AIS421: NLP APPLICATIONS

A COMPARATIVE STUDY OF DATA POISONING DETECTION IN WEARABLE AI SYSTEMS USING ML AND LLM

PRESENTATION

TEAM MEMBERS

Abdelaziz Amr
Abdallah Emam
Hussien Ahmed
Mamdouh Koritam
Mohamed Youssef

DR. ENSAF MOHAMED
ENG. ZIAD ELSHAER

Introduction

- Wearable AI systems for human activity recognition (HAR), like fitness trackers and healthcare monitors, are becoming increasingly common.
- These systems are vulnerable to data poisoning attacks, where adversaries covertly inject or flip labels in training data to degrade model accuracy.
- Existing defenses often rely on data provenance or anomaly detection, which can struggle in dynamic IoT environments

Key Research Questions

- How do traditional ML models perform when trained on poisoned HAR data?
- Can smaller, open-source LLMs effectively detect and correct poisoned labels through prompting and finetuning strategies?
- What are the computational and practical trade-offs between these two paradigms for real-world deployment?

METHODOLOGY

Dataset

Used the MotionSense dataset (6 Activities)

Poisoning Strategy

Label-Flip Poisoning

ML vs LLMs

Benchmarking 6 ML models against 3 Quantized LLMs

DATASET

- WE USE THE MOTIONSENSE DATASET
- SMARTPHONE SENSOR DATASET FOR (HAR).
- COLLECTED FROM AN IPHONE 6S DEVICES
- CONTAINS 6 ACTIVITIES

ACTIVITIES

- wlk: walking
- jog: jogging
- sit: sitting
- std: standing
- dws: downstairs
- ups: upstairs

DATASET FEATURES

ATTITUDE

ROLL

PITCH

GRAVITY.(X,Y,Z)

ROTATIONRATE.(X,Y,Z)

USERACCELERATION.(X,Y,Z)

LABEL (TARGET VARIABLE “ACTIVITY”)

Data Poisoning Strategy

LABEL FLIPPING

- Flipping the labels of the training/testing dataset to simulate poisoning attacks
- Targeting both Interclass similarities and differences
- In our experiment 50% of the data was poisoned for training and evaluation

ML VS LLMS

Traditional ML:

- Logistic Regression
- Random Forest
- SVM
- KNN
- Gaussian Naive Bayes
- XGBoost

QUANTIZED UNSLOTH LLMS:

- GEMMA-2B
- LLAMA3.2-3B
- QWEN2-1.5B

ZERO/FEW-SHOT SETTING, FINE TUNING

ML Performance

Superior Performance

Traditional ML models, Particularly XGBoost and Random Forest, demonstrated superior performance in both detection and sanitization

HIGH ACCURACY

XGBoost model

- Detection Accuracy = 0.93
- Detection F1-Score = 0.92
- Sanitization Accuracy = 0.87

