Supplementary Information for "Multimodal Foundation Models Exploit Text to Make Medical Image Predictions" by Buckley et al. 2024

Table of Contents

Section 1: Model Setup

To evaluate GPT-4V, which is no longer publicly available, all image challenge cases were run through the ChatGPT web interface using the GPT-4V model during October 2023. A new session was used for each question to prevent information leakage from previous queries. To evaluate GPT-4V Turbo, we used OpenAI's Application Programming Interface (API) (model "gpt-4-vision-preview," default temperature=1, detail="high") during late December 2023 and early January 2024. The training cutoff for this model was reported as April 2023. Llama-3.2-90B was evaluated using Google's Vertex AI platform, using the default temperature. Gemini Pro 1.0 was evaluated using Google's Gemini Pro API (models "gemini-pro" with default temperature=0.9 and "gemini-pro-vision" with default temperature=0.4) on January 3, 2024. We benchmarked the default medical instruction-tuned LLaVA-Med-v1.5 using the published model weights. Inference was performed using the same script the authors used to benchmark LLaVA-Med on downstream tasks (using the default temperature). Since LLaVA-Med was unable to follow instructions on formatting its output to be parsed directly, we needed to use GPT4 Turbo to extract the multiple choice answer from LLaVA-Med's response using careful prompting (see Section 2).

Section 2: 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.

At the end of your response, you must select exactly one answer and format it the following way: [['YOUR_ANSWER']]. For example, if your answer is Pneumonia, end your response with: [['Pneumonia']]

Question:
{question}

Select from the following options:
{options}

Zero-shot image-only prompt

You are a medical expert taking an image-based case challenge quiz. You will be given an image and must select the most likely answer or diagnosis. Please answer to the best of your abilities as a medical expert would.

At the end of your response, you must select exactly one answer and format it the following way: [['YOUR_ANSWER']]. For example, if your answer is Pneumonia, end your response with: [['Pneumonia']]

Select from the following options:
{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.

At the end of your response, you must select exactly one answer and format it the following way: [['YOUR_ANSWER']]. For example, if your answer is Pneumonia, end your response with: [['Pneumonia']]

Question:
{question}

Select from the following options:
{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

LLaVA-Med-v1.5 evaluation prompting

You must extract the answer by an AI assistant to a medical challenge problem. You will be provided the five choice options, as well as the assistants response to this question. You do not need access to the question itself, your task is to just determine if the assistant chose of one the five answers and extract that answer. Return the answer as [['ANSWER_HERE']]. For example, if the assistant claims that 'Pneumonia' is the answer and this was also one of the five multiple choice options, you should return [['Pneumonia']]. If the assistant provides an answer that is not one of the five multiple choice options, return [['NR']]. If the assistant does not provide any answer at all, also return [['NR']].

These are the five possible options:
{options}

Here is the assistants response:
{response}

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}

Robustness to incorrect text experiment prompt (Colleague version)

You are a medical expert taking an image-based case challenge quiz. At the end of your response, you must select exactly one answer and format it the following way: [['YOUR_ANSWER']]. For example, if your answer is Pneumonia, end your response with: [['Pneumonia']]

Your esteemed colleague has taken a history of present illness, suggesting that the patient's symptoms closely align with {distract_diagnosis}. Please comment and choose from the following options:
{options}

Robustness to incorrect text experiment prompt (Vignette version)

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.

At the end of your response, you must select exactly one answer and format it the following way: [['YOUR_ANSWER']]. For example, if your answer is Pneumonia, end your response with: [['Pneumonia']]

Case:
{case}

Select from the following options:
{options}

Prompt used to generate fictional patient vignettes with GPT-4
Create a fictional patient vignette that strongly suggests the following:
"{diagnosis}"

Please make it very short (around 5 sentences). Don't include any descriptions of imaging, tests, or physical exams. Only what the patient might immediately present with.

Additionally, your vignette should be suggestive of "{diagnosis}", but do not include the name itself. This will be used in a diagnostic challenge.

