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

Anonymous ACL submission

Abstract

In this paper, we introduce S-MedQA, an English medical question-answering (QA) dataset for benchmarking large language models in fine-grained clinical specialties. We use S-MedQA to check the applicability of a popular hypothesis related to knowledge injection in the knowledge-intense scenario of medical QA, and show that 1) training on data from a specialty does not necessarily lead to best performance on that specialty and 2) regardless of the specialty fine-tuned on, token probabilities of clinically relevant terms for all specialties increase consistently. Thus, we believe improvement gains come mostly from domain shifting (e.g., general to medical) rather than knowledge injection and suggest rethinking the role of fine-tuning data in the medical domain. We release S-MedQA and all code needed to reproduce all our experiments to the research community.[1]

1 Introduction

Multiple-choice question-answering (QA) datasets are widely used to benchmark large language models (LLMs) in the medical domain (Singhal et al., 2023; Labrak et al., 2024) and 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 may require LLMs to address specific clinical problems and are often interested in performance within *one or a few clinical specialties* (e.g., obstetrics or oncology). To the best of our knowledge, no open-source medical QA datasets include clinical specialty annotations, limiting research into knowledge transfer across clinical specialties.

To address this gap, we develop **S-MedQA**, the first English medical QA dataset with multiple clinical specialty annotations. We build S-MedQA based on the widely used MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) datasets, and use `gpt-3.5-turbo-0125` (GPT-3.5) and medical experts to map samples onto clinical specialties. We label QA pairs with single specialties using tailored prompts, retaining only majority-agreed annotations. Expert validation on single specialty achieves up to 97.8% accuracy (details in §2.4). S-MedQA spans 15 specialties, each with hundreds to thousands of samples (see §2 for details). We then expand these annotations to multiple specialties using conformal prediction with calibrated confidence guarantees, achieving a coverage rate of 95% and an F-1 score of 0.60 at a 95% confidence level on a curated calibration set.

We use our dataset to investigate the hypothesis that almost all knowledge in LLMs originates from pretraining, and that fine-tuning primarily serves to shift the model toward a specific knowledge domain (Zhou et al., 2024). To that regard, we first fine-tune LLMs on one specialty and evaluate on other specialties (see §3.2). Interestingly, *the best results often come from fine-tuning on unrelated specialties*. E.g., fine-tuning with *infectious disease* data performs best on *cardiology* domain, despite their knowledge being largely unrelated. Moreover, different pre-trained LLMs exhibits different knowledge transfer patterns across clinical specialties (see §A.12). In further experiments, we curate clinically relevant per-specialty terms (e.g., diseases, procedures) and analyze changes in these term probabilities before and after fine-tuning on data from different specialties. Our results suggest that performance gains are driven more by domain shifts (from *general* to *medical* domain) than by fine-grained clinical knowledge injection.

2 A Benchmark of clinical specialties

We describe the creation of S-MedQA, a high-quality benchmark for medical QA with clinical specialty annotations (overview in Fig. 1). We release multiple versions of S-MedQA with varying