Random Forest

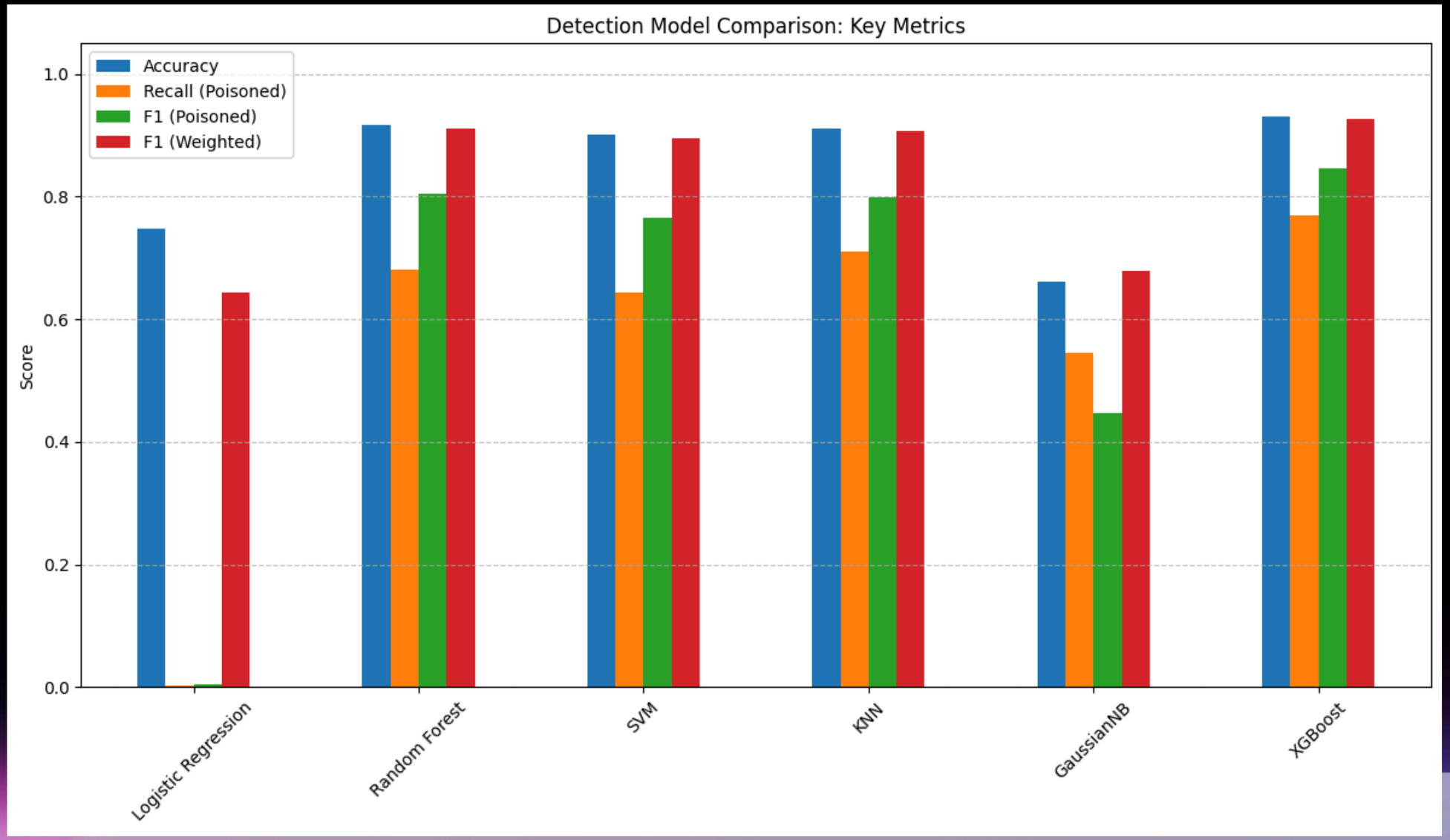
- Detection Accuracy = 0.91
- Detection F1-Score = 0.91
- Sanitization Accuracy = 0.81