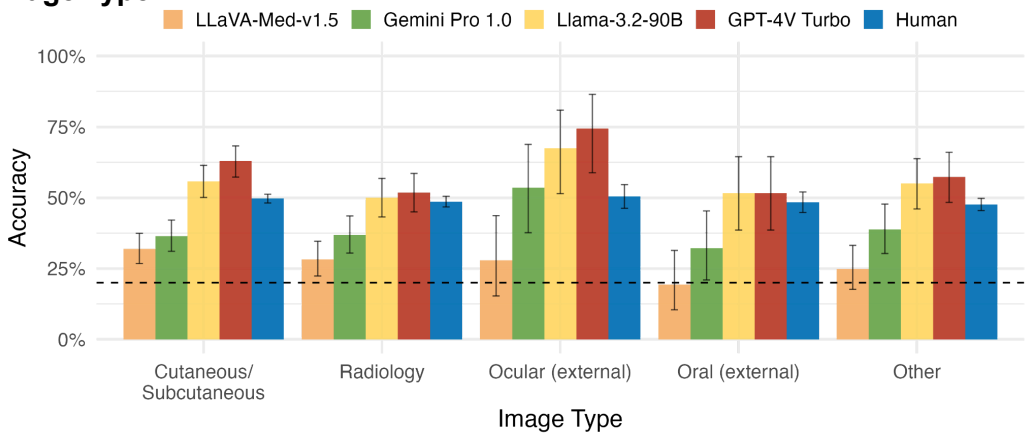
Section 3: Diagnostic Accuracy by Image Type and by Skin Color

GPT-4V Turbo and Llama-3.2-90B performed similarly across all categories of images, including including natural and dermatoscopic images of skin disease, radiographic images, external ocular images, and external oral images (Figure s1A), outperforming Gemini Pro 1.0, LLaVA-Med-v1.5, and human respondents. Skin and eye quizzes were associated with the highest diagnostic accuracy for the multimodal LLMs, with worsened performance on radiology questions. We analyzed the accuracy of models and human respondents by the Fitzpatrick skin type assigned to a quiz, categorized into “light” (1-2), “intermediate” (3-4), and “dark” (5-6) by a board-certified dermatologist in a prior study. GPT-4V Turbo and Llama-3.2-90B outperformed human respondents and other models on all FST categories. Gemini Pro 1.0 and LLaVA-Med-v1.5 performed worse than human respondents but better than random across FST groups.

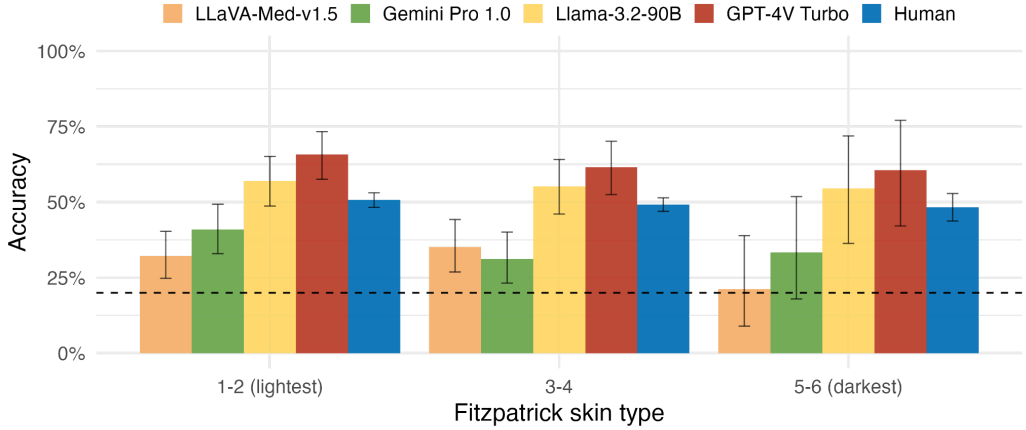
When provided image alone (Figure s1C), GPT-4V Turbo did not exhibit significant differences in accuracy by skin tone (ANOVA, p=0.59). Llama-3.2-90B (ANOVA, p=0.86) and LLaVA-Med-v1.5 (ANOVA, p=0.99) also perform similarly across skin tones. Gemini Pro 1.0 with image alone shows a borderline significant difference in accuracy between these three FST groups (ANOVA, p=0.047), with the lowest performance among FST 5-6. Differences in the reference group, where skin tone should not affect the answer, are not significant for GPT-4V (ANOVA, p=0.34), Llama-3.2-90B (ANOVA, p=0.81), LLaVA-Med-v1.5 (ANOVA, p=0.26),Gemini Pro 1.0 (ANOVA, p=0.80). LLaVA-Med-v1.5 does not appear to be affected by skin tone, performing on par with a random guess for all skin tone groups.

