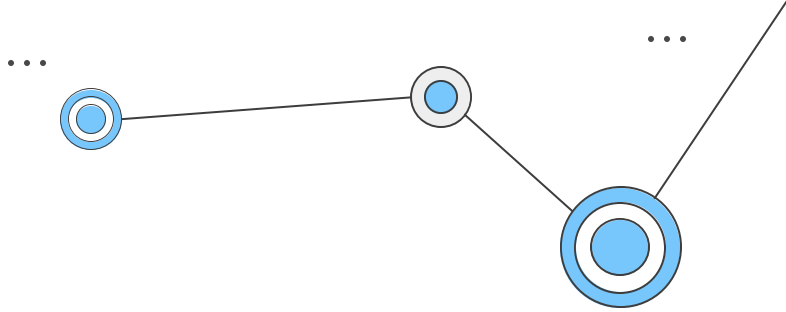
38th IEEE/ACM International Conference on Automated Software Engineering (ASE 2023): NIER Track



Towards Self-Adaptive Machine Learning-Enabled Systems Through QoS-Aware Model Switching

Authors: Shubham Kulkarni, Arya Marda, Karthik Vaidhyanathan,
Software Engineering Research Center, IIIT Hyderabad, India

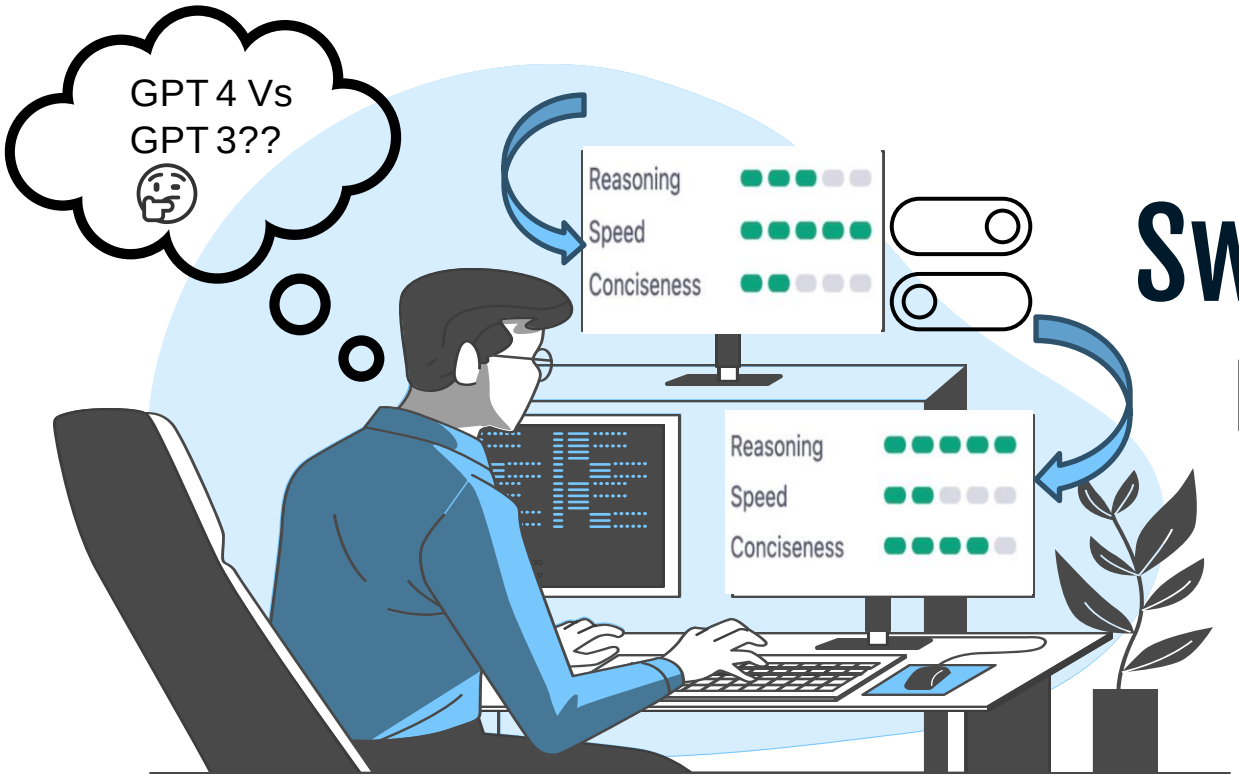
Software Engineering Research Centre

Microsoft Research

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

IIIT Hyderabad

ML Trade Off:
Speed vs. Accuracy



Can we **Switch Models** in runtime to have best of both worlds ?

