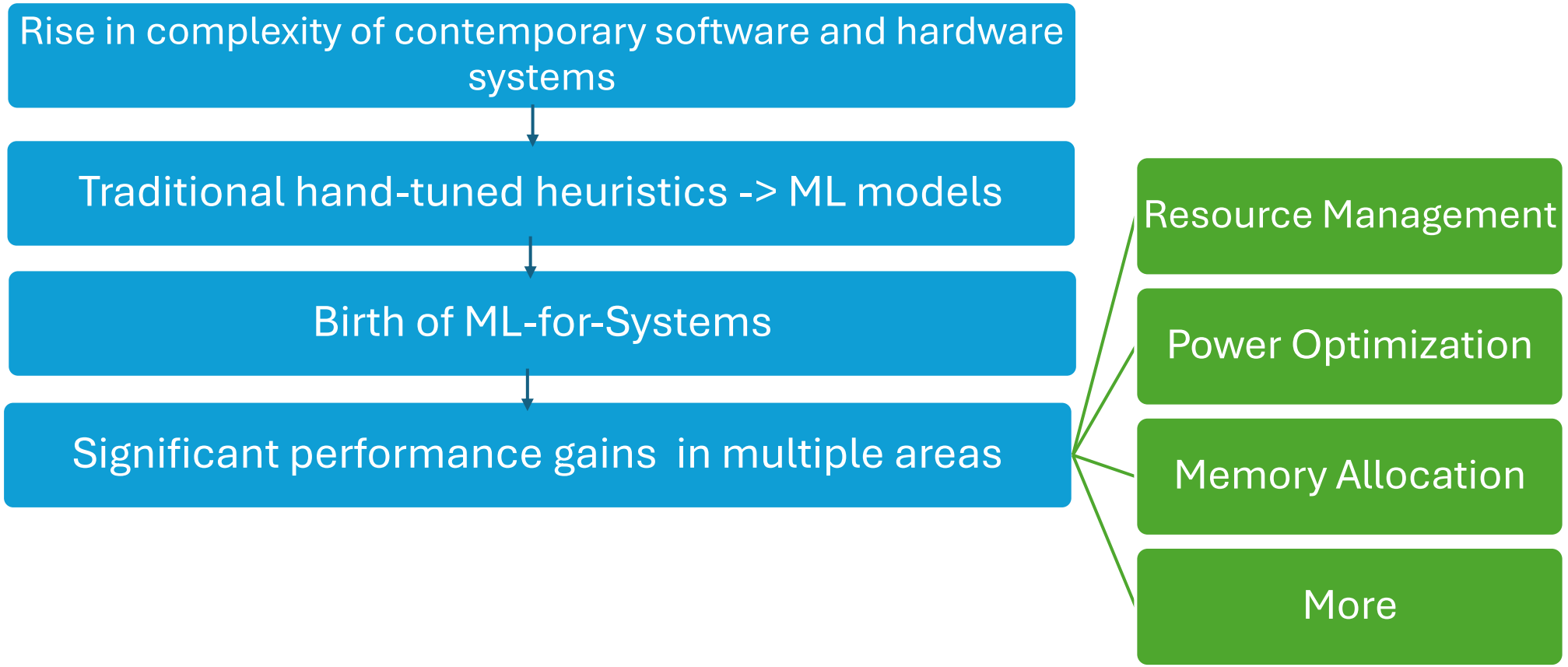
The Case for Uncertainty-Governed Predictor Hierarchies in ML for Systems