Figure s1. Accuracy by Image Type and by Fitzpatrick Skin Type

A. Performance of Multimodal Vision-Language Models versus Human Respondents by Image Type



B. Performance of Multimodal Vision-Language Models versus Human Respondents by Fitzpatrick Skin Type on Dermatology Images



C. Performance of Image Only Vision-Language Models versus Human Respondents by Fitzpatrick Skin Type on Dermatology Images

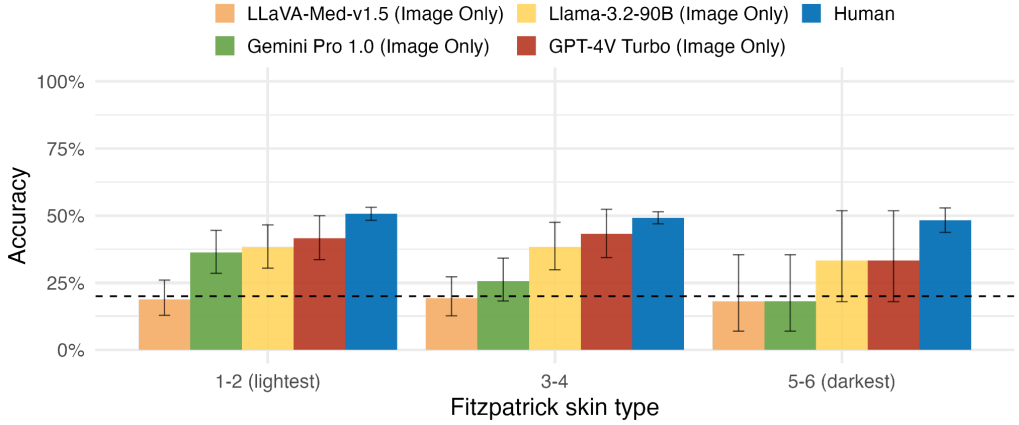


Figure s1. Accuracy by Image Type and Image Skin Color.
Accuracy of LLaVA-Med-v1.5, Gemini Pro 1.0, Llama-3.2-90B, and GPT-4V vs. human respondents in the *NEJM* Image Challenges by **A.** type of image for 764 annotated cases, and **B, C.** Fitzpatrick skin type for 307 quizzes with visible skin tone. **C.** shows the accuracy of these models when given only the image.

Image annotations were retrieved from Diao et al. *JAAD* 2020 and Fitzpatrick skin type was rated by a board-certified dermatologist, categorized into “light” (1-2), “intermediate” (3-4), and “dark” (5-6). Error bars are 95% confidence intervals and the dashed black line indicates random guesses.

Section 4: Additional Figures and Tables

Table s1: Variability of GPT-4V with no image on repeated low-text questions

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

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.

Table s2: Variability of GPT-4V on repeated question-image pairs for low, medium, and high question word counts

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 Word Count	2013-11-07 What are these crystals that were aspirated from the bursa of an elbow of a patient with rheumatoid arthritis?	0	0	0	49	0	1
	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	0	0	50	0	0	0

	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?						
	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 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?	50	0	0	0	0	0

Questions were manually selected to be representative of the overall case challenge. There were 5 questions for each word count category. 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.

