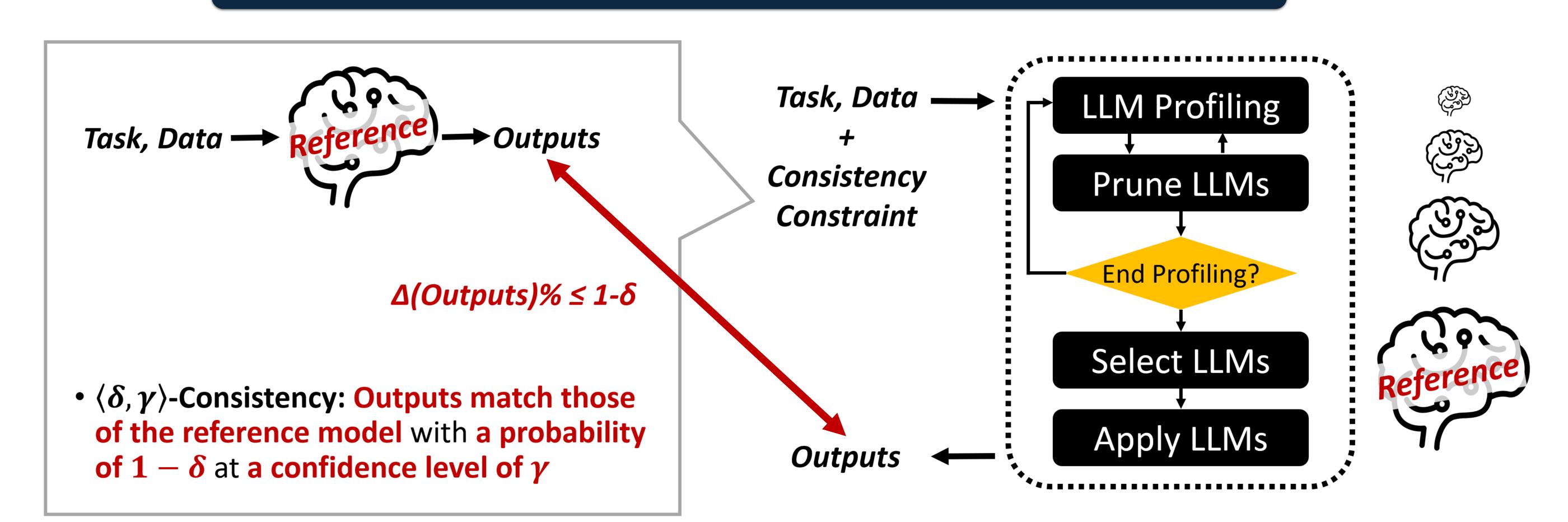


SpareLLM

Automatically Selecting Task-Specific Minimum-Cost Large Language Models under Equivalence Constraint

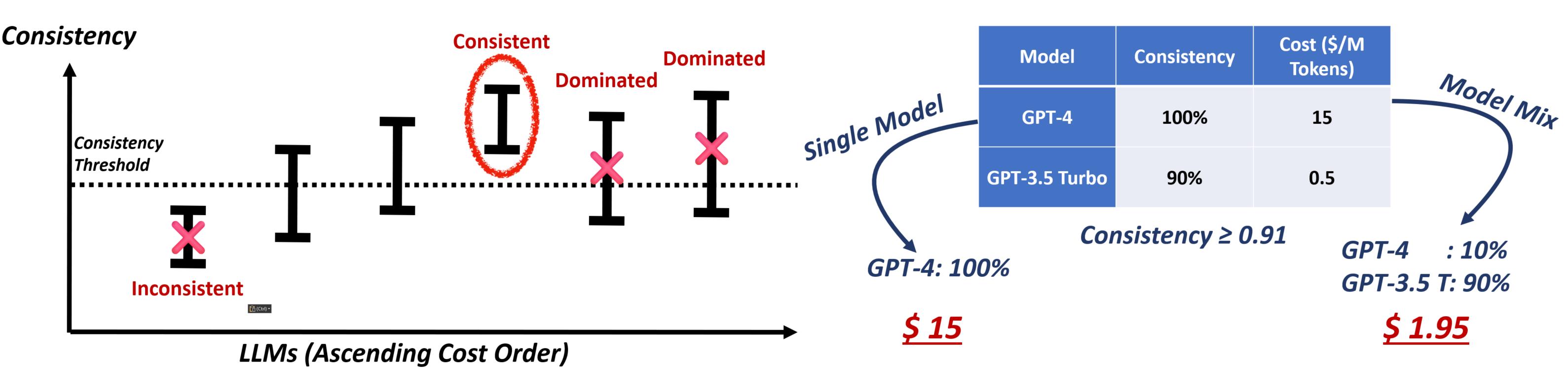
Saehan Jo and Immanuel Trummer Cornell University

