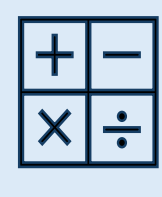


CP-Bench: Evaluating LLMs for Constraint Modelling

Kostis Michailidis, Dimos Tsouros, Tias Guns @ DTAI, KU Leuven, Belgium

Motivation

Constraint Modelling is **powerful**,
but requires **expertise** in:

-  **Domain Knowledge**
-  **Mathematical Reasoning**
-  **Solver Syntax & Semantics**

CP-Bench: Dataset & Evaluation

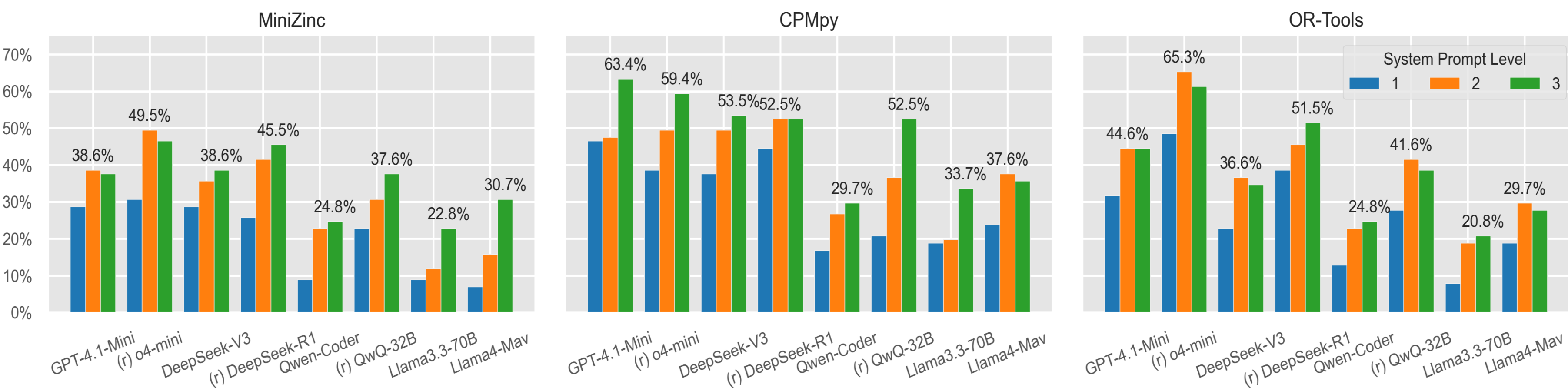
A collection of diverse combinatorial problems & models from established sources (online live leaderboard available):

- 101 problems including both satisfaction and optimization problems with their CPMpy models
- Automated Solution-based evaluation
- Dataset Problems List:

