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

Xinlan Yan[1,2] **Di Wu**[2]* **Yibin Lei**[2] **Christof Monz**[2] **Iacer Calixto**[1,2]
[1]Department of Medical Informatics, Amsterdam UMC
[2]University of Amsterdam
x.yan@amsterdamumc.nl, d.wu@uva.nl, y.lei@uva.nl,
c.monz@uva.nl, i.coimbra@amsterdamumc.nl

Abstract

In this paper, we introduce S-MedQA, an English medical question-answering (QA) dataset for benchmarking large language models (LLMs) in fine-grained clinical specialties. S-MedQA has over 24k examples, covers 15 medical specialties, and QA pairs can have multiple specialty annotations, e.g., when a question is cross-disciplinary, constructed with both machine and expert verification to maximize data availability and reliability. We use S-MedQA to investigate the role of clinical specialties in the knowledge-intensive scenario of medical QA. Our results show that 1) training on data from a clinical specialty *does not necessarily lead to best performance on that specialty*, and 2) regardless of the specialty the LLM was fine-tuned on, token probabilities of clinically relevant terms *increase consistently across all specialties*. Thus, we hypothesize improvement gains, at least in our settings, are derived mostly from *domain shifting* (e.g., general to medical) rather than specialty-specific knowledge injection, and suggest rethinking the role of fine-tuning data in the medical domain. To motivate further advancements in the clinical NLP field, 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; MASH-QA, Zhu et al., 2020; 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.

To address this gap, we propose **S-MedQA**, the first English medical QA dataset with multiple clinical specialty annotations (see Figure 1 for an overview, and §2 for details). We build S-MedQA based on the widely used MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) datasets, incorporating both machine and expert verification to semi-automatically map samples onto clinical specialties at scale, while guaranteeing the quality of the annotations. We first use multiple prompts and a majority voting mechanism to label QA pairs with a single specialty. S-MedQA includes 15 specialties, each with hundreds to thousands of samples, and single specialty expert validation shows up to 97.8% accuracy in this annotation (see §2 for details). In order to account for cross-disciplinary questions, the second step is to expand single-specialty annotations into multiple specialties. We use a multi-label conformal prediction procedure with coverage guarantees. We annotate a held-out set of QA pairs with medical experts with a Jaccard Similarity of 0.82, and build a conformal multi-label annotation procedure that leads to a precision of 0.69, recall of 0.52, 24% exactly correct predictions, and an average number of labels of 1.55.

Different clinical specialties can have very different amounts of available data (e.g., see Figure 2). We thus first use our dataset to address the following research question (§3.2): *to what extent can LLMs use the knowledge learned from a clinical specialty to answer questions about other specialties?* We investigate Zhou et al. (2024)'s 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. To that regard, we first fine-tune LLMs on one clinical specialty and evaluate on all other specialties. Interestingly, we find that *the best results often come from fine-tuning on unrelated clinical*

[*] Corresponding author.

[1] https://github.com/nlp4health-lab/S-MedQA

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8339–8358
March 24-29, 2026 ©2026 Association for Computational Linguistics

