MedBLINK: Probing Basic Perception in Multimodal Language Models for Medicine

Supplementary Material

A. MEDBLINK Curation

A.1. Prompt Details: Text and Prompt

We leverage two main types of prompts: text questions and visual prompts. The questions used for each tasks is outlined in Table 9. We use circles/points/dots for visual prompting on 3 tasks. Specifically for both Depth Estimation tasks we use 3 (red, green, blue) colored circles with 10px radius on 512x512 resized image (original 112x112). For the Histology Structure tasks we leverage points whos size depend on the size of the WSI before cropping, specifically we set the size of the circles to 1/70 the size of the max(width, height) of the WSI.

A.2. Human Evaluation Method

We obtain human evaluation scores from a pool of 4 human experts (3 co-authors, 1 independent). Each task is evaluated by at least one expert and the average score is used as the human benchmark.

A.3. Benchmark Statistics

MEDBLINK statistics can be found in Table 6, we also outline the Label distribution and count of individual tasks in Table 7

Statistics	Number
Total Questions	1429
Total Images	1605
Questions with Visual Prompts	431
Questions with Multiple Images (2)	176

Table 6. Detailed statistics of the MEDBLINK benchmark.

Task	**Label Distribution**	**Total**
Task 1	{yes: 67, no: 67}	134
Task 2	{red: 55, green: 38, blue: 51}	144
Task 3	{red: 49, green: 51, blue: 46}	146
Task 4	{red: 47, green: 53, blue: 41}	141
Task 5	{flip: 100, correct: 100}	200
Task 6	{1: 127, 2: 49}	176
Task 7	{0: 30, 2: 28, 4: 30}	88
Task 8	{pediatric: 100, adult: 100}	200

Table 7. Distribution of labels across different medical imaging tasks.

B. Baseline Model Details

We test 19 Multimodal LMs on MEDBLINK, setting the temperature of all models to 0, including:

1. GPT-4O [27] version
2. CLAUDE 3.5 SONNET [1]
3. GEMINI 1.5 PRO [52]
4. QWEN 2.5 VL [6], specifically, we leverage the 3B, and 7B parameterized models.
5. LLAVA-ONEVISION [32]. Here we use two versions as well, the 0.5B parameterized model, and 7B parameterized models
6. LLAVA-MED [33]
7. MED-FLAMINGO [41], unlike other models for MED-FLAMINGO to produce valid responses, we need to use few-shot prompting [49]. Specifically, we prompt it with three questions and answers from PMC-VQA benchmark [66] as seen in Table 8 for free-from evaluation following MediConfusion [49] setup.
8. RADFM [60]
9. AURORA [8]
10. SpatialRGPT [14]
11. LLaVA 1.5 (7B) [35]
12. INTERNVL 2.5 [13], we leverage the 4B, 8B, 26B and 38B parameterized models.
13. LLAMA 3.2 11B [21]

B.1. Small Specialized models

We train small specialized models for some of the tasks with sizeable train sets from the original dataset used to construct the task. We finetune a ResNet-50 [23] model on both the age estimation and image orientation tasks. For training we used a batch-szie of 32, using an 80/20 split we trained each model for 10 epochs and used a learning rate of 1e-3 and decay of 1e-4.

C. Qualitative Evaluation of Failure Cases

In the following section, we present case-based qualitative analysis of failure cases to better understand the pattern of failed prediction, Figures 4, 6, 5, 7, 8, 9, 10, 11, 12, 13.

Model	Prompt
Med-Flamingo	You are a helpful medical assistant. You are being provided with images, a question about each image and an answer. Follow the examples and answer the last question. <image>Question: What radiological technique was used to confirm the diagnosis? Answer: Mammography<|endofchunk|> <image>Question: What did the CT scan show? Answer: Cerebral edema<|endofchunk|> <image>Question: What is the purpose of the asterisk shown in the figure? Answer: To indicate the normal lentoid shape of hypocotyl nuclei.<|endofchunk|> <image>Question: **QUESTION** Answer:

Table 8. Medical imaging prompt template for Med-Flamingo model.