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

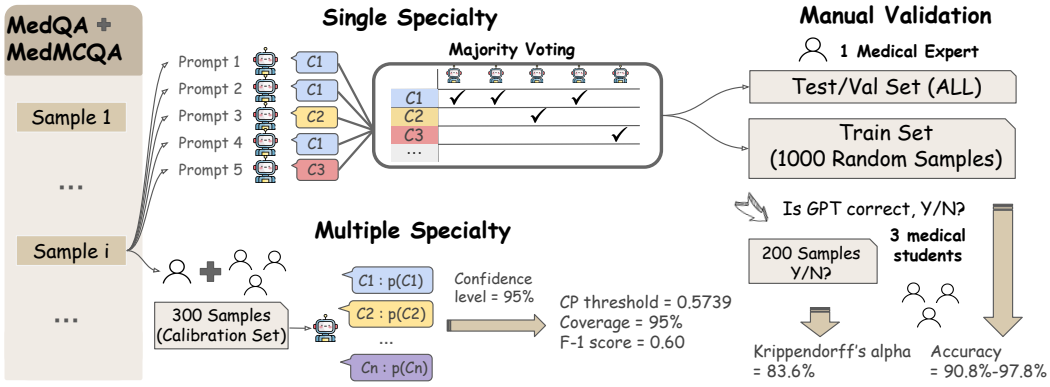

Figure 1: Overview of S-MedQA's construction process. For single specialty annotation of each sample, we generate predictions using 5 different prompts and only keep those where predictions agree (3+, 4+, or 5 times). We randomly sample $1,000$ questions from our train set and ask a medical expert to evaluate GPT-3.5's predictions, achieving an accuracy ranging from 97.8% to 90.8%. Three medical students then annotate the same 200 samples out of the original $1,000$ annotated samples for computing inter-annotator agreements (see §2.4 for details). The expert also manually annotates S-MedQA's whole validation and test set. For multi-specialty annotation, we leverage conformal prediction to assign labels across multiple clinical specialties, using a calibration set of 300 samples manually annotated by the medical expert and medical students. At a target confidence level of 90%, we determine a conformal threshold of 0.5739, which is then applied to all remaining samples for calibrated predictions.

accuracy/coverage trade-offs, controlled by majority voting thresholds for including examples. Users can opt for a *cleaner dataset with fewer samples* or a *noisier one with more samples* (details in §2.4).

Data and splits We source examples from MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022), two widely used medical QA datasets. MedQA samples follow their original train/valid/test splits, whereas we only use MedMCQA's training split as its test labels are not public.

2.1 Clinical specialty categorization

We consider the 55 medical specialties recognized in the European Union for labeling (see §A.1).

Single clinical specialties Manually labeling QA examples with clinical specialties is costly and time-consuming. To reduce manual effort, we use GPT-3.5 to annotate samples with single clinical specialties. However, preliminary experiments with a single prompt to predict a single specialty achieve low accuracy($\sim$ 75%). To improve this, we design five prompts (details in §A.2), generate single specialty predictions with GPT-3.5 for each Q&A sample, and apply majority voting to select the specialty (Ding et al., 2023; Goel et al., 2023). Human evaluation shows accuracies between 90.8%–97.8% (see §2.4 for details).

From single to multiple clinical specialties Single specialty labels do not always capture the complexity of clinical questions. To account for this, we applied conformal prediction (Angelopoulos and Bates, 2021) to generate calibrated multi-specialty labels in S-MedQA. At a 95% confidence level, we achieve an average coverage of 95% and an F1 score of 0.60 on the calibration set (see §A.5 for the methods and evaluation).

2.2 S-MedQA's predicted specialties

In §A.3 (Fig. 3) we show the distribution of predicted specialties. We exclude 2310 (6%) samples categorized as *Others* (containing clinically irrelevant information), and then focus on 15 out of 55 specialties with more than 300 samples to ensure statistical reliability. The final dataset comprises 40007 / 899 / 893 samples in train/validation/test sets after human validation (described next in §2.4). For subsequent analyses, we use the top 6 specialties with the most samples (*Cardiology*, *Gastroenterology*, *Infectious diseases*, *Neurology*, *Obstetrics and Gynecology*, and *Pediatrics*).

2.3 Term overlap analysis

We analyze the overlap of clinically relevant terms in the train/test splits of specialties (*within* and *across* specialties). We use *scispaCy*[2] to extract medical terms from S-MedQA based on the Unified Medical Language System (Bodenreider, 2004). We map each term to relevant clinical specialties by automatically linking the medical concepts to disorders via the Human Phenotype On-

[2] https://allenai.github.io/scispacy

tology (HPO; Castellanos et al., 2024) and tracing diseases to their ancestors in the SNOMED-CT[3] hierarchy to identify top-level systematic disease categories. General terms unrelated to specific specialties or shared by four or more of the six specialties are excluded from this analysis. Please refer to §A.8 for more details. In summary, the average overlap of clinically relevant terms *within specialties* is **63.4%**, whereas *across specialties* is **32.8%**, which means that term mentions are consistent within a specialty's train/test splits, whereas different specialties share limited terminology.

2.4 Manual validation for single specialty predictions

A medical expert labels all the examples in S-MedQA's validation and test sets with the *single most correct clinical specialty*. This expert also validates $1,000$ random samples from the train set, confirming whether the specialties predicted by GPT-3.5 are correct.[4] In general, when using voting with multiple prompts, we see large performance gains compared to using single prompts (e.g., from 72.8–80.2% to 90.8–97.8%; see §A.7).

