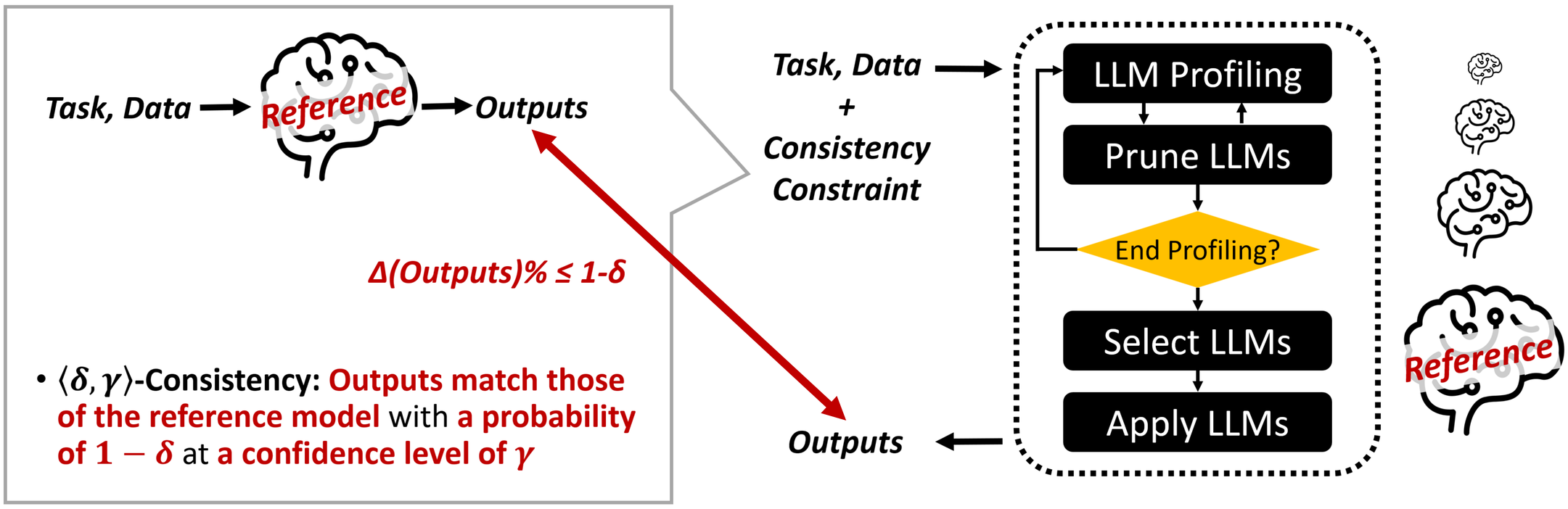
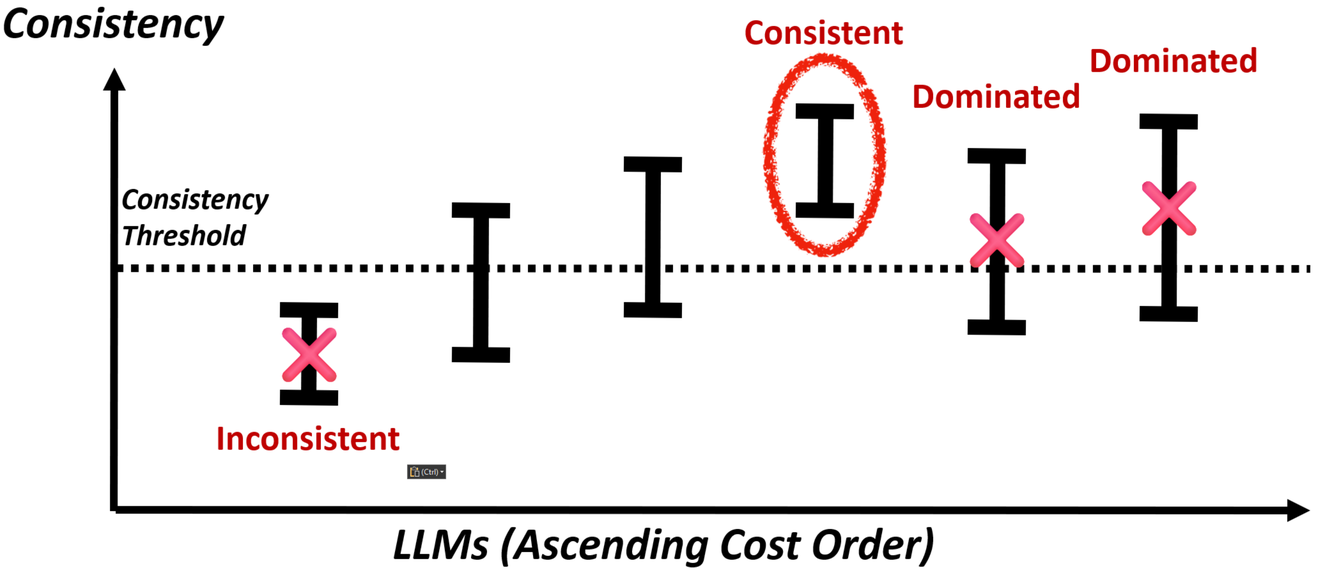
SpareLLM