Yes, we say you should switch, but why? 🤔
Because of System, Model & Environment Uncertainties!

01

Model Uncertainties

Models abstract real-world data, leading to potential inaccuracies

02

Environment Uncertainties

Varying and unpredictable incoming data requests challenge consistent performance

03

System Uncertainties

Resource constraints and latency issues challenge system performance

So

The Need for Adaptability

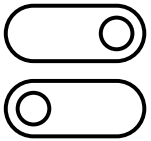
Can we self-adapt system in real-time for optimal outcomes?

Introducing : The ML Model Balancer

The Heart of Dynamic Model Switching

Dynamic Evaluation

Assessing models in real-time scenarios

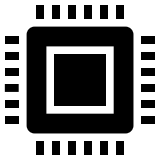


Seamless Switching

Transitions between models in real-time, minimizing latency and ensuring optimal outcomes

Overcoming Model Limitations

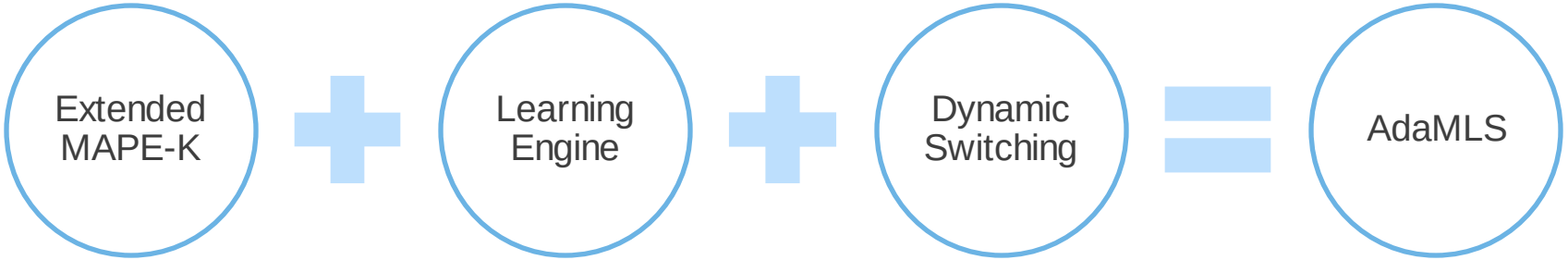
By leveraging multiple models, it mitigates the weaknesses of any single model



Prelude to AdaMLS

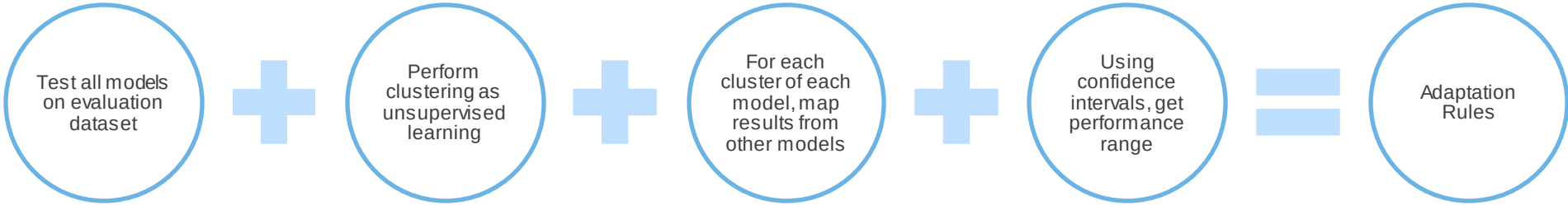
The foundational concept that **AdaMLS** builds upon for software architecture-driven adaptability