Christos Zarkos*, Nevena Stojkovic*, Varun Gohil, Christina Delimitrou

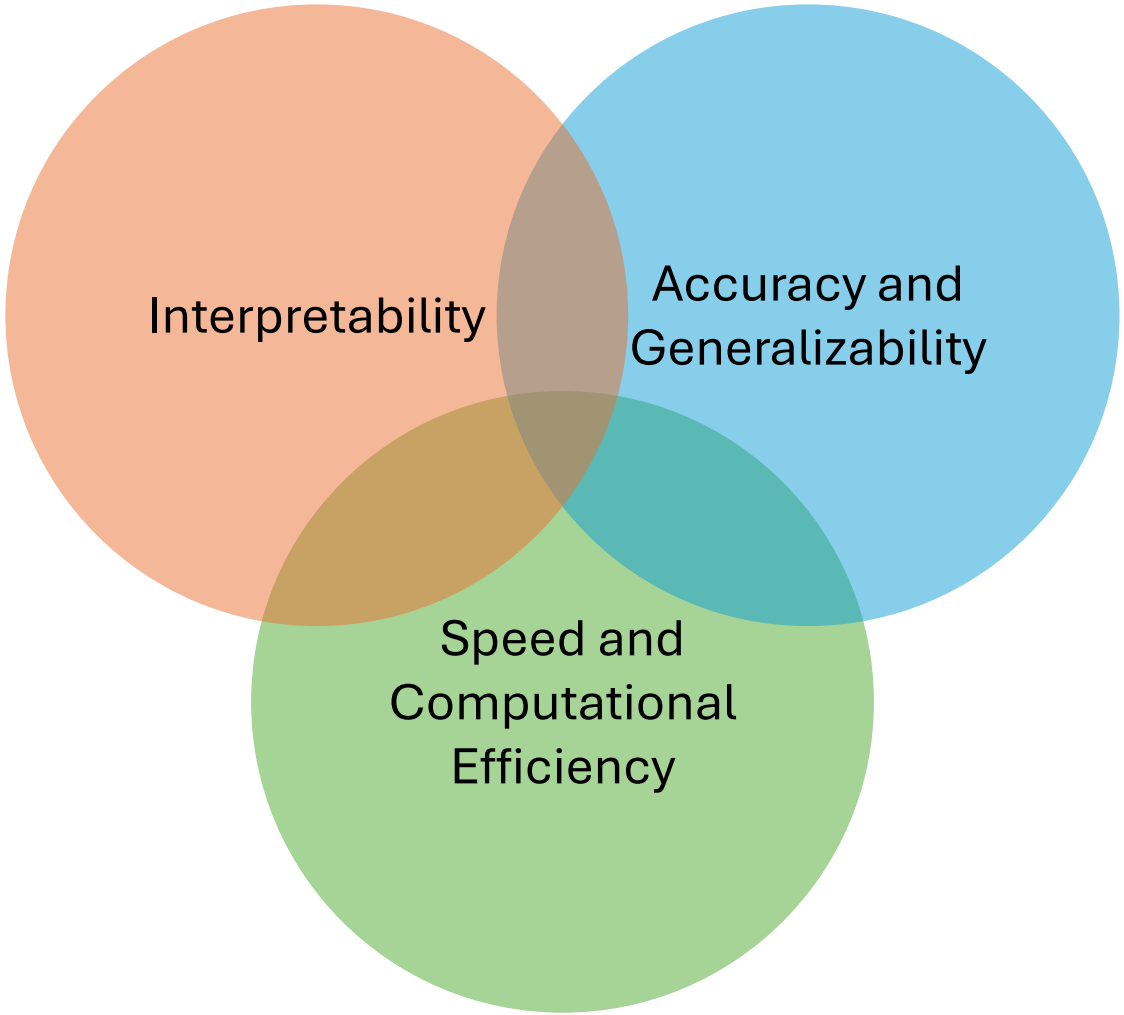
MIT

ML-for-Systems



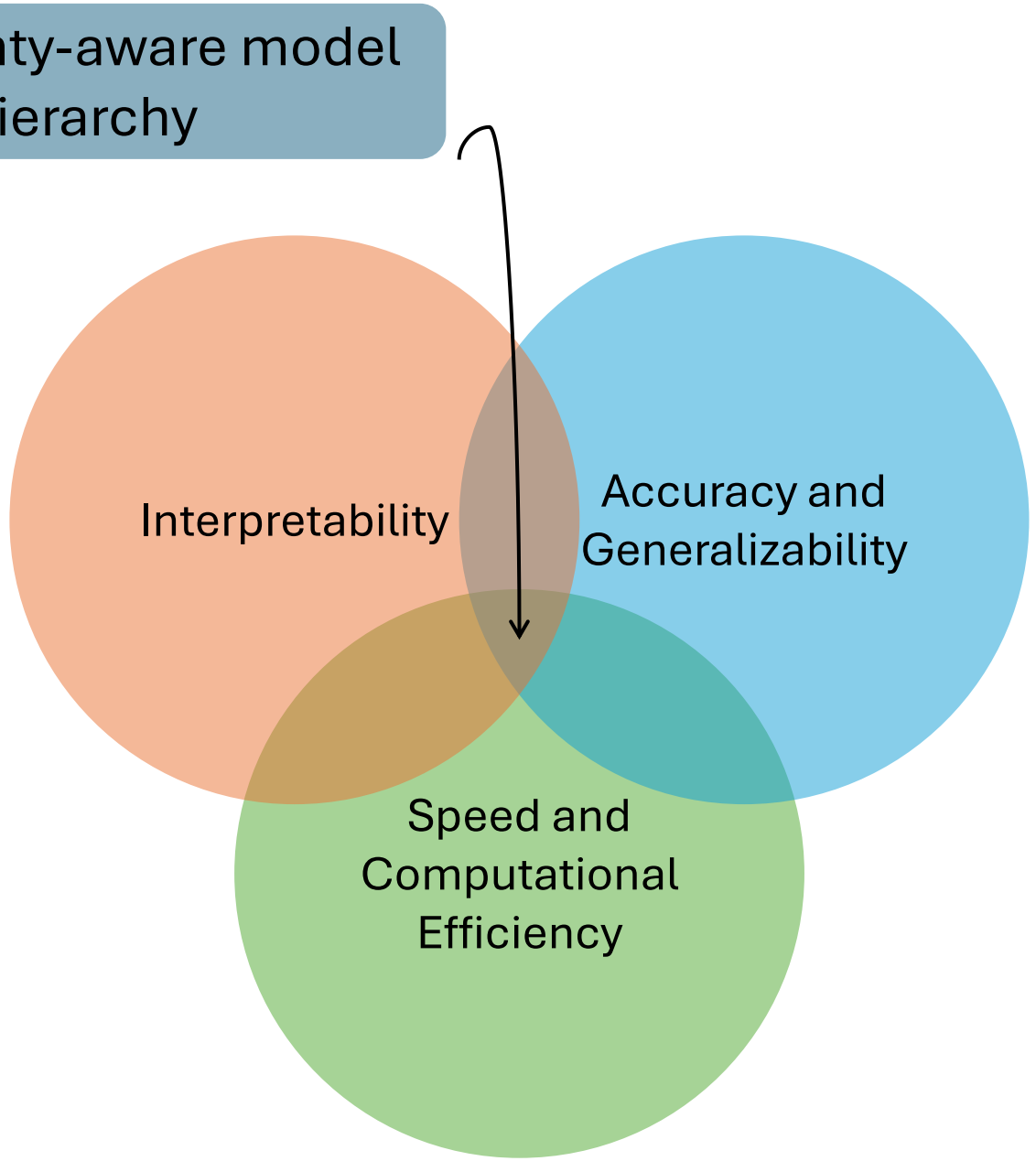
Motivation/Problem

- Traditional ML base models
  - Generalizability, accuracy
  - Interpretability, speed
- (Interpretable) Surrogate models
  - Interpretability, speed
  - Generalizability and accuracy
- Uncertainty-awareness
  - Generalizability
  - Interpretability, computational efficiency



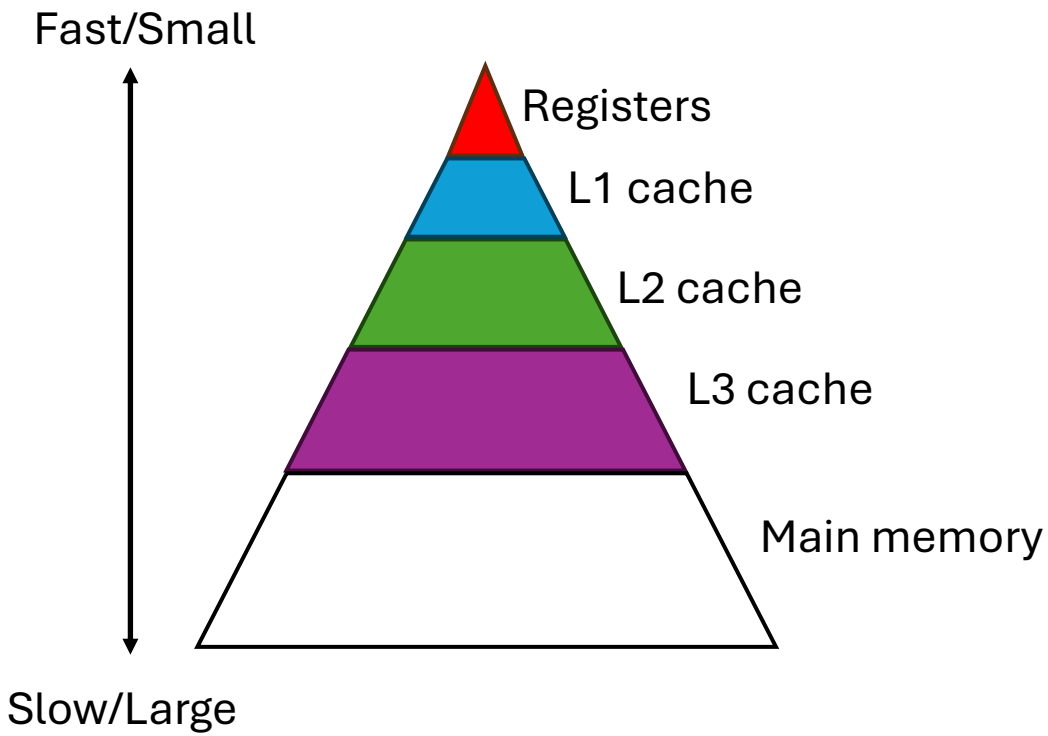
Motivation/Problem

- Traditional ML base models
  - Generalizability, accuracy
  - Interpretability, speed
- (Interpretable) Surrogate models
  - Interpretability, speed
  - Generalizability and accuracy
- Uncertainty-awareness
  - Generalizability
  - Interpretability, computational efficiency



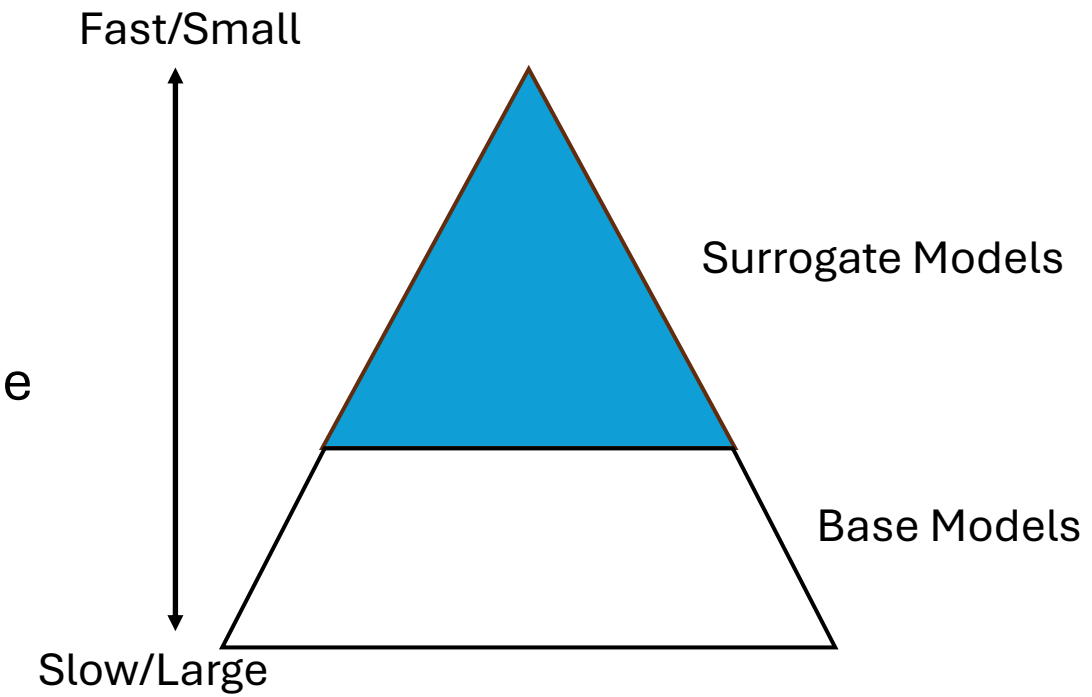
Traditional Systems Approach: (Memory) Hierarchy

- Fundamental trade-off:
  - Fast but small memory tiers vs. Large but slow memory tiers
- Solution:
  - Multi-tier memory hierarchy
- Fallback mechanism:
  - On miss query the tier right above you