In Table 1, we show the accuracy vs. coverage trade-off over the $1,000$ random samples from the train set for different requirements for majority voting. A higher quorum results in higher accuracy ($90.8 \rightarrow 97.8$) but greatly decreases the coverage ($89.1 \rightarrow 49.2$). We release per-prompt single-specialty categorizations together with votes for 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.

	# 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.

Moreover, to assess the trustworthiness of the medical expert, we randomly sample 200 from the $1,000$ examples and further ask three medical graduate students to validate the same examples in the same procedure. We use Krippendorff's alpha (Hayes and Krippendorff, 2007) to measure the inter-annotator agreement among the four annotators and obtain 83.6% (95% CI [69.0%, 93.9%]).

3 Cross-specialty evaluation

3.1 Experimental setup

We experiment on 8 open-source LLMs. Six general-purpose models: Llama2-Chat-7B, Llama2-Chat-13B, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct (Touvron et al., 2023), Mistral-Instruct-v0.1 and v0.2 (Jiang et al., 2023), one biomedical model: Bio-Medical-LLaMA-3-8B. We also include OLMo-2-1124-7B-Instruct to rule out the possible contamination of training data.

We fine-tune each LLM on the six per-specialty training sets using prompts from §A.9, and evaluate on all six test sets. Models are trained for up to 10 epochs, with selection based on per-specialty validation accuracy. We also train on the combined dataset to evaluate how exposure to larger and more diverse data affects models' performance. Additional details and hyperparameters are in §A.10.

We follow best practices to evaluate LLM performance on QA datasets by shuffling answers multiple times and adding multiple shuffled QA pairs in the test set, and by using the entire answer to a question instead of the LLM's maximum token probability among options A, B, C, D in the answer (Zheng et al., 2023; Wang et al., 2024). Please see §A.11 for details on our evaluation protocol. To verify the effectiveness of the fine-tuning process, we assess the model's performance on the training data and achieve an average accuracy of 89.3% across the fine-tuning datasets.

3.2 Results

Table 2 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 in terms of average performance, demonstrating the effectiveness of instruction fine-tuning.

Are improvements in Table 2 truly indicative of knowledge acquisition or injection? Looking at the results of models fine-tuned on individual specialties, *none of the best performing models were trained on the corresponding specialty's training*

[3] https://www.snomed.org/

[4] This is the specialty that is the most relevant to the QA pair. In this evaluation, we exclude any cases where there is ambiguity in which clinical specialty is the most relevant.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Mistral-v0.2†	52.0	45.9	48.2	37.0	52.9	43.5	46.9
Train Sets	Cardio	51.8	54.6	44.3	47.6	51.5	44.6	49.4
	Gastro	54.9	54.1	38.9	43.5	51.7	46.4	48.8
	Infect	**55.4**	53.9	43.0	**47.8**	54.3	43.5	50.1
	Neuro	54.4	**54.8**	41.9	45.7	**56.0**	**49.0**	50.8
	Obstetrics	54.4	52.4	**45.7**	41.9	51.5	44.6	48.8
	Pediatrics	50.8	53.0	42.4	35.9	48.1	46.1	46.5
	Combined‡	53.7	53.8	42.7	43.7	52.2	45.7	49.1

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. For each specialty, the best performance when fine-tuned on different specialty datasets is in **bold**, and scores for models fine-tuned on the same specialty are underlined. Surprisingly, none of the best performances come from models fine-tuned on their corresponding training sets.

data, e.g., the best performance on the *Cardio* test set (55.4%) was achieved by the model trained on *Infect*. If the improvements were due to knowledge injection, we would expect to see best performing models consistently along the diagonal (e.g., *Cardio* → *Cardio*, *Infect* → *Infect*, and so on). We report results for other pretrained LLMs in §A.12 and note that similar conclusions hold whereas with different transfer patterns across specialties.

4 Discussion

Where do improvements come from? We hypothesize that the performance improvements in Table 2 are primarily due to domain shifting from general LLMs to the clinical domain, rather than due to the injection of new clinical knowledge.