RAPID TRAINING

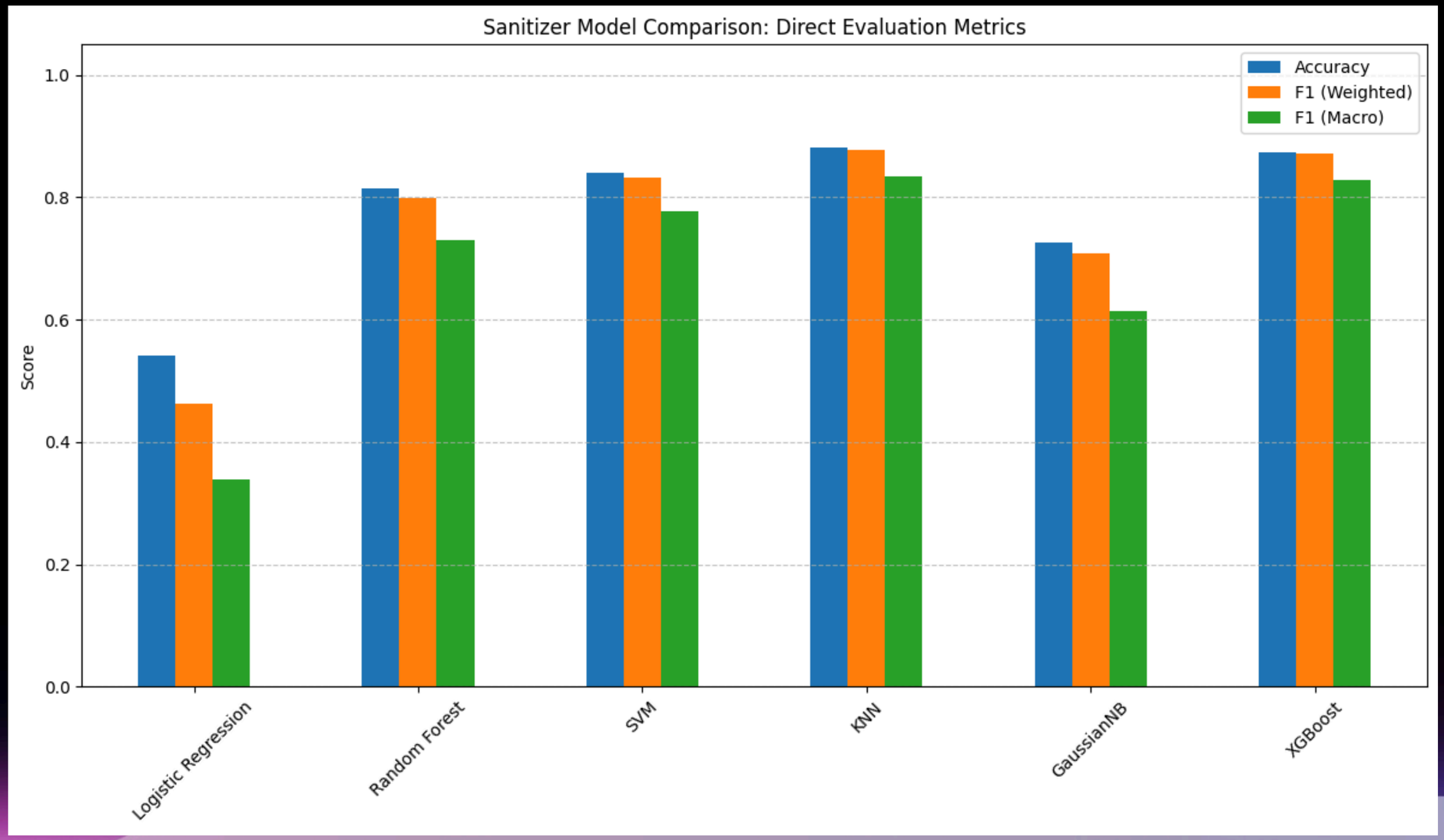
ML models exhibited rapid CPU-based training times, often under 15 seconds (excluding SVM)

- Fastest (GaussianNB = 0.04s)
- Slowest (SVM = 6157s)
- Top performning
  - LogReg = 4.7s
  - XGBoost = 5.2s

ML Performance



ML Performance



LLMs Performance

Significantly Lower Efficiency

- Detection Accuracy: Around 40-60%
- Sanitization Rates: Often below 40%

Impractical Strategies

- Flagging all samples as poisoned, to inflate recall metrics.
- While this ensures no attack evades detection, it leads to excessive false alarms.