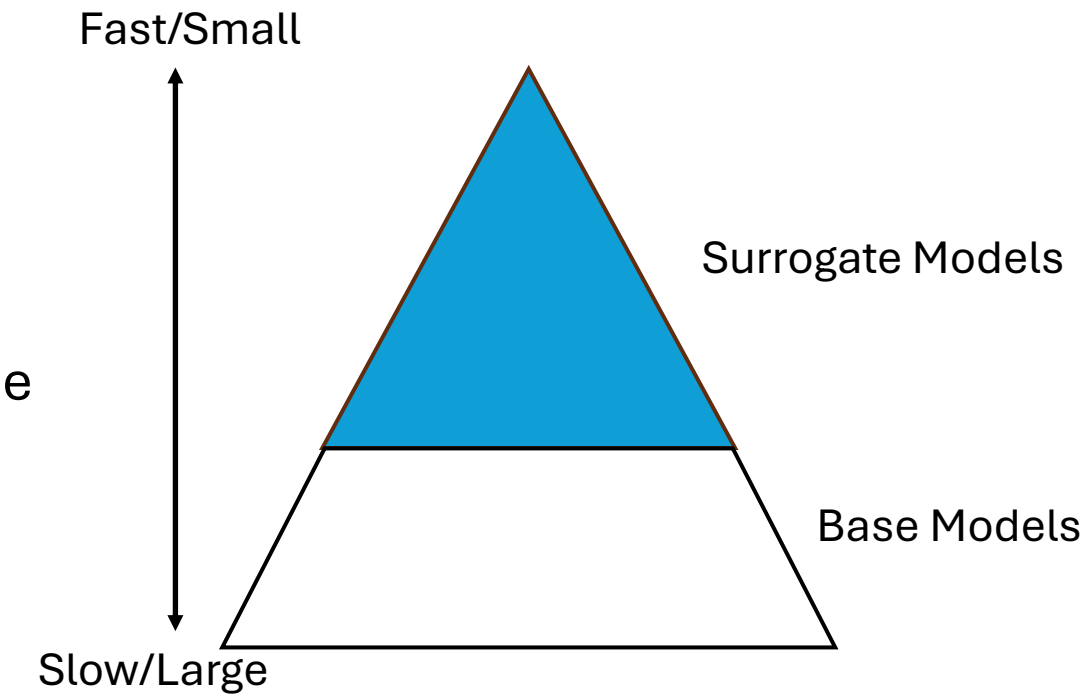
Correspondence and the Challenge

- Fundamental trade-off:
  - Fast (and interpretable) smaller (surrogate) models vs. Larger, slower but more accurate and generalizable models
- Solution:
  - Multi-tier model hierarchy
- Fallback mechanism?
  - Not straightforward
  - Ground truth is unavailable at inference time

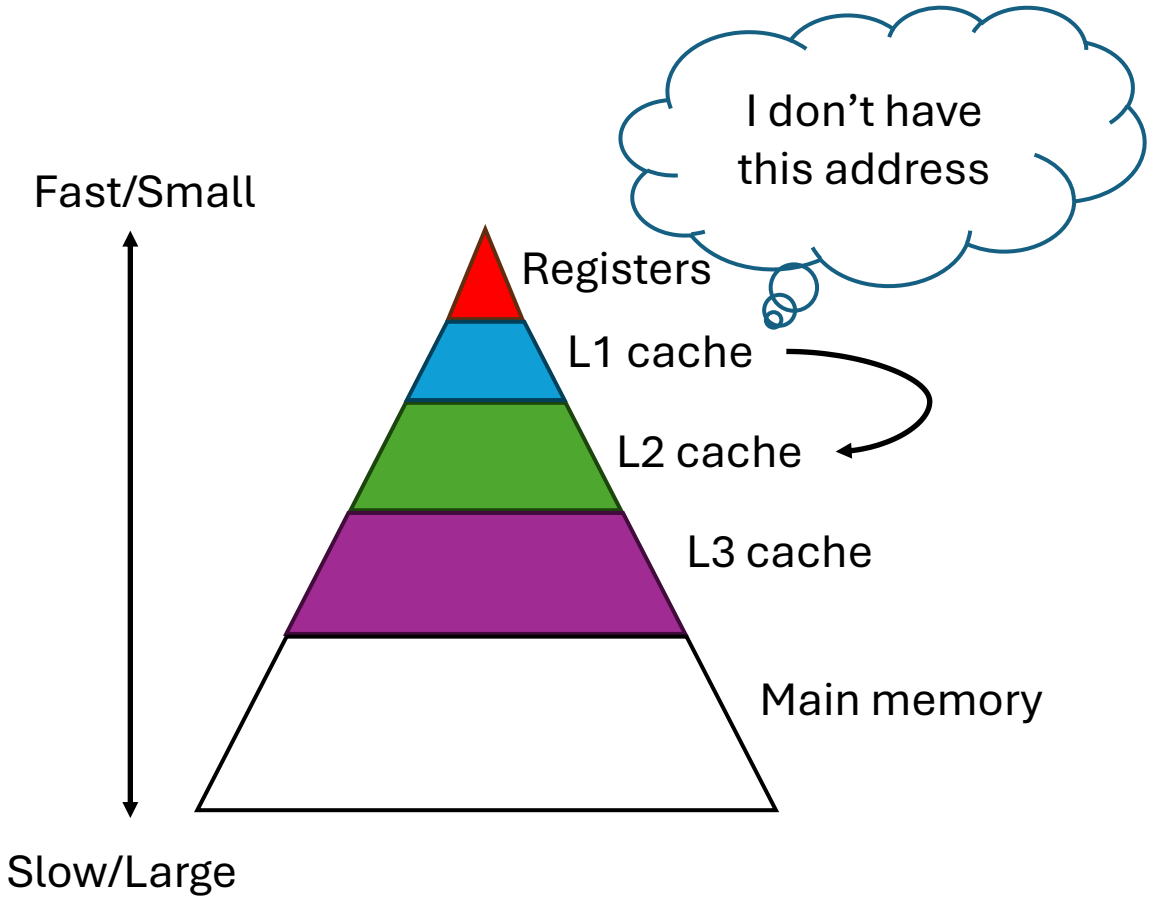


Correspondence and the Challenge

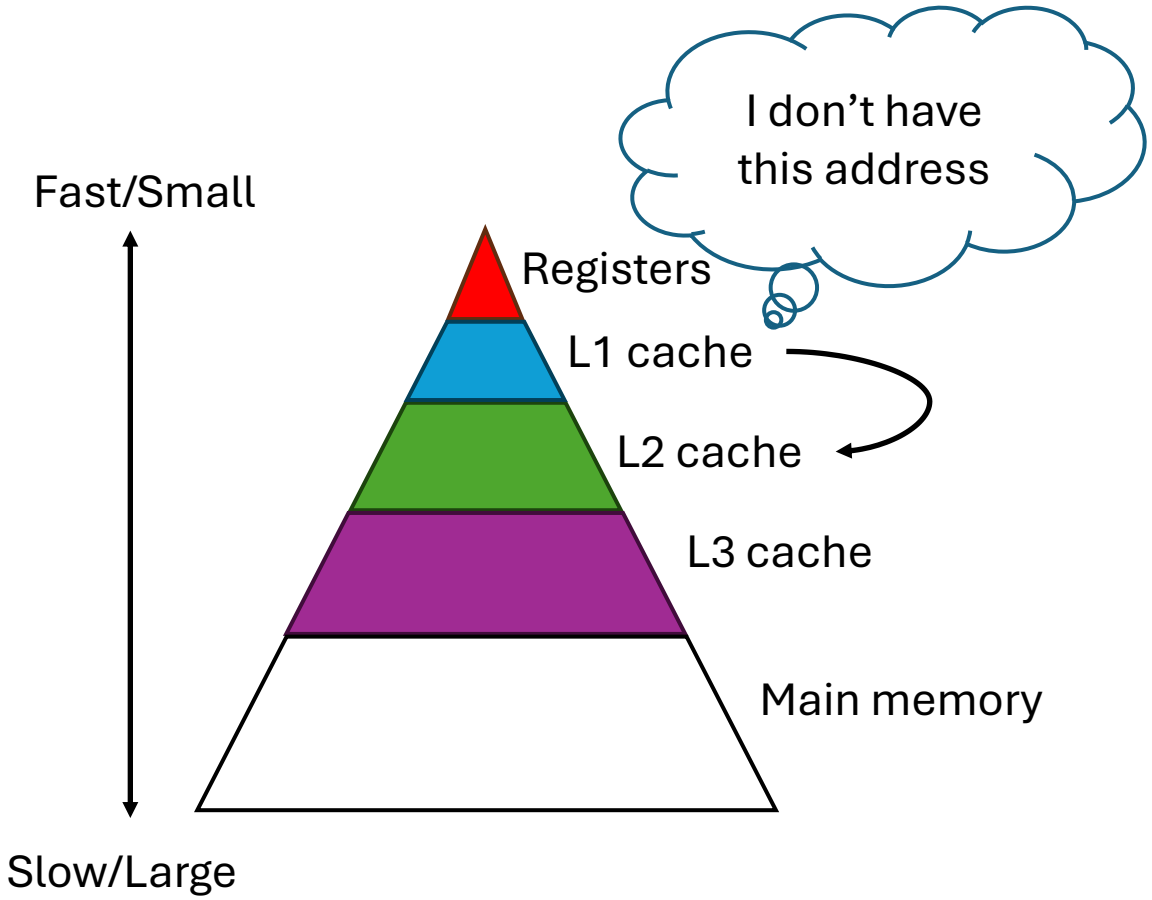
- Fundamental trade-off:
  - Fast (and interpretable) smaller (surrogate) models vs. Larger, slower but more accurate and generalizable models
- Solution:
  - Multi-tier model hierarchy
- Fallback mechanism?
  - Not straightforward
  - Ground truth is unavailable at inference time
  - What should we consider a “miss” here?

