Supplementary Appendix for "Accuracy of a Vision-Language Model on Challenging Medical Cases" by Buckley et al. 2023

Section 1: Prompts

Zero-shot multimodal prompt:
You are a medical expert taking an image-based case challenge quiz. You will be given the description of a case and must use the associated image to give the most likely answer or diagnosis. You might also be given a quiz question to test your medical knowledge. Please answer to the best of your abilities as a medical expert would.

Question:
{question text}

Select from the following options:
{multiple choice options}

Zero-shot image-only prompt:
You are a medical expert taking an image-based case challenge quiz. You will be provided an image and must give the most likely answer or diagnosis. Please answer to the best of your abilities as a medical expert would.

Select from the following options:
{multiple choice options}

Zero-shot text-only prompt:
You are a medical expert taking a text-based case challenge quiz. You will be given the description of a case and must give the most likely answer or diagnosis. You might also be given a quiz question to test your medical knowledge. Please answer to the best of your abilities as a medical expert would.

Question:
{question text}

Select from the following options:
{multiple choice options}

Zero-shot CoT prompt:
You are a medical expert taking an image-based case challenge quiz. You will be given the description of a case and must use the associated image to give the most likely answer or diagnosis. You might also be given a quiz question to test your medical knowledge. Please answer to the best of your abilities as a medical expert would.

Question:
{0}

Select from the following options:
{1}

Let's think step by step

GPT-4 Evaluation Prompting
[Instruction]Please act as an impartial judge and evaluate the response provided by an AI assistant to a multiple choice quiz question displayed below.
The question is followed by 5 possible multiple-choice options and the AI must choose the correct ground truth answer.
[Criteria]
Score 1: There was a single definitive answer provided but it does not match the reference value.
Score 2: The answer provided matches the reference value.
Score 3: The response does not contain an answer that can be compared to the reference value. A response can also receive this score if multiple values are described but none are definitively chosen as the answer.
[Ground truth]
{reference}
Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 3 by strictly following this format: "[[rating]]", for example: "Rating: [[3]]".

[Question]
{input}

[The Start of Assistant\'s Answer]
{prediction}
[The End of Assistant\'s Answer]

Zero-shot formatted-response multimodal prompt:
You are a medical expert taking an image-based case challenge quiz. You will be given the description of a case and must use the associated image to give the most likely answer or diagnosis. You might also be given a quiz question to test your medical knowledge. Please answer to the best of your abilities as a medical expert would.

You should select exactly one answer and format it the following way: ['YOUR_ANSWER']. For example, if your answer is pneumonia, please only respond with ['pneumonia']. Do not include any other text or explanation.

Question:
{0}

Select from the following options:
{1}

Section 2: Supplemental Figures and Tables

Question	Q1	Q2	Q3	Q4	Q5	NR
2005-10-20 What process is illustrated in the radiograph?	0	0	0	0	0	50
2006-08-17 This patient's appearance is a consequence of what surgery?	0	0	0	0	0	50
2006-11-02 What accounts for this patient's hand pain?	0	0	0	1	0	49
2015-02-05 What is the diagnosis?	0	0	0	0	0	50
2014-09-18 What is the most likely diagnosis in this asymptomatic male?	0	37	0	0	0	13

Table 1A: Variability of GPT-4V with no image on repeated low-text questions. Questions were manually selected to be difficult or impossible without the image to determine when the model would still try to answer. Each question was input to the model with the zero-shot formatted-response multimodal prompt (in Prompts section) 50 times, restricting the model to select a particular option each time. “NR” represents nonresponse from the model, which happened in a variety of ways. The model would often ask the user for more information about the case, or respond that it does not have enough information to give an answer. Some responses were not formatted in the way that we asked in the prompt, but we manually assigned these to the corresponding answer.

