Annotation Guidelines

Project title: Error-type annotation of GPT-4 responses to USMLE questions

Aim of the study:

Your first task is to understand the reason for the incorrect GPT-4 response and assign the response to one or more of the seven error categories. In the annotation guidelines document, we provide a formal definition (fixed rules) to represent an error category and provide one or more positive examples for each category along with an explanation. You will be asked to annotate 20 data points into the error categories provided.

Task description – Span Labeling

You are given a USMLE (United States Medical Licensing Examination) question with options. This question is then asked to GPT-4, and its response (correct option + explanation) is recorded. The following cases are the cases where the GPT-4 answer (one of five options) does not match the reference (ground-truth / correct) answer.

(i) **USMLE question given to GPT-4:** USMLE question with options + "*Please explain your answer and why the other options are wrong*" as a prompt or instruction.

(ii) **Correct answer:** Answer provided by USMLE official source.

(iii) **GPT-4 Response:** This includes a detailed response explaining the chosen option and the rationale behind considering it as the correct answer. Following this, brief justifications (1 or 2 sentences in length) will be provided to counter all the remaining options.

Instruction for participants (annotators)

1. Please go through the annotation guidelines thoroughly before attempting the annotation study.

2. Please watch the video tutorial explaining how to interact with the Potato annotation platform at: https://youtu.be/n6vDBIfDSQk.

This will help you (the annotator) understand how to interact with the Potato annotation platform and specifically perform the span labeling exercise to label the spans of text within the GPT-4 response into different error categories.

3. You do not need to download any separate software to complete the study. You will only require a browser (preferably, Google Chrome) and a stable Internet connection.

4. You are required to **annotate 20 responses** generated by GPT-4 to USMLE questions. You will be allotted a total of **100 minutes** to complete the annotation of the 20 instances and to fill out the two forms at the end. There is no specific time limit for annotating each instance.
The two forms will inquire about the following: (i) feedback regarding the annotation exercise and interaction with the platform, (ii) participant demographics including academic background, age, and country of birth, and (iii) medical background and experience with the USMLE examination.

5. At the end of the study, you will be asked to click on a link containing the completion code that will redirect you to the Prolific platform and inform us that you have completed the study.

6. Every submission will undergo manual review, and we will reject annotations that appear random or insincere. Specifically, we have incorporated a few validation questions within the set of 20 data points.

7. You may expect to receive the payment within one to two weeks of completing the study.

Error-type annotation (**7 + 2 = 9 classes**):

1. Non-medical factual error
2. Sticking with the wrong diagnosis
3. Incorrect or vague conclusion
4. Ignore missing information
5. Incorrect understanding of the task
6. Hallucination of information
7. Unsupported medical claim

We also provide **two non-error categories** which the annotators may utilize if the identified error cannot be assigned to our seven error categories:

8. Reasonable response by GPT-4: You can select this option if you find the explanation provided by GPT-4 reasonable and the explanation does not have incorrect facts or flawed reasoning. Since GPT-4 selects the incorrect option, the answer by GPT-4 is also a potential or more medically reasonable answer.

9. Cannot pick any category: You can select this option when you find certain flaws in the facts or reasoning of GPT-4 and are not fully convinced with the GPT-4 response. However, you are also not able to assign the GPT-4 response to any of the error categories.

For these non-error categories, such as "Reasonable response by GPT-4" and "Cannot pick any category", please select any sentence in the GPT-4 response. However, you cannot select any other error category after selecting a non-error category.

Background Reading Material

United States Medical Licensing Examination (USMLE)

USMLE stands for United States Medical Licensing Examination. If you are already familiar with USMLE questions, you can skip this section. Otherwise, please browse through the official USMLE website (https://www.usmle.org/). Specifically, please make yourself familiar with the following reading materials:

https://www.usmle.org/sites/default/files/2022-01/USMLE_Content_Outline_0.pdf

https://www.usmle.org/sites/default/files/2023-08/2024bulletin.pdf.pdf

Additional reading material to understand the emerging field of evaluating large language models on medical benchmarks

Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2

Potato – the annotation platform

Please read the documentation at https://potato-annotation.readthedocs.io/

Detailed Definition of Error Categories

If you observe an error in a sentence that matches one of the 7 categories or more, please highlight the sentence(s) with the corresponding class (error type).