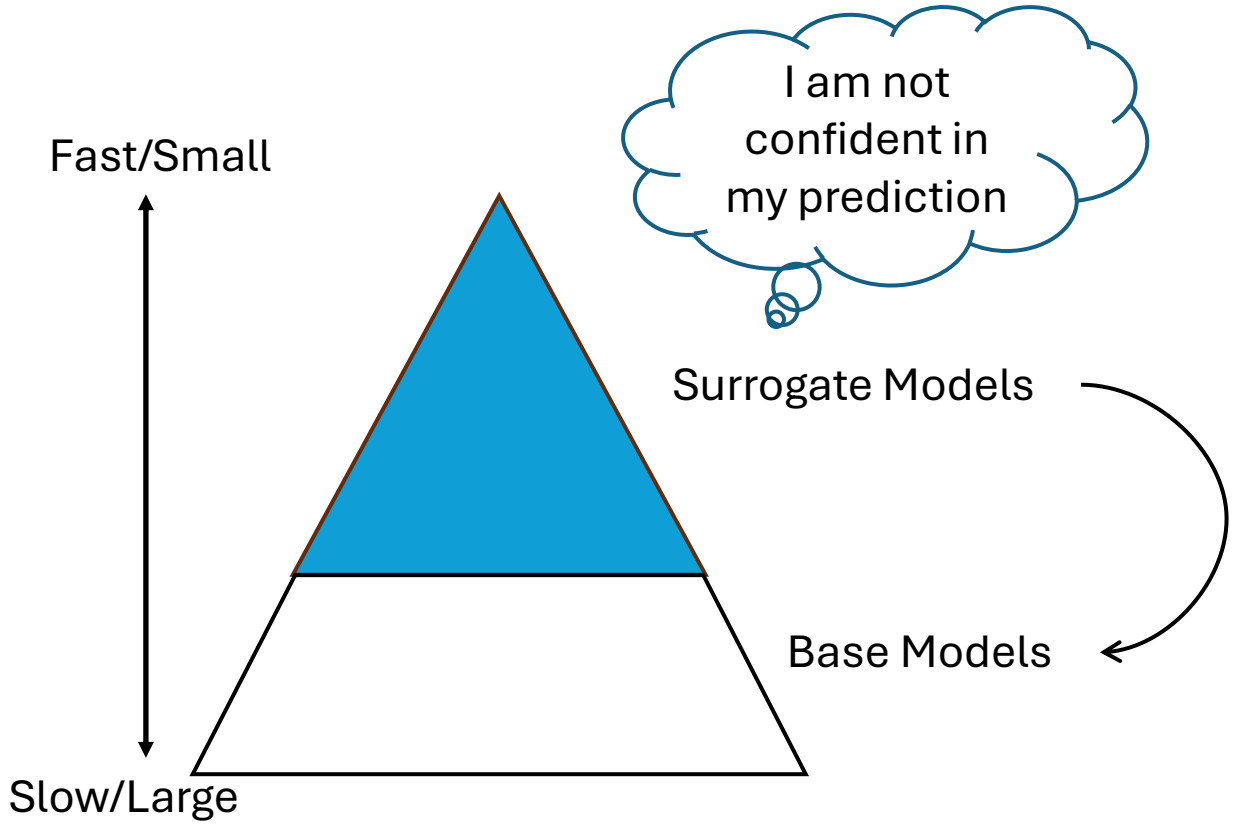


What should we consider a miss?

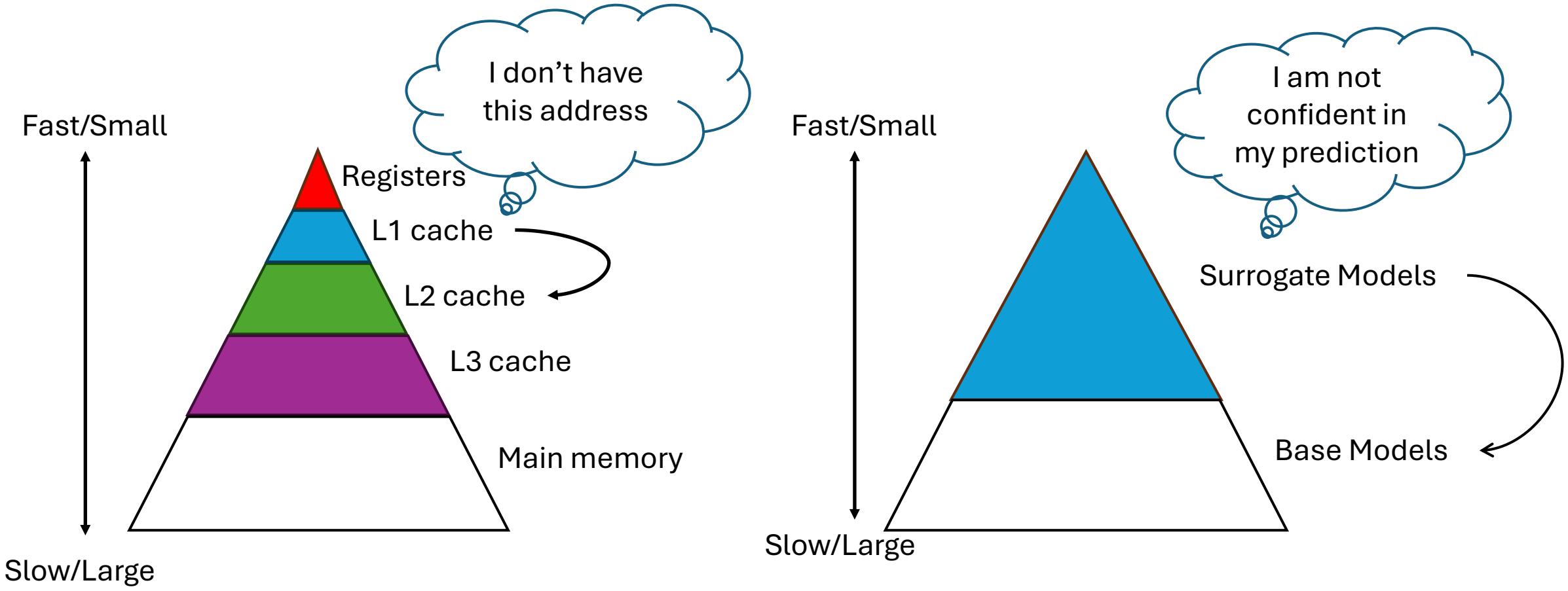


What should we consider a miss?





What should we consider a miss?