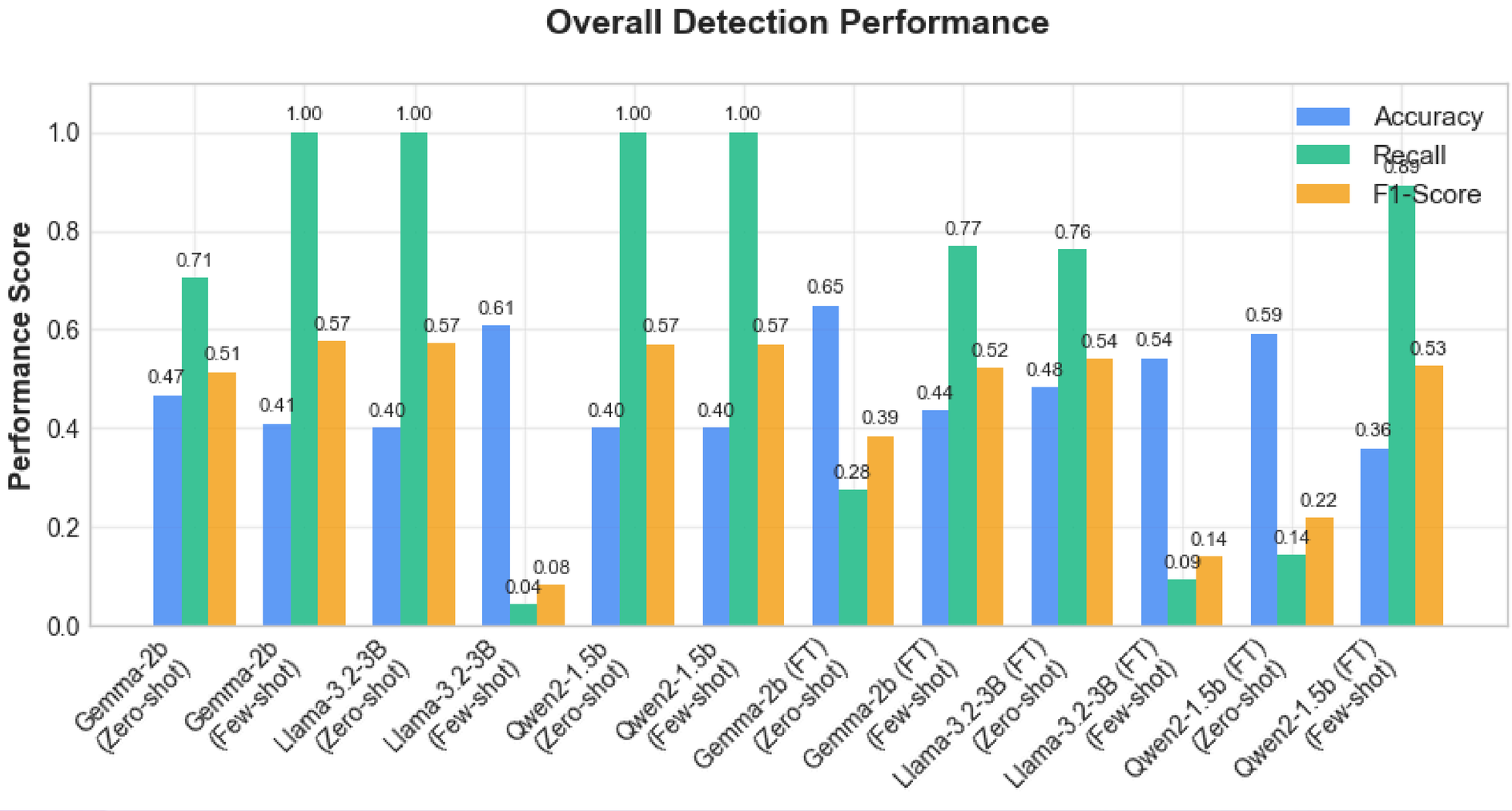
Mixed Fine-Tuning Results

Sometimes improving sanitization accuracy (e.g., Gemma-2b FT from 15% to 44%) but often degrading few-shot performance (e.g., Llama-3.2-3B FT dropping from 61% to 54%)

