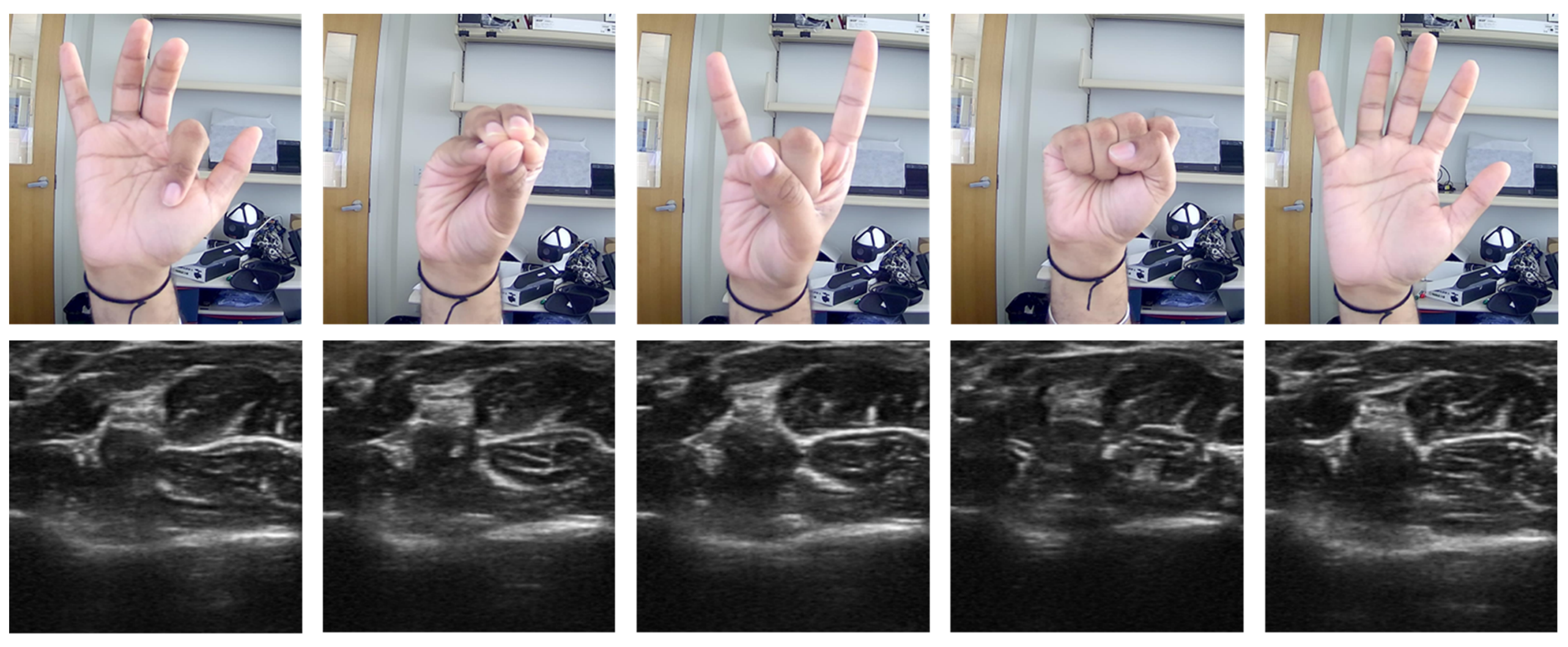
# GPT Sonograpy: Hand Gesture Decoding from Forearm Ultrasound Images via VLM

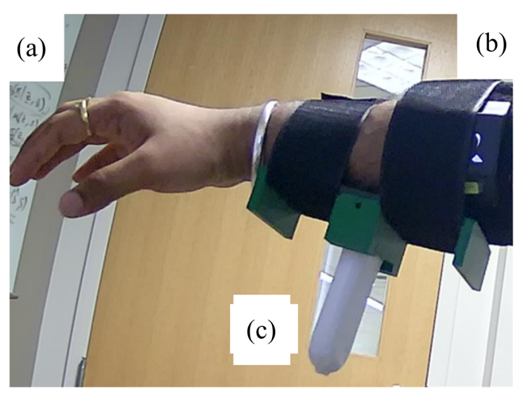
Keshav Bimbraw[1], Ye Wang[2], Jing Liu[2], Toshiaki Koike-Akino[2]
[1]Worcester Polytechnic Institute Worcester, MA 01609, USA [2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA

## Contribution

- We examine the capability of large **VLM** (Vision Language Model) for decoding **hand gestures** from **forearm ultrasound data**.
- We use **GPT-4o** to analyze gesture classification performance for 5 hand gestures based on forearm ultrasound images.
- We show that few-shot **ICL** (in-context learning) can improve the classification accuracy even without any fine-tuning.
- We demonstrate that GPT-4o can achieve greater than **70%** accuracy for hand gesture classification.
- We show that **RAG** (retrieval augmented generation) can further improve the ICL performance to near 100% accuracy.

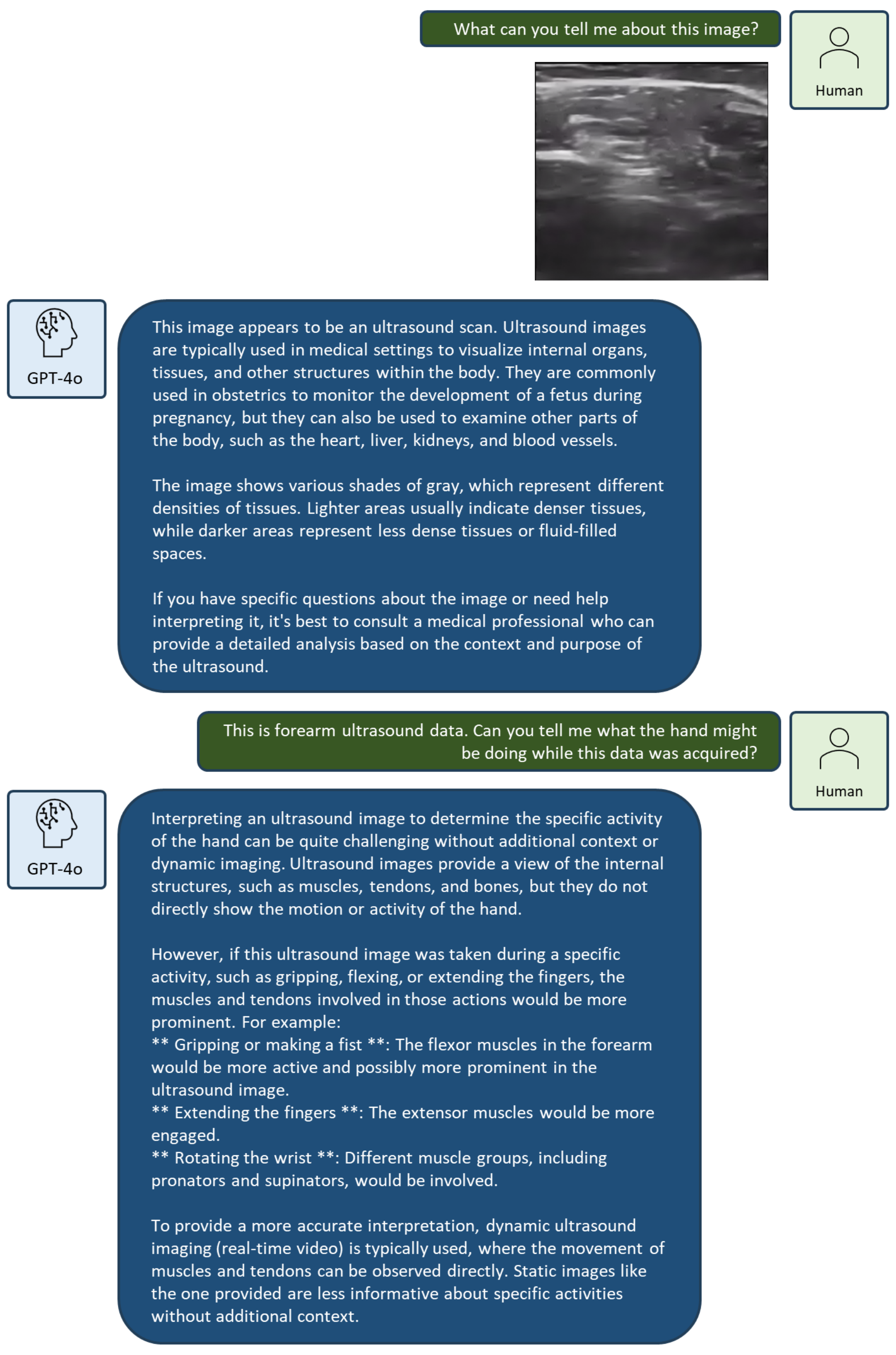
## Ultrasound Imaging for Hand Gesture





(a) Hand gesture
(b) Custom made wearable sensor armband
(c) Ultrasound probe

## VLM Sonography Capability



What can you tell me about this image?