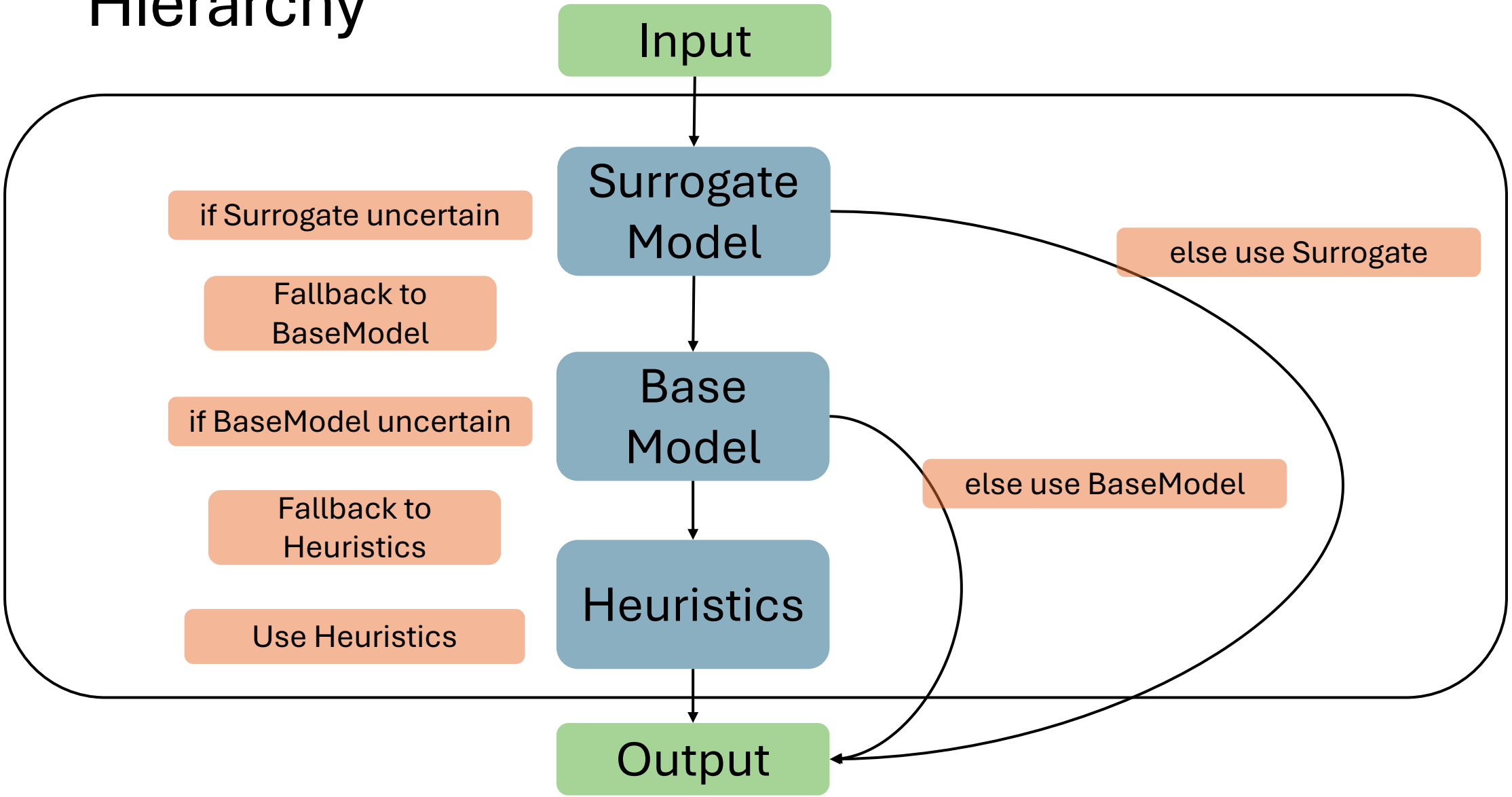


Uncertainty Awareness

Proposed Solution

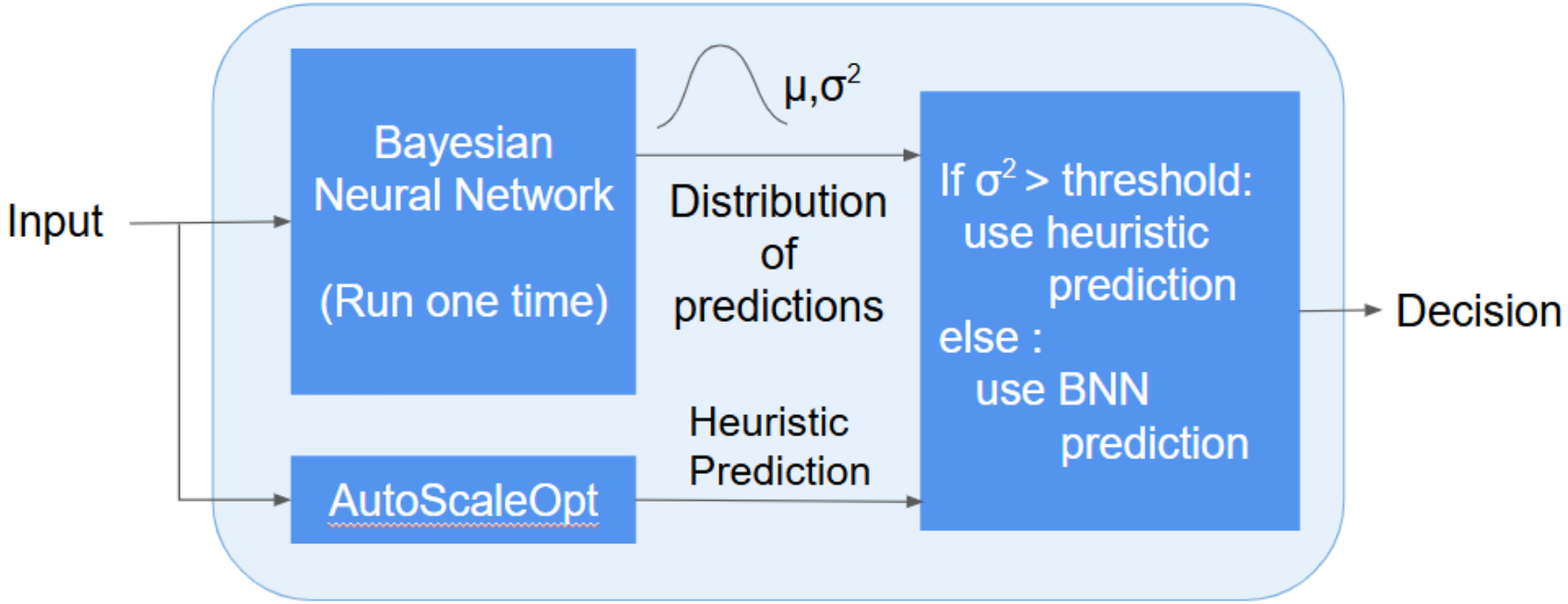
- Design an uncertainty-aware model hierarchy:
  - Accurate traditional base models
  - Fast and interpretable surrogate models
  - Uncertainty-awareness to decide fallback
- Get the best of both worlds
  - Fast, cheap, and accurate predictions in the common case
  - Do not compromise accuracy when surrogate does not work well