Task	Question Format
Task 1: Image Enhancement Detection	Is this CT scan image contrast-enhanced? (Answer with yes or no)
Task 2: Visual Depth Estimation	Which of the dots is at the greatest depth in this endoscopy image? (Answer with red or green or blue) Please put your final answer in 'boxed{}'
Task 3: Wave-Based Imaging Depth Estimation	Which of the dots is closest to the surface of the skin? (Answer with red, green, or blue)
Task 4: Histology structure	Given this melanoma biopsy, which of the dots is closest in distance to the surface of the skin (stratum corneum or epithelium)? (Answer with red or green or blue)
Task 5: Imaging Orientation	Is this X-ray image in its correct anatomical orientation or upside down? (Answer with 'correct' if it is properly oriented, or 'upside down' if it has been rotated 180 degrees.)
Task 6: Relative Position	which of these two axial slices is closer to the pelvis, 1 or 2? Please put your final answer in 'boxed{}'
Task 7: Morphology Quantification	How many wisdom teeth can be seen in this X-ray image? (A) 0 (B) 2 (C) 4
Task 8: Age Estimation	Is this a pediatric or adult chest x-ray? (Answer with pediatric or adult)

Table 9. Medical imaging tasks with corresponding question formats.

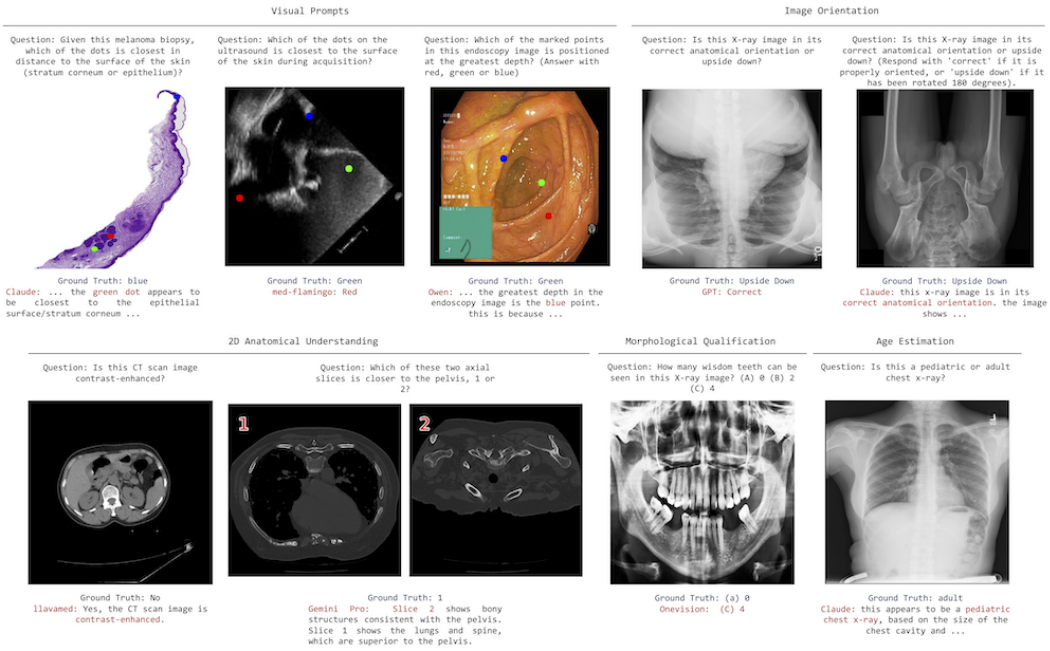
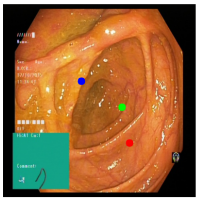


Figure 4. **Examples of *Failure* Cases in the MedBLINK Benchmark.** Examples of failure cases where LLMs struggle with MedBLINK tasks. Incorrect model responses are shown in red, highlighting the challenges LLMs face in handling these fundamental medical perception tasks.

Which of the dots is at the greatest depth in this endoscopy image?



Ground Truth: Green

Claude	looking at the endoscopic image of what appears to be the colon, the red dot appears to be positioned at the greatest depth. this can be determined by observing the concentric folds of the intestinal wall and how they create a tunnel-like perspective, with the red dot being placed furthest along this tunnel in the deepest visible fold.
GPT	the red point is positioned at the greatest depth.
Gemini-pro	the red dot is furthest along the visible tract in the image. depth perception in endoscopy images can be tricky, but the folds of tissue and the way the light falls suggest the red dot is deeper within the body
Qwen-VL	the point that appears to be positioned at the greatest depth in the endoscopy image is the blue point. this is because it is located further away from the camera and closer to the center of the image compared to the other two points, which are closer to the surface of the tissue.