To assess that, we analyze token probability changes for clinical terms linked to a single clinical specialty between a baseline model and the same model fine-tuned on data from each different specialty. In Fig. 2 we show the average log-probabilities of medical terms across specialties as predicted by **Mistral-v0.2** as the base model and its fine-tuned variants for each of the six specialties. Regardless of the specialty used for fine-tuning, we observe increased token log-probabilities for clinical terms specific to the fine-tuning specialty as well as terms associated with other specialties.

This aligns with our observations in the cross-specialty evaluation (§3.2) that, regardless of the specialty on which the model is trained and tested, consistent improvements are achieved over the base model. We note that token log-probabilities for terms from different specialties differ in range, likely reflecting the pre-trained model's existing knowledge distribution. This also seems to support our hypothesis that fine-tuning shifts the domain rather than injects new domain knowledge.

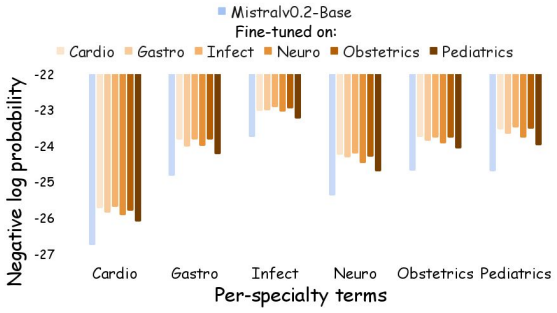

Figure 2: Negative log-probabilities for clinically relevant tokens between baseline Mistral-v0.2 and the same model further fine-tuned on each specialty data. Each group represents tokens categorized into different clinical specialties. Each color means that the same model is further fine-tuned on each specialty data.

We further fine-tune **Mistral-v0.2** on 3 Social Sciences subsets of MMLU (public relations, security studies, and sociology). There remains a gap between the average log probabilities of all six specialty terms generated by the model compared to the model trained on S-MedQA, ensuring that the observed probability increase is a result of domain-specific adaptation rather than a mere consequence of additional training (See §A.13 for detail).

5 Conclusions

In this paper, we introduce S-MedQA, the first medical question-answering dataset annotated across 15 distinct clinical specialties. Using S-MedQA, we demonstrate that fine-tuning with medical QA data enhances LLM performance, with the improvements primarily attributed to domain shifting rather than knowledge injection. However, the precise impact of different types of QA data remains unclear (e.g., complexity of the QA pair), and we recommend further research to investigate the role of fine-tuning in the medical domain.

Limitations

We limited our experiments to the medical domain. However, the findings' generalizability to other knowledge-intense domains is unknown. Further research is needed to investigate the role of fine-tuning and instruction data in other domains.

References

Anastasios N Angelopoulos and Stephen Bates. 2021. arXiv preprint arXiv:2107.07511. A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

Olivier Bodenreider. 2004. Nucleic acids research. The unified medical language system (umls): integrating biomedical terminology, 32(suppl_1):D267–D270.

Francisco Castellanos, J Caufield, Lauren Chan, Christopher Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon Davids, Maud de Dieuleveult, Vinicius de Souza, Bert de Vries, et al. 2024. Nucleic Acids Research. The human phenotype ontology in 2024: phenotypes around the world., 52(D1).

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). In Is GPT-3 a good data annotator?, 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. Machine Learning for Health (ML4H). In Llms accelerate annotation for medical information extraction, pages 82–100. PMLR.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. arXiv preprint arXiv:2106.09685. Lora: Low-rank adaptation of large language models.

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. arXiv preprint arXiv:2310.06825. Mistral 7b.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). In PubMedQA: A dataset for biomedical research question answering, 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. arXiv preprint arXiv:2402.10373. Biomistral: A collection of open-source pretrained large language models for medical domains.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Proceedings of the Conference on Health, Inference, and Learning. In Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, Proceedings of Machine Learning Researchvolume 174 of , 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. Nature. Large language models encode clinical knowledge, 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. arXiv preprint arXiv:2307.09288. Llama 2: Open foundation and fine-tuned chat models.

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

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

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

A Appendix

A.1 Specialties recognized in the European Union (EU) and European Economic Area (EEA)

According to *Directive 2005/36/EC of the European Parliament and of the Council of 7 September*

2005 on the recognition of professional qualifications,[5] the following clinical specialties are recognized in the EU and EEA: Allergist, Anaesthetics, Cardiology, Child psychiatry, Clinical biology, Clinical chemistry, Clinical microbiology, Clinical neurophysiology, Craniofacial surgery, Dermatology, Emergency medicine, 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.