Type	Question	Q1	Q2	Q3	Q4	Q5	NR
Low Word Count	2006-10-26 These lesions were neither pruritic nor painful. What is the diagnosis?	0	0	37	13	0	0
	2007-02-15 This patient presented with loss of vision. What is the diagnosis?	49	0	0	0	0	1
	2008-10-16 What diagnosis is suggested by this corneal photograph?	0	0	0	50	0	0
	2008-09-25 What is the diagnosis?	30	20	0	0	0	0
	2009-12-10 What is the diagnosis?	16	34	0	0	0	0
Medium	2013-11-07 What are these crystals that were aspirated	0	0	0	49	0	1

Word Count	from the bursa of an elbow of a patient with rheumatoid arthritis?						
	2006-06-22 A 55-year-old kidney-transplant recipient presented with headache and fever. The cerebrospinal fluid contained 84 percent neutrophils. What is the most likely diagnosis?	0	44	6	0	0	0
	2006-09-21 This plantar lesion was associated with inguinal lymphadenopathy. What is the most likely diagnosis?	0	0	0	50	0	0
	2009-07-30 What physical findings would be expected to be present in this patient?	1	0	12	37	0	0
	2019-07-25 A 48-year-old man presented to the dermatology clinic with a 6-month history of painful hand ulcerations and shortness of breath. He has no muscle weakness or arthritis. What is the diagnosis?	1	49	0	0	0	0
High Word Count	2022-12-01 A 26-year-old man presented to the outpatient clinic with a 1-month history of pain and swelling in the scrotum and low-grade fevers. On examination, there was swelling and tenderness of the right side of the scrotum. Laboratory studies showed peripheral eosinophilia. An ultrasound of the scrotum showed echogenic, linear structures moving within the lymphatic channels (arrowhead) adjacent to the epididymal head and testis (asterisk) — a finding known as “filarial dance sign.” What vector is responsible for transmitting the nematode causing this disease?	0	0	50	0	0	0
	2021-04-01 A 5-year-old girl presented with a 4-week history of painful swelling on both sides of her lower abdomen. Six weeks before presentation, her parents removed a tick they found buried in her umbilicus. Five days after this she developed fevers. What is the diagnosis?	0	0	50	0	0	0
	2017-07-20 A 28-year-old woman with vertigo, confusion, and falls 2 weeks after a surgical abortion at 11 weeks of gestation presents to the emergency department. Examination revealed spontaneous upbeat nystagmus, gaze-evoked nystagmus, and gait ataxia. What is the diagnosis?	0	0	47	3	0	0
	2017-08-24 A 43-year-old woman had an 8-month history of non-productive cough, unresponsive to antibiotic treatment. Physical exam showed bilateral wheezing, and pulmonary function tests showed obstructive disease, unresponsive to bronchodilators. Bronchoscopy revealed the following. What is the most likely diagnosis?	50	0	0	0	0	0
	2023-08-10 A 3-year-old girl presented to the emergency department with a 1-day history of dark urine and	50	0	0	0	0	0

	jaundice after the development of an upper respiratory tract infection 1 week earlier. The physical examination was notable for pharyngeal erythema and exudates, conjunctival pallor, and scleral icterus. There was no hepatosplenomegaly, rash, or acrocyanosis. Laboratory studies showed a new anemia and findings consistent with hemolysis. A direct antiglobulin test was positive for C3d and weakly positive for IgG. A subsequent peripheral blood smear is shown. What is the most likely diagnosis?						

Table 1B: Variability of GPT-4V on repeated question-image pairs for low, medium, and high question word counts (5 questions for each word count category). Questions were manually selected to be representative of the overall case challenge. Each question was input to the model with the zero-shot formatted-response multimodal prompt (in Prompts section) 50 times, restricting the model to select a particular option each time. “NR” represents nonresponse from the model, which happened when the model refused to answer the question. Some responses were not formatted in the way that we asked in the prompt, but we manually assigned these to the corresponding answer.