This image appears to be an ultrasound scan. Ultrasound images are typically used in medical settings to visualize internal organs, tissues, and other structures within the body. They are commonly used in obstetrics to monitor the development of a fetus during pregnancy, but they can also be used to examine other parts of the body, such as the heart, liver, kidneys, and blood vessels.

The image shows various shades of gray, which represent different densities of tissues. Lighter areas usually indicate denser tissues, while darker areas represent less dense tissues or fluid-filled spaces.

If you have specific questions about the image or need help interpreting it, it's best to consult a medical professional who can provide a detailed analysis based on the context and purpose of the ultrasound.

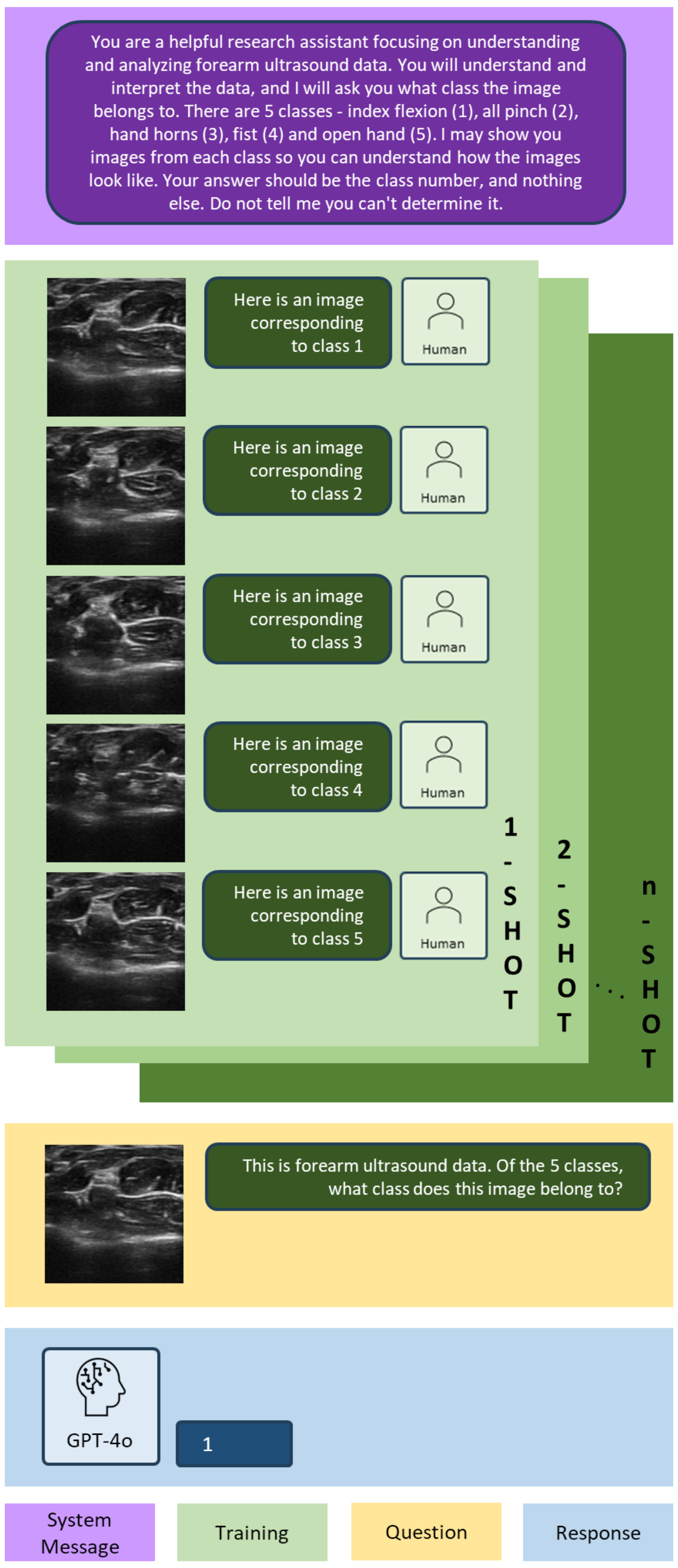
This is forearm ultrasound data. Can you tell me what the hand might be doing while this data was acquired?