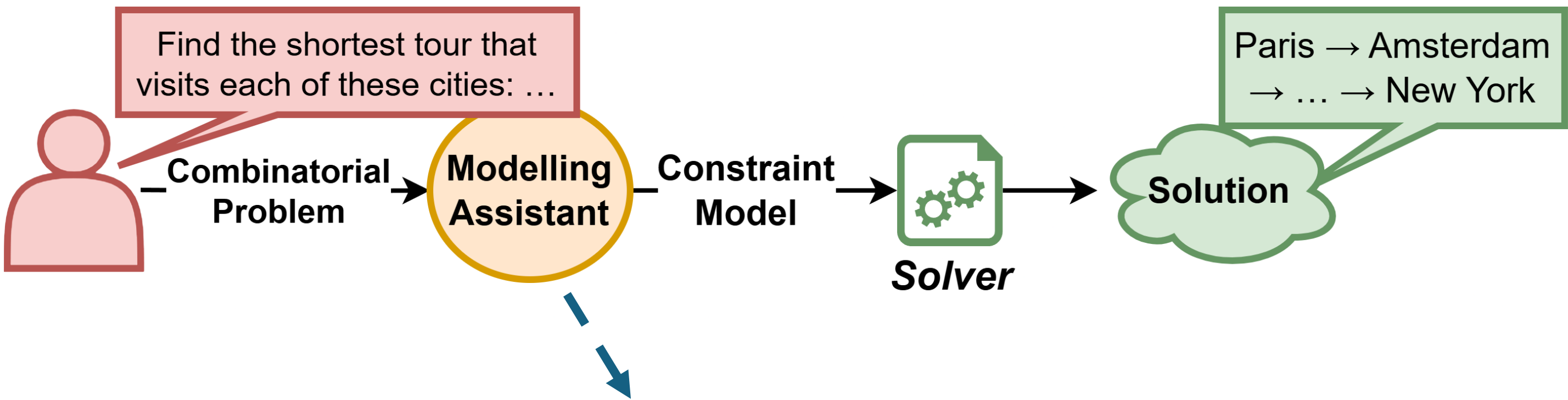
Appointment Scheduling, Bin Packing, Broken Weights, Building Blocks, Cabling, Calvin Puzzle, Capital Budget, Chess Set, Circling Squares, Clock Triplets, Crossword, Facility Location, Fifty Puzzle, Three Coins, Bank Card, Climbing Stairs, Color Simple, Exodus, Farmer And Cows, Five Floors, Grocery, Guards And Apples, Hardy 1729 Square, Kidney Exchange, Magic Square, Maximal Independent Sets, Money Change, Movie Scheduling, People In A Room, Subset Sum, Subsets 100, Bus Scheduling, Jobshop Scheduling, Knapsack, Mario, Minesweeper, N-Puzzle, Packing Rectangles, RCPSP, Room Assignment, Send More Money, Set Game, Sudoku, TSP, Wolf Goat Cabbage, Zebra, Car Sequencing, Template Design, Autocorrelation, Golomb Rules, All Interval, Vessel Loading, Perfect Square Placement, Social Golfers, Basketball Schedule, Nonogram, Progressive Party Problem, Schurs Lemma, Magic Squares And Sequences, Crossfigures, Langford, Sports Tournament Scheduling, BIBD, Word Design, Steiner, Number Partitioning, Diamond Free, Graceful Graphs, N-Queens, Costas Arrays, Hadamard Matrix and some more.

Different Modelling Frameworks

Framework	Abstraction Level	Interface Type
MiniZinc	High	Domain-Specific
CPMpy	High	Python
OR-Tools	Medium	Python



Python-based vs. Domain-Specific Frameworks
High-level modelling abstractions (e.g. global constraints)