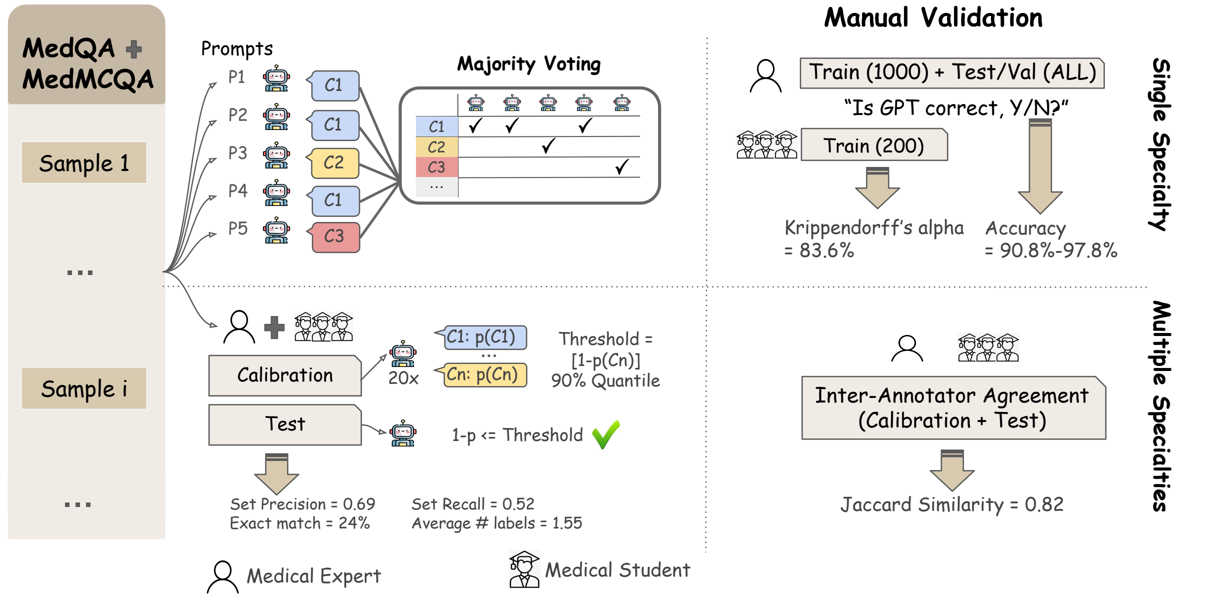


Figure 1: Overview of S-MedQA's construction process. For single specialty annotation of each sample, we generate predictions using 5 different prompts and only keep those where predictions agree (3+, 4+, or 5 times). A medical expert manually annotates S-MedQA's entire validation and test set. We randomly sample 1,000 questions from our train set and ask multiple medical experts to evaluate GPT-3.5's predictions (we detail inter-annotator agreement in §2.2), achieving accuracies ranging from 90.8–97.8%. For multi-specialty annotation, we leverage conformal prediction to assign labels across multiple clinical specialties, using a held-out set of 300 samples manually annotated by the same medical expert and three medical students as calibration/test set, with a Jaccard Similarity of 0.82, reflecting a high inter-annotation agreement. It achieves a set precision of 0.69, set recall of 0.52, with 24% exactly correct matches and an average prediction length of 1.55.

specialties. E.g., fine-tuning with *neurology* data performs best on *cardiology* domain, despite their knowledge being largely unrelated. Moreover, although different pre-trained LLMs exhibit different knowledge transfer patterns across clinical specialties, neither of the LLMs we investigated show the best knowledge transfer when training and testing within a same clinical specialty.

These results lead us to take a step back and ask: *but are QA pairs from different clinical specialties really different?* To answer this question, in §3.3 we curate the relevant clinical terms (e.g., diseases, procedures) for each clinical specialty using authoritative resources—such as the SNOMED-CT knowledge graph (Cornet and de Keizer, 2008)—to effectively estimate the clinical knowledge overlap between QA pairs across clinical specialties.

Finally, we ask *how does the probability of clinically-relevant terms for one specialty change before and after fine-tuning the LLM?* We address this question in §3.4. We again use the clinical terms curated in §3.3 and analyze changes in term probabilities before and after fine-tuning on data from *the same vs. different clinical specialties*. Our results suggest that performance gains are driven more by domain shifts (from *general* to *medical*) than by fine-grained clinical knowledge injection.

Our main contributions are:

- We introduce S-MedQA, a medical QA dataset spanning 15 clinical specialties with 24k QA pairs annotated with high quality and human validated specialty annotations.
- We consider both single-specialty annotations and multiple-specialty labels to account for cross-disciplinary QA pairs.
- We systematically evaluate the impact of fine-tuning 8 LLMs on cross-specialty medical QA performance. Our findings indicate that performance gains in medical QA tasks are primarily driven by domain shifts rather than the injection of specialty-specific knowledge.
- We release S-MedQA and all code necessary for reproduction, providing the community with a valuable resource for benchmarking and improving medical LLMs.

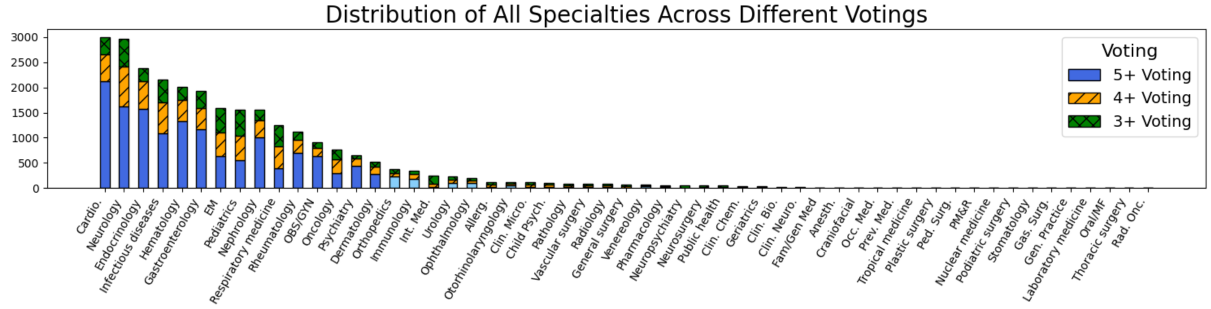


Figure 2: The distribution of all specialties classified by GPT-3.5. The dark blue specialties are the 15 we finally included in our benchmark. Note that we conduct experiments on the 6 common specialties for simplicity.

2 Creation of S-MedQA

We now describe the creation of S-MedQA, a high-quality benchmark for medical QA with clinical specialty annotations. In Figure 1, we show an overview of S-MedQA's creation process.

Creating S-MedQA involves 1) using an LLM to produce single medical specialty annotations for each QA pair (§2.1), 2) a thorough human validation step to ascertain the quality and correctness of the annotations (§2.2), and finally 3) using an LLM to extend these single specialty annotations to possibly multiple specialties whenever the question requires so drawing on ideas from conformal prediction (§2.3). We release multiple versions of S-MedQA with varying 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*.

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).

Manually labeling QA examples with clinical specialties is costly and time-consuming. We thus first use GPT-3.5 to annotate samples with single clinical specialties. Preliminary experiments using a single prompt to predict a single specialty produced low accuracy annotations ($\sim$ 75%). To improve this, we design five prompts (details in §A.2) and predict single specialties with GPT-3.5 for each QA pair independently with each prompt.