Abstract Model Hierarchy

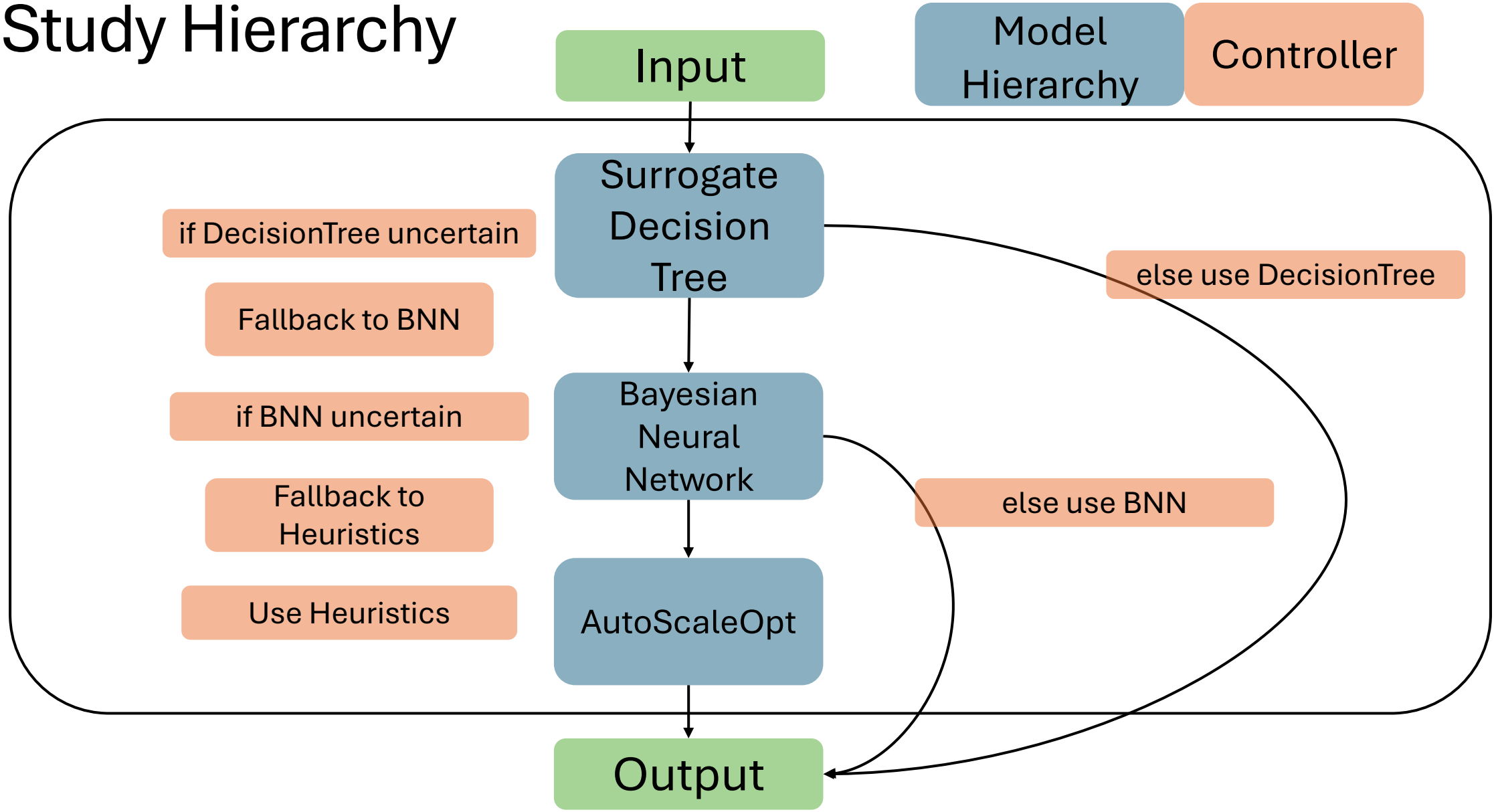


Case Study: Microservice cluster management

- ML-driven resource manager for microservices
  - Takes decisions based on the output of an ML model
  - Uncertainty-aware fallback to heuristics