Interpreting an ultrasound image to determine the specific activity of the hand can be quite challenging without additional context or dynamic imaging. Ultrasound images provide a view of the internal structures, such as muscles, tendons, and bones, but they do not directly show the motion or activity of the hand.

However, if this ultrasound image was taken during a specific activity, such as gripping, flexing, or extending the fingers, the muscles and tendons involved in those actions would be more prominent. For example:
** Gripping or making a fist **: The flexor muscles in the forearm would be more active and possibly more prominent in the ultrasound image.
** Extending the fingers **: The extensor muscles would be more engaged.
** Rotating the wrist **: Different muscle groups, including pronators and supinators, would be involved.

To provide a more accurate interpretation, dynamic ultrasound imaging (real-time video) is typically used, where the movement of muscles and tendons can be observed directly. Static images like the one provided are less informative about specific activities without additional context.

## In-Context Learning (ICL)



You are a helpful research assistant focusing on understanding and analyzing forearm ultrasound data. You will understand and interpret the data, and I will ask you what class the image belongs to. There are 5 classes - index flexion (1), all pinch (2), hand horns (3), fist (4) and open hand (5). I may show you images from each class so you can understand how the images look like. Your answer should be the class number, and nothing else. Do not tell me you can't determine it.

Here is an image corresponding to class 1

Here is an image corresponding to class 2

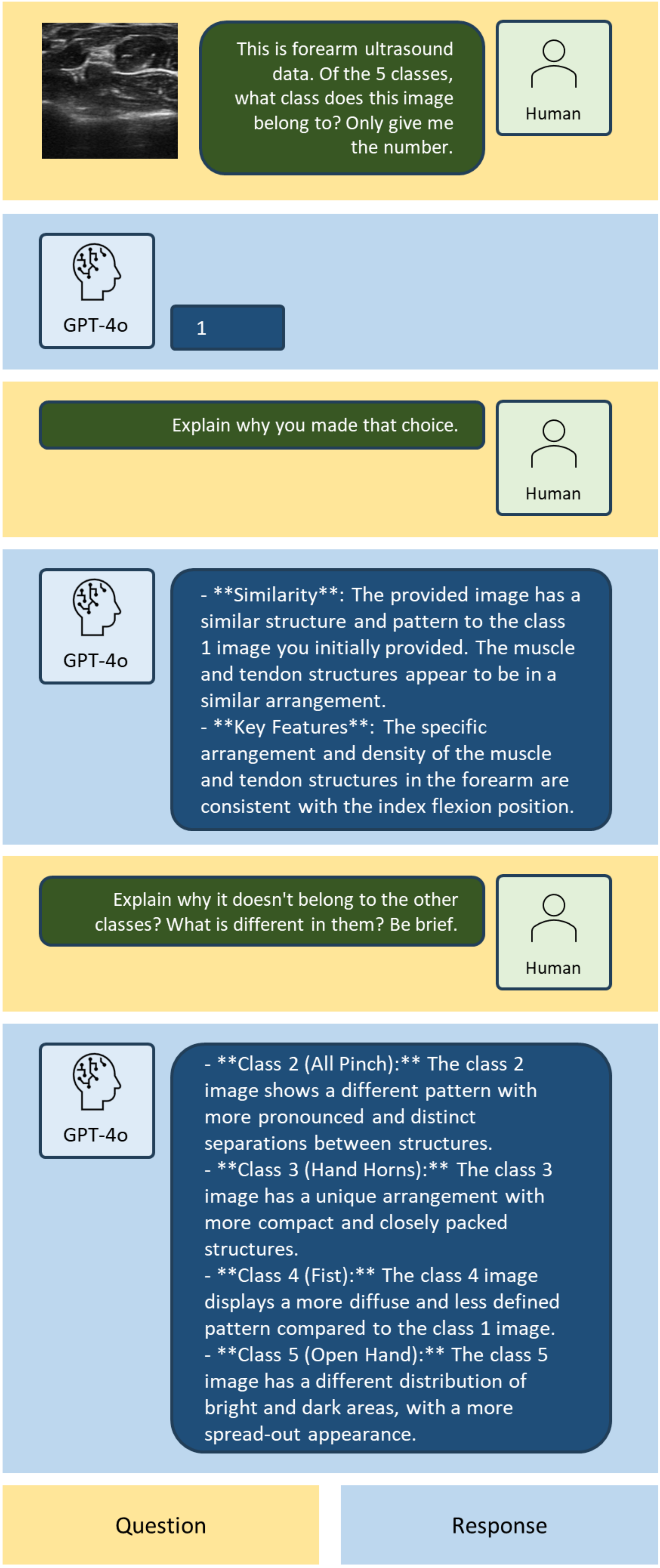
Here is an image corresponding to class 3