A.2 Prompts used for specialty classification

In Figures 5–9 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.3 Distribution of predicted specialties

In Figure 3 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.4 Excluded vs. complex examples

Excluded examples We exclude examples 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 exclude such samples from S-MedQA.

Complex examples 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

[5] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32005L0036

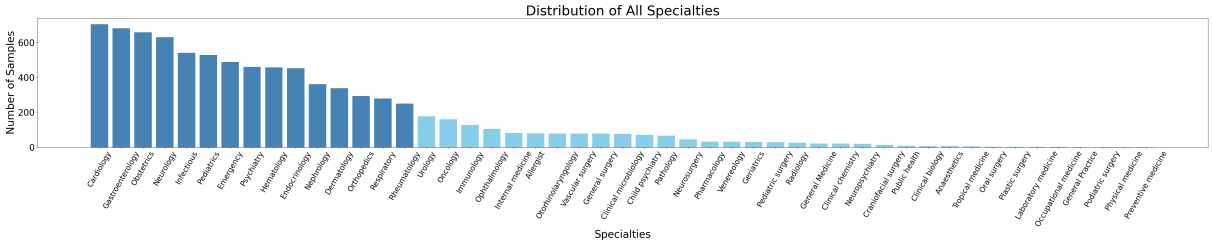

Figure 3: The distribution of all specialties classified by GPT-3.5. The dark blue specialties are the 15 we finally included in our benchmark.

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. Such complex examples were the main reason why we decided to add multiple specialty annotations per question.

A.5 Methods and evaluation of multi-specialty categorization

We curate a high-quality calibration set of 300 samples, annotated independently by a medical expert (300 samples) and three medical students (100 samples each). We define the ground truth as the union of the expert and the students' annotations to minimize the risk of missing potentially relevant specialties.

For each example in the calibration set, we compute log-probabilities for all 55 candidate specialties using GPT-3.5. This is done iteratively: at each step, we extract the top-k (with k=20) next-token predictions from the model, and extend the input with tokens that could lead to valid specialty names. This process continue until either all 55 specialties were generated — allowing us to collect their corresponding token-level log-probabilities — or the next token required for a specialty does not appear in the top-20 predictions, in which case we assign that specialty a probability of zero. For each successfully generated specialty, we compute a score as the average of its next-token log-probabilities, and normalize these scores to obtain a probability distribution over the 55 specialties. We then define the conformal score as $1 - P$, where P is the normalized probability of a given specialty, and use this to construct prediction sets with a 95% confidence level (α=0.05). Since each question may have multiple correct specialties, we select P as the highest predicted probability among the set of annotated correct specialties for evaluation.

Number of Votes (out of 5)	3+	4+	5
Cardiology	2928	2571	2090
Gastroenterology	1893	1590	1176
Obstetrics and gynecology	10671	10590	10457
Neurology	2700	2249	1495
Infectious diseases	1857	1397	845
Pediatrics	8566	8365	8181
Emergency medicine	1164	787	451
Hematology	1882	1642	1264
Endocrinology	2204	1972	1443
Nephrology	1501	1302	970
Respiratory medicine	1201	799	376
Rheumatology	1000	862	624
Dermatology	536	444	328
Psychiatry	532	466	374
Orthopedics	372	292	217
Total	40007	35328	30309

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

We report key metrics computed on the calibration dataset to evaluate the performance of our multi-specialty categorization framework. The observed coverage rate, which measures the proportion of samples where the true labels are included in the predicted label set, was 95%, indicating a good calibration. The average number of predicted labels per sample was 2.46, with an F-1 score of 0.60.

A.6 Clinical specialty benchmark description

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

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

Table 4: Accuracy of each prompt. Prompts #1 to #5 are shown in Figures 5–9 in §A.2.

A.7 Accuracy vs. coverage trade-off of GPT-3.5 single specialty predictions

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 the coverage and accuracy after applying different majority voting strategies, (i.e., at least 3, 4, or 5 responses agreed upon). Only the questions that obtain at least this number of votes are kept. We note that there is an inherent trade-off between accuracy and coverage when deciding the threshold to use for majority voting.