Figure s2: Changing word count over the years

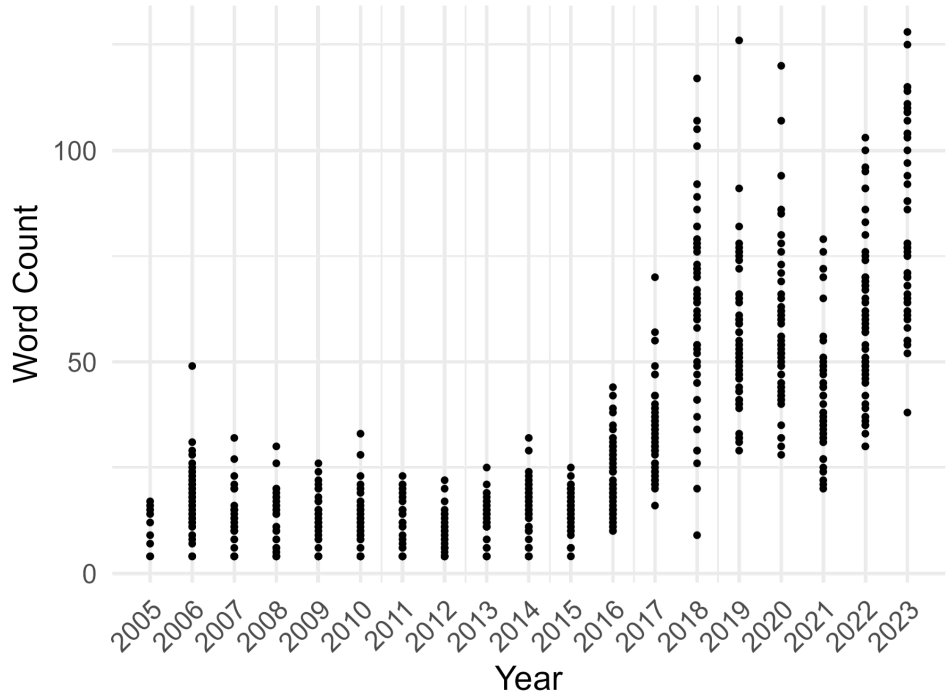


Figure s3: GPT-4 Performance by Case Text Length on Radiology Images

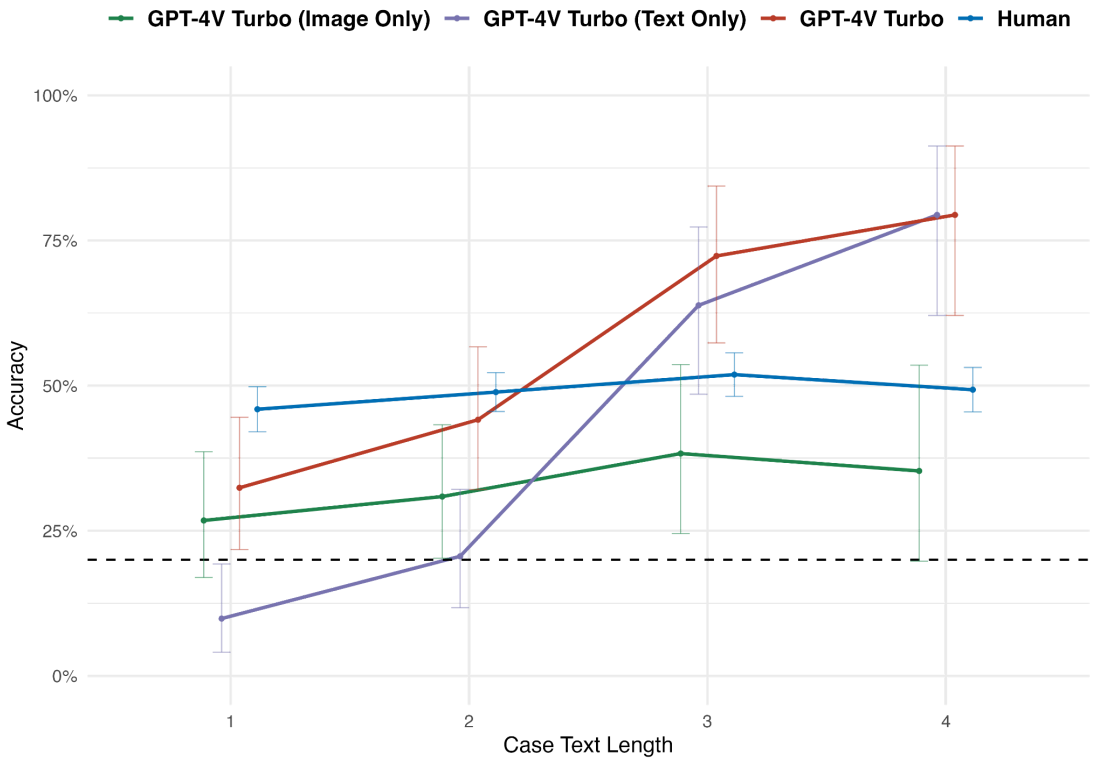


Table s3: Model Refusal Rates

GPT-4V Turbo	1.5%
GPT-4V Turbo (Image Only)	3.7%
GPT-4V Turbo (Text Only)	26.8%
Llama-3.2-90B	7.7%
Llama-3.2-90B (Image Only)	5.5%
Gemini Pro 1.0	18.6%
Gemini Pro 1.0 (Image Only)	20.5%
LLaVA-Med-v1.5	1.1%
LLaVA-Med-v1.5 (Image Only)	14.9%