Here is an image corresponding to class 4

Here is an image corresponding to class 5

This is forearm ultrasound data. Of the 5 classes, what class does this image belong to?

## VLM Reasoning Capability



This is forearm ultrasound data. Of the 5 classes, what class does this image belong to? Only give me the number.

Explain why you made that choice.

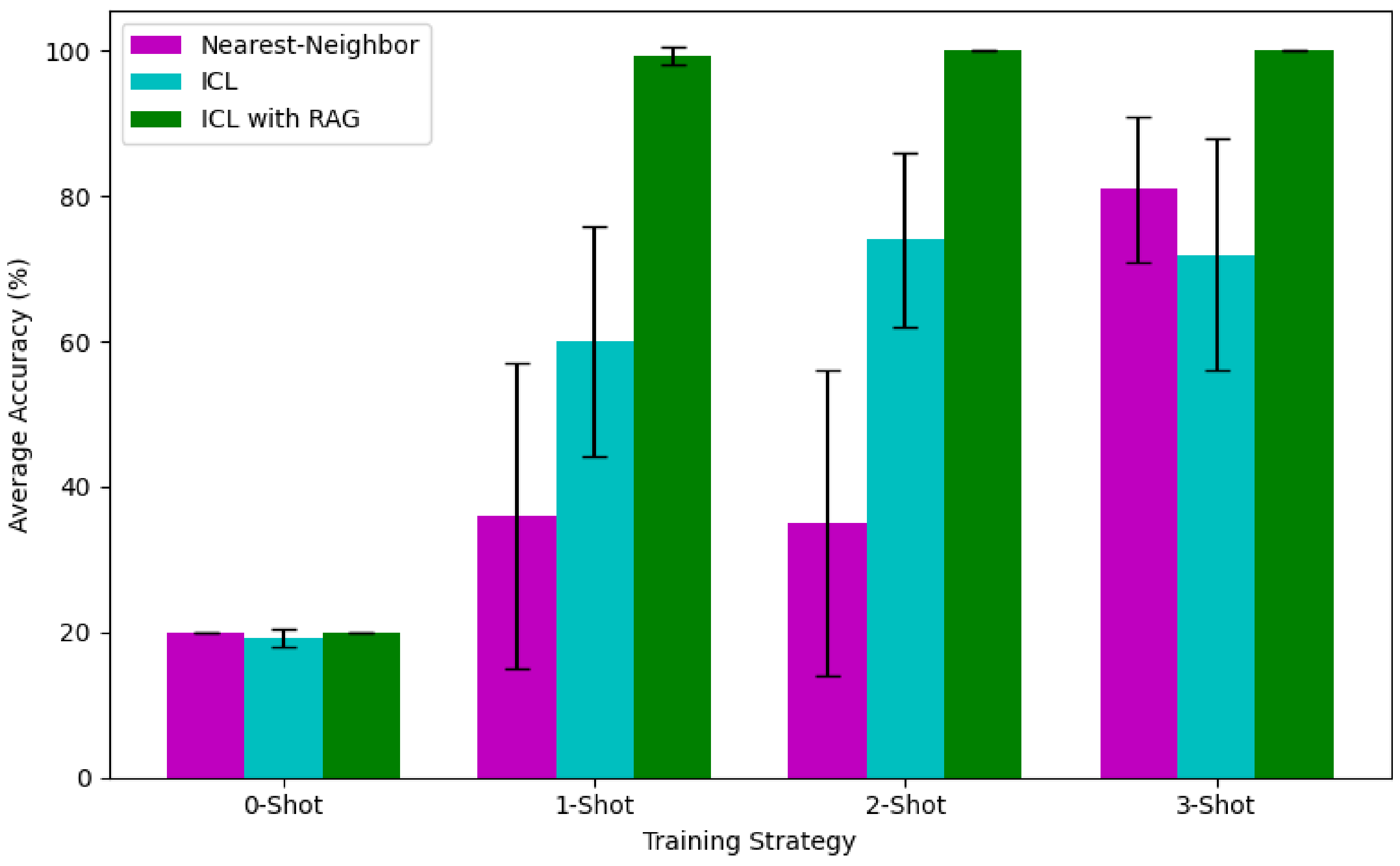
- **Similarity**: The provided image has a similar structure and pattern to the class 1 image you initially provided. The muscle and tendon structures appear to be in a similar arrangement.
- **Key Features**: The specific arrangement and density of the muscle and tendon structures in the forearm are consistent with the index flexion position.