Number of Votes (out of 5)	3+	4+	5
Cardiology	3,000	2,652	2,122
Neurology	2,954	2,419	1,627
Endocrinology	2,384	2,122	1,568
Infectious diseases	2,161	1,705	1,087
Hematology	2,005	1,746	1,332
Gastroenterology	1,934	1,594	1,176
Emergency medicine	1,597	1,098	629
Pediatrics	1,553	1,036	559
Nephrology	1,550	1,340	1,004
Respiratory medicine	1,253	825	395
Rheumatology	1,115	958	700
Obstetrics and gynecology	913	802	636
Oncology	770	570	300
Psychiatry	656	582	446
Dermatology	531	421	280
Total	**24,306**	**19,290**	**14,261**

Table 1: S-MedQA 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 then apply majority voting to decide on the specialty for that QA pair (Ding et al., 2023; Goel et al., 2023). Human evaluation shows accuracies between 90.8%–97.8% (more details next in §2.2).

We exclude $3,518$ (11.5%) samples categorized as *Others* since they mostly contain irrelevant clinical information (see Appendix A.4 for examples of exclusion), and then focus on 15 out of 55 specialties with more than 500 samples. The final dataset comprises 24306 / 899 / 899 samples in train/validation/test sets after human validation.

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 / valid / test sets. In Table 1, we show the 15 specialties included in S-MedQA with the respective numbers of samples. The columns represents 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 2: Accuracy of each single prompt. Prompts #1 to #5 are shown in Figures 6–10 in Appendix A.2.

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

Table 3: Accuracy vs. coverage for majority voting under different minimum number of votes.

2.2 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.[2] Table 2 shows the categorization accuracies when using single prompts. The accuracies when applying majority voting are shown on top in Table 3. 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%).

In Table 3, we also 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 with votes for all examples for users 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.

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%]).

2.3 From single to multiple clinical specialties

We devise a strategy based on conformal prediction to expand clinical specialty annotations to possibly multiple specialties if a QA pair requires so. Our main idea is: 1) *calibrate* an LLM to find a threshold that maximises multiple specialties' prediction accuracy in a calibration set; 2) use this threshold to build a conformal classifier that predicts multiple clinical specialties for each QA pair.

Calibration and test set. We curate a held-out set of 300 samples, annotated independently by a medical expert (300 samples) and three medical students (100 samples each). We measure the Jaccard Similarity between the expert and the students for the inter-annotator agreement, with a similarity score of 0.82. We use 200 examples (66.7%) as a calibration set, and report results on the remaining 100 examples (33.3%).

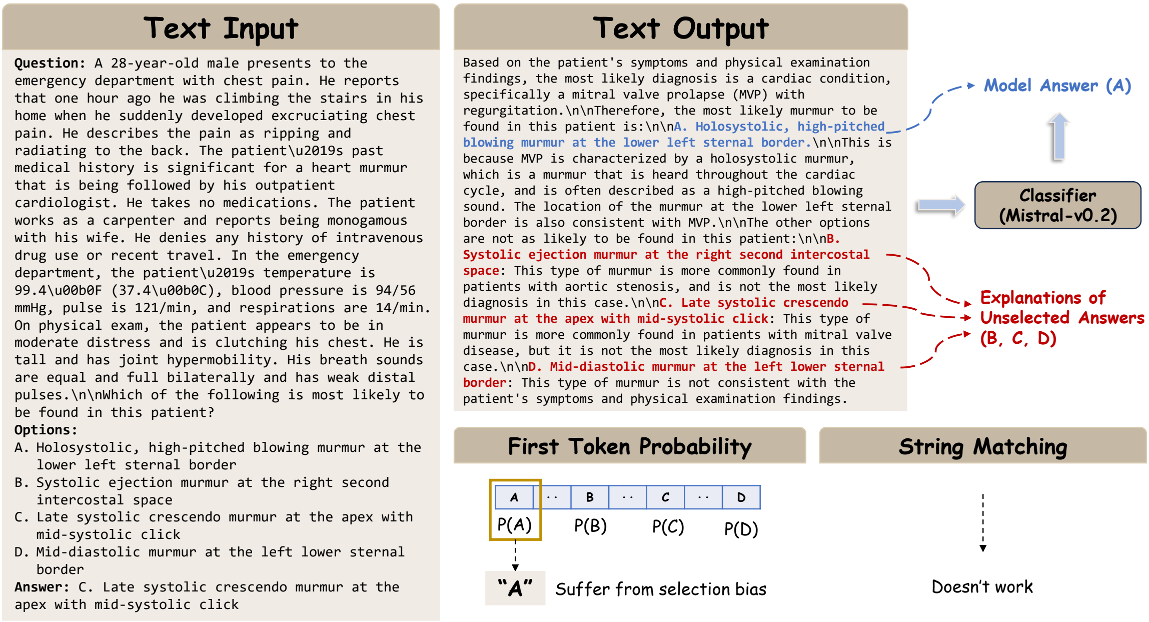
Conformal threshold and set construction. Following Ke et al. (2025), we query GPT-3.5-Turbo $n = 20$ times per input question and predict a single specialty C_i. We then get frequencies $P(C_i)$ over all possible output specialties (see §A.1). We compute the non-conformity scores as $S_i = 1 - P(C_i)$ and obtain a set of scores S_1, S_2, ..., S_n per question. We set our acceptable error to 10% ($\alpha = 0.1$) and then use the $(1 - \alpha)$ quantile of all the scores as the calibration threshold. At prediction time, we also sample 20 times single specialty categorization for each question, and compute the per specialty scores. We include all specialties that have lower score than the calibration threshold in the final multi-specialty prediction.