Figure 5. **Example Failure Case in Visual Depth Estimation** Some models incorrectly estimate the relative depth in endoscopic images.

Is this X-ray image in its correct anatomical orientation or upside down? (Respond with 'correct' if it is properly oriented, or 'upside down' if it has been rotated 180 degrees.

Ground Truth: Upside down

Claude	this x-ray image is in its correct anatomical orientation. the image shows the pelvic region and proximal femurs (upper thigh bones) in their proper orientation, with the iliac bones of the pelvis at the top and the femoral heads and shafts extending downward. the "r" marker visible in the lower left corner also confirms this is the proper orientation.
Qwen-VL	the x-ray image is correctly oriented. the pubic bones and the sacrum are positioned as they would be in a standard anteroposterior (ap) view of the pelvis, which is typically taken with the patient standing or lying in a specific position to ensure the image is correctly aligned.
LLaVA-onevision	the x-ray image is in its correct anatomical orientation.
GPT	correct

Figure 6. **Example Failure Case in Anatomical Orientation Reasoning of MLM on pelvic X-Ray Images** Some models incorrectly classify upside-down pelvic X-Ray images as correct.

Is this a pediatric or adult chest x-ray?

Ground Truth: Adult

Claude	this appears to be a pediatric chest x-ray, based on the size of the chest cavity and the characteristic appearance of pediatric chest anatomy. the heart size relative to the chest wall and the rib spacing are typical of what you would see in a pediatric patient.
GPT	Pediatric
LLaVA-Med	Pediatric

Figure 7. **Example Failure Case in Age Estimation of MLM on Chest X-Ray Images** Some models incorrectly classify adult chest X-Ray images as pediatric and vice versa.

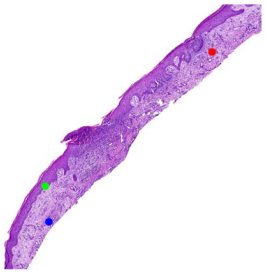
Question: Is this X-ray image in its correct anatomical orientation or upside down? (Answer with 'correct' if it is properly oriented, or 'upside down' if it has been rotated 180 degrees.)

Ground Truth: Upside Down

Claude	This chest X-ray is in its correct anatomical orientation. We can tell this because: 1.The heart apex points to the left (as it should) 2.The diaphragm curves upward at the bottom 3.The breast tissue shadows are at the top of the image 4.The vertebral column alignment is correct (superior to inferior)

Figure 8. **Example Failure Case in Anatomical Orientation Reasoning of MLM on Chest X-Ray Images** Some models incorrectly classify upside-down chest X-ray images as upright, and vice versa.

Question: 'Which of the dots is closest to the surface of the skin? (Answer with red, green, or blue)

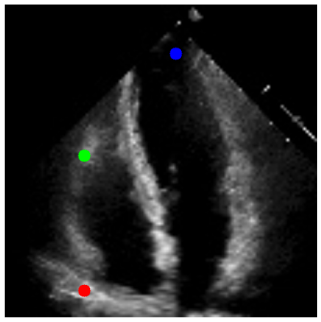


Ground Truth: Green

Gemini Pro	Red. The ultrasound probe is placed on the skin, and the red dot is closest to the edge of the image, which corresponds to the probe's location.
GPT	The red dot is closest in distance to the surface of the skin (stratum corneum or epithelium).
LLaVA-Med	The dot closest in distance to the surface of the skin (stratum corneum or epithelium) is the red dot.
LLaVA-OneVision	Blue
Med-flamingo	Blue
RadFM	The deeper melanoma cells are found in the blue dot.

Figure 9. **Example Failure Case in Distance Reasoning of MLM on pathology Images** Some models incorrectly identify one color to be the closest to the surface of skin, regardless of their location.

Question: Which of the dots on the ultrasound is closest to the surface of the skin during acquisition? (Answer with red or green)

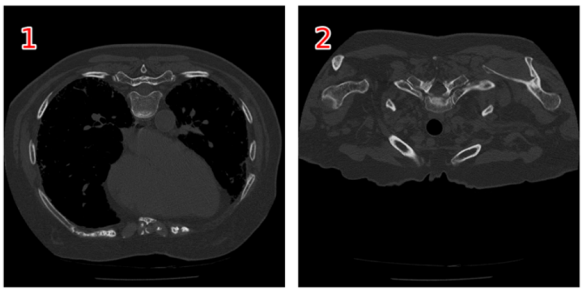


Ground Truth: Blue

Gemini Pro	Red. The image is an apical 4-chamber view of the heart taken with an ultrasound probe placed on the chest. The red dot is closest to the top of the image, which corresponds to the chest wall and skin surface.
GPT	The red dot is closest to the surface of the skin.
LLaVA-Med	The red dot is closest to the surface of the skin.
LLaVa-OneVision	Red
Qwen-VL	The dot that is closest to the surface of the skin is the green dot.

