

Supplementary Information

Article title: Generative Artificial Intelligence for Opportunistic Osteoporosis Screening Using Lumbar Spine Imaging: Diagnostic Performance, Reproducibility, Effects of Few-Shot Prompting, and Comparison with Physicians

Journal: European Spine Journal

Online Resource 1

Supplementary methods describing the zero-shot prompts and few-shot prompts and supplementary data showing diagnostic accuracy across five repeated inferences for each generative AI model

Prompts Used for Generative AI-Based Image Assessment

To ensure standardized evaluation across generative AI models, identical prompts were used for all models. Two prompting conditions were evaluated: zero-shot prompting and few-shot prompting. The prompts used in each condition are provided below.

Supplementary Method 1. Zero-shot prompting

In the zero-shot condition, the AI models evaluated the test images without being provided with reference examples or bone mineral density information. The following prompt was used:

Task: You will be shown a set of spine lateral radiographs and CT images.

Based only on visual impression, judge whether the bone appears weaker than normal.

Use the following labels only: A=normal bone appearance B=osteoporotic / weak bone appearance

Do not use any clinical information, metadata, or quantitative measurements. Do not provide any explanation or additional text.

Supplementary Method 2. Few-shot prompting

In the few-shot condition, reference images with corresponding bone mineral density Young Adult Mean (YAM) values were first provided to the AI models to establish visual criteria for bone quality. The following prompt was used:

Task: You will be provided with two distinct categories of images:

Reference Set: Images showing spine lateral radiographs and CT sagittal scans with their corresponding BMD YAM values displayed.

Use these to establish your visual criteria for “Normal” vs “Osteoporotic” bone quality.

Test set: Two sets of images for judgment (Set A: Radiograph 1-45, Set B: CT images 1-45). These images do not have YAM values displayed.

Instructions:

Phase 1 (Learning): Analyze the Reference Set to calibrate your visual assessment. Correlate the visual bone density, trabecular pattern, and cortical thickness with the provided YAM values.

Phase 2 (Judgment): Based on the visual criteria established in Phase 1, judge the Test Set (1-45). Judge whether the bone in each image appears weaker than normal.

Labeling: Use the following labels only: A=normal bone appearance B=osteoporotic / weak bone appearance

Strict Constraint: Do not provide any clinical information, metadata, or quantitative measurements. Do not provide any explanation or additional text.

Supplementary Table S1. Diagnostic accuracy of five repeated inferences for each AI model

T-score -1.0

Radiograph

Model	Run1	Run2	Run3	Run4	Run5	Mean (%)	Majority vote
ChatGPT	28/45 (62.2)	24/45 (53.3)	21/45 (46.7)	25/45 (55.6)	29/45 (64.4)	56.4 ± 7.1	25/45 (55.6)
Gemini	29/45 (64.4)	17/45 (37.8)	28/45 (62.2)	30/45 (66.7)	23/45 (51.1)	56.4 ± 12.0	29/45 (64.4)
Claude	29/45 (64.4)	30/45 (66.7)	26/45 (57.8)	35/45 (77.8)	34/45 (75.6)	68.4 ± 8.2	33/45 (73.3)

CT

Model	Run1	Run2	Run3	Run4	Run5	Mean (%)	Majority vote
ChatGPT	19/45 (42.2)	19/45 (42.2)	21/45 (46.7)	21/45 (46.7)	16/45 (35.6)	42.7 ± 4.6	18/45 (40.0)
Gemini	23/45 (51.1)	18/45 (40.0)	20/45 (44.4)	30/45 (66.7)	23/45 (51.1)	50.7 ± 10.1	21/45 (46.7)
Claude	22/45 (48.9)	19/45 (42.2)	22/45 (48.9)	25/45 (55.6)	25/45 (55.6)	50.2 ± 5.6	22/45 (48.9)

T-score -2.5

Radiograph

Model	Run1	Run2	Run3	Run4	Run5	Mean (%)	Majority vote
ChatGPT	25/45 (55.6)	23/45 (51.1)	22/45 (48.9)	26/45 (57.8)	28/45 (62.2)	55.1 ± 5.3	24/45 (53.3)
Gemini	20/45 (44.4)	26/45 (57.8)	23/45 (51.1)	27/45 (60.0)	24/45 (53.3)	53.3 ± 6.1	25/45 (55.6)
Claude	26/45 (57.8)	27/45 (60.0)	29/45 (64.4)	26/45 (57.8)	29/45 (64.4)	60.9 ± 3.4	30/45 (66.7)

CT

Model	Run1	Run2	Run3	Run4	Run5	Mean (%)	Majority vote
ChatGPT	20/45 (44.4)	22/45 (48.9)	22/45 (48.9)	22/45 (48.9)	21/45 (46.7)	47.6 ± 2.0	21/45 (46.7)
Gemini	22/45 (48.9)	23/45 (51.1)	25/45 (55.6)	15/45 (33.3)	22/45 (48.9)	47.6 ± 8.4	22/45 (48.9)
Claude	25/45 (55.6)	24/45 (53.3)	25/45 (55.6)	24/45 (53.3)	22/45 (48.9)	53.3 ± 2.7	23/45 (51.1)

Values are presented as the number of correctly classified images out of 45 images, with percentages shown in parentheses. Mean values represent the average accuracy across five independent repeated inferences.

The majority vote indicates the final classification obtained by majority voting across the five inferences.