational medicine*, *Oncology*, *Ophthalmology*, *Oral and maxillofacial surgery*, *Orthopedics*, *Otorhinolaryngology*, *Pediatric surgery*, *Pediatrics*, *Pathology*, *Pharmacology*, *Physical medicine and rehabilitation*, *Plastic surgery*, *Podiatric surgery*, *Preventive medicine*, *Psychiatry*, *Public health*, *Radiation Oncology*, *Radiology*, *Respiratory medicine*, *Rheumatology*, *Stomatology*, *Thoracic surgery*, *Tropical medicine*, *Urology*, *Vascular surgery*, *Venereology*, *Others*

Here are some examples:
Question: A 62-year-old woman presents for a regular check-up. She complains of lightheadedness and palpitations which occur episodically. Past medical history is significant for a myocardial infarction 6 months ago and NYHA class II chronic heart failure. She also was diagnosed with grade I arterial hypertension 4 years ago. Current medications are aspirin 81 mg, atorvastatin 10 mg, enalapril 10 mg, and metoprolol 200 mg daily. Her vital signs are a blood pressure of 135/90 mm Hg, a heart rate of 125/min, a respiratory rate of 14/min, and a temperature of 36.5°C (97.7°F). Cardiopulmonary examination is significant for irregular heart rhythm and decreased S1 intensity. ECG is obtained and is shown in the picture (see image). Echocardiography shows a left ventricular ejection fraction of 39%. Which of the following drugs is the best choice for rate control in this patient?
Answer: Cardiology
Question: A 68-year-old man comes to the physician because of recurrent episodes of nausea and abdominal discomfort for the past 4 months. The discomfort is located in the upper abdomen and sometimes occurs after eating, especially after a big meal. He has tried to go for a walk after dinner to help with digestion, but his complaints have only increased. For the past 3 weeks he has also had symptoms while climbing the stairs to his apartment. He has type 2 diabetes mellitus, hypertension, and stage 2 peripheral arterial disease. He has smoked one pack of cigarettes daily for the past 45 years. He drinks one to two beers daily and occasionally more on weekends. His current medications include metformin, enalapril, and aspirin. He is 168 cm (5 ft 6 in) tall and weighs 126 kg (278 lb); BMI is 45 kg/m2. His temperature is 36.4°C (97.5°F), pulse is 78/min, and blood pressure is 148/86 mm Hg. On physical examination, the abdomen is soft and nontender with no organomegaly. Foot pulses are absent bilaterally. An ECG shows no abnormalities. Which of the following is the most appropriate next step in diagnosis?
Answer: Gastroenterology
Question: A 6-year-old male who recently immigrated to the United States from Asia is admitted to the hospital with dyspnea. Physical exam reveals a gray pseudomembrane in the patient's oropharynx along with lymphadenopathy. The patient develops myocarditis and expires on hospital day 5. Which of the following would have prevented this patient's presentation and decline?
Answer: Infectious diseases
Question: A 35-year-old woman with a history of Crohn disease presents for a follow-up appointment. She says that lately, she has started to notice difficulty walking. She says that some of her friends have joked that she appears to be walking as if she was drunk. Past medical history is significant for Crohn disease diagnosed 2 years ago, managed with natalizumab for the past year because her intestinal symptoms have become severe and unresponsive to other therapies. On physical examination, there is gait and limb ataxia present. Strength is 4/5 in the right upper limb. A T1/T2 MRI of the brain is ordered and is shown. Which of the following is the most likely diagnosis?
Answer: Neurology
Question: A 23-year-old G1 at 10 weeks gestation based on her last menstrual period is brought to the emergency department by her husband due to sudden vaginal bleeding. She says that she has mild lower abdominal cramps and is feeling dizzy and weak. Her blood pressure is 100/60 mm Hg, the pulse is 100/min, and the respiration rate is 15/min. She says that she has had light spotting over the last 3 days, but today the bleeding increased markedly and she also noticed the passage of clots. She says that she has changed three pads since the morning. She has also noticed that the nausea she was experiencing over the past few days has subsided. The physician examines her and notes that the cervical os is open and blood is pooling in the vagina. Products of conception can be visualized in the os. The patient is prepared for a suction curettage. Which of the following is the most likely cause for the pregnancy loss?
Answer: Obstetrics and gynecology
Question: An 8-month-old boy is brought to a medical office by his mother. The mother states that the boy has been very fussy and has not been feeding recently. The mother thinks the baby has been gaining weight despite not feeding well. The boy was delivered vaginally at 39 weeks gestation without complications. On physical examination, the boy is noted to be crying in his mother's arms. There is no evidence of cyanosis, and the cardiac examination is within normal limits. The crying intensifies when the abdomen is palpated. The abdomen is distended with tympany in the left lower quadrant. You suspect a condition caused by the failure of specialized cells to migrate. What is the most likely diagnosis?
Answer: Pediatrics

User: Question: A 39-year-old woman comes to the physician because of an 8-month history of progressive fatigue, shortness of breath, and palpitations. She has a history of recurrent episodes of joint pain and fever during childhood. She emigrated from India with her parents when she was 10 years old. Cardiac examination shows an opening snap followed by a late diastolic rumble, which is best heard at the fifth intercostal space in the left midclavicular line. This patient is at greatest risk for compression of which of the following structures?
Answer: