To Err is Human, How About Medical Large Language Models? Comparing Pre-trained Language Models for Medical Assessment Errors and Reliability

Wen-wai Yim†, Yujuan Fu*, Asma Ben Abacha†, Meliha Yetisgen*
†Microsoft Health AI, *University of Washington
†{yimwenwai,abenabacha}@microsoft.com
*{velvinfu,melihay}@uw.edu

Abstract

Unpredictability, especially unpredictability with unknown error characteristics, is a highly undesirable trait, particularly in medical patient care applications. Although large pre-trained language models (LLMs) have been applied to a variety of unseen tasks with highly competitive and successful results, their sensitivity to language inputs and resulting performance variability is not well-studied. In this work, we test state-of-the-art pre-trained language models from a variety of families to characterize their error generation and reliability in medical assessment ability. Particularly, we experiment with multiple choice tests for general medical assessment, as well as their open-ended and true-false alternatives. We also profile model consistency, error agreements with each other and with humans; and finally, quantify their ability to recover and explain errors. The findings in this work provide new insights on medical models so that modelers can make better-informed decisions instead of relying only on standalone performance metrics.

Keywords: Medical Question Answering, Large Language Models, Evaluation

1. Introduction

In 1999, the United States Institute of Medicine published a report that at least forty thousand deaths are a result of preventable medical errors. The report, titled "To Err Is Human: Building a Safer Health System" asserted that humans – in these cases, highly trained domain experts like doctors and nurses – will make honest mistakes but a well-planned system can prevent or at least mitigate such errors (Institute of Medicine (US) Committee on Quality of Health Care in America, 2000). While human error is a well-studied topic, it is less clear what types of errors are produced and how to mitigate the role of errors from pre-trained generalized large language models (LLMs), especially in the medical domain.

Unlike previous models, pre-trained LLMs (e.g., GPT-3.5, PALM) are being applied as generalized AI models, "instructed" to solve a variety of tasks, previously unseen, and doing so with competitive performances. Such tasks can be as diverse as multiple choice reading comprehension, summarization, machine translation, text classification, named entity recognition, and relation extraction. However, early findings reveal several unprecedented challenges: (a) performance inconsistency for the exact same inputs processed at different times and (b) performance inconsistency for semantically equivalent but variant surface form prompts. Specifically the same task prompt reworded slightly differently or with additional context may lead to wildly different results. Even the same prompt with inference performed at a different time may result in different outputs – a phenomenon previously not characteristic of artificial intelligence machine learning models. Because of these model behaviors, it is difficult to predict when these LLMs will fail and if so, how ungrounded errors are.

In this work, we survey models applied to the medical domain, framed in the context of general medical assessment tests, including its multiple choice, binary, and free-text variants. The objective is to understand the stability of LLMs as applied to medical problems and compare their errors to that of smaller pre-trained language models competitive in the same task. We further quantify differences in model errors and compare them with possible human errors to assess differences.

Our contributions are as follows:

- An annotated dataset of approximated question errors by medical students and by medical doctors, as well as a dataset of model outputs with different configurations which may be used to develop new consistency metrics.
- A framework for measuring consistency for non-deterministic LLMs (e.g., GPT-3.5, PALM) in the medical domain; and results for the consistency assessment for current state-of-the-art models. [1]

[1] https://github.com/velvinnn/LLM_error_study

LREC-COLING 2024, pages 16211–16223
20-25 May, 2024.

Original Multi-choice Question

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 10-year-old Caucasian female with Turner's syndrome underwent an abdominal imaging study and was discovered that the poles of her kidneys were fused inferiorly. Normal ascension of kidney during embryological development would be prevented by which of the following anatomical structures?

Options: (A) Inferior vena cava (B) Superior mesenteric artery (C) Inferior mesenteric artery (D) Celiac artery (E) Splenic artery

GPT-4 response: Explanation: We refer to Wikipedia articles on medicine for help. The superior mesenteric artery (SMA) is the structure that the kidneys must ascend past during embryological development. If the kidneys fail to ascend past the SMA, they may become fused at the lower poles, resulting in a horseshoe kidney.

Answer: (B) Superior mesenteric artery.

Open-ended Question Version Example

Instructions: Answer the question concisely based on your medical knowledge. Solve them in a step-by-step fashion.

Question: ... Normal ascension of kidney during embryological development would be prevented by what anatomical structures?

GPT-4 response: The normal ascension of the kidneys during embryological development can be prevented by the inferior mesenteric artery and the gonadal vessels. These structures can act as physical barriers to the upward migration of the kidneys.

Binarized Question Version Example

Instructions: Decide if the answer is true to the question based on medical knowledge. Solve them in a step-by-step fashion.

Question: ... Normal ascension of kidney during embryological development would be prevented by what anatomical structures?

Answer: Inferior vena cava

GPT-4 response: We refer to Wikipedia articles on medicine for help. The answer is wrong because the inferior vena cava is not an anatomical structure that would prevent the normal ascension of the kidneys during embryological development. The kidneys ascend from the pelvic region to the lumbar region during development, and this process is not hindered by the inferior vena cava.

Table 1: Example of a medical assessment exam multiple choice question with its true/false and open-ended variants and GPT-4 response.

- Quantifiable measurements for how often the incorrect answers from different medical assessments pre-trained language models are consistent with each other and to humans.

- A study of generalized LLMs' ability to differentiate and give explanations for medical assessment errors.

We suggest that traditional measures of quality, such as accuracy, are only one facet in this new AI modeling landscape; understanding a model's consistency and error profiles compared to domain experts are important considerations. We envision this study can be used to inform modelers on expected results and failures of state-of-the-art medical models and inspire ways to emphasize strengths and manage risks - much in the way that the original Institute of Medicine report spurred the creation of checklists and safety measures for doctors and nurses.

2. Background and Related Work

Medical assessment testing is a hallmark of medical education worldwide. These tests include combinations of multiple choice (MC), free written, and oral examinations (Price et al., 2018; Li et al., 2021). In the United States, the United States Medical Licensing Examination (USMLE) medical students take several tests during their student careers including the UMSLE Step 1, 2, and 3, prior to clinical rotations, after clinical rotations, and during the first year of residency (USMLE, 2023; Association, 2022). An example MC question is shown in Table 1.

As well as testing humans, medical assessment multiple choice questions have been one choice for automated benchmarking of medical assessment ability progress. For example, the MedQA dataset includes a dataset of medical practice exam questions for USMLE questions and Chinese exams (both Mainland and Taiwan), with several retrieval- and BERT-based baselines(Jin et al., 2020). The authors, (Li et al., 2021), provide BERT and RoBERTa as baselines for a variety of biomedical question-answer (QA) scenarios with both text and other data (e.g., images or tables) in Chinese. In (Nori et al., 2023), GPT-4 is shown to score an 86.70 accuracy score on average across Steps 1, 2, and 3. The work of (Singhal et al., 2023) reveals FlanPalm accuracies of 57.6 on the MedMCQA dataset, 67.6 on the MedQA (USMLE type) dataset, and 79.0 on PubMedQA.

The advent of instruction-tuned LLM (Ouyang et al., 2022) has spurred home-grown small expert-validated datasets that have been used to quantify the general skills of these models, such as ChatGPT. For example, in (Johnson et al., 2023) 284 domain-expert-created medical questions are used to test ChatGPT and GPT-4 with answers graded by physicians using a Likert-scale rubric. (Al-Dujaili et al., 2023) studied ChatGPT performance for 20 cases of pharmacotherapy questions; the answers for these questions were assessed at different time points with the highest score of 87.5% accuracy.

In (Schubert et al., 2023), GPT-4 achieves 85% accuracy for a 1,956 neurology board exam questions. To test reproducibility, the same prompt was repeated multiple times. GPT-4 was given a score of 81.3% (if defining 75% same output across 50 independent queries). The authors also found that correct answers tended to be on average more consistent.

In (Liévin et al., 2022), the authors tested GPT-3.5 on USMLE, MedMCQA, and PubMedQA and analyzed if errors came from incorrect reasoning, insufficient knowledge, or reading comprehension problems. The authors suggested that when not

sure, LLMs can make choices by heuristics and be inclined to a certain option such as A or D.

In the general domain, there has been limited study on LLM consistency. Primarily the work of (Jang et al., 2022) and (Jang and Lukasiewicz, 2023) test consistency of LLM models with respect to various 2-input classification tasks for models such as RoBERTa, ChatGPT, BART, and T5. Specifically, score consistency by measuring model output when the order of the two inputs was switched and the answer didn't expect output (e.g., in a paraphrase classification task, the answer to if "sentence1 is a paraphrase of sentence2" should be equivalent to if "sentence2 is a paraphrase of sentence1").

In this work, like (Jang and Lukasiewicz, 2023), we apply a suite of systematic inconsistency tests, however, we construct our experiments as related to medical questions given a clinical narrative context. Different from (Al-Dujaili et al., 2023; Schubert et al., 2023), we extend the consistency and reliability tests by additionally studying surface and semantic variations; as well as compare results across multiple model families.

3. Dataset

In this work, we study the US portion of the MedQA dataset (Jin et al., 2020) Typically a question will contain a clinical context, a question, and several options, as shown in Table 1. In the real USMLE test the number of multiple choice options can range from 5-11. The MedQA dataset includes both a simplified 4-option version as well as the full-un-edited version. While most work report on the 4-option version, preliminary experiments from Microsoft/OpenAI's study of GPT-4 revealed a drop in performance when considering a 5-option multiple choice versus 4-options(Nori et al., 2023).

We further extend the dataset by creating an open-ended and true-false variant of the multiple choice questions in the test set using a mixture of automated replacements and human evaluation. For example, we replace 'which' with 'what', so that questions such as 'which of the following is the best treatment for ... ' will be converted to 'what is the best treatment for ...'. The True/False variant further extends these questions by appending each possible answer in the option pool and querying for veracity. Examples are shown in Table 1; Full details are described in the Appendix 8.1.

To understand whether models would make the same mistakes as human domain experts, 200 randomly chosen questions out of the standard test set were selected and annotated. To create a proxy for human domain expert error, domain experts were asked to identify the next best answer if the correct answer was unavailable. We assume that although medical question and answer datasets, by design, will seek to provide one best answer they would also provide other plausible answers; the annotators' function is not only to screen away trivially incorrect answers but also to find a credible incorrect answer if available. The inter-annotator agreement (IAA) between two doctors (Dct), a family medicine and emergency medicine medical doctor, and four medical students (Std) on 25 questions for the error approximation is shown in Table 2.

	$\textbf{Dct}_1$	$\textbf{Dct}_2$	$\textbf{Std}_1$	$\textbf{Std}_2$	$\textbf{Std}_3$	$\textbf{Std}_4$
$\textbf{Dct}_1$	100	40	24	40	20	20
$\textbf{Dct}_2$		100	36	40	24	28
$\textbf{Std}_1$			100	32	48	56
$\textbf{Std}_2$				100	28	52
$\textbf{Std}_3$					100	44
$\textbf{Std}_4$						100

Table 2: Pairwise IAA (accuracy, %) on the next best answer on 25 questions.

Based on the IAA experiments, individual pairwise agreements for the next best answer were not high, signaling the difficulty of the task and the subjective nature of the problem. The medical student and doctor annotators had different thought processes in their reasoning for selecting their next-best answer. Despite this, we found that though individual pairs may not always agree on the same labels per question, in a small "crowd" a clear favorite label by majority vote will emerge. In fact, 21/25 (84%) of questions had at least 3 annotators agree on the same next best answer.

The rest of the randomly sampled 200 questions were annotated by one medical doctor and at least three medical students. 28 out of the 200 questions were marked as invalid due to missing essential information, i.e. an image is needed to answer the question. The medical doctor also provided a free-text explanation of why the chosen error option is incorrect.

4. Methods

4.1. Evaluated Models

For our study, we selected several publicly available state-of-the-art (SOTA) LLMs from different model families, either proven to show good performance on the MedQA-USMLE or widely used for a variety of tasks.

Instruction-tuned LLM's model families include:

- PALM (Anil et al., 2023) (Palm2), a large general artificial intelligence model developed by Google. MedPalm or MedPalm2 was not available at the time of this work.

- GPT-3.5 (ChatGPT) (Brown et al., 2020b) and GPT-4 created by OpenAI and Microsoft (Achiam et al., 2023; Nori et al., 2023)
- Llama (Llama2-70b-chat) an open source instruction-tuned large language model developed by Meta AI (Touvron et al., 2023).

Fine-tuning-based task-specific LLMs include:

- BioMedLM (Bolton et al., 2022) (previously known as PubMedGPT), a 2.7B parameter language model trained on biomedical literature out of Stanford University
- Dragon (Yasunaga et al., 2022) a QA-specific model pretrained with text and knowledge graph information, with proven succesful results on the MedQA dataset

4.2. LLM Consistency Experiments

We propose measuring consistency by providing alternate forms of the original medical assessment questions with surface and semantic variations, as well as typical LLM configuration differences such as changing temperature and priming with N-shot examples. The 5-option multiple choice (with 5-shot examples for instruction-based LLMs, temperature set to 0[2]) version is used as a basis of comparison unless otherwise stated – we picked this setting to be more realistic to actual testing and running settings. LLM experiments settings included fixed prompts with inspiration from previous work (Singhal et al., 2023).

In the following text, we describe our experimental settings as well as our measures of consistency.

Experiments with typical LLM configuration, optimization, and superficial differences:

1. **Repeated runs**: For the instruction-tuned LLMs, we used the same prompt repeated at temperature 0. Fine-tuning-based models repeat the same experiments conducted with the same fine-tuning hyperparameters but different random seeds.
2. **Temperature**: In these experiments, temperature is varied [0,0.2,0.4,0.6,0.8,1.0].
3. **N-shot**: For instruction-based models, providing examples to prompt the model has experimentally been shown to improve results. Here we quantify the variations of 1-5 shot examples.
4. **MC re-order**: Here the same MC options are presented however the order of the multiple choice is shuffled. We experimented with three different settings: (1) random shuffling keeping the same answer ID (e.g., (a) can move to the second position), (2) random shuffling with answer ID's renamed such they are alphabetical (e.g., first option is (a), second option (b), etc), (3) randomly shuffling the order of the 5-shot examples.

To measure consistency, we take the standard deviation of accuracy across runs in an experiment group.

Experiments where the original problem is altered:

1. **4-option version**: Compared to the baseline system, the 4-option version effectively removes one additional wrong option. Thus the problem is simplified.
2. **Open-ended**: In these experiments, questions are changed to their open-ended forms. For the evaluation of open-ended questions of correctness, an annotator manually evaluated 50 random questions accuracy. We additionally calculated typical natural language generation (NLG) metrics such as ROUGE.
3. **Binarization**: Here questions are changed to their True/False variation. This is created by using the open-ended variant and concatenating each possible option to the end - with instructions for a model to assess truth-hood.

To measure consistency, we measure the number of questions for which answers are changed as compared to the baseline setting.

4.3. Error Comparison

Previous work studies overall model performances in terms of accuracy, direct human assessment (e.g., rating the same samples of output), or isolated case studies. However, these assessments do not provide information on whether models may get the same questions correct or incorrect and if the provided incorrect answers may match each other.

In this analysis, two system outputs (e.g., two different models or a model and a human), may be compared according to the equation:

$$error\_overlap_{ij} = \frac{|\{q_{error}\}_i \cap \{q_{error}\}_j|}{|\{q_{error}\}_j|} \quad (1)$$

where $\{q_{error}\}_x$ is the set of all questions answered incorrectly by system *x*. We also use an exact answer variant that requires each question must overlap and have the same answer.

[2]For Llama with temperature 0, results are documented to be highly unstable because of division by 0 calculations. Instead, we use a small temperature of 0.01 to approximate this.

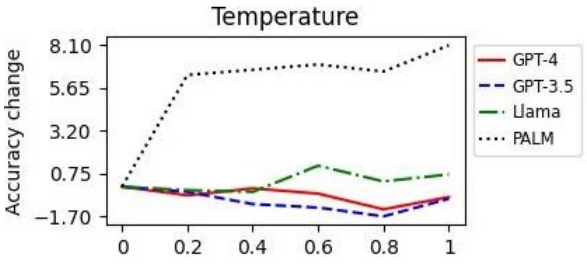

Figure 1: LLM prediction accuracy vs temperature.

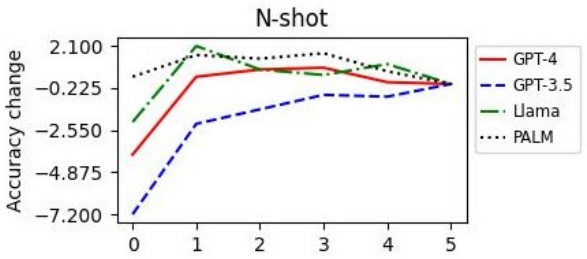

Figure 2: LLM prediction accuracy change from the baseline experiments, for N-shot experiments.

4.4. Error Recovery and Explanation

Besides the ability to recognize a correct answer, recognizing an error is another test of medical assessment ability. In these experiments, we supply both right and wrong answers (annotated by a medical doctor) to a model and ask for explanations on why there is an error. The explanations are then manually compared to those provided by the medical doctor annotator. We measure: (a) percent accuracy on providing reasonable explanations for errors, as assessed by an annotator compared to doctor explanation annotation, and (b) the percentage of questions for which models are able to recognize there is no errors are present even if asked to give an error explanation.

This provides a quantification of the possible ability of LLM models to "recover" from errors.

5. Results

5.1. Overall results

Table 3 outlines the summary of results from our general accuracy and consistency experiments. We highlight several findings here:

- **Most LLMs are relatively stable across superficial differences:** Most LLMs have relatively low standard deviations across repeated runs and temperature experiments. The exception was PALM. We further plot relative differences for LLMs across temperatures (Figure 1), revealing significantly different behavior for PALM.

- **Aggregated N-shot experiments reveal slightly higher standard deviation and clear differences related to number of examples:** Of LLM optimization settings, n-shot experiments showed higher deviations than the next-highest MC reorder experiments. To further investigate, we plotted performances across different N-shot experiments for LLM models (Figure 2). This revealed sizeable performance differences between 0-5 shots as well as different rates of change across models. In general, the performance improvement is large when adding the first few examples, and gradually diminishes until a certain threshold number of examples. We believe that the first increase comes from the LLMs' in-context learning ability. Meanwhile, when there are enough in-context examples, adding additional examples does not necessarily bring about more information. This observation is consistent with previous work(Brown et al., 2020a).

- **Number of options severely degrades model performance:** Comparing the mean accuracy on the 5-option dataset with the 4-option dataset, we observe a performance drop in accuracy ranging from 3.1 to 7.4 across models (example in Appendix 8.2).

LLM	Baseline		Temp		N-shot		MC reorder		4-option		Binarization		Open-ended		Error Recovery & Explanation		
	score	std	score	std	score	std	score	std	score	cons	score	cons	score	cons	HumAgree	recovery	explain
GPT-4	**80.2**	0.5	**80.5**	0.5	**80.7**	0.8	**80.6**	1.0	**83.6**	89.6	**43.4**	43.8	**54.0**	50.0	**82.4**	25.1	**70.0**
GPT-3.5	64.1	0.1	63.5	0.6	61.9	2.5	63.3	1.3	67.2	78.1	14.2	15.8	52.0	48.0	71.4	**33.5**	68.0
Llama	41.5	0.1	41.9	0.5	42.1	1.4	41.5	1.1	48.9	70.5	6.0	6.0	20.0	22.0	74.1	9.9	34.0
PALM	40.3	0.1	45.9	2.6	41.1	0.5	40.6	0.7	46.3	59.6	1.5	1.2	30.0	18.0	77.0	19.7	34.0
BioMedLM	44.6	0.8	-	-	-	-	-	-	50.1	66.7	-	-	-	-	-	-	-
Dragon	41.6	0.9	-	-	-	-	-	-	44.5	59.1	-	-	-	-	-	-	-

Table 3: Summary of LLM Accuracy, Consistency, and Error Experiments. (All scores are measured in terms of accuracy unless otherwise specified. std=standard deviation among those accuracy measures, cons=consistency measured through percent agreement with baseline), HumAgree=human agreement with error labels (error overlap), recovery=detect of no errors in correct text, explain=give reasonable explanation of errors in error text. Baseline is 5-option with 5-shot examples, CoT reasoning

- **GPT 4 shows strongest accuracy performance across reasonable consistency across different metrics:** Though in repeated runs at a higher standard deviation of 0.5 than GPT3.5, Llama, and PALM, GPT-4 gives across-the-board higher performance across different types of perturbations. When changing to binarized and open-ended variations, the percentage overlap as related to the original is also the highest.

- **Semantic differences lead to large differences in LLM performance outcomes:** As evidenced by large score differences, especially when changing to open-ended and binarization alternatives, certain models may not give consistent answers if the same query is adapted in a different way. These seem to affect both model ranking as well as absolute scores.

- **LLMs cannot tell if the right answer is given as wrong most of the time; however can give reasonable explanations of errors if truly wrong:** We applied a rule-based keyword algorithm to detect the LLMs' claim on the answer's correctness. We observed that in more than 66% of correct text for which we prompted models to explain the alleged error, models still generated a reason for an imaginary error. That said, we see some moderately high percentage of models with explanations consistent with humans (e.g., 68% for GPT3.5).

5.2. Studying Instability vs Inaccuracy

Consistency and accuracy are not necessarily correlated. However, understanding whether or not consistent answers may be more associated with correctness per model is useful information.

In this analysis, we first categorize questions into varying degrees of consistencies based on the number of values ever given per question (e.g., question id 1, which gives one answer value in all five runs is more consistent than question id 2 which as (b) or (c) chosen during the five runs). To understand if accuracies are different across consistency behaviors, we measure accuracy for all runs related across the questions in a category.

Thus, as shown in Table 4, in 6 temperature runs, we observe GPT-4 give one answer in 78.6% of questions – however within those 90.1% is correct. When at least two selected options occur across all runs (for 15.4% of questions), the accuracy among this population drops to 48.6% (a change of 41.4%).

These results confirm that for the most part in all LLMs, higher consistency is related to higher accuracies; though this behavior is less pronounced in Llama and PALM. For example, in Llama temperature experiments the accuracy from 1-observed-option questions (the most consistent) drops from 46.8% to 29.0% for 2-observed-options, with a difference of 24.3% smaller than GPT-4 in the same experiments. Thus, in practice, if changing your query across several temperatures or n-shot cases, leads to the same answer in the multiple choice setting, it is likely the result is accurate. Moreover, interestingly the profiles of consistency will change among the different experiments - where N-shot differences are likely to lead to larger consistency differences (e.g., the percentage of GPT-4 consistent questions goes to 77.7%).

5.3. Qualitative Analysis with Open-ended Variation Questions

We experimented with both automatic and human evaluations for the open-ended question variation. However, while the gold answer is usually a concise phrase, with an average of 4 words, the LLM outputs were 10-45 times the average length of the gold answer. Thus, we found traditional automatic NLG metrics unreliable for this task.

Qualitatively in this set, we found that the LLMs struggled with questions related to describing mechanism-related expectations. In one example, a question was "What explains the pathophysiology underlying this patient's disorder?" where the expected answer was "Iso-

	# options ever chosen and its percentage under each consistency testing experiment																							
	Repeated Runs (5 runs)						Temperature (6 runs)						MC reorder (4 runs)						N-shot (5 runs)					
# options	1		2		≥3		1		2		≥3		1		2		≥3		1		2		≥3	
	Acc	%	Acc	%	Acc	%	Acc	%	Acc	%	Acc	%	1	%	Acc	%	Acc	%	Acc	%	Acc	%	Acc	%
GPT-4	85.2	89.6	39.2	9.4	21.7	0.9	90.1	78.6	48.6	17.4	31.4	4.1	87.6	84.4	43.6	14.3	25.0	1.3	89.8	77.7	52.2	18.9	30.0	3.5
GPT-3.5	65.2	97.3	25.3	2.7	N/A	0.0	79.1	59.5	45.0	29.7	28.3	10.8	74.8	67.0	40.2	28.8	28.9	4.2	79.7	52.7	46.4	35.1	29.3	12.2
Llama	41.5	99.5	40.0	0.5	N/A	0.0	46.8	74.5	29.0	20.4	21.3	5.1	53.5	47.4	31.9	42.7	25.9	9.8	49.3	61.2	32.8	29.6	24.1	9.2
PALM	40.3	98.6	36.7	1.4	N/A	0.0	56.5	52.0	38.5	33.9	24.7	14.1	49.9	53.1	32.2	37.6	24.0	9.3	46.8	63.2	32.7	32.3	25.9	4.6

Table 4: Observations of each LLM's answer consistency accuracy and frequency under different experimental settings. For example, with simple repeated runs, GPT-4 will give exactly one answer for 89.6% of questions at 85.2% accuracy, but in 9.4% of questions will have at least two different answers at a 39.2% accuracy. N/A refers to the situation where LLM has 0 questions with such number options.

lated gonadotropin-releasing hormone (GnRH) deficiency", and the GPT 3.5 answer was "Kallmann syndrome". Though the latter is associated with the former in some cases (which may be true in this clinical scenario), the idea of the question is for the test-taker to understand what type of biological mechanisms may be interrupted rather than jumping to a diagnosis. In another example, a question was "The anatomic structure that was most likely injured in this patient has what characteristics?" with an expected answer of "Runs posteriorly from the medial femoral condyle", and the answers of "ACL" or "PCL" from different models.

This suggests at least one of two ideas: (a) as much as humans do, instead of critical thinking on expected findings and perhaps updating source attributions and hypotheses, LLMs memorize a frequent conclusion, which is a problem if there are incomplete or inaccurate assumptions at play; and (b) multiple choice wording and expectations of certain types of answers are specific to a limited set of scenarios (e.g., standardized testing), but may not be as prevalent (and therefore less learnable) or universally-understood in all contexts.

5.4. Further Study in the Binarized Variation

To be comparable to the baseline version, we changed the prompt to be a four-shot learning (1 right and 3 wrong options from one question), both with and without chain-of-thought (CoT) reasoning taking an example from Wei et al. (2022).

LLM	%Parsed	%Valid	Acc
GPT-4	98.6	47.8	43.4
GPT-3.5	76.0	22.2	14.2
Llama	47.8	11.5	6.0
PALM	29.6	3.8	1.5

Table 5: Binarization experiment: percentage among all questions in the test set.

Table 5 presents the results of the accuracies in the converted binarized questions. Because every question in the original dataset was converted to five binary instances (True/False for each available option) - to make scores comparable, we convert the binary instances back by collecting all True/False decisions per the same original question. Only questions with only predicted true out of the 5 original options were considered valid.

While GPT-4 results were parsable (e.g., a rule-based regex was able to collect True/False answers) at greater than 90%, this wasn't the case for GPT-3.5, Llama, and, PALM variants. Most un-parsable results are because the LLM failed to understand the task as True/False classification, but explains why the answer is correct. Noticeably, the percentage of valid questions was modest across all model groups, with the exception of GPT-4. We observed a great performance drop from parsed to valid, especially for Llama and PALM, due to the LLM predicting true for multiple options. Finally, we observe significant drops in performance, with the best accuracies at 43.4%, whereas the highest open-ended variant scores were at 58%. This could be attributable to the stricter requirement to be able to answer five classification instances correctly instead of just one.

5.5. LLM and Human Error Agreement

Figure 3 details error question overlap agreements according to equation (1). Figure 4 adds the additional constraint of requiring both question ID and answer label to overlap. We additionally provide a breakdown of the agreement of LLM with different numbers of annotators in Table 6. Table 7 provides both automated and human evaluation scores when models are tasked with explaining errors (even when none exist). We highlight several key findings:

- **When wrong, models differ about their preferred answer:** The drop between Figure 3 and 4 show that even for the same questions for which two models commit the same error, the actual value of the error is often different. For example between GPT-4 and GPT-3, there is a 78.9% question error overlap if taking the GPT-4 number of errors as denominator (Figure 3). When factoring the answer label, the overlap score drops to 50%(Figure 4).

- **Models will pick an answer consistent with a human, some with higher percentages:** Except for Dragon, at least 70% of all other LLMs' wrong predictions agree with at least one human annotator (Figure 6). This suggests that the options confusing to LLMs can confuse humans as well. Among all the LLMs, GPT-4's wrong predictions agree with most human annotators, as well as the doctor, as seen in the last column of Figure 4.

- **Automated NLG evaluation metrics give rankings comparable to humans:** Although the measures of Rouge and BLEURT show modest performances, in general, the rankings based on these metrics and BERTScore are consistent with humans. That said more study needs to go into developing better metrics for this task as well as multiple reference gold standards.

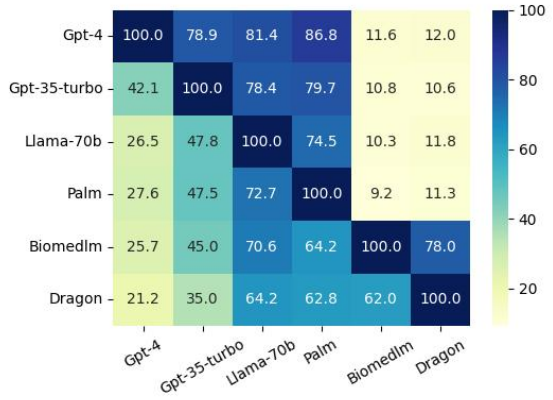

Figure 3: Pairwise error question overlap (%) using Equation (1). (Denominator of incorrect answers are indexed by rows.)

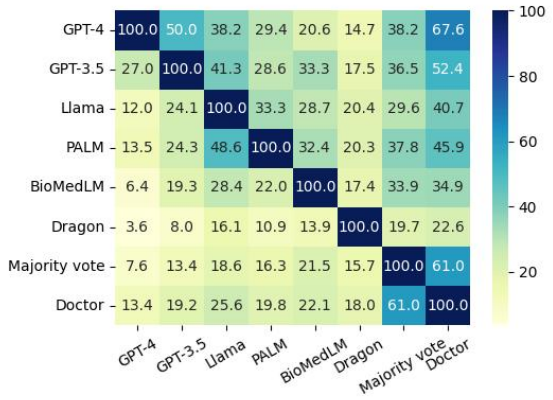

Figure 4: Pairwise error overlap agreement (%). Same as Figure 4, however the same questions and the same incorrect answer must agree. (Denominator of wrong answers are indexed by rows.)