We report key metrics computed on the testing data. The number of instances where the predicted label set exactly matched the true gold label set was 24. Precision, defined as the proportion of predicted labels that were correct, was 0.69. Recall, which measures the proportion of gold labels correctly included in the prediction sets, was 0.52. The average number of predicted labels was 1.55.

3 Experiments

We first investigate how training LLMs on data from a clinical specialty impacts their performance on other specialties (§3.2). Our findings show that training on one clinical specialty does not necessarily lead to the best performance in that specialty. We hypothesize whether QA pairs from different clinical specialties are indeed very different from

[2] 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.



Question: A 28-year-old male presents to the emergency department with chest pain. He reports that one hour ago he was climbing the stairs in his home when he suddenly developed excruciating chest pain. He describes the pain as ripping and radiating to the back. The patient\u2019s past medical history is significant for a heart murmur that is being followed by his outpatient cardiologist. He takes no medications. The patient works as a carpenter and reports being monogamous with his wife. He denies any history of intravenous drug use or recent travel. In the emergency department, the patient\u2019s temperature is 99.4\u00b0F (37.4\u00b0C), blood pressure is 94/56 mmHg, pulse is 121/min, and respirations are 14/min. On physical exam, the patient appears to be in moderate distress and is clutching his chest. He is tall and has joint hypermobility. His breath sounds are equal and full bilaterally and has weak distal pulses.\n\nWhich of the following is most likely to be found in this patient?

Options:

A. Holosystolic, high-pitched blowing murmur at the lower left sternal border

B. Systolic ejection murmur at the right second intercostal space

C. Late systolic crescendo murmur at the apex with mid-systolic click

D. Mid-diastolic murmur at the left lower sternal border

Answer: C. Late systolic crescendo murmur at the apex with mid-systolic click

Based on the patient's symptoms and physical examination findings, the most likely diagnosis is a cardiac condition, specifically a mitral valve prolapse (MVP) with regurgitation.\n\nTherefore, the most likely murmur to be found in this patient is:\n\nA. Holosystolic, high-pitched blowing murmur at the lower left sternal border.\n\nThis is because MVP is characterized by a holosystolic murmur, which is a murmur that is heard throughout the cardiac cycle, and is often described as a high-pitched blowing sound. The location of the murmur at the lower left sternal border is also consistent with MVP.\n\nThe other options are not as likely to be found in this patient:\n\nB. Systolic ejection murmur at the right second intercostal space: This type of murmur is more commonly found in patients with aortic stenosis, and is not the most likely diagnosis in this case.\n\nC. Late systolic crescendo murmur at the apex with mid-systolic click: This type of murmur is more commonly found in patients with mitral valve disease, but it is not the most likely diagnosis in this case.\n\nD. Mid-diastolic murmur at the left lower sternal border: This type of murmur is not consistent with the patient's symptoms and physical examination findings.

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 the latter suffers heavily from selection bias in multiple-choice QA (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%.

each other. That leads us to quantify the overlap of clinically relevant terms in train/test splits *within* and *across* specialties (§3.3). Finally, we investigate to what extent token probabilities of clinically relevant tokens change when fine-tuning LLMs in S-MedQA compared to non-clinical data (§3.4).

3.1 Experimental setup

To minimize potential evaluation bias, we follow best practices to evaluate LLM performance on QA datasets (1) by shuffling the order of the answers multiple times and adding multiple shuffled QA pairs in the test set, and (2) 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).

More specifically, 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, we follow Wang et al. (2024) to 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 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 3 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.

In all subsequent analyses below, we use the 6 most common specialties (*Cardiology*, *Gastroenterology*, *Infectious diseases*, *Neurology*, *Obstetrics and Gynecology*, and *Pediatrics*). The number of test samples in each specialty is 80, 83, 102, 74, 88, and 90, respectively. The final number of samples in the test set is 400, 415, 510, 370, 440, and 450 (i.e., times 5 shuffles).

3.2 Cross-specialty evaluation

We experiment with 8 open-source LLMs: **Llama2-Chat-7B**, **Llama2-Chat-13B**, **Llama-3.1-8B-Instruct**, **Llama-3.2-3B-Instruct** (Touvron et al., 2023), **Mistral-Instruct-v0.1**, **Mistral-Instruct-v0.2** (Jiang et al., 2023) , **Bio-Medical-LLaMA-3-8B** (ContactDoctor, 2024), and **OLMo-2-1124-7B-Instruct** (OLMo et al., 2025). These include six general-purpose models (Llama-2, Llama-

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 4: Accuracy matrix for Llama-3.1-8B-Instruct. †Model is applied without finetuning. ‡Model is trained on the combination of all 6 specialty train sets. For robustness, accuracies are computed as the average over the test set with answer options shuffled five times. 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.

3, and Mistral-based models), one biomedical Llama model, and a fully open-source model (OLMo).

We fine-tune each LLM on the six per-specialty training sets using prompts from §A.3, 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. We use LoRA (Hu et al., 2021) on all projection layers for the fine-tuning process in all experiments with the following hyperparameters: learning rate 2e-5, rank 32, alpha 16, dropout rate 0.1, batch size 8.