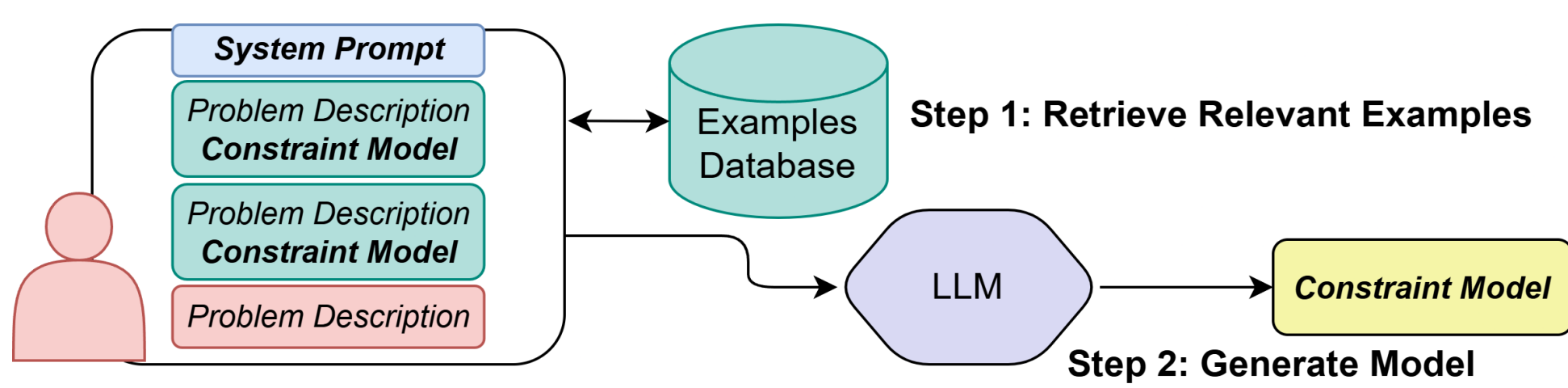


Inference-Time Compute Methods for Modelling

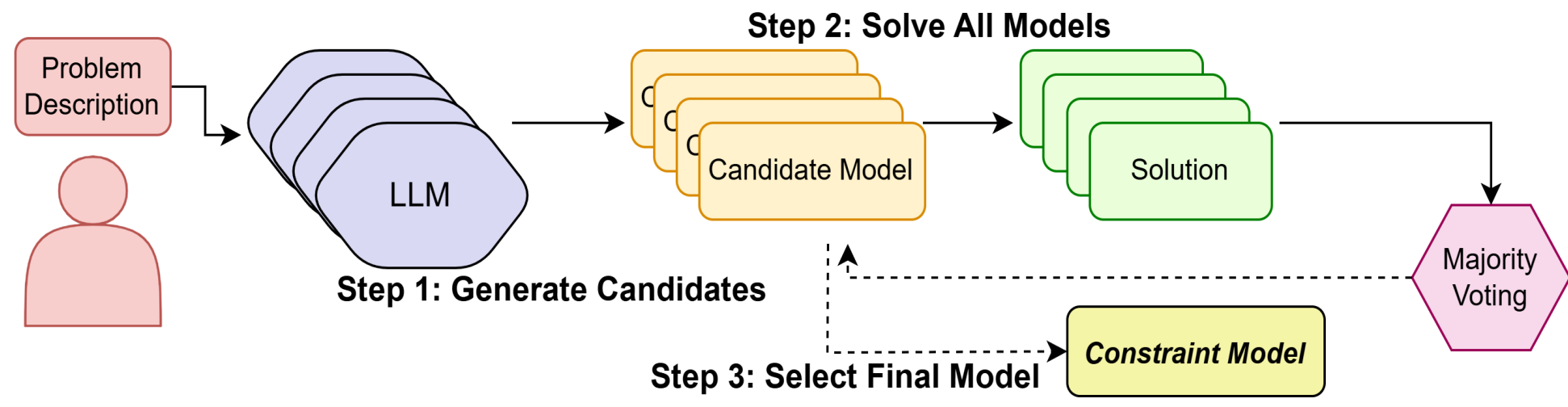
1. Prompt Engineering:

Basic (Level 1): You are an expert in modelling...
Guidelines (Level 2): Define decision variables...
Documentation (Level 3): Here is the API of the modelling framework...

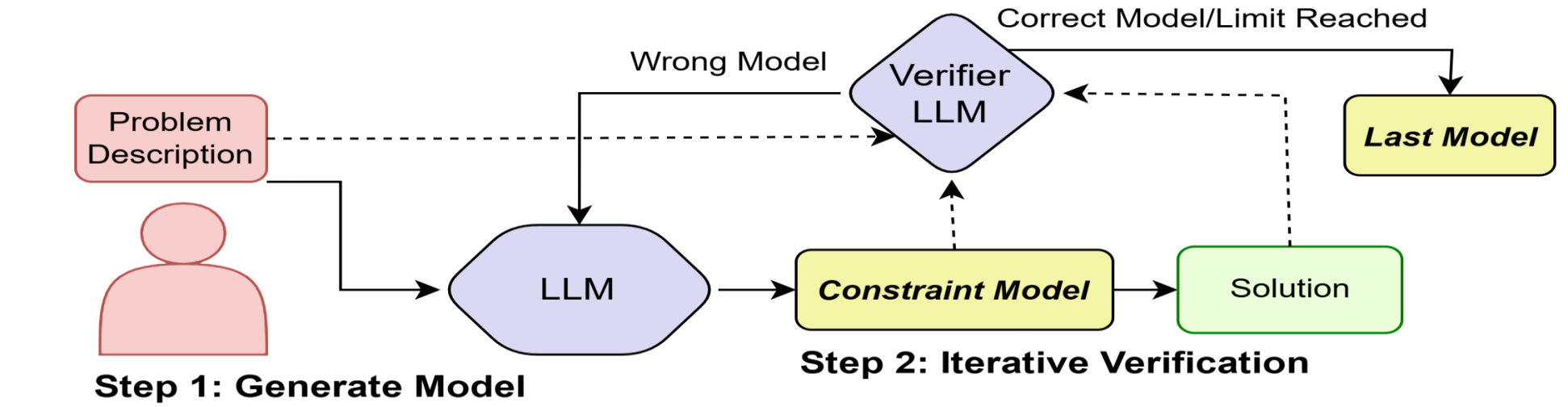
2. In-Context Learning:



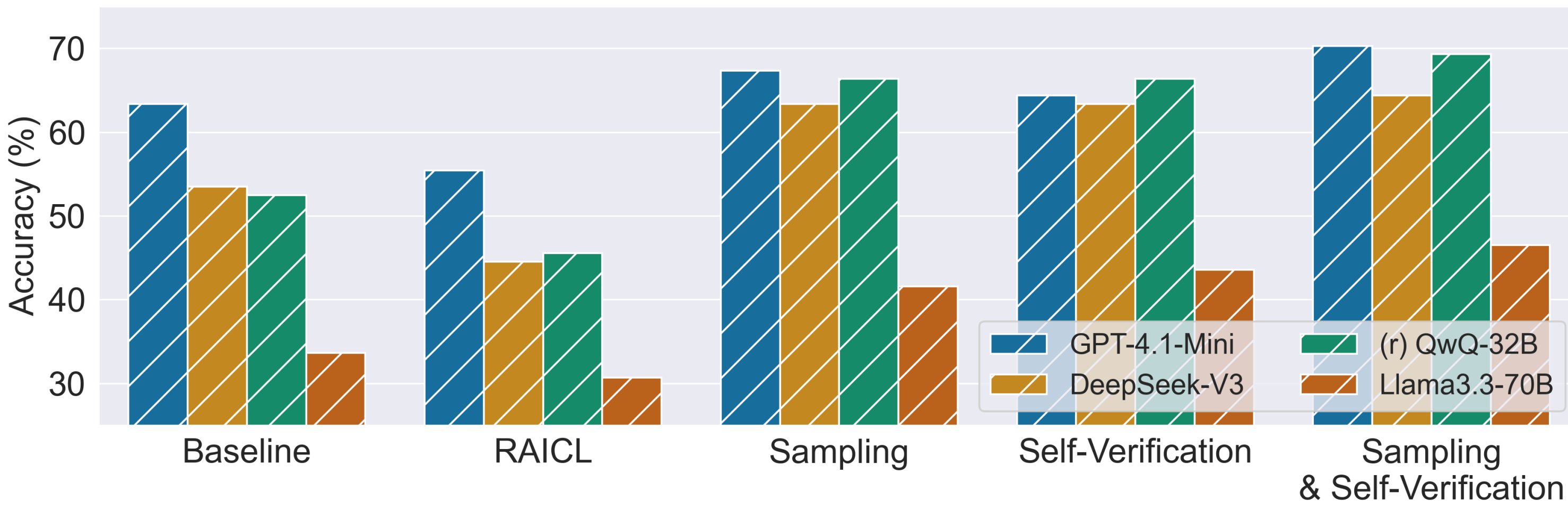
3. Repeated Sampling:



4. Self-Verification:



Modelling Framework: CPMpy, Baseline: Documentation Prompt



- Documentation Prompt vs. In-Context Learning**
- Sampling & Self-Verification boost performance**



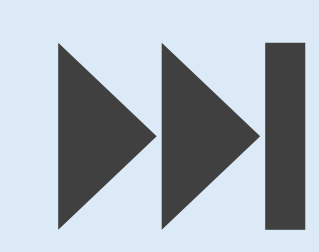
A diverse benchmark for evaluating LLM-driven CP modelling.



Systematic evaluation of LLMs across common modelling frameworks



Tailoring Inference-Time Compute methods for constraint modelling



Beyond correctness: Modelling efficiently?

