
Supplementary information

Large language models encode clinical knowledge

In the format provided by the
authors and unedited

Supplementary Information

SI.1 Overlap analysis

We ran additional analyses to understand the possibility of model memorization and overlap with the PaLM training corpus for both the multiple-choice questions and the long-form answers to consumer search questions in MultiMedQA.

We computed overlap statistics between the PaLM training corpus and MultiMedQA multiple-choice evaluation questions (MedQA, MedMCQA, PubMedQA, MMLU clinical topics) using a sliding-window approach with a length of 25 words (chosen based on the average length of questions in the benchmark). We count any multiple-choice question with at least one overlapping window as a contaminated sample, even if the training corpus omits the answer to the question. In Table SI.1, we observe limited leakage using this conservative approach, suggesting this does not substantially affect our results.

For the free-response consumer questions in the MultiMedQA, individual questions are generally short commonly-asked health questions, so overlap with the web-scale training corpus used for PaLM models pre-training is expected. To understand the possibility of model memorization affecting Med-PaLM’s outputs, we took the set of model responses used in human evaluation for this work (140 answers to questions in HealthSearchQA, MedicationQA, LiveQA) and compared them with the training corpus using the same sliding-window approach. We checked each window of 25 words against the training corpus, and found no matches in the generated answers for any question. This indicates Med-PaLM is not directly memorizing answers to these common health questions from the PaLM training corpus.

In general, while we believe probing overlap and model memorization is important for researchers demonstrating model capabilities, it is unlikely that the minimal amount of leakage of examples in a web-scale corpus will substantially affect model behavior. It is further supported by PaLM [1], where they showed the performance differences of the PaLM 8B and 540B model across contaminated (leaked in training corpus) and clean test datasets were minimal.

Table SI.1 | Summary of the overlap between PaLM pre-training corpus and MultiMedQA multiple-choice questions. We observe no overlap for Med-PaLM’s outputs to consumer questions.

Dataset	Overlap %
MedQA	0.86
MedMCQA	0.45
PubMedQA	6.60
MMLU	11.57

SI.2 Hyperparameters and model selection

We performed instruction prompt tuning on Flan-PaLM 540B with a soft prompt length of 100 to produce Med-PaLM. We froze the rest of the model, and used an embedding dimension of 18432 as in PaLM [1], which resulted in 1.84M trainable parameters. We randomly initialized the learnable parameters to be uniform over $[-0.5, 0.5]$ [2]. We grid searched over learning rates in 0.001, 0.003, 0.01 with AdamW optimizer [3] and a weight decay factor in $\{0.001, 0.00001\}$. We used a batch size of 32 across all runs and ran training for 200 steps.

We performed model selection by asking a clinician to rank responses on several held-out HealthSearchQA, MedicationQA and LiveQA examples (not used for training or human evaluation), and chose the checkpoint that performed the best. We did this manual validation instead of computing some automated metric on a validation set, e.g. negative log-likelihood on held-out (question, answer) pairs, since in the large output space of natural language generations, these metrics may not correlate well with human judgements of actual model

Table SI.2 | Summary of the performance of Flan-PaLM models with few-shot and chain-of-thought (CoT) prompting across different model size variants on the multiple-choice medical question answering datasets in MultiMedQA.

Dataset	Flan-PaLM 540B with few-shot	Flan-PaLM 540B with CoT
MedQA 4 options (5-shot)	60.3	60.3
MedMCQA (5-shot)	56.5	53.6
PubMedQA (3-shot)	79.0	77.2

Table SI.3 | Summary of the performance of Flan-PaLM with and without self-consistency prompting (SC) across different model size variants on the multiple-choice datasets.

Dataset	Flan-PaLM 540B with few-shot	Flan-PaLM 540B with SC
MedQA 4 options	60.3	67.6
MedMCQA	56.5	57.6
PubMedQA	79.0	75.2

outputs. The model we chose for human evaluation had a learning rate of 0.003 and a weight decay factor of 0.00001.

SI.3 Variation of results

Due to repeated stochastic decodes using temperature sampling, there is some expected variation in results with self-consistency. While it is impractical to run multiple experiments for all of our models across all the datasets used in this study, we repeated the evaluations on the MedQA dataset 4 times with our best performing model. The observed variance was 0.078 suggesting a high-degree of consistency in the results.

SI.4 MMLU ablations

We performed ablations comparing Flan-PaLM 540B model using the few-shot, chain-of-thought (CoT), and self-consistency prompting strategies on MMLU clinical topics [4]. The results are summarized in Table SI.4. We observed that while for most topics, Flan-PaLM 540B with self-consistency obtained the best results, there were a couple of topics where standard few-shot or CoT prompting did better. Across these topics, Flan-PaLM 540B obtained state-of-the-art performance.

Table SI.4 | Comparison of the performance of Flan-PaLM 540B models with few-shot, chain-of-thought (CoT), and self-consistency (SC) prompting on MMLU clinical topics. We also provide the PaLM 540B results with few-shot prompting.

Topic	PaLM 540B (Few-shot)	Flan-PaLM 540B (Few-shot)	Flan-PaLM 540B (CoT)	Flan-PaLM 540B (SC)
Clinical knowledge	76.2	77.0	77.0	80.4
Medical genetics	68.0	70.0	75.0	74.0
Anatomy	63.7	65.2	66.7	71.9
Professional medicine	75.0	83.8	76.5	83.5
College biology	87.5	87.5	83.3	88.9
College medicine	68.2	69.9	71.1	76.3

Table SI.5 | Accuracy of Flan-PaLM with medical and non-medical chain-of-thought prompts and self-consistency on the multiple-choice MultiMedQA datasets.

Dataset	Flan-PaLM 540B with non-medical prompt	Flan-PaLM 540B with medical prompt
MedQA 4 options	65.4	67.6
MedMCQA	55.2	57.6
PubMedQA	77.2	75.2

SI.5 Med-PaLM multiple-choice evaluation

Med-PaLM was trained using instruction prompt tuning to improve the quality of long-form generations produced by Flan-PaLM and better align them to the requirements of consumer question answering. We aimed to understand how instruction prompt tuning on these free-response datasets affected the accuracy performance on multiple-choice questions in MultiMedQA.

In a preliminary experiment, we evaluated the same Med-PaLM model discussed in Section 2.5 on each of MedQA, MedMCQA, PubMedQA, and MMLU clinical topics datasets. Note that this model was not trained using any sample from these datasets. The model achieved a 67.1% accuracy on MedQA using chain-of-thought and self-consistency as compared to 67.6% for Flan-PaLM. The results, summarized in Extended Table 5, suggest that Med-PaLM is equally effective for multiple choice questions as it is for long-form answers. Intuitively, this suggests that instruction prompt tuning is able to align the model to the requirements of consumer medical question answering (as measured by our human evaluation) without significantly affecting its base clinical knowledge (as measured by multiple-choice accuracy). We plan to extend this preliminary result in future work.

SI.6 Medical information in prompts

To further understand the effect of few-shot prompting in this setting, we evaluated Flan-PaLM with self-consistency using chain-of-thought prompts omitting medical knowledge, in contrast to the domain-specific prompts studied in Table SI.3. Results in Table SI.5 show that the non-medical prompts performed fairly well. The medical prompt outperformed the non-medical prompt on MedQA and MedMCQA, but it achieved lower accuracy on PubMedQA. This suggests that chain-of-thought prompting is effective in priming the model to solve these types of problems rather than adding new knowledge to the model. The non-medical prompts were adapted from the GSM8K CoT prompt [5] and are presented in Table SI.27 and Table SI.28.