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.8 Term overlap analysis

To perform an overlap analysis of medical terms, we first extract clinically relevant terms in S-MedQA, map these terms onto all of their relevant clinical specialties within the scope of the top 6 specialties we selected for further experiments. We leverage SNOMED-CT,[6] a medical ontology system with all medical concepts, and Human Phenotype Ontology (HPO; Castellanos et al., 2024), which provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. We include 4 types of medical concepts in SNOMED-CT: *disorder*, *finding*, *procedure*, and *observable entity*.

[6] https://www.snomed.org/

The high-level SNOMED-CT concepts we use for each speciality is: Disorder of cardiovascular system (disorder), Disorder of digestive system (disorder), Infectious disease (disorder), Disorder of female reproductive system (disorder), Disorder of fetus and/or mother during labor (disorder), Disorder of nervous system (disorder), Behavioral and emotional disorder with onset in childhood (disorder), Disorder of fetus and/or newborn (disorder), Developmental disorder (disorder).

We directly map disorders onto one or more specialties by traversing through their ancestors until they match any high-level specialty-specific concepts as above. We found no simple mechanism to map other types of terms (i.e., findings, procedures, observable entities) to specialties. We thus first search these terms in HPO to identify their relevant disorders, and then used these disorders similarly to map onto clinical specialties.

The average overlap in train/test splits across specialties (e.g., *Cardio*'s train vs. *Neuro*'s test) is 32.8%, indicating that different specialties only share limited common specialty-specific terminology, hence there will be minor "knowledge leakage" when evaluating cross-specialty performances. Meanwhile, the average overlap between a same speciality's train/test sets is 63.4%. This shows that the training sets contains sufficient knowledge coverage to solve the questions in the test sets.

A.9 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

Train Sets \ Test Sets	Cardio(%)	Gastro(%)	Infect(%)	Neuro(%)	Obstetrics(%)	Pediatrics(%)
Cardio	67.5	31.4	38.3	31.0	25.7	31.5
Gastro	30.8	64.3	43.4	32.1	24.8	36.7
Infect	26.1	30.9	64.1	27.5	28.4	40.8
Neuro	27.5	27.4	41	62.3	23.5	33.8
Obstetrics	25.1	26.9	56.5	27.2	62.9	40.8
Pediatrics	32.8	33.4	44.8	28.8	34.7	59.2

Table 5: Medical term overlap ratios between train/test sets across specialties and within the same specialty (in percentage). Each column represents a test set as the denominator for overlap ratio calculations, while each row represents the test sets used for matching overlapping terms. Underlined scores indicate overlaps between the training and test sets within the same specialty, while other scores represent overlaps between different specialties.

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.10 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 rate=0.1, batch size=8.

A.11 LLM evaluation protocol

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.

Figure 4 illustrates the approach (classifier) we use to evaluate the performance of the models and addresses the issues with using the first token probability or simple string matching (Wang et al., 2024). The classifier is trained based on **Mistral-v0.2** and applied in all experiments.

A.12 Cross-specialty evaluation results for Mistral-7B-v0.1, Llama-2-7B and -13B, Llama-3.1-8B, Llama-3.2-3B, Bio-Medical-LLaMA-3-8B, and OLMo-2-1124-7B-Instruct

In Tables 7, 8, 9, 10, 11, 12, 13 we show cross-specialty evaluation matrices for Llama2-7b-chat, Llama2-13b-chat, Mistral-7b-instruct-v0.1, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Bio-Medical-LLaMA-3-8B, and and OLMo-2-1124-7B-Instruct in addition to our main results (in §3.2). 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.

A.13 Average token log probabilities gap between model fine-tuned on Social Science subsets vs. S-MedQA

We fine-tune Mistral-v0.2 on three Social Sciences subsets of MMLU—public relations, security studies, and sociology—which are entirely unrelated to the medical domain. We compare the average token log-probabilities of six specialty medical terms produced by this model with those obtained from models fine-tuned on the corresponding S-MedQA datasets. As shown in Table 6, there is a significant drop in token probabilities when fine-tuned on out-of-domain data, confirming that the change in token probabilities is brought by domain shifting rather than only addition training.