6. Discussion and Conclusions

In our results, we quantified different accuracy and consistency profiles across LLMs. GPT-4 remains empirically the most robust when measuring across changing answer options, superficial differences, semantic, consistency across variations, and error overlap with human annotation. The drops in performances between 5-options to 4-options, and to open-ended/binarized variations, suggest that the GPT family works well if given a closed set of possibilities; however, if requiring applications toward open-ended questions or problems with multiple good solutions, there may be challenges. This implies that one important research direction is the generation of a small closed set of optimal choices to reduce errors.

Our analysis measuring consistency versus accuracy revealed that, across all LLMs, consistency and accuracy are well-correlated. Thus, if certain confidence thresholds are required in a system, it is recommended to do some simple surface form variations to identify possible uncertain answers. As models have great overlap in incorrect answers but different answers, ensembling efforts – while useful in understanding which questions may be incorrect – may not lead to additional gains for correcting wrong assessments. That said performance is still less than ideal for medical applications and would still require additional review by experts.

LLMs show remarkable performances considering their generalized training regiment that is not tuned for specific tasks. However, the differences in scores when given larger semantic variations in querying for the same knowledge as well as their inability to recognize false presuppositions (e.g., tasked with finding an error in a correct medical narrative) suggests unexpected tasks and unseen set-ups are still challenging even when they are related to very visible and well-studied datasets. One future direction in dataset creation is to construct benchmarks that query the same knowledge but from different aspects; or add different related challenges built upon the same base tasks. Such a direction may give a better overall sense of a model's robustness for medical assessment than isolated publicly available task performances.

LLM	Overlap with # medical annotators				Overlap with Doctor
	≥ 4	≥ 3	≥ 2	≥ 1	
GPT-4	14.7	41.2	58.8	82.4	67.6
GPT-3.5	12.7	36.5	44.4	71.4	52.4
Llama	7.4	29.6	50.9	74.1	40.7
PALM	13.5	36.5	50.0	77.0	45.9
BioMedLM	11.0	22.0	37.6	71.6	34.9
Dragon	3.6	14.6	26.3	56.9	22.6

Table 6: Percentage of wrong predictions' overlap with number of human annotators' approximated errors.

LLM	Rouge			BERT Score	BLEURT	Human Eval (%)		
	-1	-2	-L			contradicts	non-overlap	consistant
GPT-4	18.4	5.9	12.2	84.5	48.2	**10.0**	20.0	70.0
+ 4shot	**21.5**	**6.9**	**14.7**	**85.2**	45.5	12.0	12.0	**76.0**
GPT-3.5	14.4	4.5	10	83.4	**48.8**	14.0	18.0	68.0
+ 4shot	19.3	5.7	13	84.6	46.0	34.0	24.0	42.0
Llama	10.5	3.1	7.6	82	51.4	22.0	44.0	34.0
+ 4shot	14.9	4.2	10.5	83.6	47.3	50.0	26.0	24.0
PALM	11	3	7.9	82.3	50.3	26.0	40.0	34.0
+ 4shot	12.4	3.3	8.9	82.8	46.6	26.0	40.0	34.0

Table 7: Comparison between the medical doctor's explanation with the LLMs' on why an option is wrong. The human evaluation is the percentage of LLMs' responses in each sub-category, among the 50 randomly selected questions. (Baseline is 0-shot)

7. Ethics and Broader Impact Statement

No use of patient information was used in the creation of this dataset. All labeled data was created by hired domain expert workers, fairly compensated as consistent with their state.

The medical assessment testing here was developed in by United States content creators and therefore may not represent all the possible schools of thought on medical knowledge. Content creators themselves may have differing opinions on correct assessments which is not always reflected in the questions datasets.

We provide an analysis related to specific models however, these trends may not hold for other models or for future versions of current models. The analysis here includes some but not all possible ways of testing accuracies, errors, and consistency in the medical domain.

Bibliography

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zahraa Al-Dujaili, Sarah Omari, Jey Pillai, and Achraf Al Faraj. 2023. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. pages S1551–7411(23)00365–0.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

American Medical Association. 2022. What is the USMLE®?

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Institute of Medicine (US) Committee on Quality of Health Care in America. 2000. *To Err is Human: Building a Safer Health System*. National Academies Press (US).

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696. International Committee on Computational Linguistics.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of ChatGPT.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Douglas Johnson, Rachel Goodman, J. Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, Elizabeth Scoville, Tyler Reese, Debra Friedman, Julie Bastarache, Yuri van der Heijden, Jordan Wright, Nicholas Carter, Matthew Alexander, Jennifer Choe, Cody Chastain, John Zic, Sara Horst, Isik Turker, Rajiv Agarwal, Evan Osmundson, Kamran Idrees, Colleen Kiernan, Chandrasekhar Padmanabhan, Christina Bailey, Cameron Schlegel, Lola Chambless, Mike Gibson, Travis Osterman, and Lee Wheless. 2023. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the chat-GPT model. pages rs.3.rs–2566942.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A chinese multi-choice biomedical

question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874. Association for Computational Linguistics.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions?

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Tristan Price, Nick Lynn, Lee Coombes, Martin Roberts, Tom Gale, Sam Regan de Bere, and Julian Archer. 2018. The international landscape of medical licensing examinations: A typology derived from a systematic review. 7(9):782–790.

Marc Cicero Schubert, Wolfgang Wick, and Varun Venkataramani. 2023. Evaluating the performance of large language models on a neurology board-style examination. ISSN: 2329-2598 Pages: 2023.07.13.23292598.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. 620(7972):172–180. Number: 7972 Publisher: Nature Publishing Group.

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

USMLE. 2023. About the USMLE | USMLE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*.