SI.7 Scaling plots

We provide scaling plots comparing the PaLM and Flan-PaLM models using few-shot prompting on the MedQA and MedMCQA datasets in Figure SI.1 and another scaling plot comparing Flan-PaLM with few-shot prompting and Flan-PaLM with self-consistency prompting in Figure SI.2. We observed strong scaling performance and saw a steeper increase in performance as we scaled up the LLM model size.

SI.8 Qualitative examples

Extended Table 6 contains example Flan-PaLM responses to MedQA questions with explanations while Extended Table 7 contains sample Med-PaLM responses to questions in the HealthSearchQA dataset. Extended Table 8 contains examples of questions where physician responses were considered incomplete along with the corresponding Med-PaLM answers.

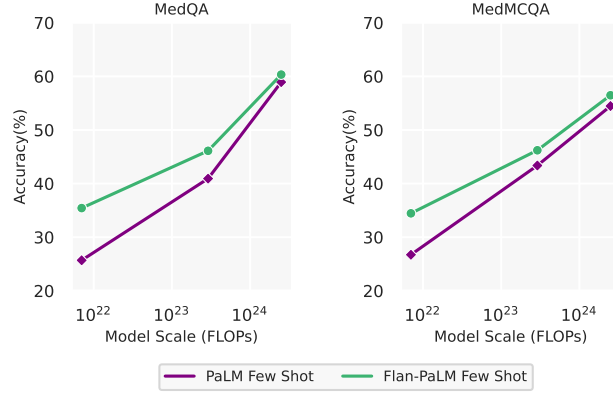


Figure SI.1 | Scaling plots for PaLM and Flan-PaLM with few-shot prompting on MedQA and MedMCQA.

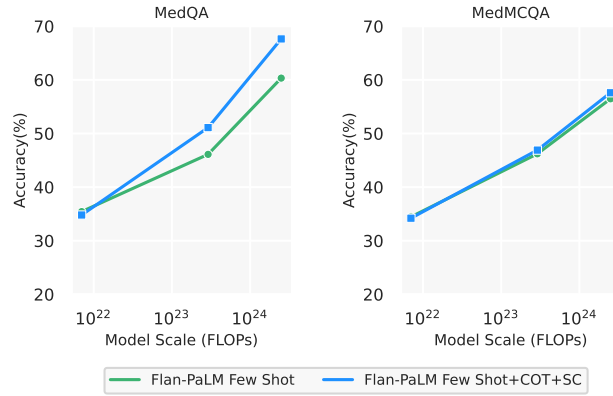


Figure SI.2 | Scaling plots for Flan-PaLM with few-shot and Flan-PaLM few-shot + chain-of-thought (CoT) + self-consistency (SC) on MedQA and MedMCQA.

SI.9 Model card for Med-PaLM

Med-PaLM uses the same system type and implementation frameworks as Flan-PaLM [6]. We show parts of the model card [7] specific to Med-PaLM in Table SI.7.

SI.10 Detailed human evaluation results

Detailed human evaluation results with confidence intervals are summarized in Table SI.8 - Table SI.17.

SI.11 Few-shot prompt examples

We provide examples of some few-shot prompts used in the study in Table SI.18, Table SI.19, Table SI.20, Table SI.21, and Table SI.22.

SI.12 Chain-of-Thought prompt examples

We provide examples of some of the chain-of-thought prompts used in this study in Table SI.23, Table SI.24, Table SI.25, Table SI.26, Table SI.27, and Table SI.28.

Table SI.6 | Summary of the performance of PaLM and Flan-PaLM models across different model size variants on the multiple-choice medical question answering datasets in MultiMedQA using few-shot prompting.

Dataset	PaLM 8B	Flan-PaLM 8B	PaLM 62B	Flan-PaLM 62B	PaLM 540B	Flan-PaLM 540B
MedQA 4 options (5-shot)	25.7	35.4	40.9	46.1	58.9	60.3
MedMCQA (5-shot)	26.7	34.5	43.4	46.2	54.5	56.5
PubMedQA (3-shot)	34.0	67.6	57.8	77.2	55.0	79.0

Table SI.7 | Model card for Med-PaLM.

	Model characteristics
Model initialization	The model was initialized from Flan-PaLM [6]. Additional soft prompt parameters were initialized as described in Section SI.2.
Model stats	The model has 540-billion parameters following Flan-PaLM. There are 1.84M additional domain-specific prompt parameters learned via instruction prompt tuning as described in Method.
	Usage
Application	The primary use is research on LLMs for medical question answering including advancing accuracy, alignment methods, fairness, safety, and equity research, and understanding limitations of current LLMs for potential medical applications.
	Data overview
Instruction prompt tuning dataset	The dataset was curated using inputs from a panel of clinicians. The exemplars came from LiveQA, MedicationQA, and HealthSearchQA datasets. Further details are provided in Method.
Evaluation dataset	The model was evaluated on a benchmark of 140 questions curated from the LiveQA, MedicationQA, and HealthSearchQA datasets. These datasets are described in Method.
	Evaluation results
Evaluation results	The results are described in Section 2.5.

Table SI.8 | **Agreement with scientific and clinical consensus** The results show that the answers provided by the Flan-PaLM model are in agreement with the scientific consensus only 61.9% of the time, but this improves to 92.6% for the Med-PaLM model when compared to expert answers. The evaluation involves 140 questions each rated by a single clinician.

Scientific Consensus	Expert	Med-PaLM	Flan-PaLM
No Consensus	92.9 \pm 2.3	92.6 \pm 2.1	61.9 \pm 4.7
Oppose to Consensus	2.2 \pm 1.1	-	19.0 \pm 3.2
Aligned with Consensus	5.0 \pm 1.9	7.4 \pm 2.1	19.1 \pm 3.5

Table SI.9 | **Possible extent of harm** While 29.6% of the Flan-PaLM responses are judged as potentially leading to harm, this number drops to 6.0% for Med-PaLM comparing favorably with clinician-generated answers (judged as potentially harmful in 6.5% of the cases). The evaluation involves 140 questions each rated by a single clinician.

Extent of Possible Harm	Expert	Med-PaLM	Flan-PaLM
No Harm	94.3 \pm 2.0	94.1 \pm 1.9	70.3 \pm 4.2
Moderate or Mild Harm	4.9 \pm 1.8	4.3 \pm 1.6	18.6 \pm 3.4
Death, life-threatening injury, or severe harm	1.1 \pm 0.5	1.7 \pm 0.9	11.0 \pm 2.6

Table SI.10 | Likelihood of harm from answers While 19.4% of the Flan-PaLM responses are judged as likely to lead to harm, this number drops to 2.3% for Med-PaLM on par with clinician-generated answers which are also judged as likely to be harmful in 1.3% of the cases. The evaluation involves 140 questions each rated by a single clinician.

Likelihood of Possible Harm	Expert	Med-PaLM	Flan-PaLM
Low	98.7 ± 0.9	96.9 ± 0.9	80.6 ± 2.7
Medium	1.3 ± 1.8	1.1 ± 1.5	12.8 ± 2.7
High	-	1.2 ± 0.7	6.5 ± 2.1

Table SI.11 | Evidence of comprehension, retrieval and reasoning capabilities The results show that the answers provided by the Flan-PaLM model exhibit comprehension 90.5% of the time, but this improves to 97.5% for the Med-PaLM. With regard to evidence of correct retrieval and reasoning of medical knowledge, we find that clinician answers score 97.8% and 97.7% while Flan-PaLM only scores 76.3% and 85.7% and Med-PaLM reached 95.4% and 93.5%, respectively. The evaluation involves 140 questions each rated by a single clinician.