Error type 1: Non-medical Factual Error

A statement in the explanations of the answer is factually wrong (e.g. 1+1 = 3). These involve non-medical facts and includes computational errors, terminology errors, and can be reliably annotated by a non-medical expert. This does not require verification through any external resource or inline reference. Based on our experience, this is the most infrequent form of error category.

Error type 2: Sticking with Wrong Diagnosis

The model confidently states its (wrong) answer early in the answer. Its reasoning when explaining this choice is sound and based in fact with regards to the diagnosis, but the model still makes the wrong choice instead of correcting it based on its explanations.

Positive Example:

USMLE question: A previously healthy 10-year-old boy is brought to the emergency room by his mother 5 hours after the onset of abdominal pain and nausea. Over the past 2 weeks, he has also had progressive abdominal pain and a 4-kg (8.8-lb) weight loss. The mother reports that her son has been drinking more water than usual during this period. Last week he wet his bed three times despite being completely toilet-trained since 3 years of age. His temperature is 37.8°C (100°F), pulse is 128/min, respirations are 35/min, and blood pressure is 95/55 mm Hg. He appears lethargic. Physical examination shows deep and labored breathing and dry mucous membranes. The abdomen is soft, and there is diffuse tenderness to palpation with no guarding or rebound. Serum laboratory studies show:
Na+ 133 mEq/L
K+ 5.9 mEq/L
Cl- 95 mEq/L
HCO3- 13 mEq/L
Urea nitrogen 25 mg/dL
Creatinine 1.0 mg/dL
Urine dipstick is positive for ketones and glucose. Further evaluation is most likely to reveal which of the following?"

Option: {'A': 'Decreased total body potassium', 'B': 'Increased total body sodium', 'C': 'Increased arterial pCO2', 'D': 'Hypervolemia', 'E': 'Serum glucose concentration > 600 mg/dL'}
Please explain your answer and why the other options are wrong

Correct answer: A

GPT-4 Response:
The boy's polyuria (as evidenced by bedwetting), polydipsia (increased water drinking), and unexplained weight loss, along with the presence of glucose and ketones in the urine, suggests new-onset type 1 diabetes mellitus (DM1). Prolonged hyperglycemia leads to osmotic diuresis, dehydration, and eventually diabetic ketoacidosis (DKA) if not treated. DKA is characterized by hyperglycemia (serum glucose >250 mg/dL), ketosis, and metabolic acidosis. The clinical features of DKA include polyuria, polydipsia, nausea/vomiting, abdominal pain, Kussmaul respirations (ie, rapid, deep breathing), and lethargy to coma. The diagnosis of DKA is confirmed by arterial blood gas (ABG) analysis showing a low pH (acidemia) and low bicarbonate concentration due to the metabolic acidosis.

Option E, 'Serum glucose concentration > 600 mg/dL', is the correct answer.

Options A, B, D could be seen in different conditions but not in DKA.
Option A: Total body potassium is usually high in DKA as a result of the acidosis, not decreased.
Option B: Total body sodium can be low, normal, or slightly high during DKA, but not significantly increased.
Option D: Hypervolemia is not a feature of DKA which usually present with dehydration due to osmotic diuresis.

Option C, 'Increased arterial pCO2', is incorrect because the body would be compensating for the metabolic acidosis by hyperventilating (Kussmaul respirations) in an attempt to lower the CO2 level, not increase it. Hence the pCO2 would be expected to be low, not high.

Explanation behind assignment of error category: GPT-4 identifies factually correct disease symptoms but introduces the new (noisy) idea of prolonged hyperglycemia, i.e., if the symptoms are allowed to persist for a long time, the patient may eventually develop diabetic ketoacidosis (DKA) in future. However, the reasoning provided by GPT-4 is entirely (overly) centered around DKA, which may potentially be a wrong diagnosis because it is based on the hypothesis that the symptoms go untreated for some time. Therefore, we think that GPT-4 selected the wrong answer because initial disease identification is incorrect and there exists better alternatives that GPT-4 did not consider or entertain in later parts of the response.