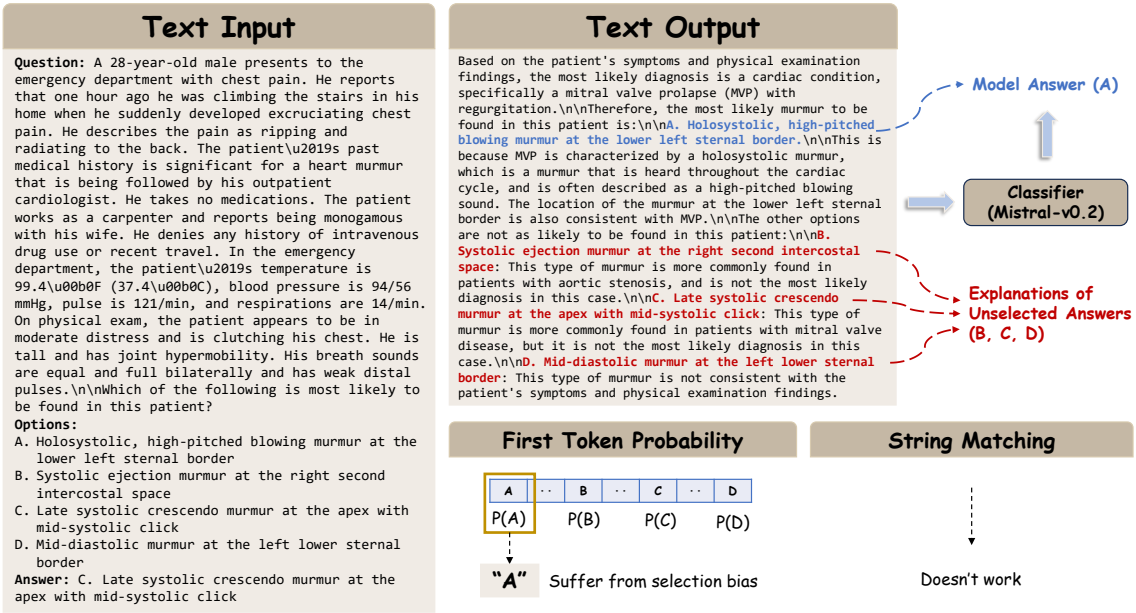

Figure 4: The illustration of first token probability, string matching, and our approach (classifier) to evaluating LLMs performance on S-MedQA. We use text output instead of first token probability for evaluation because first token probability suffers heavily from selection bias in multiple-choice 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%.

Figure 5: 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 6: 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 7: 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 8: 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 9: 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:

Table 6: Token probability increase by clinical specialty after fine-tuning.

Specialty	**Increase (%)**
Cardio	10.5
Gastro	10.7
Infect	9.5
Neuro	8.4
Obstetrics	7.2
Pediatrics	7.5

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Llama2-7b	36.0	36.3	36.7	34.6	40.6	41.4	37.7
Train Sets	Cardio	34.3	29.1	32.6	28.3	31.5	29.5	31.0
	Gastro	31.3	**32.5**	30.7	28.8	**38.3**	30.1	32.0
	Infect	**35.0**	31.3	32.8	31.0	33.3	29.0	32.1
	Neuro	32.8	26.3	**35.4**	**31.3**	33.5	**36.1**	32.7
	Obstetrics	34.5	30.0	34.6	30.7	37.9	33.2	33.6
	Pediatrics	30.5	27.8	34.1	25.8	33.5	31.0	30.7
	Combined	42.0	40.1	38.5	37.2	42.5	36.1	39.4

Table 7: Cross-specialty accuracy matrix of Llama2-7b.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Llama2-13b	43.3	34.9	40.6	36.1	45.4	39.2	40.0
Train Sets	Cardio	38.3	34.7	**41.1**	28.5	**38.8**	31.8	35.8
	Gastro	**38.5**	**35.8**	31.8	31.3	36.0	32.7	34.3
	Infect	36.0	31.3	36.5	32.9	39.2	32.7	34.9
	Neuroy	31.3	29.3	36.2	28.8	35.4	33.5	32.7
	Obstetrics	33.5	29.5	36.5	30.4	36.0	32.7	33.3
	Pediatrics	37.5	34.9	37.0	**33.2**	38.5	**36.1**	36.3
	Combined	44.0	45.9	42.4	40.2	45.8	42.9	43.6