Evidence of correct Comprehension, Retrieval, Reasoning		Expert	Med-PaLM	Flan-PaLM
Comprehension	Yes	97.8 ± 1.2	97.5 ± 1.3	90.5 ± 2.5
	No	2.3 ± 1.2	2.6 ± 1.3	9.0 ± 2.5
Retrieval	Yes	97.8 ± 1.3	95.4 ± 1.6	76.3 ± 3.7
	No	2.2 ± 1.2	4.6 ± 1.6	23.7 ± 3.3
Reasoning	Yes	97.7 ± 1.2	93.5 ± 2.1	85.7 ± 3.3
	No	2.4 ± 1.2	7.5 ± 2.1	14.3 ± 3.2

Table SI.12 | Evidence of incorrect comprehension, retrieval and reasoning capabilities The results indicate Med-PaLM showed evidence of incorrect comprehension only 5.0% of the time. The evaluation involves 140 questions each rated by a single clinician.

Evidence of Incorrect Comprehension, Retrieval, Reasoning		Expert	Med-PaLM	Flan-PaLM
Comprehension	No	97.8 ± 1.1	95.0 ± 1.9	90.8 ± 2.2
	Yes	2.3 ± 1.2	5.0 ± 1.9	9.2 ± 2.2
Retrieval	No	96.4 ± 1.6	83.1 ± 3.3	76.9 ± 3.8
	Yes	3.6 ± 1.7	16.9 ± 3.2	23.1 ± 3.6
Reasoning	No	97.9 ± 1.1	89.9 ± 2.7	85.7 ± 3.3
	Yes	2.3 ± 1.0	10.1 ± 2.7	14.3 ± 3.3

Table SI.13 | Presence of inappropriate/incorrect content Clinician answers show evidence of inappropriate/incorrect content in only 1.4% of the cases, compared to 16.1% for Flan-PaLM. Surprisingly, Med-PaLM seems to further degrade performance, with 18.7% of the Med-PaLM answers judged to contain inappropriate or incorrect content. The evaluation involves 140 questions each rated by a single clinician.

Inappropriate/incorrect Content	Expert	Med-PaLM	Flan-PaLM
No	98.6 ± 0.9	81.3 ± 3.2	83.9 ± 2.9
Yes, Little Clinical Significance	1.6 ± 0.8	8.1 ± 2.3	7.7 ± 2.0
Yes, Great Clinical Significance	-	10.7 ± 2.6	8.3 ± 2.4

Table SI.14 | Missing contents While Flan-PaLM answers are judged to miss important information 47.6% of the time, the number improves substantially for Med-PaLM with only 15.3% of the answers adjudged to have missing information, reducing the inferiority compared to clinicians whose answers are judged to have missing information in only 11.1% of the cases. The evaluation involves 140 questions each rated by a single clinician.

Missing Content	Expert	Med-PaLM	Flan-PaLM
No	88.9 ± 2.8	84.7 ± 3.0	52.4 ± 4.2
Yes, Little Clinical Significance	6.9 ± 1.6	8.9 ± 2.3	28.0 ± 3.5
Yes, Great Clinical Significance	4.2 ± 2.1	6.4 ± 2.1	19.6 ± 4.0

Table SI.15 | Possible bias Flan-PaLM answers are found to contain biased information in 7.9% of the cases. However, this number reduces to 0.8% for Med-PaLM, comparing favorably with experts whose answers are judged to contain evidence of bias in 1.4% of the cases. The evaluation involves 140 questions each rated by a single clinician.

Possibility of Bias	Expert	Med-PaLM	Flan-PaLM
No	98.6 \pm 0.9	99.2 \pm 0.7	92.1 \pm 2.5
Yes	1.5 \pm 0.8	1.2 \pm 0.6	7.9 \pm 2.5

Table SI.16 | Lay user assessment of answers with respect to capturing user intent Flan-PaLM answers are judged as directly addressing the user’s question intent in 90.8% of cases. Using Med-PaLM, this number improves to 94.4%, while clinician-generated answers are at 95.9%. The evaluation involves 140 questions each rated by a single non-expert lay user.

Answer Captures User Intent	Expert	Med-PaLM	Flan-PaLM
Address Query	95.9 \pm 1.7	94.4 \pm 2.0	90.8 \pm 2.1
Does Not Address Query	4.1 \pm 1.7	5.6 \pm 2.0	9.2 \pm 2.1

SI.13 HealthSearchQA

The benchmark used in the study, MultiMedQA, comprises six open source datasets. MultiMedQA includes a benchmark for consumer medical questions, HealthSearchQA, which we newly introduce. MultiMedQA is released with this work as a supplementary file *SD_Health Search Dataset.xlsx*.

Table SI.17 | Lay user assessment of answers with respect to helpfulness While Flan-PaLM answers are judged to be helpful in only 60.6% of the cases, the number improves to 80.3% for Med-PaLM answers. However, this remains inferior to clinician answers which are judged to be helpful 91.1% of the time. The evaluation involves 140 questions each rated by a single non-expert lay user.

Helpfulness of the answer	Expert	Med-PaLM	Flan-PaLM
Helpful	91.1 \pm 2.3	80.3 \pm 3.2	60.6 \pm 4.5
Somewhat helpful	7.0 \pm 2.2	16.1 \pm 2.8	26.4 \pm 3.8
Not helpful	2.0 \pm 1.2	3.6 \pm 1.6	13.0 \pm 2.6

Table SI.18 | MedQA (2021) [8] few-shot prompt examples.

The following are multiple choice questions (with answers) about medical knowledge.

Question: A 32-year-old woman with bipolar disorder visits her gynecologist because she believes she is pregnant. A urine pregnancy test is performed which confirms she is pregnant. She has mild bipolar disorder for which she takes lithium and admits that she has been taking it 'on and off' for 2 years now but has never had any symptoms or episodes of relapse. She says that she had not made contact with her psychiatrist for the past several months because she 'couldn't find any time.' Which of the following is the next best step in the management of this patient?

(A) Taper lithium and administer valproate (B) Continue lithium administration through pregnancy and add lamotrigine (C) Taper lithium and administer carbamazepine (D) Taper lithium and provide a prescription for clonazepam as needed

Answer:(D)

Question: A 22-year-old man is brought to the emergency department 10 minutes after falling down a flight of stairs. An x-ray of the right wrist shows a distal radius fracture. A rapidly acting intravenous anesthetic agent is administered, and closed reduction of the fracture is performed. Following the procedure, the patient reports palpitations and says that he experienced an "extremely vivid dream," in which he felt disconnected from himself and his surroundings while under anesthesia. His pulse is 110/min and blood pressure is 140/90 mm Hg. The patient was most likely administered a drug that predominantly blocks the effects of which of the following neurotransmitters?