Explain why it doesn't belong to the other classes? What is different in them? Be brief.

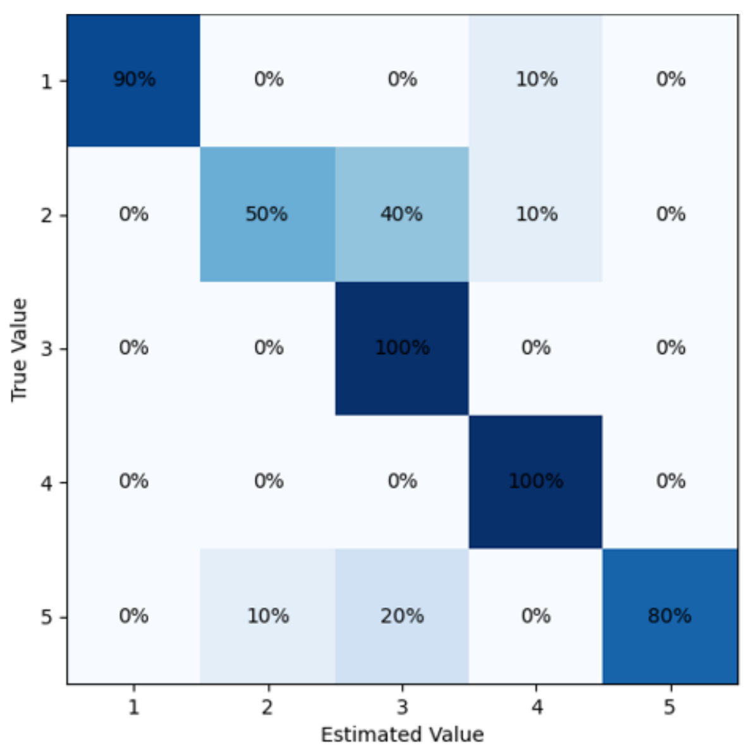
- **Class 2 (All Pinch):** The class 2 image shows a different pattern with more pronounced and distinct separations between structures.
- **Class 3 (Hand Horns):** The class 3 image has a unique arrangement with more compact and closely packed structures.
- **Class 4 (Fist):** The class 4 image displays a more diffuse and less defined pattern compared to the class 1 image.
- **Class 5 (Open Hand):** The class 5 image has a different distribution of bright and dark areas, with a more spread-out appearance.

