Katherine Chen

1. Is server's confidence always higher than wand's confidence from your observations? What is your hypothetical reason for the observation?

Not always, but in many cases, the server shows higher confidence than the wand. A hypothetical reason could be that the server model is larger and trained with more resources, possibly with more complex architecture or data augmentation. On the other hand, the wand's edge model is smaller and optimized for speed and low power, which might reduce its confidence or accuracy.

2. Sketch the data flow of this lab

Data Flow:

[Sensor (MPU6050)]

 ↓

[Wand device collects gesture data]

 ↓

[Run inference locally (Edge Impulse model)]

 ↓

[If confidence < threshold]

 ↓   ↘

[Send data to Flask Server] [If confident → Display result on LED]

 ↓

[Server runs larger model and returns prediction]

 ↓

[Wand receives prediction and displays result]

3. Edge-first, fallback-to-server approach — Pros and Cons

Aspect	Pros	Cons
Reliance on Connectivity	Works even when offline (edge), only needs a connection when uncertain	Still needs internet for full accuracy when confidence is low
Latency	Fast local inference	Slower when falling back to the server due to network delay
Prediction Consistency	Balances speed and accuracy. Fallback can correct edge errors	Edge and server models might differ slightly in prediction behavior
Data Privacy	Most data stays on the device. Only uncertain cases are sent	Some user data is still sent over the network (potential privacy risk)

4. Name a strategy to mitigate at least one limitation named in question 3

A strategy to mitigate the connectivity limitation is to improve the edge model using model compression or distillation. By training a smaller edge model to mimic the performance of a larger server model, the edge device can make more confident and accurate predictions without needing to send data to the server, thus reducing dependency on internet access.