Positive example:

USMLE question: A 30-year-old man is brought into the emergency room for complaints of acute onset chest pain and shortness of breath. He has a history of mental retardation and lives at home with his adoptive parents. His parents inform you that he has not seen a doctor since he was adopted as child and that he currently takes no medications. The patient's temperature is 99.1°F (37.3°C),pulse is 108/min, blood pressure is 125/70 mmHg, respirations are 25/min, and oxygen saturation is 92% on 2L nasal canula. Physical exam is notable for a tall, thin individual with high-arched feet and mild pectus excavatum. There is mild asymmetry in the lower extremities with discomfort to dorsiflexion of the larger leg. Lung auscultation reveals no abnormalities. What is the next step in the diagnosis of this patient's underlying disorder?
Option: {'A': 'Genetic testing', 'B': 'Chest radiograph', 'C': 'Angiogram', 'D': 'Electrocardiogram', 'E': 'Serum blood test'}
Please explain your answer and why the other options are wrong

Correct answer: E

GPT-4 response:
The picture provided here is suggestive of Marfan syndrome, a connective tissue disorder characterized by skeletal abnormalities (tall stature, high-arched feet, pectus excavatum), ocular abnormalities, and most importantly cardiovascular complications, most commonly aortic root dilatation and dissection. Patients may present with chest pain suggestive of acute aortic dissection, a life-threatening emergency that requires prompt diagnosis and treatment. The initial diagnostic test for suspected aortic dissection is a chest imaging study to visualize the aorta. Therefore, the correct answer is 'B': Chest radiograph.

Let's review why the other options are incorrect:

'A': Genetic testing - Although Marfan syndrome is a genetic disorder caused by a defect in the fibrillin-1 gene, genetic testing is not the most appropriate next step given the patient's acute presentation. It may be necessary for confirming the diagnosis of Marfan syndrome but is not the first step when an aortic dissection is suspected.

'C': Angiogram - While this procedure can provide a detailed view of the blood vessels and may identify an aortic dissection, it is invasive and not usually the first step in the diagnostic process.

'D': Electrocardiogram - An EKG is useful for diagnosing cardiac conditions such as myocardial infarction, arrhythmias, and other heart abnormalities but is not the most appropriate test for evaluating a suspected aortic dissection.

'E': Serum blood test - Blood tests can provide a lot of useful information and may be used as part of the evaluation of a patient with suspected Marfan syndrome, especially to rule out other conditions, but they would not provide direct evidence of an aortic dissection.

Therefore, given the patient's presentation and the likelihood of aortic dissection, a chest radiograph would be the most appropriate next step in this patient's care.

Explanation behind assignment of error category: Since GPT-4 suggested Marfan syndrome (most commonly aortic root dilatation and dissection) as the only underlying disorder, it refuted all the other options solely on that ground; it found Serum blood test (option E, the ground-truth correct answer) as useful but not most appropriate. We believe that GPT-4 chose the wrong option because its initial diagnosis of Marfan syndrome may be wrong. We selected this option because we found the associated statements to be factually correct (limitation: as we are from the computer science background, but our annotators will be from medical background, so it will not be a limitation then).

Error Type 3: Incorrect or Vague Conclusion

The explanations given by the model are factually correct and support the proposed answer, but the statements drawing a conclusion about the correctness of option(s) are not definitive. Instead, vague terms are used (e.g. usually, most of the time) and the model prioritizes only a few symptoms, leading to the wrong answer. Another scenario (Question id 82) can be that the difference in reasoning in the response between the GPT-4 chosen option and the ground-truth reference answer, is not clearly mentioned and sufficient explanation is not provided in this context.

Positive example:

"USMLE question: A 53-year-old man comes to the emergency department because of severe right-sided flank pain for 3 hours. The pain is colicky, radiates towards his right groin, and he describes it as 8/10 in intensity. He has vomited once. He has no history of similar episodes in the past. Last year, he was treated with naproxen for

swelling and pain of his right toe. He has a history of hypertension. He drinks one to two beers on the weekends. Current medications include amlodipine. He appears uncomfortable. His temperature is 37.1°C (99.3°F), pulse is 101/min, and blood pressure is 130/90 mm Hg. Examination shows a soft, nontender abdomen and right costovertebral angle tenderness. An upright x-ray of the abdomen shows no abnormalities. A CT scan of the abdomen and pelvis shows a 7-mm stone in the proximal ureter and grade I hydronephrosis on the right. Which of the following is most likely to be seen on urinalysis?
Option: {'A': 'Urinary pH: 7.3', 'B': 'Urinary pH: 4.7', 'C': 'Hexagon shaped crystals', 'D': 'Positive nitrites test', 'E': 'Largely positive urinary protein'}
Please explain your answer and why the other options are wrong"