Final diagnosis	GPT-4 vision differential	Quality Score	GPT-4 no vision Differential	Quality Score
Infection with Blastomyces dermatitidis.	1. Ectopic Cushing's syndrome due to metastatic pancreatic neuroendocrine tumor (PNET). 2. Cushing's disease due to pituitary adenoma. 3. Adrenal adenoma or carcinoma. 4. Exogenous corticosteroid use. 5. Pseudo-Cushing's syndrome (e.g., from chronic alcoholism or depression). 6. ACTH-independent macronodular adrenal hyperplasia 7. Cushing's syndrome due to other ectopic ACTH production (e.g., from small cell lung cancer or carcinoid tumors). 8. Paraneoplastic syndrome from other types of tumors. 9. Iatrogenic Cushing's from medications. Familial Cushing's syndrome (e.g., Carney complex, MEN1).	0	1. Ectopic Cushing's syndrome secondary to metastatic pancreatic neuroendocrine tumor with pulmonary histoplasmosis. 2. Pulmonary histoplasmosis. Corticotroph adenoma causing Cushing's disease. 3. Secondary adrenal insufficiency from pituitary surgery. 4. Pneumonia due to methicillin-sensitive Staphylococcus aureus. 5. Metastatic disease in the lung causing the cavitary lesion. 6. Other pulmonary infections: Tuberculosis, Nocardiosis, Aspergillosis. 7. Drug-induced pneumonitis (e.g., from everolimus). 8. Pulmonary embolism. 9. Other fungal infections: Cryptococcosis, Blastomycosis, Coccidioidomycosis. 10. Non-tuberculous mycobacterial infection. 11. Lymphangitic carcinomatosis. 12. Autoimmune or inflammatory lung disease.	5

Lead poisoning	1. Pancreatic adenocarcinoma 2. Gastric carcinoma 3. Chronic pancreatitis 4. Cholangiocarcinoma 5. Hepatocellular carcinoma 6. Peptic ulcer disease Gastrointestinal lymphoma 7. Inflammatory bowel disease (e.g., 8. Crohn's disease or ulcerative colitis) 9. Autoimmune pancreatitis 10. Gastrointestinal stromal tumor (GIST) 11. Metastatic cancers to the abdomen 12. Amyloidosis 13. Gallbladder disease or cholecystitis 14. Mesenteric ischemia	0	1. Acute intermittent porphyria (AIP) 2. Lead poisoning 3. Variegate porphyria or Hereditary coproporphyria 4. Chronic mesenteric ischemia 5. Small bowel obstruction secondary to adhesions 6. Thiamine deficiency (Wernicke's encephalopathy) 7. Hypothyroidism 8. Celiac disease 9. Medication-induced cognitive impairment or electrolyte abnormalities 10. Non-convulsive status epilepticus 11. Chronic pancreatitis 12. Gastric carcinoma	5
Vascular Ehlers–Danlos syndrome.	1. Lateral medullary syndrome (Wallenberg syndrome) due to another cause, like thrombosis or embolism. 2. Herpes zoster reactivation (as suggested by vesicular appearance lesions on the lips and oral cavity). 3. Multiple sclerosis. 4. Brainstem encephalitis. 5. Basilar migraine. 6. Cerebral vasculitis. 7. Antiphospholipid syndrome (given the elevated anticardiolipin IgM). 8. Intracranial mass or tumor causing compression. 9. Inflammatory conditions, such as sarcoidosis or Sjogren's syndrome, although the patient's test results for these conditions were negative.	2	1. Ehlers-Danlos syndrome, vascular type (Type IV). 2. Marfan syndrome. Loeys-Dietz syndrome. 3. Fibromuscular dysplasia. 4. Cystic medial necrosis. 5. Collagen vascular diseases (e.g., systemic lupus erythematosus). Arteritis (e.g., Takayasu's arteritis). 6. Antiphospholipid syndrome 7. Inherited thrombophilia disorders. 8. Spontaneous artery dissection from other non-specific causes.	5

Table 2. A comparison of the differing quality of diagnosis outputs from GPT-4v and GPT-4. Scale from Kanjee, et al. 5 = the actual diagnosis was suggested in the differential; 4 = the suggestions included something very close, but not exact; 3 = the suggestions included

something closely related that might have been helpful; 2 = the suggestions included something related, but unlikely to be helpful; 0 = no suggestions close to the target diagnosis.