Results. Table 4 shows the performance of Llama-3.1-8B-Instruct fine-tuned independently on each specialty and tested on all six specialty test sets, as well as after fine-tuning on a combination of all six specialties. **Are improvements truly indicative of knowledge acquisition or injection?** Notably, for models fine-tuned on individual specialties, *none of the best performing models were trained on the corresponding specialty's training data*, e.g., the best performance on the *Cardio* test set (69.9%) was achieved by the model trained on *Neuro*. 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 LLMs in §A.5 and note that similar inconsistencies hold, whereas with different transfer patterns across specialties. This pattern suggests that performance gains from fine-tuning primarily arise from factors other than knowledge injection, partially supporting Zhou et al. (2024)'s hypothesis in the clinical domain.

The number of per-specialty examples used to train Llama-3.1-8B-Instruct is imbalanced (e.g., see Table 1). To tackle this imbalance, we randomly select 913 samples from each of the six specialties' training data, which corresponds to the least number of samples among the six specialties. We fine-tune Llama-3.1-8B-Instruct on this data and we show the results in §A.6. We also conduct 20 bootstrap resampling experiments on the test sets using the Llama-3.1-8B-Instruct model fine-tuned on the whole (i.e., imbalanced) training data to tackle the issue of limited testing data. We randomly sample the same number of samples as the test sizes from the original test sets with replacement. We show the results for this experiment in §A.6. Both experiments show that most of the best performing models on each per-specialty's test set is not the same model tuned on the training data from the same specialty. This demonstrates the robustness of our main results, i.e., best models are almost always off-diagonal.

3.3 Term overlap analysis

Clinical questions from different specialties can differ substantially in terms of knowledge; for example, one cannot assume that *neurology* expertise can directly apply to *cardiology*. Here, we thus quantify the overlap of clinically relevant terms in train/test splits *within* and *across* specialties. We do this by extracting clinically relevant terms in each question of S-MedQA, by mapping these terms onto relevant clinical specialties within the scope of the top 6 specialties we selected for further experiments, and finally by quantifying the difference in clinical terminology across different specialties.

We map each term in a question to relevant clinical specialties using SNOMED-CT (Cornet and

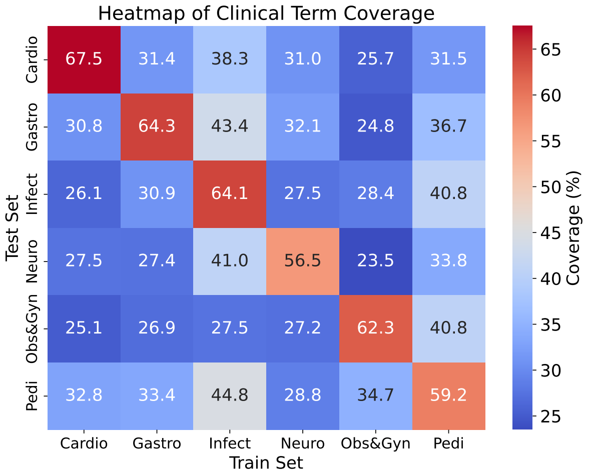


Figure 4: Heatmap showing the overlap of clinical terms within and across clinical specialties between train and test QA pairs in S-MedQA.

de Keizer, 2008)[3] and optionally the Human Phenotype Ontology (HPO; Castellanos et al., 2024). 1) We first manually identify the top-level *disorder* concepts in SNOMED-CT that correspond to each medical specialty, and use the SNOMED hierarchy to link specific diseases or concepts to this top-level category. E.g., we link any disorder under *Disorder of cardiovascular system* as part of *cardiology*.[4] We found no simple mechanism to map *findings*, *procedures*, and *observable entities* concepts to disorders in SNOMED-CT directly. We thus first search *findings*, *procedures*, and *observable entities* in HPO, identify the related disorders, which we then use to map onto clinical specialties similarly with SNOMED-CT. 2) For each question, we then use SciSpacy (Neumann et al., 2019) to perform entity linking of *disorder*, *finding*, *procedure*, and *observable entity* mentions in the question to SNOMED-CT. 3) Finally, we use the top-level concepts identified in (1) and the SNOMED-CT hierarchy to link each mention to a clinical specialty.

We exclude from this analysis general terms unrelated to specific specialties or shared by four or more of the six specialties.

Results. Figure 4 illustrates the overlap percentages of specialty-specific terminology between training and test sets for the six specialties. The average overlap in train/test splits within the same specialty is 63.4%, indicating that training sets cover well clinical terminologies necessary for solving corresponding test set questions. In contrast, the average overlap across different specialties is only 32.8%, showing the domain-specific nature of medical terminology and the limited commonality between specialties. These results confirm that "knowledge leakage" during cross-specialty evaluations is minimal, as different specialties share limited common specialty-specific terminology.

3.4 Token probability

We hypothesize that the performance improvements in Table 4 are primarily due to domain shifting from general LLMs to the clinical domain, rather than the injection of new clinical knowledge. We demonstrate this point through a detailed analysis of token probability shifts for clinical terms before and after fine-tuning.

The impact of medical data fine-tuning. To assess that, we analyze token probability changes for clinical terms in the questions from test sets linked to a single clinical specialty between a baseline model and the same model fine-tuned on data from each different specialty. Clinically relevant terms are extracted and mapped to specialties using the method described in §3.3, and each term's probability is obtained by summing the probabilities of its constituent subword tokens.