Correct answer: B

GPT-4 Response: The correct answer is A: Urinary pH: 7.3

This patient's severe, colicky, right-sided flank pain that radiates to his right groin is characteristic of nephrolithiasis (kidney stones). His history of being treated with naproxen for swelling and pain in his right toe indicates gout, which often is associated with uric acid nephrolithiasis, as both conditions are due to overproduction or underexcretion of uric acid. However, the most common type of nephrolithiasis is calcium oxalate, which can occur in patients with hypercalciuria from any cause. Uric acid stones are radiolucent and may not be seen on an x-ray, as in this patient. Calcium oxalate stones are more likely to be seen on x-ray, but a negative x-ray does not rule out the presence of a stone, as shown in this patient's CT scan. Patients with uric acid stones often have acidic urine (pH < 5.5). In contrast, calcium oxalate stones are more likely to form in alkaline urine (pH > 7.0).

Option B: Urinary pH: 4.7 is incorrect because it is associated with uric acid stones, which are less likely given this patient's history and CT findings.

Option C: Hexagon shaped crystals are associated with cystinuria, a genetic disorder that results in defective renal tubular transport of cystine, leading to recurrent nephrolithiasis. This patient has no history of recurrent stones, making this option less likely.

Option D: Positive nitrites test indicates a urinary tract infection, which typically presents with dysuria, urgency, and frequency, symptoms that this patient does not have.

Option E: Largely positive urinary protein can be seen in nephrotic syndrome, which typically presents with edema and hyperlipidemia, signs absent in this patient.

Explanation behind assignment of error category: Here, GPT-4 (correctly) identifies kidney stones as the underlying condition of the patient, and the history indicates gout, which results from an imbalance in uric acid production. “No abnormalities are shown in the x-ray" as stated in the question, GPT-4 states that calcium oxalate stones are more likely to be seen in an x-ray as compared to uric acid stones; thus suggestive that uric acid stones are more likely. However, a CT scan of abdomen and pelvis shows presence of stones. Here, GPT-4 vaguely concludes without clear justification as to why calcium oxalate stones are more likely in this case. If it had gone for uric acid stones, GPT-4 based on its own logic, would have opted for the correct option (Urinary pH: 4.7).

Positive example 2:

USMLE question: A 41-year-old African American woman presents with her husband to her primary care doctor for evaluation of depression and anxiety. She reports a 2-week history of rapid onset sadness with no clear inciting factor. She is accompanied by her husband who notes that she has had at least three similar episodes that have occurred over the past two years. He also notes that she has been “more emotional” lately and seems confused throughout the day. She has had to leave her job as a librarian at her child’s elementary school. Her past medical history is notable for two diagnostic laparoscopies for recurrent episodes of abdominal pain of unknown etiology. Her family history is notable for psychosis in her mother and maternal grandfather. Her temperature is 99°F (37.2°C), blood pressure is 125/75 mmHg, pulse is 75/min, and respirations are 17/min. On exam, she is disheveled and appears confused and disoriented. Her attention span is limited and she exhibits emotional lability. This patient’s condition is most likely due to a defect in an enzyme that metabolizes which of the following compounds?
Option: {'A': 'Aminolevulinic acid', 'B': 'Coproporphyrinogen III', 'C': 'Hydroxymethylbilane', 'D': 'Porphobilinogen', 'E': 'Protoporphyrin IX'}
Please explain your answer and why the other options are wrong

Correct answer: D

GPT-4 Response:
The correct answer is A. Aminolevulinic acid.

This patient's periodic abdominal pain and neurologic signs (e.g., emotional lability, confusion) are suggestive of acute intermittent porphyria. Acute intermittent porphyria is caused by a deficiency of the enzyme porphobilinogen deaminase (also known as hydroxymethylbilane synthase). This enzyme defect leads to an accumulation of aminolevulinic acid and porphobilinogen, precursors in the heme synthesis pathway. An attack can be precipitated by certain medications (e.g., cytochrome P450 inducers), smoking, alcohol, and dietary restriction (e.g., fasting, low carbohydrate intake). Manifestations of acute porphyria include abdominal pain (often severe), neurologic symptoms (e.g., peripheral neuropathy, psychiatric disturbances), and autonomic dysregulation (e.g., tachycardia, hypertension). The abdominal pain is thought to be due to porphyrin-induced autonomic neuropathy resulting in intestinal dysmotility.