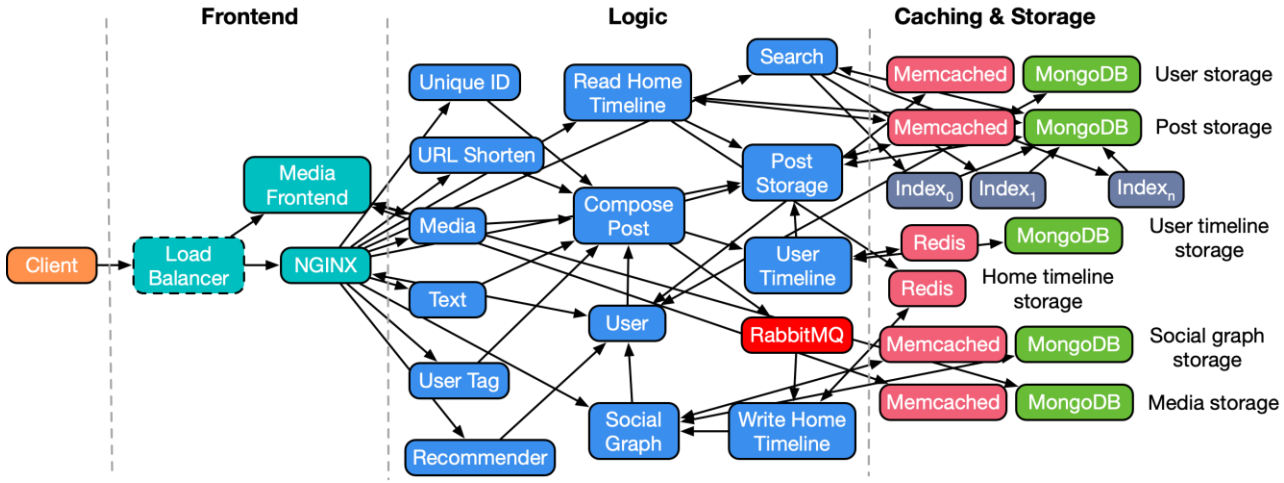


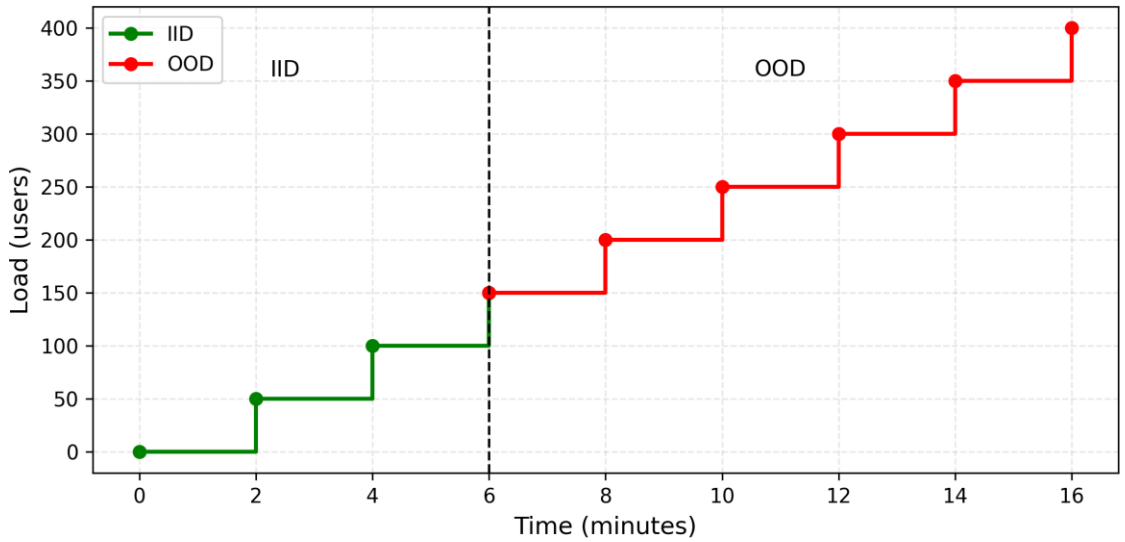
Implemented Case Study Hierarchy



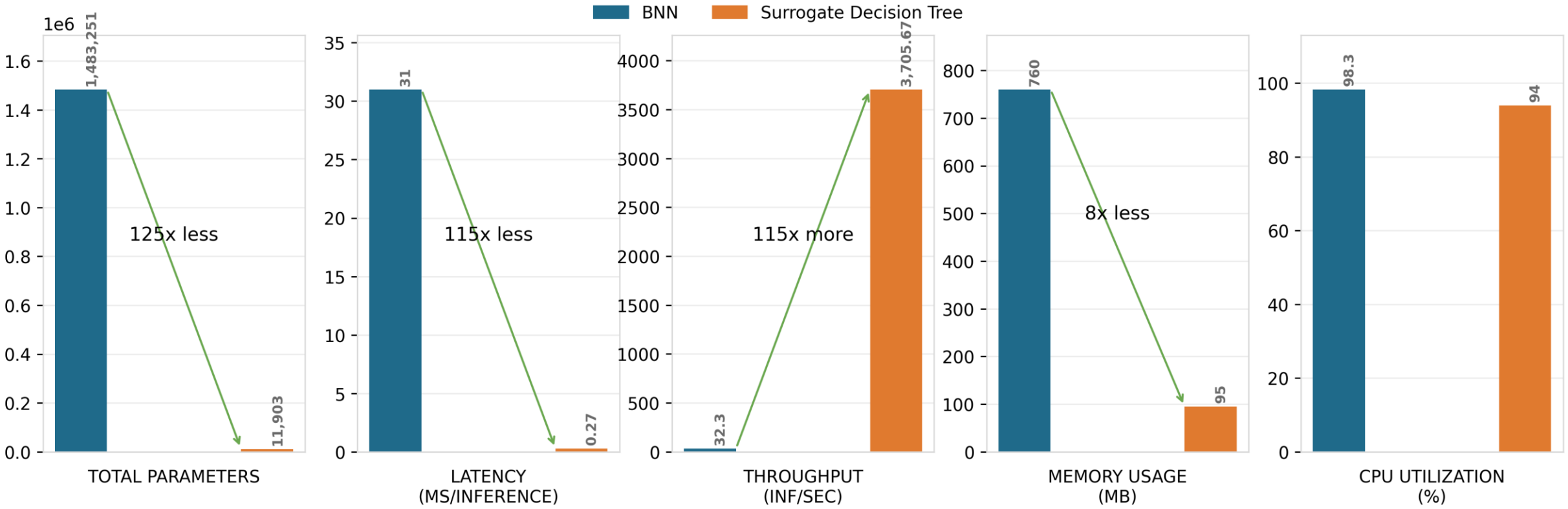
Evaluation Setup

- SocialNetwork application from DeathStarBench



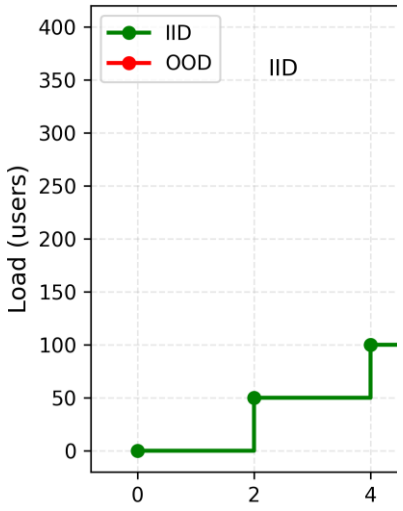


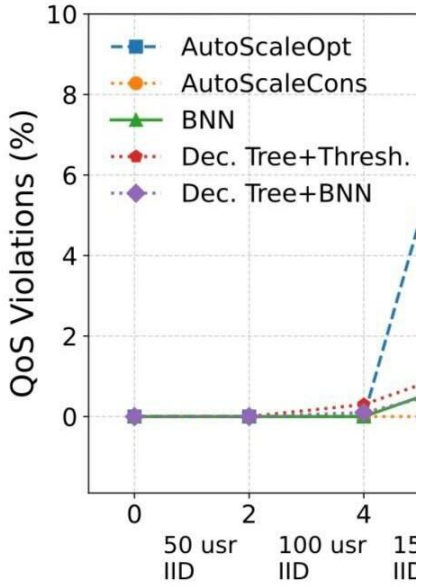
BNN and Surrogate Decision Tree Comparison

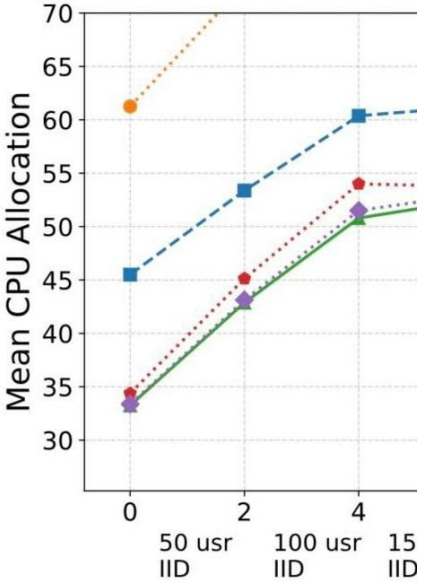


- Decision tree achieves 94% fidelity!
- Decision tree used for 87%-96% of decisions
- Average hierarchy decision latency: 4.27ms (from 31ms)