In Fig. 5 we show the average log-probabilities of medical terms across specialties as predicted by Llama-3.1-8B-Instruct (base and fine-tuned) for each of the six specialties. Regardless of the specialty used for fine-tuning, we observe similarly increased token log-probabilities for clinical terms specific to the fine-tuning specialty, as well as for terms associated with other specialties.

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.

Are improvements due to additional training steps? To rule out the possibility that the similarly increased probabilities in all clinical specialties are due simply to the additional training steps, we also fine-tune our baseline Llama-3.1-8B-Instruct on three social sciences subsets of MMLU (Hendrycks et al., 2021b,a)—public relations, security studies, and sociology—which are entirely unrelated to the medical domain. Again, we compare the average token log-probabilities of the six specialty medical terms produced by this

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

[4] For all high-level SNOMED-CT concepts we use for each clinical specialty, please refer to § A.7.

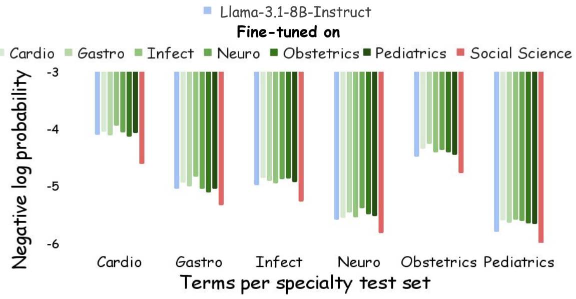


Figure 5: Negative log-probabilities for clinically relevant tokens between baseline Llama-3.1-8B-Instruct and the same model further fine-tuned on each specialty data. Each group represents tokens categorized into different clinical specialties. Each color means that the same model is further fine-tuned on each specialty data.

model with those obtained from models fine-tuned on the corresponding S-MedQA datasets. When fine-tuning on MMLU out-of-domain datasets there is a significant drop in token probabilities compared to models trained on S-MedQA data (red bars in Figure 5). We thus confirm that this drop is caused by domain shifting rather than purely the effect of additional training. See Appendix A.8 for details of the token probabilities for models trained on S-MedQA and MMLU social sciences.

4 Discussion

The effectiveness of lightweight fine-tuning on strong base models has been widely recognized in NLP, largely following the pretraining paradigm introduced in the BERT era (Devlin et al., 2019). This strength is often attributed to domain shifting, especially for linguistically focused tasks such as sentiment analysis (Zhang et al., 2020), natural language inference (Koroteev, 2021), and semantic matching (Gao et al., 2021).

In knowledge-intensive domains such as healthcare, the role of fine-tuning remains underexplored. Clinical questions across specialties often pose substantial knowledge barriers, and the capabilities of current LLMs on specialty-specific tasks are still unclear. Moreover, it remains uncertain whether these tasks benefit more from explicit knowledge injection or from effective knowledge transfer within the base model. This paper benchmarks these real-world questions in a fine-grained manner and provides preliminary evidence to explain the role of fine-tuning in the clinical domain.

Moreover, current LLMs are trained on massive datasets and raise concerns about benchmark contamination. Our experiments span a wide range of models, including open-source ones such as OLMo (Groeneveld et al., 2024), which is trained without clinical data. Similar findings on these models strengthen the reliability of our results.

5 Related Work

Medical QA datasets. Medical QA datasets have played a pivotal role in benchmarking and advancing the capabilities of LLMs in the healthcare domain. Early datasets such as PubMedQA (Jin et al., 2019) and MedQA (Jin et al., 2021) introduced multiple-choice QA tasks derived from medical board exams, focusing on assessing clinical reasoning and factual knowledge. MedMCQA (Pal et al., 2022) later expanded the scope with a larger dataset covering a wide range of medical subjects. These datasets have been instrumental in developing models like BioGPT (Luo et al., 2022) and MedPaLM (Singhal et al., 2025), which leverage domain-specific pretraining to achieve human-like performance on medical tasks.

However, these datasets lack annotations for clinical specialties, limiting their applicability in scenarios where specialty-specific knowledge is crucial. Recent efforts, such as BioMistral (Labrak et al., 2024), have focused on pretraining biomedical LLMs on medical corpora but have not addressed the need for fine-grained specialty-level evaluation. S-MedQA fills this gap by introducing per-specialty annotations and enabling the study of knowledge transfer across specialties.

Knowledge injection vs. Domain shifting. The effectiveness of fine-tuning on task-specific datasets has been extensively studied in NLP (Devlin et al., 2019; Mayfield and Black, 2020; Limsopatham, 2021). Broadly speaking, this effectiveness is often understood to either introduce new knowledge within the external corpus into pretrained models (Fu et al., 2023; Mecklenburg et al., 2024), especially in knowledge-intensive tasks, or to adapt a model's general capabilities to more specialized domains (Limsopatham, 2021; Garrido-Merchán et al., 2023; Wu et al., 2024), highlighting effective knowledge sharing across tasks.