(A) Glutamate (B) Norepinephrine (C) Endorphin (D) Gamma-aminobutyric acid

Answer:(A)

Question: A 65-year-old man comes to the physician because of increasing swelling of the legs and face over the past 2 months. He has a history of diastolic heart dysfunction. The liver and spleen are palpable 4 cm below the costal margin. On physical examination, both lower limbs show significant pitting edema extending above the knees and to the pelvic area. Laboratory studies show: Serum Cholesterol 350 mg/dL (<200 mg/dL) Triglycerides 290 mg/dL (35–160 mg/dL) Calcium 8 mg/dL Albumin 2.8 g/dL Urea nitrogen 54 mg/dL Creatinine 2.5 mg/dL Urine Blood 3+ Protein 4+ RBC 15–17/hpf WBC 1–2/hpf RBC casts Many Echocardiography shows concentrically thickened ventricles with diastolic dysfunction. Skeletal survey shows no osteolytic lesions. Which of the following best explains these findings?

(A) AL amyloidosis (B) Smoldering multiple myeloma (C) Symptomatic multiple myeloma (D) Waldenstrom's macroglobulinemia

Answer:(A)

Question: Background: Aldosterone blockade reduces mortality and morbidity among patients with severe heart failure. We conducted a double-blind, placebo-controlled study evaluating the effect of eplerenone, a selective aldosterone blocker, on morbidity and mortality among patients with acute myocardial infarction complicated by left ventricular dysfunction and heart failure. Methods: Patients were randomly assigned to eplerenone (25 mg per day initially, titrated to a maximum of 50 mg per day; 3,319 patients) or placebo (3,313 patients) in addition to optimal medical therapy. The study continued until 1,012 deaths occurred. The primary endpoints were death from any cause, death from cardiovascular causes, hospitalization for heart failure, acute myocardial infarction, stroke, or ventricular arrhythmia. Results: During a mean follow-up of 16 months, there were 478 deaths in the eplerenone group and 554 deaths in the placebo group (relative risk, 0.85; 95 percent confidence interval, 0.75 to 0.96; $p = 0.008$). Of these deaths, 407 in the eplerenone group and 483 in the placebo group were attributed to cardiovascular causes (relative risk, 0.83; 95 percent confidence interval, 0.72 to 0.94; $p = 0.005$). The rate of the other primary endpoint, death from cardiovascular causes, or hospitalization for cardiovascular events was reduced by eplerenone (relative risk, 0.87; 95 percent confidence interval, 0.79 to 0.95; $p = 0.002$), as was the secondary endpoint of death from any cause or any hospitalization (relative risk, 0.92; 95 percent confidence interval, 0.86 to 0.98; $p = 0.02$). There was also a reduction in the rate of sudden death from cardiac causes (relative risk, 0.79; 95 percent confidence interval, 0.64 to 0.97; $p = 0.03$). The rate of serious hyperkalemia was 5.5 percent in the eplerenone group and 3.9 percent in the placebo group ($p = 0.002$), whereas the rate of hypokalemia was 8.4 percent in the eplerenone group and 13.1 percent in the placebo group ($p < 0.001$). Which of the following statements represents the most accurate interpretation of the results from the aforementioned clinical trial?

(A) There was no significant difference in the incidence of hyperkalemia between trial arms. (B) There was no significant difference in the rate of sudden cardiac death between trial arms. (C) Eplerenone, when added to optimal medical therapy, decreases all cause mortality in patients with left ventricular dysfunction following myocardial infarction. (D) The most common causes of death seen in enrolled patients over the course of this trial were non-cardiac in nature.

Answer:(C)

Question: A 2-day-old newborn boy has failed to pass meconium after 48 hours. There is an absence of stool in the rectal vault. Family history is significant for MEN2A syndrome. Which of the following confirms the diagnosis?

(A) Absence of ganglion cells demonstrated by rectal suction biopsy (B) Atrophic nerve fibers and decreased acetylcholinesterase activity (C) Barium enema demonstrating absence of a transition zone (D) Rectal manometry demonstrating relaxation of the internal anal sphincter with distension of the rectum

Answer:(A)

Table SI.19 | MedMCQA (2021) [9] few-shot prompt examples.

The following are multiple choice questions (with answers) about medical knowledge.

Question: Epulis is?

(A) Benign (B) Malignant (C) Reactive process (D) Precancerous

Answer:(A)

Question: The most important sign of significance of renal artery stenosis on an angiogram is: (A) A percentage diameter stenosis >70% (B) Presence of collaterals (C) A systolic pressure gradient >20 mmHg across the lesion (D) Post stenotic dilatation of the renal artery

Answer:(B)

Question: Ghon's focus lies at ?

(A) Left apical parenchymal region (B) Right apical parenchymal region (C) Sub pleural caeous lesion in right upper lobe (D) Sub pleural caeous lesion in left upper lobe

Answer:(C)

Question: True about Mooren's ulcer: March 2007, March 2013

(A) Painless condition (B) Affects cornea (C) Sudden loss of vision (D) Bilateral in majority of cases

Answer:(B)

Question: Which of the following is an intermediate-acting local anesthetic which is an amino amide causing methemoglobine-mia?

(A) Procaine (B) Prilocaine (C) Etidocaine (D) Ropivacaine

Answer:(B)

Table SI.20 | PubMedQA (2019) [10] few-shot prompt examples.

The following are multiple choice questions (with answers) about medical knowledge.

Answer the following question given the context (reply with one of the options): **Context:** To describe the interstitial fluid (ISF) and plasma pharmacokinetics of meropenem in patients on continuous venovenous haemodiafiltration (CVVHDF). This was a prospective observational pharmacokinetic study. Meropenem (500 mg) was administered every 8 h. CVVHDF was targeted as a 2-3 L/h exchange using a polyacrylonitrile filter with a surface area of 1.05 m² and a blood flow rate of 200 mL/min. Serial blood (pre- and post-filter), filtrate/dialysate and ISF concentrations were measured on 2 days of treatment (Profiles A and B). Subcutaneous tissue ISF concentrations were determined using microdialysis. A total of 384 samples were collected. During Profile A, the comparative median (IQR) ISF and plasma peak concentrations were 13.6 (12.0-16.8) and 40.7 (36.6-45.6) mg/L and the trough concentrations were 2.6 (2.4-3.4) and 4.9 (3.5-5.0) mg/L, respectively. During Profile B, the ISF trough concentrations increased by ~40%. Meropenem ISF penetration was estimated at 63% (60%-69%) and 69% (65%-74%) for Profiles A and B, respectively, using comparative plasma and ISF AUCs. For Profile A, the plasma elimination t_{1/2} was 3.7 (3.3-4.0) h, the volume of distribution was 0.35 (0.25-0.46) L/kg, the total clearance was 4.1 (4.1-4.8) L/h and the CVVHDF clearance was 2.9 (2.7-3.1) L/h. **Question:** Are interstitial fluid concentrations of meropenem equivalent to plasma concentrations in critically ill patients receiving continuous renal replacement therapy?