8. Appendix

8.1. Open-ended Question Generation

A free-text, open-ended version of the MedQA dataset is created to further challenge generative LLMs' consistency. This version is applied for the binarization (Section. 5.4), open-ended QA (Section. 5.3), and error explanation (Section. 5.5). This was done by a rule-based algorithm that converts the question from 'which' format to 'what', i.e. 'which of the following is the best treatment for ... ' to 'what is the best treatment for ...'. A human review is conducted to improve the questions after this rule-based conversion, as well as to remove those questions that do not make sense in the open-ended format, i.e. 'What medication is NOT likely to be the cause?'

The open-ended questions can be found in our GitHub page.

8.2. Error analysis

As shown in Table 8, including one additional MC option can confuse the model. Though the MC reorder changes the GPT-4 CoT reasoning, it does not impact prediction greatly.

Table 9 demonstrates an example of open-ended MedQA question without offering the MC options. All models make plausible reasoning, but Llama 2 and PALM 2 fail to match the correct answer.

Table 10 shows model explanations on why a certain option is wrong, compared with the physician's response, with 4-shot examples. GPT-4 and PALM successfully provide explanations. GPT-3.5 does not agree with the instruction, while Llama 2 fails to follow the instructions and answers as a traditional MC question.

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 72-year-old woman is brought to the physician by her daughter because of a 6-month history of worsening short-term memory deficits and social withdrawal. Treatment with galantamine is initiated. Two weeks later, the patient develops vomiting, mild crampy abdominal pain, and watery, nonbloody diarrhea. Which of the following is the most appropriate pharmacotherapy?

Number of options severely degrades model performance

4-option multiple choices: (A) Bethanechol (B) Metoclopramide (C) Atropine (D) Oxybutynin
GPT-4 response:
Explanation: We refer to Wikipedia articles on medicine for help. The symptoms suggest that the patient has developed cholinergic side effects from the galantamine. **Atropine** is an anticholinergic drug that can be used to treat these side effects.
Answer: **(C)**
5-option multiple choices: (A) Bethanechol (B) Metoclopramide (C) Loperamide (D) Atropine (E) Oxybutynin
GPT-4 response:
Explanation: The patient's symptoms are likely side effects of the galantamine treatment, which is a cholinesterase used to treat Alzheimer's disease. The side effects are due to increased cholinergic activity and can include gastrointestinal symptoms like those the patient is experiencing. Loperamide is an opioid that decreases gastrointestinal motility and can help manage these symptoms.
Answer: **(C) Loperamide**

Random shuffling while keeping the same answer ID

5-option multiple choices: (C) Loperamide (A) Bethanechol (E) Oxybutynin (B) Metoclopramide (D) Atropine
GPT-4 response:
Explanation: We refer to Wikipedia articles on medicine for help. The symptoms suggest that the patient has diarrhea. Loperamide is an opioid that decreases the activity of the myenteric plexus, which decreases the tone of the longitudinal and circular smooth muscles of the intestinal wall. This increases the amount of time substances stay in the intestine, allowing for more water to be absorbed out of the fecal matter.
Answer: **(C) Loperamide**

Random shuffling with answer ID's renamed such they are alphabetical

5-option multiple choices: (A) Metoclopramide (B) Oxybutynin (C) Bethanechol (D) Loperamide (E) Atropine
GPT-4 response:
Explanation: We refer to Wikipedia articles on medicine for help. The symptoms suggest that the patient has diarrhea. Loperamide is an opioid that decreases the activity of the myenteric plexus which decreases the tone of the longitudinal and circular smooth muscles of the intestinal wall. This increases the amount of time substances stay in the intestine, allowing for more water to be absorbed out of the fecal matter
Answer: **(D) Loperamide**

Randomly shuffling the order of the 5-shot examples, with 5-option MC unchanged

5-option multiple choices: (A) Bethanechol (B) Metoclopramide (C) Loperamide (D) Atropine (E) Oxybutynin
GPT-4 response:
Explanation: We refer to Wikipedia articles on medicine for help. The patient is experiencing side effects from the galantamine, which is a cholinesterase inhibitor used to treat Alzheimer's disease. The side effects are due to increased cholinergic and include nausea, vomiting, and diarrhea. **Loperamide** is an opioid that decreases gastrointestinal motility and can be used to treat diarrhea.
Answer: **(C)**

Table 8: Example GPT-4 prediction on the same question, under 4-option and 5-option multiple choices, and different MC reorders.

Instructions: Answer the question concisely based on your medical knowledge. Solve them in a step-by-step fashion.
Question: A 16-year-old female presents to her physician's office after noticing a round lump in her left breast 2 months ago. She reports that the lump seemed to enlarge and became tender just preceding her last 2 menses. It is otherwise painless, and the patient denies any discharge or skin changes. She has no past medical history but her grandmother, age 72, was just diagnosed with invasive ductal carcinoma of the breast. The patient is an avid softball player at her high school and denies alcohol, smoking, or illicit drug use. On exam, the breasts appear symmetric and normal. A 3-cm round, mobile mass is palpated in the upper outer quadrant of the left breast. There is slight tenderness to deep palpation of the mass. There is no axillary lymphadenopathy on either side. What is the most likely outcome of this patient's condition?

Options: (A) This mass will likely require excision. (B) This mass will decrease in size if the patient starts oral contraceptives. (C) This mass slightly increases this patient's risk of breast cancer in the future. (D) This mass will most likely decrease in size or disappear over time. (E) If this mass grows rapidly to greater than 5 cm, radiation and chemotherapy are indicated.
Answer: (D) This mass will most likely decrease in size or disappear over time

GPT-4 response:
The patient's symptoms and age suggest that she likely has a fibroadenoma, which is a benign breast tumor common in young women. These tumors can fluctuate in size with menstrual cycles due to hormonal influences. The most likely outcome for this condition is that the fibroadenoma will remain stable or possibly regress over time. It does not increase the risk of breast cancer. However, regular monitoring is usually recommended to ensure no significant changes occur.
Human label: The open-ended response is consistent with the correct option.

GPT-3.5 response:
Explanation: We refer to Wikipedia articles on medicine for help. The patient's age, the presence of a round, mobile mass, and the absence of axillary lymphadenopathy suggest a benign breast condition. Given the tenderness that worsens before her menses, the most likely outcome is fibrocystic changes or fibroadenoma.
Answer: Benign breast condition (fibrocystic changes or fibroadenoma).
Human label: The open-ended response is consistent with the correct option.