SpareLLM Architecture

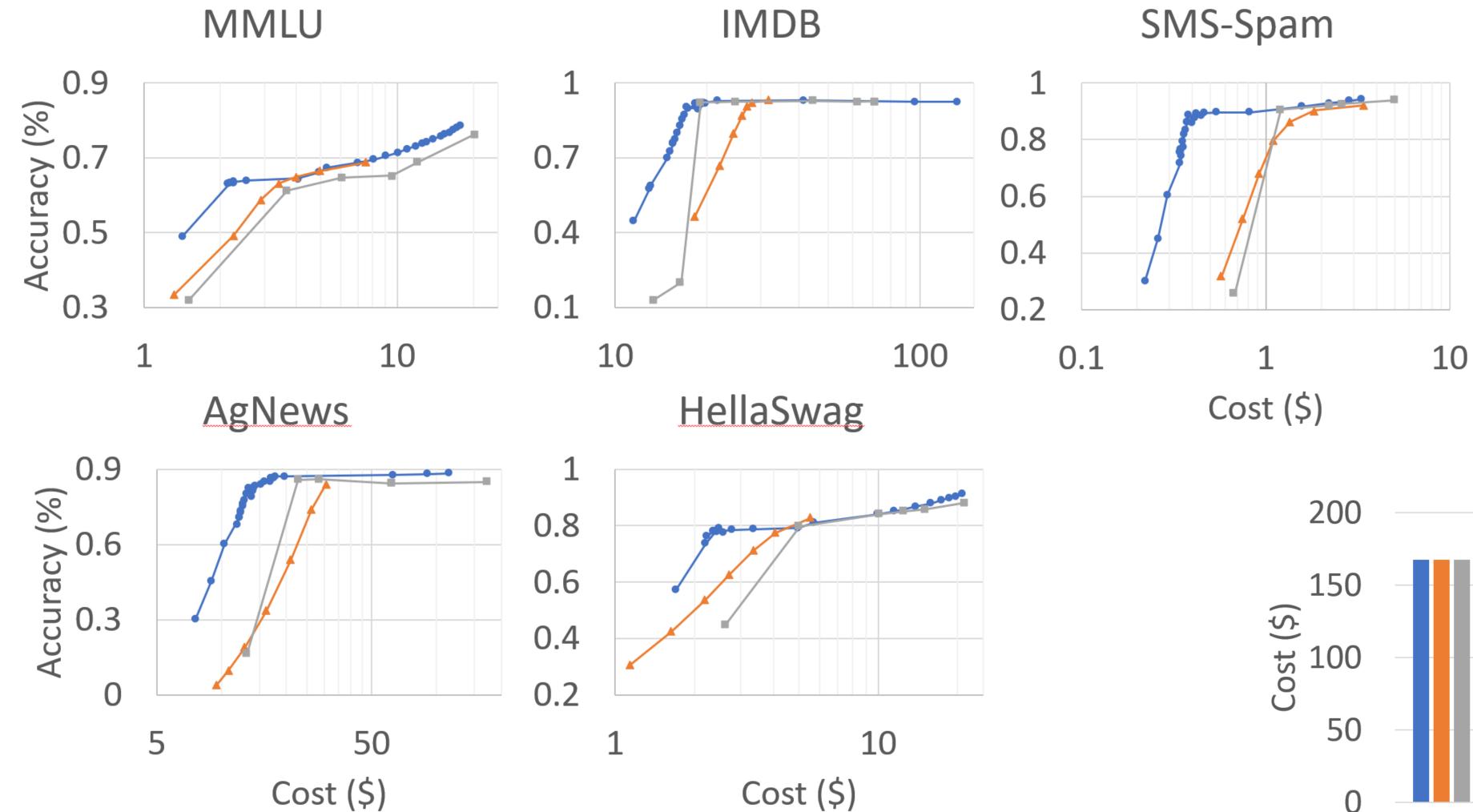


Profiling Phase

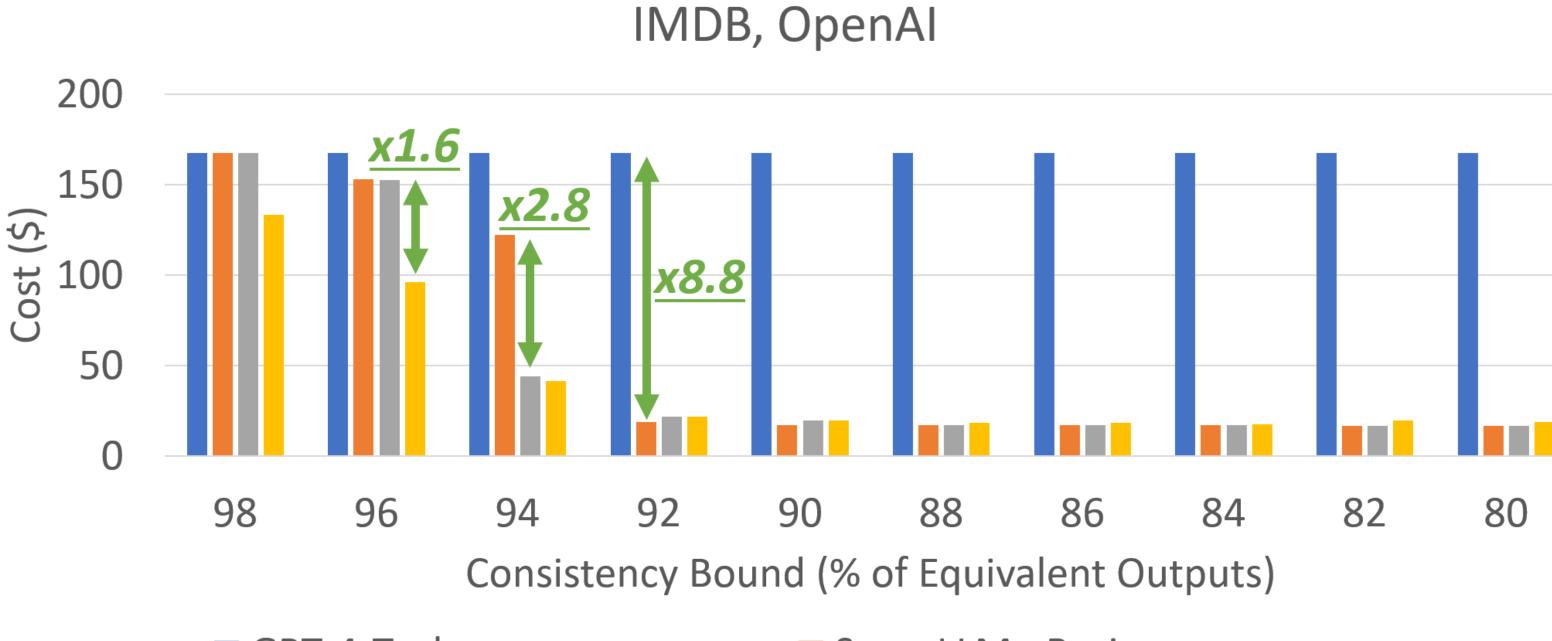
Application Phase



Experiments



- Comparison against Baselines
 (Cost Accuracy Tradeoffs)
- Comparison against GPT-4-Turbo (Costs under Consistency Constraint)



LLMCascadeFrugalGPT

SpareLLM

■ GPT-4-Turbo ■ SpareLLM - Basic ■ SpareLLM + Dynamic Profiling ■ SpareLLM + Model Mix