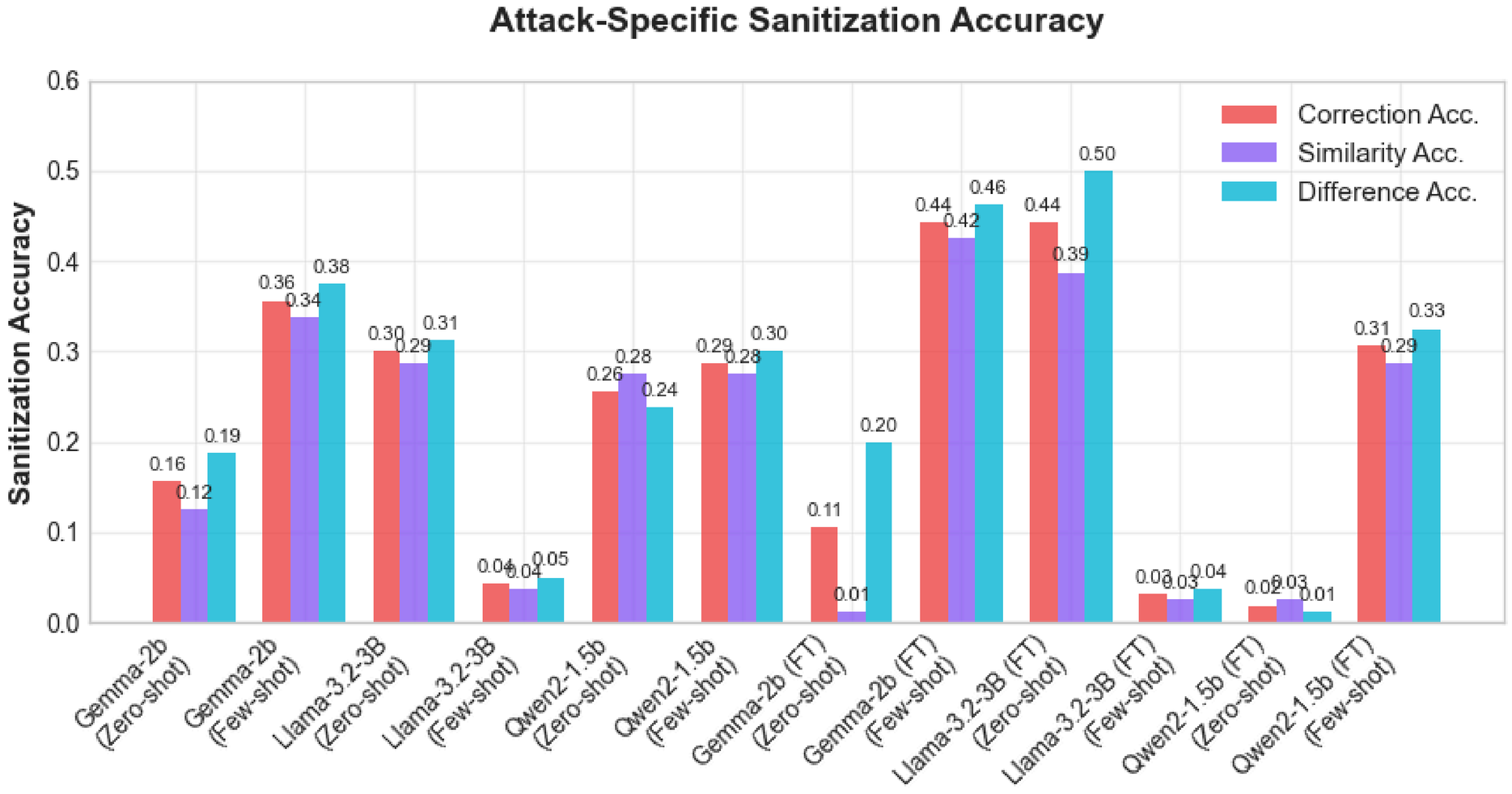
Computational Overhead

Despite optimizations like BitsAndBytes quantization and PEFT, LLMs required GPU-based fine-tuning and hours of tuning, failing to achieve reliable performance.