AdaMLS : Our Novel Self Adaptive approach



AdaMLS : Design and Working

AdaMLS : Our Novel Self Adaptive approach

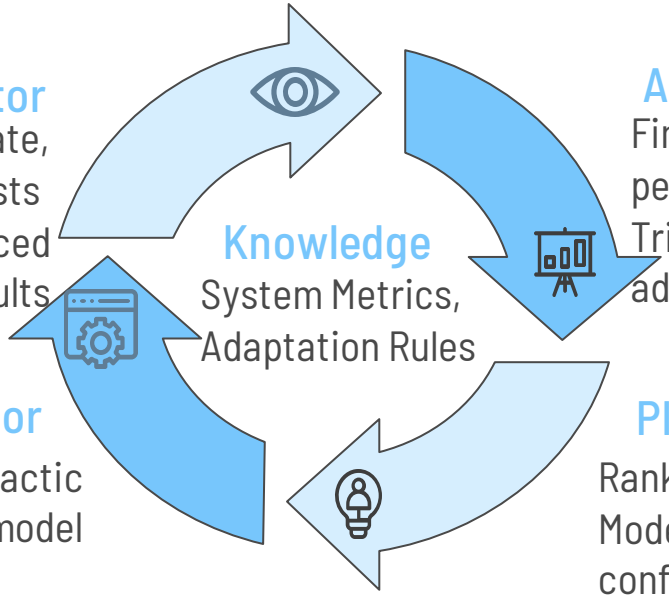


Learning Engine Functioning : Extracting Adaptation Rules

AdaMLS : Our Novel Self Adaptive approach:
Extended MAPE-K + Lightweight Unsupervised Learning + Dynamic Switching

Learning Engine

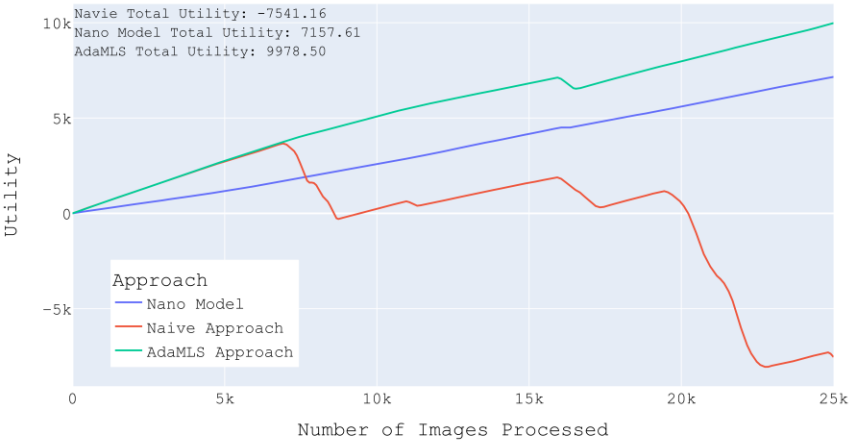
Unsupervised Learning Abstracting Performance



Machine Learning Enabled System

Dynamic Model Switching in Run-Time

Demonstrated Results: Object Detection Service



Utility : Way to represent Quality of Service, a function of speed and accuracy

AdaMLS vs. Others

Surpasses naive approach & single models

Utility Achiever

Up to 39% improvement over Yolov5n (Second Best)

Fast Model Transitions

Transition time < 0.01 s

Broad Applicability & Future Promise

Universal Fit

Not just object detection; AdaMLS fits any ML system

...

Addressing Real-World Challenges

Tackles uncertainties in dynamic environments.

...

Setting New Benchmarks

Redefining QoS in ML-driven systems, Also making then Sustainable

...

The Future is Adaptive

Embrace change; enhance performance

...

Thanks!

Do you have any questions?

Shubham Kulkarni
shubham.kulkarni@research.iiit.ac.in
Software Engineering Research Center,
IIIT Hyderabad, India

Microsoft Research

SERC
Software Engineering Research Centre

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIIT Hyderabad