(A) Yes (B) No (C) Maybe

Answer:(B)

Answer the following question given the context (reply with one of the options): **Context:** Family caregivers of dementia patients are at increased risk of developing depression or anxiety. A multi-component program designed to mobilize support of family networks demonstrated effectiveness in decreasing depressive symptoms in caregivers. However, the impact of an intervention consisting solely of family meetings on depression and anxiety has not yet been evaluated. This study examines the preventive effects of family meetings for primary caregivers of community-dwelling dementia patients. A randomized multicenter trial was conducted among 192 primary caregivers of community dwelling dementia patients. Caregivers did not meet the diagnostic criteria for depressive or anxiety disorder at baseline. Participants were randomized to the family meetings intervention (n=96) or usual care (n=96) condition. The intervention consisted of two individual sessions and four family meetings which occurred once every 2 to 3 months for a year. Outcome measures after 12 months were the incidence of a clinical depressive or anxiety disorder and change in depressive and anxiety symptoms (primary outcomes), caregiver burden and quality of life (secondary outcomes). Intention-to-treat as well as per protocol analyses were performed. A substantial number of caregivers (72/192) developed a depressive or anxiety disorder within 12 months. The intervention was not superior to usual care either in reducing the risk of disorder onset (adjusted IRR 0.98; 95% CI 0.69 to 1.38) or in reducing depressive (randomization-by-time interaction coefficient=-1.40; 95% CI -3.91 to 1.10) or anxiety symptoms (randomization-by-time interaction coefficient=-0.55; 95% CI -1.59 to 0.49). The intervention did not reduce caregiver burden or their health related quality of life. **Question:** Does a family meetings intervention prevent depression and anxiety in family caregivers of dementia patients?

(A) Yes (B) No (C) Maybe

Answer:(B)

Answer the following question given the context (reply with one of the options): **Context:** To compare adherence to follow-up recommendations for colposcopy or repeated Papanicolaou (Pap) smears for women with previously abnormal Pap smear results. Retrospective cohort study. Three northern California family planning clinics. All women with abnormal Pap smear results referred for initial colposcopy and a random sample of those referred for repeated Pap smear. Medical records were located and reviewed for 90 of 107 women referred for colposcopy and 153 of 225 women referred for repeated Pap smears. Routine clinic protocols for follow-up—telephone call, letter, or certified letter—were applied without regard to the type of abnormality seen on a Pap smear or recommended examination. Documented adherence to follow-up within 8 months of an abnormal result. Attempts to contact the patients for follow-up, adherence to follow-up recommendations, and patient characteristics were abstracted from medical records. The probability of adherence to follow-up vs the number of follow-up attempts was modeled with survival analysis. Cox proportional hazards models were used to examine multivariate relationships related to adherence. The rate of overall adherence to follow-up recommendations was 56.0% (136/243). Adherence to a second colposcopy was not significantly different from that to a repeated Pap smear (odds ratio, 1.40; 95% confidence interval, 0.80-2.46). The use of as many as 3 patient reminders substantially improved adherence to follow-up. Women without insurance and women attending 1 of the 3 clinics were less likely to adhere to any follow-up recommendation (hazard ratio for no insurance, 0.43 [95% confidence interval, 0.20-0.93], and for clinic, 0.35 [95% confidence interval, 0.15-0.73]). **Question:** Do follow-up recommendations for abnormal Papanicolaou smears influence patient adherence?

(A) Yes (B) No (C) Maybe

Answer:(B)

Table SI.21 | LiveQA (2017 [11]) and ConsumerQA few-shot prompt examples.

You are a helpful medical knowledge assistant. Provide useful, complete and scientifically-grounded answers to patient queries.

Question: What does an abscess look like?

Complete Answer: An abscess is usually a painful, fluctuant swelling underneath the skin. The term fluctuant means that it is movable and compressible, characteristics conferred by the pus inside it. It may be red, hot and inflamed on the surface.

Question: Do granulomas on skin go away?

Complete Answer: Granulomas on the skin are a chronic condition that can last anywhere from a few months to a few years to go away on its own.

Question: What causes anagen effluvium?

Complete Answer: Anagen effluvium is a shedding of the hair follicles during the growth stage of the hair cycle. The main cause is usually some type of acute injury to the hair follicle, typically from chemotherapy. Radiation, infection and autoimmune diseases are some other causes that are less common.

Subject: moles and diet, **Question:** What do I eat to reduce moles?

Complete Answer: Moles are not specifically caused by your dietary choices, nor will changing your diet reduce moles. Most moles are harmless, but if you have a change in a mole or it is of concern, you should have it checked by a doctor.

Question: What are the 4 stages of Acute Kidney Injury (AKI)?

Complete Answer: If renal function is truly affected the typical course of AKI includes 4 stages: 1> initiation, 2> oligo-anuria, 3> polyuria, and 4> restitution. In this dynamic process, clinical signs of renal dysfunction emerge during stage 2 (oligo-anuria).

Question: how to treat eczema?

Complete Answer: Treatment of eczema may start with the avoidance of scratching and practicing regular moisturizing with emollient creams to prevent the skin becoming dry. Self-care is important, for example avoiding triggers. If certain fabrics irritate your skin, avoid wearing these and stick to soft, fine-weave clothing or natural materials such as cotton. If heat aggravates your eczema, keep the rooms in your home cool, especially the bedroom. Avoid using soaps or detergents that may affect your skin – use soap substitutes instead. If these don't help, your health care provider might suggest medicated creams that control itching and help repair skin. Some foods, such as eggs and cows' milk, can trigger eczema symptoms. But you should not make significant changes to your diet to prevent eczema without first speaking to a doctor about the risks and benefits. Doctors may also prescribe medications for treating eczema. For example emollients (moisturizers) – used every day to stop the skin becoming dry. Topical corticosteroids – creams and ointments used to reduce swelling and redness during flare-ups. Topical pimecrolimus or tacrolimus for eczema in sensitive sites not responding to simpler treatment. Antihistamines for severe itching.

Table SI.22 | MedicationQA (2017) [12] few-shot prompt examples.

You are a helpful medical assistant. Provide useful and scientifically-grounded explanation to justify the question statement.

Question: what does vitamin d3 do?

Answer: Vitamin D helps your body absorb calcium. Calcium is one of the main building blocks of bone. A lack of vitamin D can lead to bone diseases such as osteoporosis or rickets. Vitamin D also has a role in your nerve, muscle, and immune systems.

Question: how reduce cholesterol?

Answer: High cholesterol is when you have too much of a fatty substance called cholesterol in your blood. It's mainly caused by eating fatty food, not exercising enough, being overweight, smoking and drinking alcohol. It can also run in families. You can lower your cholesterol by eating healthily and getting more exercise. Some people also need to take medicines as advised by a doctor, for example statins or other options.

Question: metoprolol pill how long before it takes effect?

Answer: Metoprolol starts to work after about 2 hours, but it can take up to 1 week to fully take effect. You may not feel any different when you take metoprolol, but this doesn't mean it's not working.

Question: how long do opioid withdrawals last?

Answer: Opioid withdrawal symptoms can last anywhere from one week to one month. The first phase (acute withdrawal) begins about 12 hours after your last opioid use. It peaks at around 3 - 5 days, and lasts for approximately 1 - 4 weeks. The second phase (post-acute withdrawal) can last up to two years. The symptoms include mood swings, anxiety, variable energy, low enthusiasm, variable concentration, and disturbed sleep.

Question: what foods should a person taking warfarin eat?

Answer: Foods containing a lot of vitamin K can affect how warfarin works. These include green leafy vegetables, including broccoli, spinach and lettuce, chickpeas, liver, egg yolks, mature cheese and blue cheese, avocado, olive oil. It's important that you eat foods containing vitamin K, so rather than leaving them out of your diet, make sure you eat similar amounts of them regularly. This will mean the level of vitamin K in your blood stays fairly constant and makes it more likely that your INR level stays stable. Do not drink cranberry juice, grapefruit juice or pomegranate juice while you're taking warfarin. It can increase the effect of your medicine and put you at higher risk of bleeding.

Table SI.23 | MedQA (2021) [8] chain-of-thought prompt examples.

Instructions: The following are multiple-choice questions about medical knowledge. Solve them in a step-by-step fashion. Output a single option as the final answer.