Table 8: Cross-specialty accuracy matrix of Llama2-13b.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Mistral-v0.1	41.0	39.0	40.0	30.7	42.7	37.5	38.9
Train Sets	Cardio	**52.8**	46.1	47.7	39.9	47.9	44.3	46.6
	Gastro	48.3	**50.4**	41.1	40.2	47.5	44.3	45.2
	Infect	51.5	42.7	**49.0**	**44.3**	47.5	43.5	46.5
	Neuro	50.7	46.3	47.1	44.0	49.6	**48.9**	47.8
	Obstetrics	45.8	44.2	44.3	41.3	**53.8**	46.3	46.1
	Pediatrics	48.8	45.5	42.7	35.1	51.0	**48.9**	45.5
	Combined	52.8	47.6	46.4	49.5	56.3	47.2	49.9

Table 9: Cross-specialty accuracy matrix of Mistral-v0.1

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Llama-3.1	64.0	67.5	65.3	66.8	65.0	64.5	65.5
Train Sets	Cardio	69.3	**76.5**	**73.5**	76.2	70.5	70.9	72.8
	Gastro	68.2	73.3	71.8	74.6	69.3	69.2	71.1
	Infect	65.8	74.2	69.5	72.5	66.6	66.7	69.3
	Neuro	**69.9**	76.4	72.3	**76.9**	69.5	**71.7**	72.8
	Obstetrics	69.8	75.0	72.9	75.4	68.7	69.4	71.9
	Pediatrics	66.9	75.0	72.1	75.9	68.7	70.2	71.4
	Combined	72.9	79.5	75.4	77.8	70.8	74.1	75.1

Table 10: Cross-specialty accuracy matrix of Llama-3.1-8B-Instruct.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Llama-3.2	54.0	60.0	58.5	61.0	57.0	55.5	57.7
Train Sets	Cardio	59.8	**68.2**	**67.9**	**70.6**	**62.5**	**59.0**	64.7
	Gastro	58.0	65.3	66.2	68.7	60.2	56.5	62.5
	Infect	59.5	63.9	65.7	69.0	**62.5**	57.8	63.1
	Neuro	59.8	64.6	66.9	69.8	62.2	58.6	63.6
	Obstetrics	60.1	65.6	67.1	70.2	61.6	58.9	63.9
	Pediatrics	**61.0**	66.3	67.7	70.3	62.0	58.6	64.3
	Combined	65.4	69.9	71.3	75.5	66.9	65.0	68.9

Table 11: Cross-specialty accuracy matrix of Llama-3.2-3B-Instruct.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	Bio-Llama	70.1	72.3	71.0	72.8	70.2	69.4	71.0
Train Sets	Cardio	**79.7**	82.7	80.3	82.8	76.4	**75.3**	79.6
	Gastro	77.7	**83.1**	**81.5**	82.8	76.6	74.9	79.5
	Infect	76.8	82.2	78.9	82.2	75.1	72.8	78.1
	Neuro	77.1	80.7	79.7	**83.8**	76.3	74.0	78.6
	Obstetrics	78.0	81.7	81.2	81.8	76.6	73.7	78.9
	Pediatrics	79.3	82.3	81.3	82.8	**78.2**	75.3	79.9
	Combined	82.7	84.3	82.0	85.5	79.6	79.5	82.3

Table 12: Cross-specialty accuracy matrix of Bio-Medical-Llama-3-8B.

	Test Sets (Sample size)	Cardio (80)	Gastro (83)	Infect (102)	Neuro (74)	Obstetrics (88)	Pediatrics (90)	avg.
	OLMo	36.0	39.5	40.3	38.8	37.2	35.9	38.0
Train Sets	Cardio	40.7	46.0	47.2	44.4	**41.6**	38.2	43.1
	Gastro	39.6	46.4	47.7	**45.7**	41.0	35.7	42.8
	Infect	38.4	45.5	**49.0**	45.7	40.2	37.2	42.7
	Neuro	39.0	46.4	46.5	45.4	40.7	**38.4**	42.8
	Obstetrics	**41.5**	44.5	46.1	45.3	39.5	37.6	42.4
	Pediatrics	39.6	**46.9**	48.5	43.5	39.7	37.8	42.7
	Combined	48.7	54.1	55.6	55.5	48.1	42.2	50.8

Table 13: Cross-specialty accuracy matrix of OLMo-2-1124-7B-Instruct.