Options B, C, D and E are incorrect because these compounds are not the ones that accumulate due to the enzyme defect in acute intermittent porphyria. Instead, it is aminolevulinic acid and porphobilinogen that accumulate and cause the symptoms of this condition.

Explanation behind assignment of error category: Although, GPT-4 clearly states that the porphobilinogen deaminase enzyme defect leads to accumulation of aminolevulinic acid (option A) and porphobilinogen (option D), it provides no reason as to why it prefers aminolevulinic acid over porphobilinogen, thus leading to the "Incorrect or Vague Conclusion" category. Additionally, while refuting the remaining options it incorrectly states that option D does not accumulate due to enzyme defect.

Positive example:

USMLE question: A 20-year-old man comes to the clinic complaining of shoulder pain for the past week. He is a pitcher for the baseball team at his university and reports that the pain started shortly after a game. The pain is described as achy and dull, intermittent, 7/10, and is concentrated around the right shoulder area. He denies any significant medical history, trauma, fever, recent illness, or sensory changes but endorses some difficulty lifting his right arm. A physical examination demonstrates mild tenderness of the right shoulder. When the patient is instructed to resist arm depression when holding his arms parallel to the floor with the thumbs pointing down, he reports significant pain of the right shoulder. Strength is 4/5 on the right and 5/5 on the left with abduction of the upper extremities. What nerve innervates the injured muscle in this patient?
Option: {'A': 'Accessory nerve', 'B': 'Axillary nerve', 'C': 'Long thoracic nerve', 'D': 'Subscapular nerve', 'E': 'Suprascapular nerve'}
Please explain your answer and why the other options are wrong

Correct answer: E

GPT-4 Response: The correct answer is B. Axillary nerve.

This patient's presentation is consistent with an injury to the supraspinatus muscle. The supraspinatus is one of the four muscles of the rotator cuff and contributes to shoulder stability. It is primarily responsible for the first 15 degrees of arm abduction. Injuries are common in athletes who perform overhead throwing movements (e.g., baseball pitchers, tennis players). The "empty can" test, in which the patient resists downward pressure on the arm while it is elevated to 90 degrees and internally rotated (thumb pointed down), is commonly used to evaluate for supraspinatus muscle injury. The supraspinatus muscle is innervated by the suprascapular nerve.

Option A, Accessory nerve, is incorrect because this nerve innervates the trapezius and sternocleidomastoid muscles, not the muscles of the rotator cuff. Option C, Long thoracic nerve, is incorrect because this nerve innervates the serratus anterior muscle, not the muscles of the rotator cuff. Option D, Subscapular nerve, is incorrect because this nerve innervates the subscapularis muscle, not the supraspinatus muscle. Option E, Suprascapular nerve, is incorrect because this nerve innervates the supraspinatus and infraspinatus muscles.

However, in this case the injury seems to be related to the axillary nerve as it is responsible for deltoid muscle (arm abduction from 15-90 degrees) which seems more consistent with this patient's difficulty on lifting his arm.

Also the axillary nerve innervates the teres minor which is also part of the rotator cuff and can be damaged due to repetitive overhead movements.

Explanation behind assignment of error category: GPT-4 identifies the cause as an injury to the supraspinatus muscle and further explains more in the first paragraph of the response, and finally states that this muscle is innervated by the suprascapular nerve (which is the ground-truth correct answer). However, without any justification it refutes option E stating the same statement that “this nerve innervates the supraspinatus and infraspinatus muscles”, and thus contradicting its own response. Finally, while justifying option B, it introduces that it may be due to injury in the deltoid muscle, and thus axillary nerve is more likely. This vague conclusion and contradictory statements lead to its assignment to the “Incorrect or Vague Conclusion” category.

Error Type 4: Ignore Missing Information

The model recognizes that a critical piece of information is missing that would be needed to answer the question (e.g. a missing CT scan) and states so. Nevertheless, it continues to try and answer the question with incomplete information. Another scenario is when it does not recognize (mention) that a resource mentioned in the question is missing.