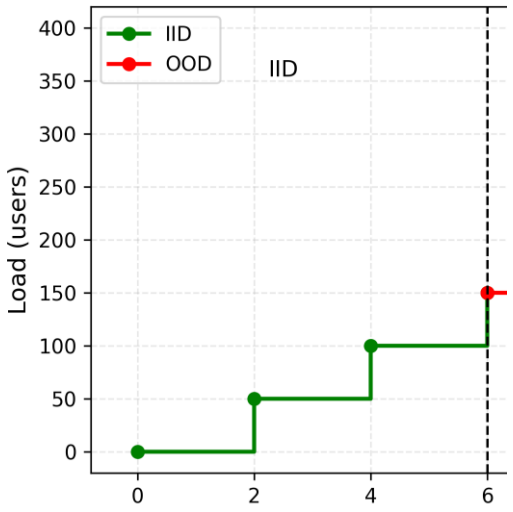
Evaluation Results

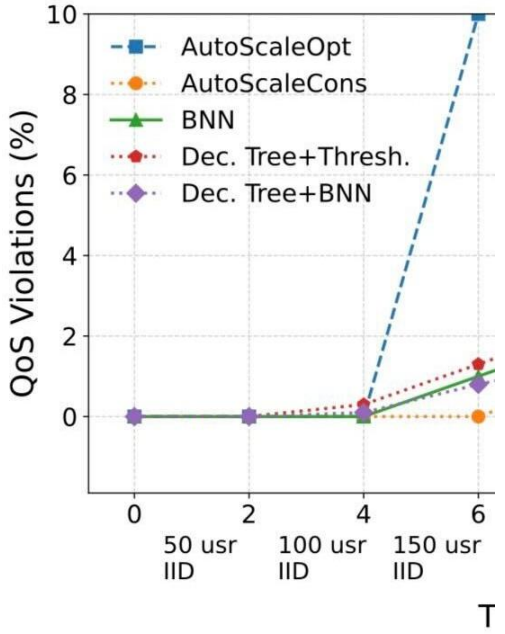


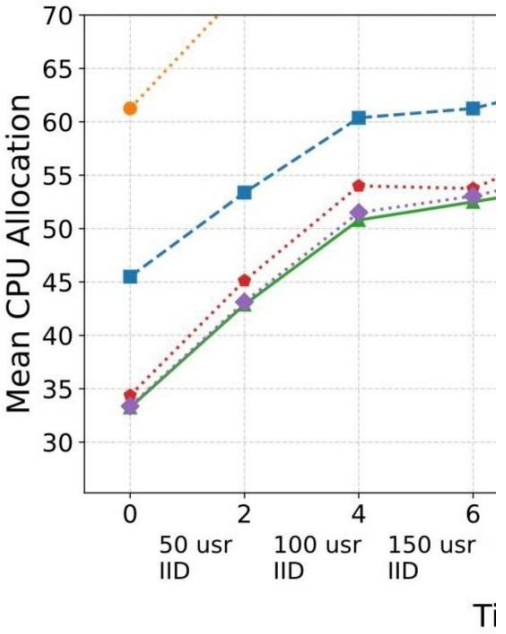




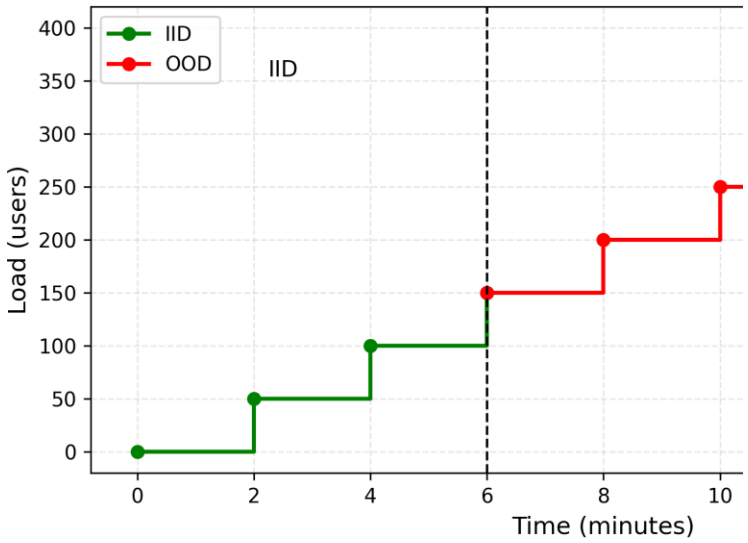
Evaluation Results

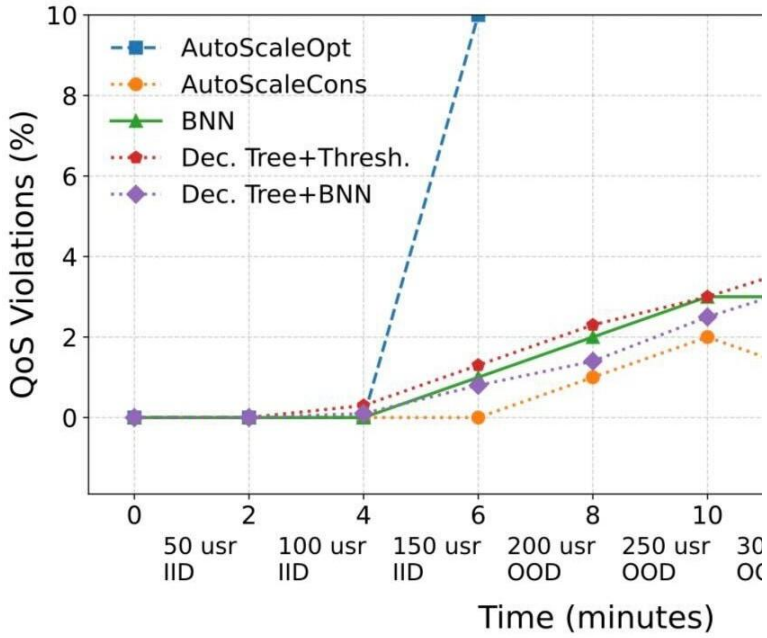


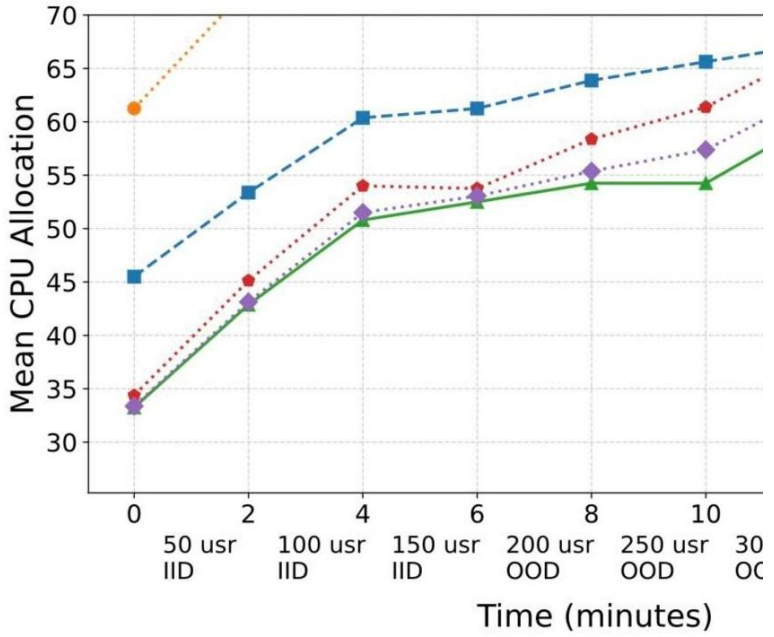




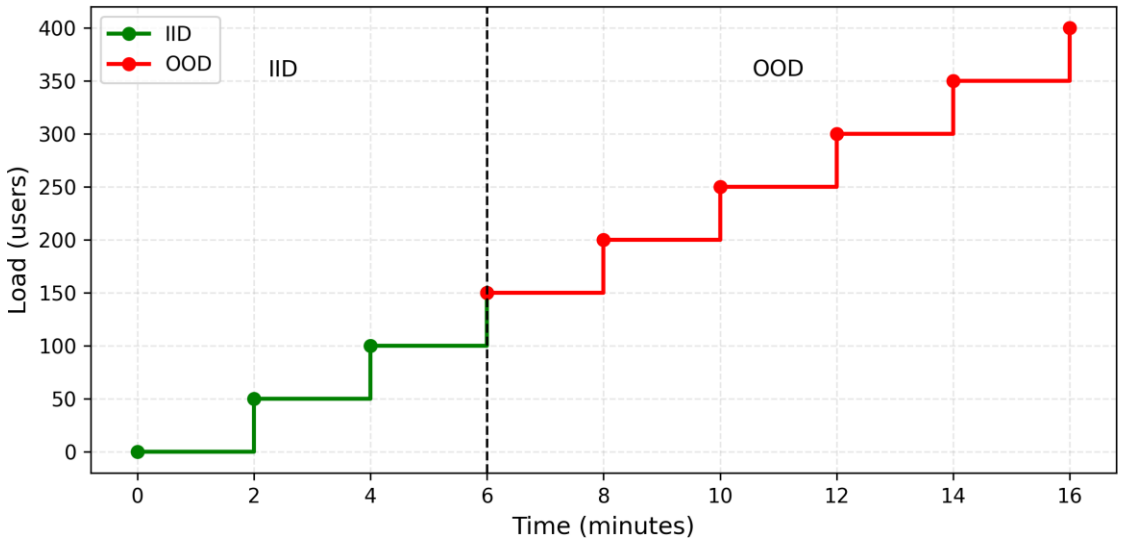
Evaluation Results

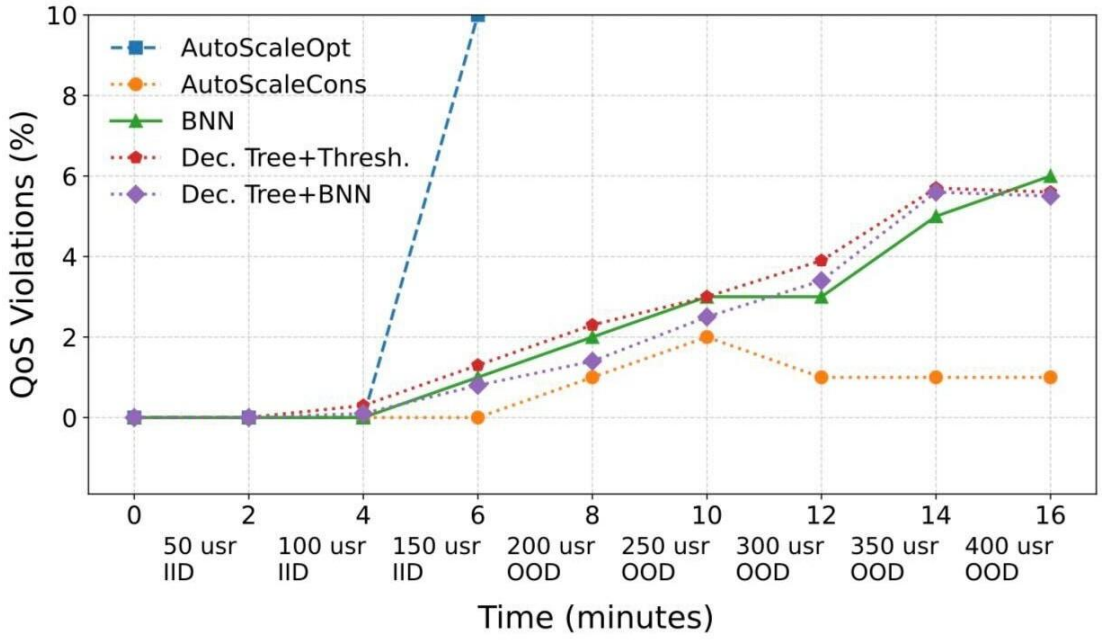


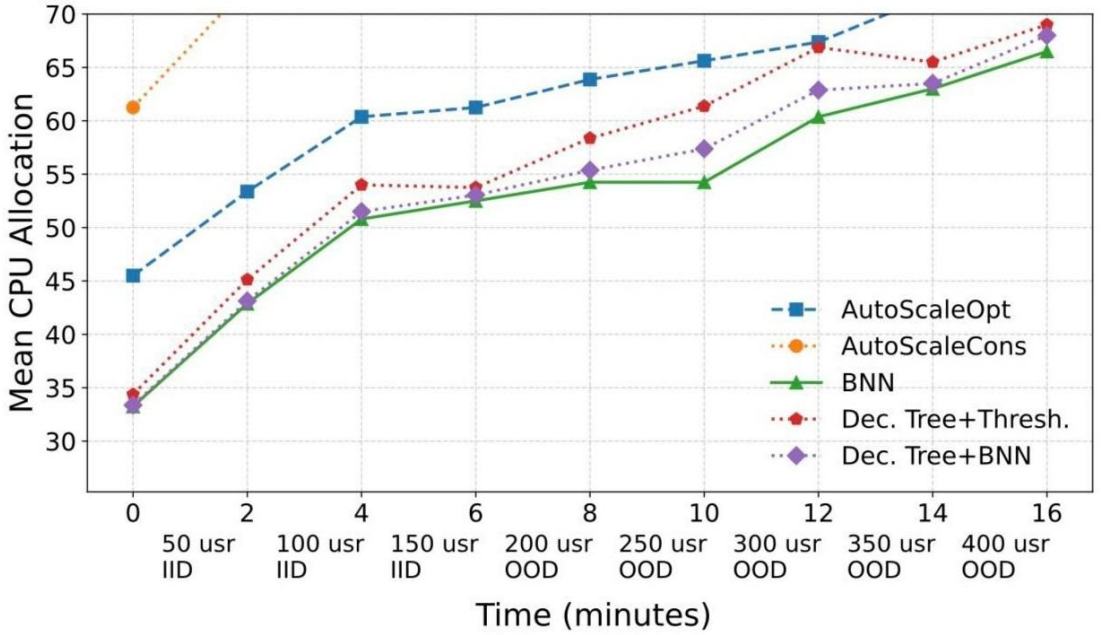




Evaluation Results







Key Takeaways

- Uncertainty-aware model hierarchies can balance the trade-offs between interpretability, generalizability and computational cost
- In our case study:
  - Inference latency: 31ms -> 4.27ms
  - Up to 96% of the decisions are interpretable and *more* computationally efficient
  - +0.5% QoS violations in total