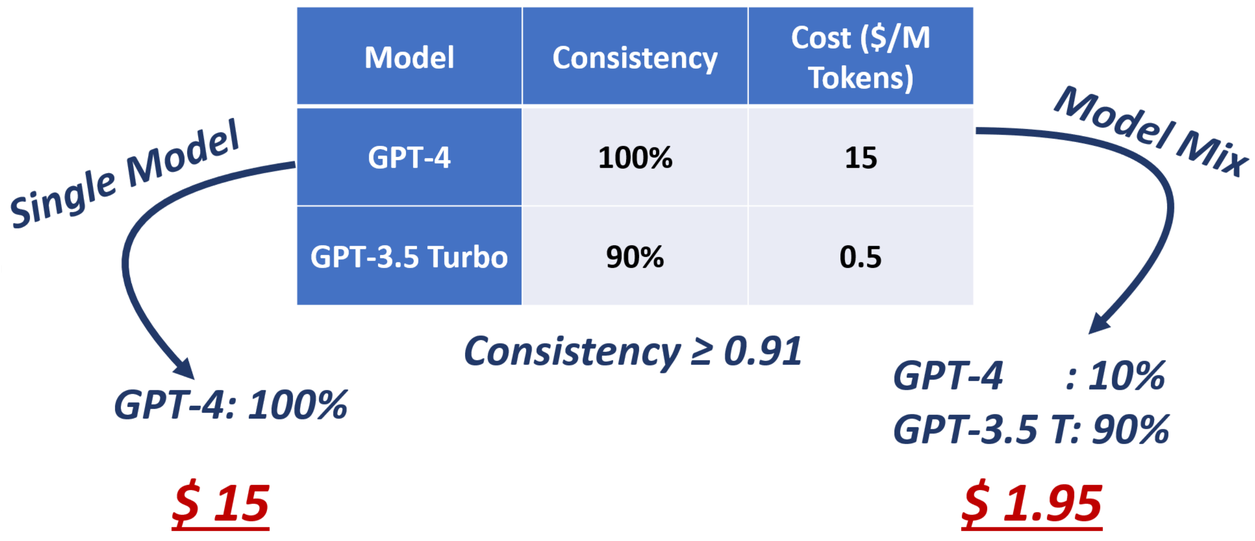
Automatically Selecting Task-Specific Minimum-Cost Large Language Models under Equivalence Constraint