Positive example 1:

USMLE question: A previously healthy 36-year-old man comes to the physician for a yellow discoloration of his skin and dark-colored urine for 2 weeks. He does not drink any alcohol. Physical examination shows jaundice. Abdominal and neurologic examinations show no abnormalities. Serum studies show increased levels of alanine aminotransferase (ALT) and aspartate aminotransferase (AST). A liver biopsy is performed and a photomicrograph after periodic acid-Schiff-staining is shown. Which of the following is the most likely additional finding in this patient?
Option: {'A': 'Tropheryma whipplei-specific RNA on PCR', 'B': 'Bullous changes of the lung bases on chest CT', 'C': 'Beading of intra- and extrahepatic bile ducts on ERCP', 'D': 'Myocardial iron deposition on cardiovascular MRI', 'E': 'Dark corneal ring on slit-lamp examination'}
Please explain your answer and why the other options are wrong

Correct answer: B

GPT-4 Response:
Without the actual image of the liver biopsy and further clinical information, it is challenging to definitively answer this question. However, basing the answer on the given details and the options provided, the most likely answer would be 'E': 'Dark corneal ring on slit-lamp examination

The patient's age, his symptoms (jaundice, dark urine), his lab findings (elevated ALT and AST), and the use of periodic acid-Schiff-staining in his liver biopsy suggest liver disease, possibly due to a metabolic disorder such as Wilson's disease.

Wilson's disease is a rare autosomal recessive disorder characterized by defective copper transport, leading to copper accumulation in various tissues, including the liver, brain, and cornea. The dark corneal ring (Kayser-Fleischer ring) is a classic ocular manifestation of this disease.

Now, let's discuss why the other choices are unlikely:

Option 'A' refers to Whipple disease, a systemic disorder caused by the bacterium Tropheryma whipplei. While it can cause malabsorption and weight loss, it is not typically associated with liver disease or the findings described.

Option 'B' refers to conditions such as COPD or Alpha-1 Antitrypsin Deficiency, which would typically present with respiratory symptoms, not just liver disease.

Option 'C' describes a cholangiogram seen in primary sclerosing cholangitis (PSC), a disease that typically presents with a history of inflammatory bowel disease, which is not mentioned here.

Option 'D' refers to iron overload conditions like Hemochromatosis. While this condition can cause liver disease, myocardial iron deposition would be an uncommon initial presentation, and the patient’s described presentation does not align well with hemochromatosis.

Remember that while the most likely answer based on the provided information is Wilson's disease, the diagnosis would typically require additional tests like serum/urinary copper and ceruloplasmin levels. Also, the liver biopsy in Wilson's disease would usually show steatosis, glycogenated nuclei, and copper deposits.

Explanation behind assignment of error category: GPT-4 understands that the actual image of the liver biopsy and further clinical information are missing. However, it continues to provide an answer and justification. The relevant portion is highlighted in red.

Error Type 5: Incorrect Understanding of Task

The model did not choose a single option as the correct answer, instead it either considers multiple answers to be correct (or most likely), states all options are wrong or does not decide overall. Another scenario is when the response does not contain any explanation as to why the other options are wrong.

Positive example:

USMLE question: A 42-year-old woman is in the hospital recovering from a cholecystectomy performed 3 days ago that was complicated by cholangitis. She is being treated with IV piperacillin-tazobactam. She calls the nurse to her room because she says that her heart is racing. She also demands that someone come in to clean the pile of garbage off of the floor because it is attracting flies. Her pulse is 112/min, respiratory rate is 20/min, temperature is 38.0°C (100.4°F), and blood pressure is 150/90 mm Hg. On physical examination, the patient appears sweaty, distressed, and unable to remain still. She is oriented to person, but not place or

time. Palpation of the abdomen shows no tenderness, rebound, or guarding. Which of the following is the most likely diagnosis in this patient?
Option: {'A': 'Acute cholangitis', 'B': 'Alcoholic hallucinosis', 'C': 'Delirium tremens', 'D': 'Hepatic encephalopathy', 'E': 'Thyroid storm'}
Please explain your answer and why the other options are wrong

Correct answer: C

GPT-4 response:
This patient's fever, tachycardia, hypertension, agitation, and visual hallucinations (garbage and flies in the room) in the setting of recent hospitalization and surgery suggest delirium, a fluctuating disturbance in attention and cognition. This patient's delirium is most likely due to her infection and use of the antipseudomonal antibiotic piperacillin-tazobactam, which can cause neurologic side effects including confusion and hallucinations, especially in older patients and those with renal impairment (Option B).