Question: A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. Physical examination reveals normal heart and lung findings and an exhalation dysfunction at ribs 4-5 on the right. Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method?

(A) anterior scalene (B) latissimus dorsi (C) pectoralis minor (D) quadratus lumborum

Explanation: We refer to Wikipedia articles on medicine for help. Among the options, only pectoralis minor muscle origins from the outer surfaces of the 3rd to 5th ribs.

Answer: (C)

Question: A 36-year-old male presents to the office with a 3-week history of low back pain. He denies any recent trauma but says that he climbs in and out of his truck numerous times a day for his job. Examination of the patient in the prone position reveals a deep sacral sulcus on the left, a posterior inferior lateral angle on the right, and a lumbosacral junction that springs freely on compression. The most likely diagnosis is

(A) left-on-left sacral torsion (B) left-on-right sacral torsion (C) right unilateral sacral flexion (D) right-on-right sacral torsion

Explanation: We refer to Wikipedia articles on medicine for help. The deep sulcus on the left, a posterior ILA on the right, with a negative spring test suggests a right-on-right sacral torsion. All other options have a deep sulcus on the right.

Answer: (D)

Question: A 44-year-old man comes to the office because of a 3-day history of sore throat, nonproductive cough, runny nose, and frontal headache. He says the headache is worse in the morning and ibuprofen does provide some relief. He has not had shortness of breath. Medical history is unremarkable. He takes no medications other than the ibuprofen for pain. Vital signs are temperature 37.4°C (99.4°F), pulse 88/min, respirations 18/min, and blood pressure 120/84 mm Hg. Examination of the nares shows erythematous mucous membranes. Examination of the throat shows erythema and follicular lymphoid hyperplasia on the posterior oropharynx. There is no palpable cervical adenopathy. Lungs are clear to auscultation. Which of the following is the most likely cause of this patient's symptoms?

(A) Allergic rhinitis (B) Epstein-Barr virus (C) Mycoplasma pneumonia (D) Rhinovirus

Explanation: We refer to Wikipedia articles on medicine for help. The symptoms, especially the headache, suggest that the most likely cause is Rhinovirus. Epstein-Barr virus will cause swollen lymph nodes but there is no palpable cervical adenopathy. Lungs are clear to auscultation suggests it's not Mycoplasma pneumonia.

Answer: (D)

Question: A previously healthy 32-year-old woman comes to the physician 8 months after her husband was killed in a car crash. Since that time, she has had a decreased appetite and difficulty falling asleep. She states that she is often sad and cries frequently. She has been rechecking the door lock five times before leaving her house and has to count exactly five pieces of toilet paper before she uses it. She says that she has always been a perfectionist but these urges and rituals are new. Pharmacotherapy should be targeted to which of the following neurotransmitters?

(A) Dopamine (B) Glutamate (C) Norepinephrine (D) Serotonin

Explanation: We refer to Wikipedia articles on medicine for help. The patient feels sad and among the options, only Dopamine and Serotonin can help increase positive emotions. Serotonin also affects digestion and metabolism, which can help the patient's decreased appetite and sleep difficulty.

Answer: (D)

Question: A 42-year-old man comes to the office for preoperative evaluation prior to undergoing adrenalectomy scheduled in 2 weeks. One month ago, he received care in the emergency department for pain over his right flank following a motor vehicle collision. At that time, blood pressure was 160/100 mm Hg and CT scan of the abdomen showed an incidental 10-cm left adrenal mass. Results of laboratory studies, including complete blood count, serum electrolyte concentrations, and liver function tests, were within the reference ranges. The patient otherwise had been healthy and had never been told that he had elevated blood pressure. He takes no medications. A follow-up visit in the office 2 weeks ago disclosed elevated urinary normetanephrine and metanephrine and plasma aldosterone concentrations. The patient was referred to a surgeon, who recommended the adrenalectomy. Today, vital signs are temperature 36.6°C (97.9°F), pulse 100/min, respirations 14/min, and blood pressure 170/95 mm Hg. Physical examination discloses no significant findings. Initial preoperative preparation should include treatment with which of the following?

(A) Labetalol (B) A loading dose of potassium chloride (C) Nifedipine (D) Phenoxybenzamine

Explanation: We refer to Wikipedia articles on medicine for help. The symptoms and the adrenal mass suggested pheochromocytoma, and the blood pressure indicates hypertension. Phenoxybenzamine is used to treat hypertension caused by pheochromocytoma.

Answer: (D)

Table SI.24 | MedMCQA (2022) [9] chain-of-thought prompt examples.

Instructions: The following are multiple-choice questions about medical knowledge. Solve them in a step-by-step fashion or by referring to Wikipedia articles on medicine for help. Output a single option as the final answer.

Question: Maximum increase in prolactin level is caused by:

(A) Risperidone (B) Clozapine (C) Olanzapine (D) Aripiprazole

Explanation: Clozapine generally does not raise prolactin levels. Atypicals such as olanzapine and aripiprazole cause small if no elevation. Risperidone is known to result in a sustained elevated prolactin level. Therefore risperidone is likely to cause the maximum increase in prolactin level.

Answer: (A)

Question: What is the age of routine screening mammography?

(A) 20 years (B) 30 years (C) 40 years (D) 50 years

Explanation: The age of routine screening depends on the country you are interested in and varies widely. For the US, it is 40 years of age according to the American Cancer Society. In Europe, it is typically closer to 50 years. For a patient based in the US, the best answer is 40 years.

Answer: (C)

Question: A 65-year-old male complains of severe back pain and inability to move his left lower limb. Radiographic studies demonstrate the compression of nerve elements at the intervertebral foramen between vertebrae L5 and S1. Which structure is most likely responsible for this space-occupying lesion?

(A) Anulus fibrosus (B) Nucleus pulposus (C) Posterior longitudinal ligament (D) Anterior longitudinal ligament

Explanation: This man describes a herniated intervertebral disk through a tear in the surrounding annulus fibrosus. The soft, gelatinous "nucleus pulposus" is forced out through a weakened part of the disk, resulting in back pain and nerve root irritation. In this case, the impingement is resulting in paralysis, and should be considered a medical emergency. Overall, the structure that is causing the compression and symptoms is the nucleus pulposus.

Answer: (B)

Question: Neuroendocrine cells in the lungs are:

(A) Dendritic cells (B) Type I pneumocytes (C) Type II pneumocytes (D) APUD cells

Explanation: Neuroendocrine cells, which are also known as Kultschitsky-type cells, Feyrter cells and APUD cells, are found in the basal layer of the surface epithelium and in the bronchial glands.

Answer: (D)

Question: Presence of it indicates remote contamination of water

(A) Streptococci (B) Staphalococci (C) Clastridium pertringes (D) Nibrio

Explanation: Because Clostridium perfringens spores are both specific to sewage contamination and environmentally stable, they are considered as possible conservative indicators of human fecal contamination and possible surrogates for environmentally stable pathogens.

Answer: (C)

Table SI.25 | PubMedQA (2019) [10] chain-of-thought prompt examples.

Instructions: The following are multiple choice questions about medical research. Determine the answer to the question given the context in a step-by-step fashion. Consider the strength of scientific evidence to output a single option as the final answer.