PALM 2 response:
The most likely outcome of this patient's condition is that she has a fibroadenoma, which is a benign (noncancerous) tumor of the breast. Fibroadenomas are the most common type of breast tumor in women under 30 years old. They are usually round, smooth, and rubbery, and they can range in size from a few millimeters to several centimeters. Fibroadenomas are often caused by hormonal changes, such as those that occur during puberty or pregnancy. They are usually not painful, but they can become tender before or during menstruation. Fibroadenomas are usually diagnosed by a physical examination and confirmed by a mammogram or ultrasound. If a fibroadenoma is small and not causing any problems, it may be simply monitored over time. If it is larger or causing symptoms, it may be removed surgically. The removal of a fibroadenoma is a simple procedure that is usually done on an outpatient basis. The patient will usually be able to go home the same day as the surgery. The recovery time is usually short, and the patient can usually resume normal activities within a few days.
The patient's grandmother's diagnosis of invasive ductal carcinoma of the breast is concerning, but it is important to remember that breast cancer is not hereditary. Only about 5-10% of breast cancers are caused by inherited genetic mutations. The patient's family history does increase her risk of developing breast cancer, but she is still at a relatively low risk. She should continue to have regular mammograms and breast exams to screen for breast cancer.
Human label: The open-ended response is wrong but can be mapped to one of options.

Llama 2 response:
Based on the information provided, the most likely outcome of this patient's condition is a benign breast tumor. The patient's age, lack of significant medical history, and tender swelling in the breast suggest a high likelihood of a benign breast tumor, such as a fibroadenoma. The fact that the patient's grandmother was recently diagnosed with invasive ductal carcinoma of the breast may increase the patient's risk for breast cancer, but it is still unlikely given the patient's young age.
The patient's symptoms, such as the round lump in her breast that seemed to enlarge and became tender just preceding her last 2 menses, are consistent with a benign breast tumor. The absence of any discharge or skin changes also supports this diagnosis. The patient's physical examination, which revealed a 3-cm round, mobile mass in the upper outer quadrant of the left breast, with slight tenderness to deep palpation, further supports the diagnosis of a benign breast tumor. The absence of axillary lymphadenopathy on either side also suggests a benign condition.
Therefore, based on the information provided, it is likely that the patient's condition is a benign breast tumor, such as a fibroadenoma, and further diagnostic testing, such as a breast ultrasound or biopsy, may be necessary to confirm the diagnosis.
Human label: The open-ended response is wrong and cannot be mapped to any options

Table 9: Model prediction on the open-ended, no MC version of the MedQA question, with CoT reasoning.

Instructions: Explain why the answer is wrong, based on your medical knowledge. Solve them in a step-by-step fashion.
Question: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions?
Options: (A) Inhibition of thymidine synthesis (B) Inhibition of proteasome (C) Hyperstabilization of microtubules (D) Generation of free radicals (E) Cross-linking of DNA
Wrong answer: (D) Generation of free radicals.

Physician' explanation:
Free radical generation is not typically associated with the ototoxicity seen in drugs used to treat transitional cell carcinoma of the bladder, like cisplatin, which primarily act via DNA cross-linking.

GPT-4:
The given answer is incorrect. The drug that likely caused this patient's symptoms is cisplatin, a chemotherapy agent commonly used in the treatment of transitional cell carcinoma of the bladder. Cisplatin is known to cause ototoxicity, which can manifest as tinnitus (ringing in the ears) and sensorineural hearing loss.
The beneficial effect of cisplatin is not due to the generation of free radicals. Instead, cisplatin works by forming covalent bonds with the purine bases on the DNA, which leads to the formation of DNA adducts. These adducts distort the DNA helix and prevent DNA replication and transcription, leading to cell death.
Therefore, the beneficial effect of cisplatin is due to its ability to interfere with DNA replication and transcription, not the generation of free radicals.
Human label: the model's explanation is **consistent with** the physician's explanation

GPT-3.5:
The answer is correct. The expected beneficial effect of the drug that caused the patient's symptoms is most likely due to the generation of free radicals. Neoadjuvant chemotherapy drugs, such as cisplatin, commonly used in the treatment of transitional cell carcinoma of the bladder, can cause ototoxicity, which is damage to the inner ear resulting in hearing loss. This ototoxicity is thought to be caused by the generation of free radicals, which can damage the delicate structures of the inner ear and lead to sensorineural hearing loss.
Human label: the model's explanation **contradicts** the physician's explanation

PALM 2:
The answer is wrong because generation of free radicals is a harmful effect of chemotherapy. Free radicals are unstable molecules that can damage cells, including healthy cells. This damage can lead to a variety of side effects, including hearing loss.
The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to its ability to kill cancer cells. Chemotherapy drugs work by damaging the DNA of cancer cells, which prevents them from dividing and growing. This damage can also kill healthy cells, which is why chemotherapy often causes side effects.
In this case, the patient's hearing loss is likely due to damage to the hair cells in the inner ear. Hair cells are responsible for converting sound waves into electrical signals that are sent to the brain. When hair cells are damaged, they can no longer function properly, which can lead to hearing loss.
There is no cure for hearing loss caused by chemotherapy, but there are treatments that can help to improve hearing. These treatments include hearing aids, cochlear implants, and assistive listening devices.
Human label: the model's explanation **does not overlap** with the physician's explanation

Llama 2:
We refer to the Wikipedia article on Cisplatin, a chemotherapy drug commonly used to treat bladder cancer, for help.
Cisplatin can cause ototoxicity, which is damage to the inner ear and auditory nerve, leading to hearing loss. The generation of free radicals is one of the proposed mechanisms for cisplatin ototoxicity. Free radicals are highly reactive molecules that can damage cellular components, including the delicate structures of the inner ear.
Therefore, the answer is (D) Generation of free radicals.
Human label: the model's explanation **contradicts** the physician's explanation

Table 10: Model explanation on why a certain option is wrong, compared with the physician's explanations, under 4-shot examples. The 4-shot examples come from the 4 options from one MedQA question, where three examples are from the wrong options and start with 'The answer is wrong because', and one example is from the right option and starts with 'The answer is right because'.