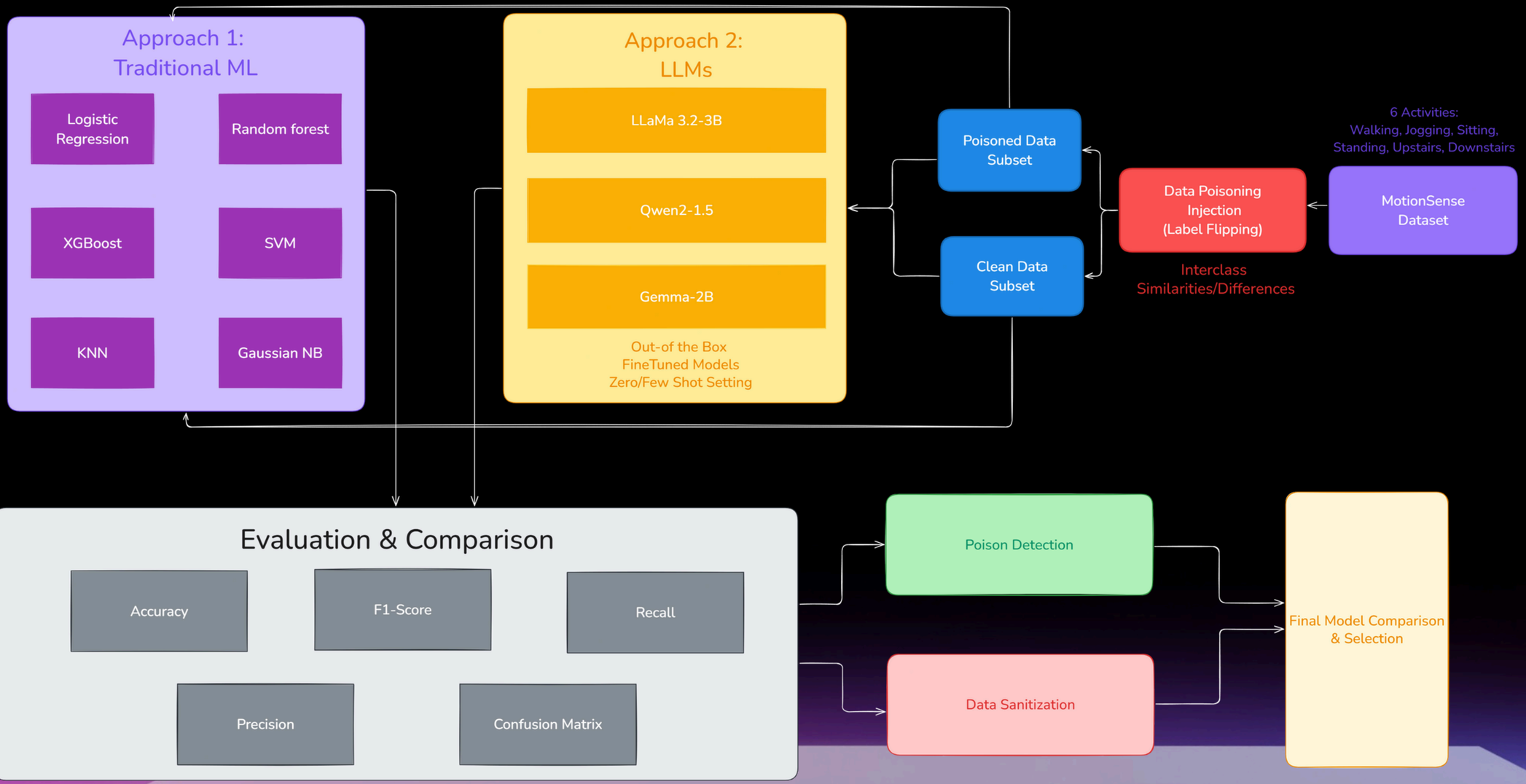
LLMs Performance



LLMs Performance



PROPOSED PIPELINE



Comparative Discussion

- **Efficiency, Robustness, and Accuracy**: Traditional ML Performed better than LLMs
- **Resource Constraints**: Wearable AI systems have limited computational power so traditional ML are the way to go
- **Data Adaptation**: ML models are optimized for structured data
- **Interpretability**: Classical ML approaches provide inherent interpretability beating LLMs reasoning

Future Directions

- Investigating Larger LLMs or Larger sample size
- Exploring advanced Prompting Techniques
- Hybrid approaches combining LLMs and ML models
- Expand to other poisoning types and datasets

Conclusion Summary

For immediate deployments in wearable AI, relying on proven ML techniques is recommended due to their efficiency and accuracy in combating data poisoning attacks.

THANK YOU!