## Experiments

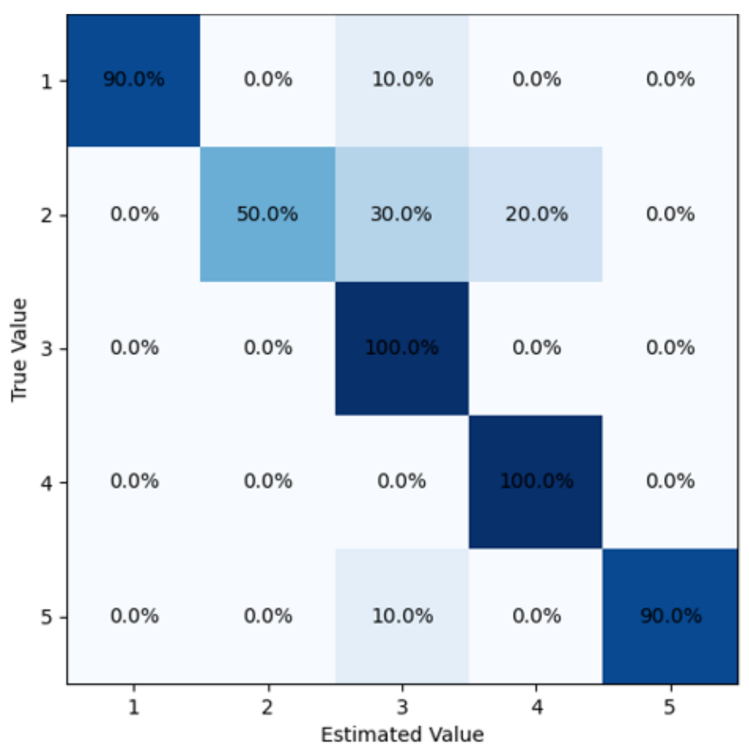
- RAG: Top-$K$ images in cosine-similarity scores.



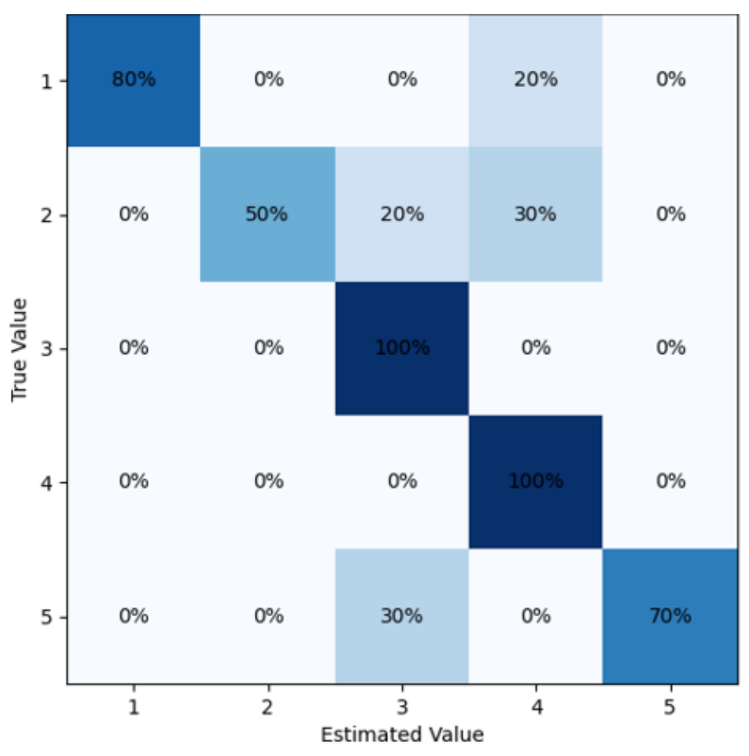
- Low-descriptive prompt: No system prompt; no gesture category; "What class does the image belong to? Only give the class number."
- High-descriptive prompt: More contextual information; to focus on brightness, anatomical and physiological properties.



(a) Low-Descriptive Prompt (82% Acc.)



(b) Mid-Descriptive Prompt (86% Acc.)



(c) High-Descriptive Prompt (80% Acc.)

**Email: koike@merl.com**