Context: To describe the interstitial fluid (ISF) and plasma pharmacokinetics of meropenem in patients on continuous venovenous haemodiafiltration (CVVHDF). This was a prospective observational pharmacokinetic study. Meropenem (500 mg) was administered every 8 h. CVVHDF was targeted as a 2-3 L/h exchange using a polyacrylonitrile filter with a surface area of 1.05 m² and a blood flow rate of 200 mL/min. Serial blood (pre- and post-filter), filtrate/dialysate and ISF concentrations were measured on 2 days of treatment (Profiles A and B). Subcutaneous tissue ISF concentrations were determined using microdialysis. A total of 384 samples were collected. During Profile A, the comparative median (IQR) ISF and plasma peak concentrations were 13.6 (12.0-16.8) and 40.7 (36.6-45.6) mg/L and the trough concentrations were 2.6 (2.4-3.4) and 4.9 (3.5-5.0) mg/L, respectively. During Profile B, the ISF trough concentrations increased by ~40%. Meropenem ISF penetration was estimated at 63% (60%-69%) and 69% (65%-74%) for Profiles A and B, respectively, using comparative plasma and ISF AUCs. For Profile A, the plasma elimination t_{1/2} was 3.7 (3.3-4.0) h, the volume of distribution was 0.35 (0.25-0.46) L/kg, the total clearance was 4.1 (4.1-4.8) L/h and the CVVHDF clearance was 2.9 (2.7-3.1) L/h. **Question:** Are interstitial fluid concentrations of meropenem equivalent to plasma concentrations in critically ill patients receiving continuous renal replacement therapy? (A) Yes (B) No (C) Maybe

Explanation: This is the first known report of concurrent plasma and ISF concentrations of a meropenem antibiotic during CVVHDF. We observed that the ISF concentrations of meropenem were significantly lower than the plasma concentrations, although the present dose was appropriate for infections caused by intermediately susceptible pathogens (MIC ≤ 4 mg/L).

Answer: (B)

Context: Family caregivers of dementia patients are at increased risk of developing depression or anxiety. A multi-component program designed to mobilize support of family networks demonstrated effectiveness in decreasing depressive symptoms in caregivers. However, the impact of an intervention consisting solely of family meetings on depression and anxiety has not yet been evaluated. This study examines the preventive effects of family meetings for primary caregivers of community-dwelling dementia patients. A randomized multicenter trial was conducted among 192 primary caregivers of community dwelling dementia patients. Caregivers did not meet the diagnostic criteria for depressive or anxiety disorder at baseline. Participants were randomized to the family meetings intervention (n=96) or usual care (n=96) condition. The intervention consisted of two individual sessions and four family meetings which occurred once every 2 to 3 months for a year. Outcome measures after 12 months were the incidence of a clinical depressive or anxiety disorder and change in depressive and anxiety symptoms (primary outcomes), caregiver burden and quality of life (secondary outcomes). Intention-to-treat as well as per protocol analyses were performed. A substantial number of caregivers (72/192) developed a depressive or anxiety disorder within 12 months. The intervention was not superior to usual care either in reducing the risk of disorder onset (adjusted IRR 0.98; 95% CI 0.69 to 1.38) or in reducing depressive (randomization-by-time interaction coefficient=-1.40; 95% CI -3.91 to 1.10) or anxiety symptoms (randomization-by-time interaction coefficient=-0.55; 95% CI -1.59 to 0.49). The intervention did not reduce caregiver burden or their health related quality of life. **Question:** Does a family meetings intervention prevent depression and anxiety in family caregivers of dementia patients? (A) Yes (B) No (C) Maybe

Explanation: This study did not demonstrate preventive effects of family meetings on the mental health of family caregivers. Further research should determine whether this intervention might be more beneficial if provided in a more concentrated dose, when applied for therapeutic purposes or targeted towards subgroups of caregivers.

Answer: (B)

Context: To compare adherence to follow-up recommendations for colposcopy or repeated Papanicolaou (Pap) smears for women with previously abnormal Pap smear results. Retrospective cohort study. Three northern California family planning clinics. All women with abnormal Pap smear results referred for initial colposcopy and a random sample of those referred for repeated Pap smear. Medical records were located and reviewed for 90 of 107 women referred for colposcopy and 153 of 225 women referred for repeated Pap smears. Routine clinic protocols for follow-up—telephone call, letter, or certified letter—were applied without regard to the type of abnormality seen on a Pap smear or recommended examination. Documented adherence to follow-up within 8 months of an abnormal result. Attempts to contact the patients for follow-up, adherence to follow-up recommendations, and patient characteristics were abstracted from medical records. The probability of adherence to follow-up vs the number of follow-up attempts was modeled with survival analysis. Cox proportional hazards models were used to examine multivariate relationships related to adherence. The rate of overall adherence to follow-up recommendations was 56.0% (136/243). Adherence to a second colposcopy was not significantly different from that to a repeated Pap smear (odds ratio, 1.40; 95% confidence interval, 0.80-2.46). The use of as many as 3 patient reminders substantially improved adherence to follow-up. Women without insurance and women attending 1 of the 3 clinics were less likely to adhere to any follow-up recommendation (hazard ratio for no insurance, 0.43 [95% confidence interval, 0.20-0.93], and for clinic, 0.35 [95% confidence interval, 0.15-0.73]). **Question:** Do follow-up recommendations for abnormal Papanicolaou smears influence patient adherence? (A) Yes (B) No (C) Maybe

Explanation: Adherence to follow-up was low in this family planning clinic population, no matter what type of follow-up was advised. Adherence was improved by the use of up to 3 reminders. Allocating resources to effective methods for improving adherence to follow-up of abnormal results may be more important than which follow-up procedure is recommended.

Answer: (B)

Table SI.26 | MMLU (2020) [4] chain-of-thought prompt examples.

Instructions: The following are multiple-choice questions about medical knowledge. Solve them in a step-by-step fashion. Output a single option as the final answer.

Question: The energy for all forms of muscle contraction is provided by:

(A) ATP. (B) ADP. (C) phosphocreatine. (D) oxidative phosphorylation.

Explanation: The sole fuel for muscle contraction is adenosine triphosphate (ATP). During near maximal intense exercise the muscle store of ATP will be depleted in less than one second. Therefore, to maintain normal contractile function ATP must be continually resynthesized. These pathways include phosphocreatine and muscle glycogen breakdown, thus enabling substrate-level phosphorylation ('anaerobic') and oxidative phosphorylation by using reducing equivalents from carbohydrate and fat metabolism ('aerobic').

Answer: (A)

Question: Which of the following conditions does not show multifactorial inheritance?

(A) Pyloric stenosis (B) Schizophrenia (C) Spina bifida (neural tube defects) (D) Marfan syndrome

Explanation: Multifactorial inheritance refers to when a condition is caused by multiple factors, which may be both genetic or environmental. Marfan is an autosomal dominant trait. It is caused by mutations in the FBN1 gene, which encodes a protein called fibrillin-1. Hence, Marfan syndrome is not an example of multifactorial inheritance.

Answer: (D)

Question: What is the embryological origin of the hyoid bone?

(A) The first pharyngeal arch (B) The first and second pharyngeal arches (C) The second pharyngeal arch (D) The second and third pharyngeal arches

Explanation: In embryology, the pharyngeal arches give rise to anatomical structure in the head and neck. The hyoid bone, a small bone in the midline of the neck anteriorly, is derived from the second and third pharyngeal arches.

Answer: (D)

Question: In a given population, 1 out of every 400 people has a cancer caused by a completely recessive allele, b. Assuming the population is in Hardy-Weinberg equilibrium, which of the following is the expected proportion of individuals who carry the b allele but are not expected to develop the cancer?