Particularly, recent studies (Zhou et al., 2024) in LLMs suggest that fine-tuning primarily serves to shift the model's focus toward a specific domain, leveraging knowledge already encoded during pretraining, rather than injecting new knowledge into the model. Zhou et al. (2024) introduced the hypothesis that alignment tuning—e.g., supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF)—primarily teaches large language models (LLMs) to select a subdistribution of response styles and formats. This hypothesis was supported by their study on the LIMA model, where fine-tuning with as few as 1,000 examples yielded strong alignment. This study demonstrated that the success of alignment tuning lies in its ability to adjust stylistic tokens, transitional phrases, and disclaimers, while the majority of the knowledge required to answer user queries remains unchanged from the base model. Our work is a first step towards investigating the role of fine-tuning in knowledge-intensive and finely segmented domains, since we specifically address the medical domain with different clinical specialties.

6 Conclusions

In this paper, we introduce S-MedQA, the first English medical QA dataset annotated for 15 distinct clinical specialties including single- and multi-specialty annotations for cross-disciplinary questions. We investigate different questions using S-MedQA and find that questions from different clinical specialties have significant differences in terms of clinical terminology. Despite that, we also find that the best QA performance is almost always observed for LLMs fine-tuned on unrelated clinical specialties. Moreover, we also show that fine-tuning the same base LLM on high-quality non-medical data leads to decreased token probabilities for clinically relevant terms. Our results suggest that Zhou et al. (2024)'s hypothesis apply to the medical domain and that improvements can be primarily attributed to domain shifting rather than knowledge injection. However, the precise impact of different types of QA data (e.g., complexity or difficulty of the QA pair) remains unclear. Finally, we recommend further research to investigate the role of fine-tuning in the medical domain.

Limitations

In the following, we discuss some limitations of our work.

Experiments on knowledge transfer patterns. Although we conduct experiments investigating how knowledge transfers within and across clinical specialties, we limited our experiments to the medical domain. That means that validating to what extent (Zhou et al., 2024)'s hypothesis generalises to other knowledge-intensive domains is not something we address in this work. Further research is needed to investigate the role of fine-tuning and instruction data in other domains.

Clinical specialties. We have included 15 clinical specialties in our dataset due to their availability in the source datasets we used. This means that possibly relevant clinical specialties have been left out of our dataset (i.e., see the long tail in Figure 2). Future work should aim to include QA pairs from more varied and possibly rarer clinical specialties (e.g., speech language pathology; Kim et al. 2024).

Language and healthcare system. Our dataset draws on examples in English extracted mostly from US-centric datasets. We believe that important future work lies in extending QA datasets to other languages and healthcare systems.

Acknowledgment

XY and IC are funded by the project CaRe-NLP with file number NGF.1607.22.014 of the research programme AiNed Fellowship Grants which is (partly) financed by the Dutch Research Council (NWO).

References

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).

ContactDoctor. 2024. Contactdoctor-bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B.

Ronald Cornet and Nicolette de Keizer. 2008. BMC Med Inform Decis Mak. Forty years of SNOMED: a literature review, 8 Suppl 1(Suppl 1):S2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). In BERT: Pre-training of deep bidirectional transformers for language understanding, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Peng Fu, Yiming Zhang, Haobo Wang, Weikang Qiu, and Junbo Zhao. 2023. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. In Revisiting the knowledge injection frameworks, pages 10983–10997, Singapore. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. arXiv preprint arXiv:2104.08821. Simcse: Simple contrastive learning of sentence embeddings.

Eduardo C Garrido-Merchán, Cristina González-Barthe, and María Coronado Vaca. 2023. arXiv preprint arXiv:2303.13373. Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related financial risks.

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.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). In OLMo: Accelerating the science of language models, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Proceedings of the International Conference on Learning Representations (ICLR). Aligning ai with shared human values.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Proceedings of the International Conference on Learning Representations (ICLR). Measuring massive multitask language understanding.

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.

Yusong Ke, Hongru Lin, Yuting Ruan, Junya Tang, and Li Li. 2025. Mathematics. Correctness coverage evaluation for medical multiple-choice question answering based on the enhanced conformal prediction framework, 13(9):1538.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. In MedExQA: Medical question answering benchmark with multiple explanations, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

Mikhail V Koroteev. 2021. arXiv preprint arXiv:2103.11943. Bert: a review of applications in natural language processing and understanding.

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.

Nut Limsopatham. 2021. Proceedings of the Natural Legal Language Processing Workshop 2021. In Effectively leveraging BERT for legal document classification, pages 210–216, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Briefings in bioinformatics. Biogpt: generative pretrained transformer for biomedical text generation and mining, 23(6):bbac409.

Elijah Mayfield and Alan W Black. 2020. Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. In Should you fine-tune BERT for automated essay scoring?, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, et al. 2024. arXiv preprint arXiv:2404.00213. Injecting new knowledge into large language models via supervised fine-tuning.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Proceedings of the 18th BioNLP Workshop and Shared Task. In ScispaCy: Fast and robust models for biomedical natural language processing, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Nature Medicine. Toward expert-level medical question answering with large language models, 31(3):943–950.

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.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. Findings of the Association for Computational Linguistics: ACL 2024. In How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Proceedings of the AAAI conference on artificial intelligence. In Semantics-aware bert for language understanding, 05, pages 9628–9635.

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.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Findings of the Association for Computational Linguistics: EMNLP 2020. In Question answering with long multiple-span answers, pages 3840–3849, Online. Association for Computational Linguistics.