Saehan Jo and Immanuel Trummer
Cornell University

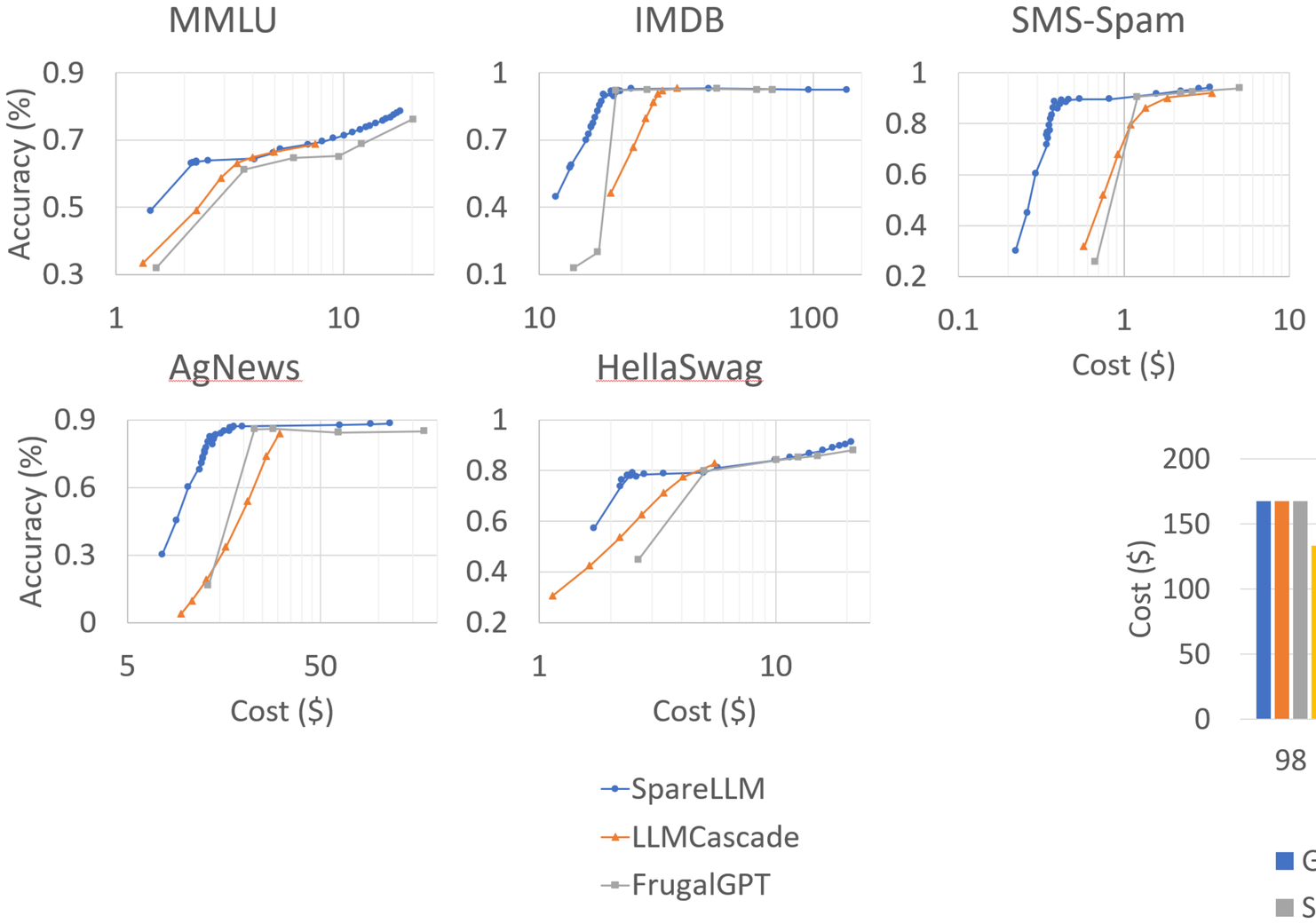
SpareLLM Architecture







Experiments



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