Acute cholangitis (Option A) usually presents with Charcot triad (fever, abdominal pain, jaundice) or Reynold pentad (Charcot triad plus hypotension and altered mental status). However, this patient has no abdominal pain and has been on appropriate antibiotics therapy for her infection.

Alcohol hallucinosis (Option C) is a syndrome of sustained auditory hallucinations occurring in clear consciousness that develops within 48 hours of reduction or cessation of heavy alcohol intake. This patient does not have a reported history of alcohol abuse and her hallucinations are visual, not auditory.

Delirium tremens (Option D) is a severe, life-threatening form of alcohol withdrawal that includes delirium, hallucinations, severe autonomic hyperactivity (eg, fever, tachycardia, hypertension), and seizures. However, it typically occurs 48-96 hours after the last drink. This patient does not have a reported history of alcohol abuse and her symptoms developed in the hospital, not after cessation of alcohol.

Hepatic encephalopathy (Option D) presents with confusion, asterixis, and hyperreflexia due to hyperammonemia in the setting of acute or chronic liver disease. This patient has no known liver disease.

Thyroid storm (Option E) is a life-threatening exacerbation of hyperthyroidism that usually occurs in patients with known disease. Manifestations include fever, tachycardia, hypertension, agitation, and delirium. However, other signs or symptoms of hyperthyroidism (eg, weight loss, heat intolerance, goiter, ophthalmopathy) are typically present, none of which are seen in this patient.

Explanation behind assignment of error category: GPT-4 refutes all the options. Although it initially selects option B, it later also refutes option B. Therefore, it did not follow the instruction of “Which of the following is the most likely diagnosis in this patient? Please explain your answer and why the other options are wrong.”

Error Type 6. Hallucination of Information

The model makes up some information (e.g. a symptom not stated in the question) to justify why an option is or is not correct.

Positive example:

USMLE question: A 5-year-old male visits his pediatrician for a check-up. His height corresponds to the 99th percentile for his age, and pubic hair is present upon physical examination. Serum renin and potassium levels are high, as is 17-hydroxyprogesterone. Which of the following is likely deficient in this patient?
Option: {'A': '17a-hydroxylase', 'B': '11ß-hydroxylase', 'C': '21-hydroxylase', 'D': 'Aromatase', 'E': '5a-reductase'}
Please explain your answer and why the other options are wrong

Correct answer: C

GPT-4 Response: The correct answer is B: 11ß-hydroxylase.

This patient's findings suggest 11ß-hydroxylase deficiency, a form of congenital adrenal hyperplasia (CAH). This condition results in excessive production of adrenal androgens, causing virilization (as evidenced by the presence of pubic hair) and tall stature in childhood (due to early epiphyseal plate closure). The deficiency also leads to overproduction of 11-deoxycorticosterone, a mineralocorticoid precursor, which leads to hypertension and hypokalemia (hence the high renin and potassium levels). Lastly, the high 17-hydroxyprogesterone is a common finding in all forms of CAH due to feedback upregulation of ACTH.

Option A, 17a-hydroxylase deficiency, presents as female phenotype in both genetic males and females (due to lack of sex hormone production) and hypertension (due to increased mineralocorticoid production).

Option C, 21-hydroxylase deficiency, is the most common form of CAH, which usually presents with salt wasting and virilization, but not with hypertension.

Option D, Aromatase deficiency, results in masculinization of external genitalia in genetic females and tall stature in both males and females, but it does not cause hypertension.

Option E, 5a-reductase deficiency, would result in ambiguous genitalia (i.e., male pseudohermaphroditism) in genetic males, but would not cause hypertension.

Explanation behind assignment of error category: The ground-truth correct answer is C whereas GPT-4 selects option B. While explaining why Option C is incorrect, GPT-4 introduces a new symptom "hypertension" that is not explicitly mentioned in the USMLE question. The additional information is what we refer to as hallucination.

Error type 7: Unsupported Medical Claim

One or multiple claims in the model's answer are not supported by proper evidence from textbooks or other online knowledge sources. We understand this may be quite time-consuming, so we propose the following steps. You may ignore these steps if you can directly find an unsupported claim.