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,

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

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 6–10 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 Prompts used 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 36.50̆0b0C (97.70̆0b0F). 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.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 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 Additional cross-specialty evaluation results

In Tables 4, 5, 6, 7, 9, 10, and 11 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 OLMo-2-1124-7B-Instruct in addition to our main results (in §3.2). Here we also observe that in $\sim$70% of the cases 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.

A.6 Robustness of cross-specialty evaluation

We carry out 2 experiments to show the robustness of our cross-specialty evaluation results in Table 4. We show the results of our cross specialty evaluation fine-tuned on balanced training set sizes in Table 12, and the results of the 20 bootstrap resampling experiments on the test sets in Table 13. Both experiments show that most of the best performing models on each per-specialty's test set is not the same model tuned on the training data from the same specialty. This demonstrates the robustness of our main results, i.e., best models are almost always off-diagonal.

A.7 High-level Concepts in SNOMED-CT

The high-level SNOMED-CT *disorder* concepts we use when mapping questions to clinical specialities are: *cardiology* (Disorder of cardiovascular system, `SCTID: 49601007`), *gastroenterology* (Disorder of digestive system, `SCTID: 53619000`), *infectious diseases* (Infectious disease, `SCTID: 40733004`), *obstetrics* (Disorder of female reproductive system, `SCTID: 363124003`; Disorder of fetus and/or mother during labor, `SCTID: 1269083008`), *neurology* (Disorder of nervous system, `SCTID: 118940003`), *pediatrics* (Behavioral and emotional disorder with onset in childhood, `SCTID: 231538003`; Disorder of fetus and/or newborn, `SCTID: 414025005`; Developmental disorder, `SCTID: 5294002`).

A.8 Token probability change

Table 14 shows the token negative log-likelihoods (the higher the better) for terms from each clinical specialty before and after fine-tuning on the six S-MedQA specialty data and MMLU social sciences data. We note that, apart from a few exceptions, training on S-MedQA leads to consistently improved token probabilities for clinically relevant terms, whereas training on MMLU makes clinical relevant terms less likely.

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 5: Cross-specialty accuracy matrix of Llama2-7b.

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 6: Cross-specialty accuracy matrix of Llama2-13b.

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 7: Cross-specialty accuracy matrix of Mistral-v0.1

	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
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 8: Cross-specialty accuracy matrix of Mistral-v0.1

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 9: Cross-specialty accuracy matrix of Llama-3.2-3B-Instruct.

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 10: Cross-specialty accuracy matrix of Bio-Medical-Llama-3-8B.

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	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 11: Cross-specialty accuracy matrix of OLMo-2-1124-7B-Instruct.

Figure 6: 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 7: 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?

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
	Llama-3.1	64.0	67.5	65.3	66.8	65.0	64.5	65.5
Train Sets	Cardio	**71.8**	**77.4**	72.3	76.1	**69.9**	71.0	73.0
	Gastro	69.6	77.1	72.4	77.2	69.1	71.9	72.8
	Infect	68.8	74.5	73.3	**77.4**	69.8	71.3	72.4
	Neuro	70.8	77.1	**73.4**	76.9	68.6	**72.1**	73.0
	Obstetrics	69.8	75.0	72.9	75.4	68.7	69.4	71.9
	Pediatrics	68.9	75.3	72.8	77.0	69.1	70.4	72.2

Table 12: Cross specialty evaluation results fine-tuned on training data with equal sizes.

Figure 8: 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?

	Test Sets	Cardio	Gastro	Infect	Neuro	Obstetrics	Pediatrics	avg.
	Llama-3.1	70.1	72.3	71.0	72.8	70.2	69.4	71.0
Train Sets	Cardio	**79.7**	75.4	72.3	**77.6**	**73.1**	71.1	73.6
	Gastro	70.0	**77.4**	72.2	77.2	73.0	**72.7**	73.6
	Infect	70.4	73.8	72.7	77.3	71.0	70.3	72.6
	Neuro	72.0	76.0	72.7	77.3	72.0	71.3	73.4
	Obstetrics	70.3	75.4	71.9	75.7	72.5	69.5	72.4
	Pediatrics	71.2	74.8	**72.9**	76.4	71.4	70.8	72.8
	Combined	63.1	81.4	77.5	78.2	73.3	76.6	76.7

Table 13: Cross specialty evaluation results of 20 resamplings of the test sets.

Figure 9: 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 10: 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:

Models	**Test Sets** (↑ is better)					
	Cardio	**Gastro**	**Infect**	**Neuro**	**Obstetrics**	**Pediatrics**
Mistralv0.2-Base	-4.11	-5.05	-4.99	-5.59	-4.49	-5.80
Cardio	-4.06	-4.94	-4.86	-5.56	-4.35	-5.60
Gastro	-4.12	-5.01	-4.91	-5.46	-4.27	-5.65
Infect	-3.95	-4.84	-4.96	-5.55	-4.41	-5.59
Neuro	-4.07	-5.05	-4.88	-5.39	-4.37	-5.61
Obstetrics	-4.14	-5.11	-4.87	-5.50	-4.41	-5.66
Pediatrics	-4.08	-5.05	-4.93	-5.53	-4.46	-5.67
MMLU	-4.62	-5.34	-5.27	-5.83	-4.78	-6.01

Table 14: Average token negative log-likelihood for terms from each clinical specialty before and after fine-tuning on the six S-MedQA specialty data and MMLU social sciences data. Best viewed in colour.