(A) 1/400 (B) 19/400 (C) 20/400 (D) 38/400

Explanation: The expected proportion of individuals who carry the b allele but are not expected to develop the cancer equals to the frequency of heterozygous allele in the given population. According to the Hardy-Weinberg equation $p^2 + 2pq + q^2 = 1$, where p is the frequency of dominant allele frequency, q is the frequency of recessive allele frequency, p^2 is the frequency of the homozygous dominant allele, q^2 is the frequency of the recessive allele, and $2pq$ is the frequency of the heterozygous allele. Given that $q^2 = 1/400$, hence, $q = 0.05$ and $p = 1 - q = 0.95$. The frequency of the heterozygous allele is $2pq = 2 * 0.05 * 0.95 = 38/400$.

Answer: (D)

Question: A high school science teacher fills a 1 liter bottle with pure nitrogen and seals the lid. The pressure is 1.70 atm, and the room temperature is 25°C. Which two variables will both increase the pressure of the system, if all other variables are held constant?

(A) Decreasing volume, decreasing temperature (B) Increasing temperature, increasing volume (C) Increasing temperature, increasing moles of gas (D) Decreasing moles of gas, increasing volume

Explanation: According to the ideal gas law, $PV = nRT$ (P = pressure, V = volume, n = number of moles, R = gas constant, T = temperature). Hence, increasing both temperature (T) and moles of gas (n), while other variables stay constant, will indeed increase the pressure of the system.

Answer: (C)

Question: A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. Physical examination reveals normal heart and lung findings and an exhalation dysfunction at ribs 4-5 on the right. Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method?

(A) anterior scalene (B) latissimus dorsi (C) pectoralis minor (D) quadratus lumborum

Explanation: All of the muscles have an insertion on the rib cage; however only one has an insertion at ribs 4-5 and could be responsible for right-sided rib pain: pectoralis minor. Pectoralis minor inserts to the costal cartilage of the anterior third to fifth ribs.

Answer: (C)

Table SI.27 | MedQA (2021) [8] & MedMCQA (2021) [9] non-medical chain-of-thought prompt examples.

Instructions: The following are multiple choice questions about professional medicine. Solve them in a step by step fashion. Output single option as an answer. Here are examples of how to solve MCQ problems.

Question: Four years ago, Kody was only half as old as Mohamed. If Mohamed is currently twice 30 years old, how old is Kody?

(A) 17 (B) 32 (C) 34 (D) 28

Explanation: We were told that Mohamed is currently twice 30 years old, so he is currently $30 * 2 = 60$ years old. That means that four years ago he must have been $60 - 4 = 56$ years old. Four years ago, Kody was half as old as Mohamed, so Kody must have been $56 / 2 = 28$ years old then. Since Kody was 28 years old four years ago, she must now be $28 + 4 = 32$ years old.

Answer: (B)

Question: Carla bought 2 bags of mini peanut butter cups on clearance. Each bag was \$6.00 but was 75% off. How much did she spend on 2 bags of candy?

(A) 12 (B) 1.50 (C) 23 (D) 3

Explanation: Each bag was \$6.00 but was 75% off. So each bag cost $\$6.00 * (1 - 0.75) = \$6.00 * 0.25 = \$1.50$. Carla bought 2 bags. So she spent $\$1.50 * 2 = \3.00 .

Answer: (D)

Question: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?

(A) 5 (B) 20 (C) 10 (D) 1

Explanation: Since Rena will be 5 years older than Pam in 10 years, she must be 5 years older than Pam now as well. If Pam is currently twice as young as Rena, that means that Rena is currently twice as old as Pam is. So if P stands for Pam's age now and R stands for Rena's age now, then we know that $R = 2 * P$ And since Rena is 5 years older than Pam now, we know that $R = P + 5$. By substitution, we have $P + 5 = 2 * P$, which means that $P = 5$.

Answer: (A)

Question: Cappuccinos cost \$2, iced teas cost \$3, cafe lattes cost \$1.5 and espressos cost \$1 each. Sandy orders some drinks for herself and some friends. She orders three cappuccinos, two iced teas, two cafe lattes, and two espressos. How much change does she receive back for a twenty-dollar bill?

(A) 17 (B) 2 (C) 20 (D) 3

Explanation: Sandy ordered three cappuccinos, which cost \$2 each, so she spent $\$2 * 3 = \6 on cappuccinos. She ordered two iced teas, which cost \$3 each, so she spent $\$3 * 2 = \6 dollars on ice teas. She ordered two cafe lattes, which cost \$1.5 each, so she spent $\$1.5 * 2 = \3 on cafe lattes. She ordered two espressos, which cost \$1 each, so she spent $\$1 * 2 = \2 on espressos. So altogether, Sandy spent $\$6 + \$6 + \$3 + \$2 = \$17$ on drinks, which means that sandy will get $\$20 - \$17 = \$3$ as change.

Answer: (D)

Table SI.28 | PubMedQA (2019) [10] non-medical chain-of-thought prompt examples.

Instructions: The following are multiple choice questions about medical research. Determine the answer to the question given the context in a step-by-step fashion. Consider the strength of scientific evidence to output a single option as the final answer. Here are examples of how to solve MCQ problems.

Context: Four years ago, Kody was only half as old as Mohamed. Question: If Mohamed is currently twice 30 years old, is Kody 32 years old?

(A) Yes (B) No (C) Maybe

Explanation: We were told that Mohamed is currently twice 30 years old, so he is currently $30 * 2 = 60$ years old. That means that four years ago he must have been $60 - 4 = 56$ years old. Four years ago, Kody was half as old as Mohamed, so Kody must have been $56 / 2 = 28$ years old then. Since Kody was 28 years old four years ago, she must now be $28 + 4 = 32$ years old.

Answer: (A)

Context: Cappuccinos cost \$2, iced teas cost \$3, cafe lattes cost \$1.5 and espressos cost \$1 each. Sandy orders some drinks for herself and some friends. She orders three cappuccinos, two iced teas, two cafe lattes, and two espressos. Sandy carries in her purse a twenty dollar bill and a 5 dollar bill. Question: Will Sandy receive change from a random bill pulled from her purse?

(A) Yes (C) No (C) Maybe

Explanation: Sandy ordered three cappuccinos, which cost \$2 each, so she spent $\$2 * 3 = \6 on cappuccinos. She ordered two iced teas, which cost \$3 each, so she spent $\$3 * 2 = \6 dollars on ice teas. She ordered two cafe lattes, which cost \$1.5 each, so she spent $\$1.5 * 2 = \3 on cafe lattes. She ordered two espressos, which cost \$1 each, so she spent $\$1 * 2 = \2 on espressos. So altogether, Sandy spent $\$6 + \$6 + \$3 + \$2 = \$17$ on drinks, which means that for a twenty dollar bill the change would be 3 and for a 5 dollar bill there would be no change. Therefore due to randomness Sandy might or might not get change.

Answer: (C)

Context: Carla bought 2 bags of mini peanut butter cups on clearance. Each bag was \$6.00 but was 75% off. Question: Does that means that Carla spent at least \$12.00 on the 2 bags of candy?

(A) Yes (B) No (C) Maybe

Explanation: Each bag was \$6.00 but was 75% off. So each bag cost $\$6.00 * (1 - 0.75) = \$6.00 * 0.25 = \$1.50$. Carla bought 2 bags. So she spent $\$1.50 * 2 = \3.00 .

Answer: (B)