Figure 10. **Example Failure Case in Depth Reasoning of MLM on ultrasound Images** Similar to cases in Figure 9, models incorrectly identify one color to be the closest to the probe, regardless of their location.

which of these two axial slices is closer to the pelvis, 1 or 2?



Ground Truth: 1

Gemini-pro	Slice 2 shows bony structures consistent with the pelvis. Slice 1 shows the lungs and spine, which are superior to the pelvis.
GPT	2
Claude	Looking at these two CT scan axial slices: Image 1 shows a chest CT scan, as evidenced by the clear visualization of: - Both lungs (the dark/black areas) - Heart in the middle (the large white structure) - Thoracic spine and ribs Image 2 shows a pelvic CT scan, as evidenced by: - The characteristic pelvic bone structure - Hip joints - Absence of lung tissue Therefore, 2 is closer to the pelvis.

Figure 11. **Example Failure Case in Relative Position of MLM on axial CT slices** Some models incorrectly orient the anatomical location of CT slices, or confuse the location relationship of two CT slices.

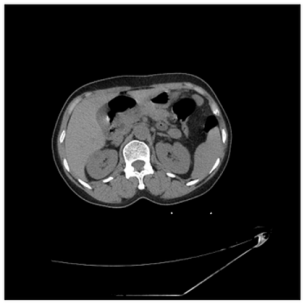
How many wisdom teeth can be seen in this X-ray image? (A) 0 (B) 2 (C) 4

Ground Truth: (A) 0

LLaVA-Med	In this X-ray image, there are 4 wisdom teeth visible.
GPT	The X-ray image shows (C) 4 wisdom teeth.
Claude	Looking at this dental panoramic X-ray image, I can see 4 wisdom teeth (option C). The wisdom teeth are visible in all four corners of the jaw - two in the upper jaw (maxilla) and two in the lower jaw (mandible). They appear as the last molars in each quadrant of the mouth. This is a typical presentation of wisdom teeth in a dental X-ray, though their position and alignment can vary between individuals.

Figure 12. **Example Failure Case in Morphology Quantification** Some models fail to count the number of wisdom teeth, which are distinguishable morphological features.

Is this CT scan image contrast-enhanced? (Answer with yes or no)



Ground Truth: No

Claude	Yes, this is a contrast-enhanced CT scan, as evidenced by the bright (white) appearance of blood vessels and certain structures in the abdomen. The vertebral body (spine) also shows the typical bright appearance of contrast in the surrounding vessels.

Figure 13. **Example Failure Case in Perceptual Reasoning in Image Enhancement Detection** Some models fail to detect enhanced regions and misinterpret whether the CT slice is contrast-enhanced.

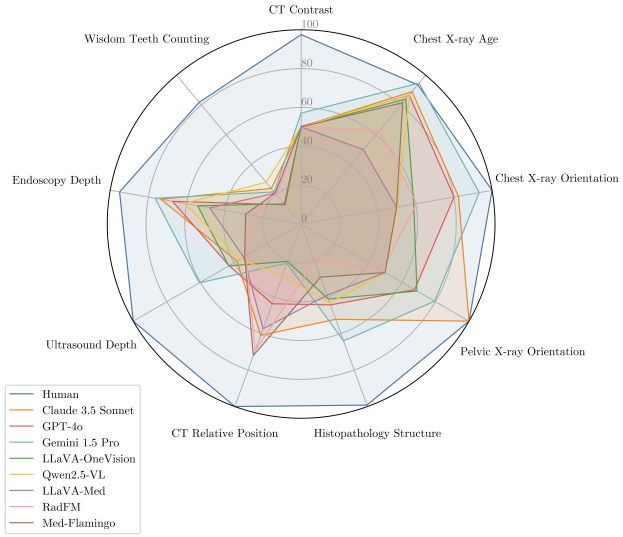


Figure 14. Accuracies of multimodal LMs on MEDBLINK.