(i) First check whether the sentences are supported by the points mentioned in the question. No need to check any external resource at this step.
(ii) If no sentences are identified by step (i), then you need to use external resource:
  a) Use the reference (ground-truth) correct answer and find the explanation in the GPT-4 response, where it explains why that option is incorrect. These explanations are usually 2 to 3 sentences long. Based on your medical expertise, please try to find external resources from MedScape (https://www.medscape.com/) and UpToDate (https://www.wolterskluwer.com/en-in/solutions/uptodate) that refutes the claim.
  b) If you are not able to refute the claim as shown in (ii).a, then select the sentences where GPT-4 explains the reason for choosing an option. It is usually the first or last paragraph. Based on your medical expertise, please try to find external resources that refutes the claim.

Please note, only select this option if you can identify one or more sentences that contains an unsupported claim as well as a resource (publication link or website, we recommend MedScape and UpToDate) that refutes the claim

Positive example:

USMLE question: A 68-year-old man comes to the physician because of recurrent episodes of nausea and abdominal discomfort for the past 4 months. The discomfort is located in the upper abdomen and sometimes occurs after eating, especially after a big meal. He has tried to go for a walk after dinner to help with digestion, but his complaints have only increased. For the past 3 weeks he has also had symptoms while climbing the stairs to his apartment. He has type 2 diabetes mellitus, hypertension, and stage 2 peripheral arterial disease. He has smoked one pack of cigarettes daily for the past 45 years. He drinks one to two beers daily and occasionally more on weekends. His current medications include metformin, enalapril, and aspirin. He is 168 cm (5 ft 6 in) tall and weighs 126 kg (278 lb); BMI is 45 kg/m2. His temperature is 36.4°C (97.5°F), pulse is 78/min, and blood pressure is 148/86 mm Hg. On physical examination, the abdomen is soft and nontender with no organomegaly. Foot pulses are absent bilaterally. An ECG shows no abnormalities. Which of the following is the most appropriate next step in diagnosis?
Option: {'A': 'Esophagogastroduodenoscopy', 'B': 'CT scan of the abdomen', 'C': 'Hydrogen breath test', 'D': 'Cardiac stress test', 'E': 'Abdominal ultrasonography of the right upper quadrant'}
Please explain your answer and why the other options are wrong

Correct answer: D

GPT-4 response: This patient's abdominal discomfort with exertion and postprandial discomfort may suggest "abdominal angina" due to chronic mesenteric ischemia (CMI). CMI is most frequently observed in patients with generalized atherosclerotic cardiovascular disease which can be surmised from this patient's history of smoking, peripheral arterial disease, hypertension, and type 2 diabetes mellitus. The splanchnic circulation has a rich collateral flow. Consequently, at least two of the three major splanchnic vessels (celiac trunk, superior mesenteric artery, inferior mesenteric artery) must be significantly stenotic before symptoms of CMI develop. This patient's postprandial abdominal pain is due to an inability to augment splanchnic blood flow in response to increased metabolic demand after meals. As a result, patients may eat less and lose weight. The pain after a "big meal" and his inability to gain symptomatic relief despite a significant decrease in food intake due to fear of provoking the pain support this diagnosis. Atherosclerosis in other vascular beds can cause symptoms (eg, claudication), but this patient's discomfort with walking is more likely to be due to his deconditioning and obesity than to his stage 2 peripheral arterial disease, which is typically asymptomatic. Computed tomographic angiography (CTA) of the abdomen is the best initial diagnostic test for CMI. Therefore, Answer 'B' (CT scan of the abdomen) is the most likely answer.

Option 'A' (Esophagogastroduodenoscopy) is not indicated as this patient doesn't show any alarming symptoms like weight loss, anemia, or dysphagia that are typical for upper GI malignancies.

Option 'C' (Hydrogen breath test) is used to diagnose conditions like lactose intolerance and bacterial overgrowth, which are not likely in this case.

Option 'D' (Cardiac stress test) is not necessary since there's no evidence of cardiac disease.

Option 'E' (Abdominal ultrasonography of the right upper quadrant) is usually used to evaluate hepatobiliary disease, which this patient doesn't seem to have.

Explanation behind assignment of error category: The part highlighted in light green refers to the portion where GPT-4 refutes the reference (ground-truth) answer, i.e. point ii.a of error category description. The part highlighted in yellow refers to the portion where GPT-4 provides explanation for its predicted option, i.e., point (ii.b) of error category description.

The lines highlighted in red contradict the point that there are strong indications or possibilities of cardiac disease.