Ph.D. Thesis Defence

Exploring the Synergies Between Operations Research and Machine Learning

Xinglu Wang

Examining Committee:

Dr. Jian Pei, Co-supervisor
Dr. Jiannan Wang, Co-supervisor
Dr. Tianzheng Wang, Committee Member
Dr. Wuyang Chen, Internal Examiner
Dr. Wei Wang, External Examiner
Dr. Jiangchuan Liu, Chair

SFU

School of Computing Science
Simon Fraser University

Two powerful problem-solving tools

Figure 1 Production planning

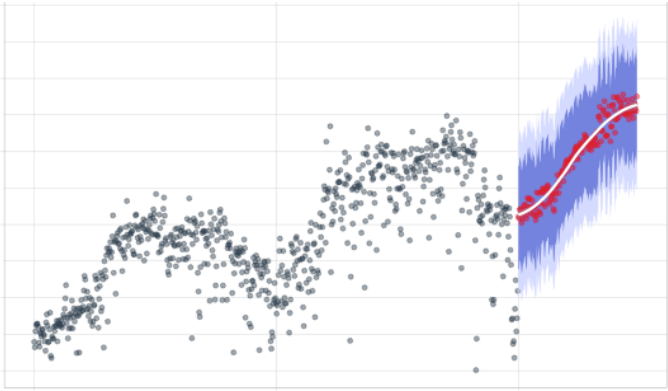


Figure 2 Time series prediction

Operations Research (OR)

OR is a scientific approach to decision-making that seeks to operate a system ***optimally***, usually under ***constraints*** and scarce resources.
(Winston, 2004)

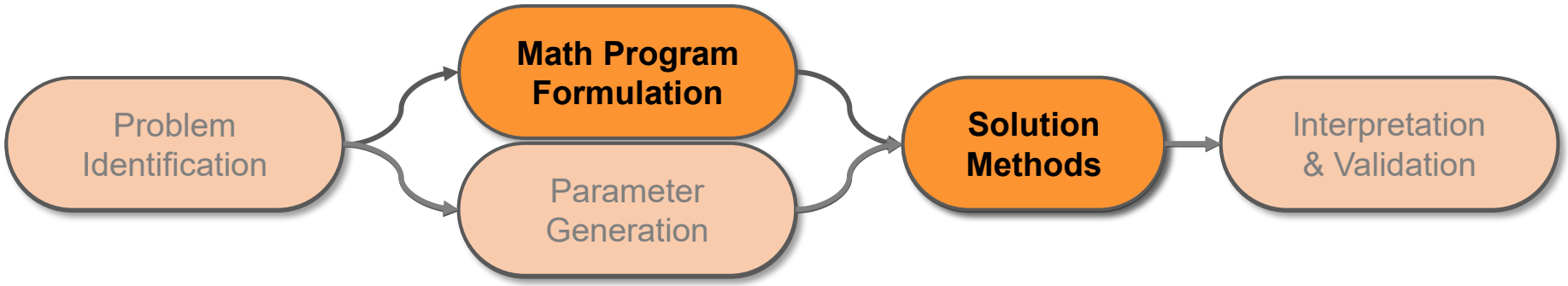
Machine Learning (ML)

ML allows computers to ***learn from data*** without being explicitly programmed.
(Goodfellow et al., 2016)

Bottlenecks in the OR process?

OR follows a standard pipeline widely used in industrial applications

A typical OR process (Rajgopal et al., 2024)



- Real-world problems are increasingly complex.
- Complexity may introduce inefficiencies in certain stages.
- Can ML be leveraged to enhance efficiency of these stages?

Inefficiency in the *Formulation* stage

Problem: **Product Planning**

A factory can produce up to 30 boxes of apple slices per day. **Given** user demands over a planning horizon, please **decide** daily production quantities to **satisfy** all demands on time, while **minimizing** total machine setup, production, and inventory carryover costs.

Known parameters:

Demands on 3 days **20, 30, 1**	Daily Capacity **30 Boxes**
Setup Cost **$100 / Day**	Production Cost **$2 / Box**
Inventory Carryover **Allowed**	Carryover Cost **$1 / Box / Day**

Challenge in modeling efficiency

Math Program

Parameters: d_t: demand on day t

Decision variables:

$x_t \geq 0$: units produced on day t

$y_t \in \{0,1\}$: whether produce on day t

$$\min_{x \in \mathbb{R}^n} f(x)$$

s.t. $x \in$ feasible set $\mathcal{F}$

$$\min \quad \sum_{t=1}^{3} (100\, y_t + 2\, x_t) + \sum_{t=1}^{2} \left(\sum_{\tau=1}^{t} x_\tau - \sum_{\tau=1}^{t} d_\tau \right)$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} x_\tau \geq \sum_{\tau=1}^{t} d_\tau, \qquad t = 1,2,3$$

$$0 \leq x_t \leq 30\, y_t, \qquad t = 1,2,3$$

$$y_t \in \{0,1\}, \qquad t = 1,2,3$$

Opportunity:

Formulation experience

Descriptions of problem instances

Inefficiency in the *Solution* stage

Math Program

Parameters: d_t: demand on day t
Decision variables:
$x_t \geq 0$: units produced on day t
$y_t \in \{0,1\}$: whether produce on day t

$$\min \quad \sum_{t=1}^{3} (100\, y_t + 2\, x_t) + \sum_{t=1}^{2} \left(\sum_{\tau=1}^{t} x_\tau - \sum_{\tau=1}^{t} d_\tau \right)$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} x_\tau \geq \sum_{\tau=1}^{t} d_\tau, \qquad t = 1,2,3$$

$$0 \leq x_t \leq 30\, y_t, \qquad t = 1,2,3$$

$$y_t \in \{0,1\}, \qquad t = 1,2,3$$

Program size

#Time periods: T
#Products: P

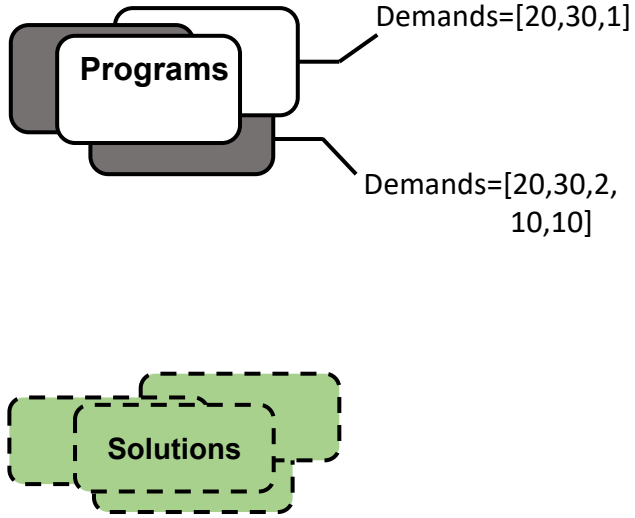
#Decision vars:
$2PT$

#Constraints:
$2PT + T$

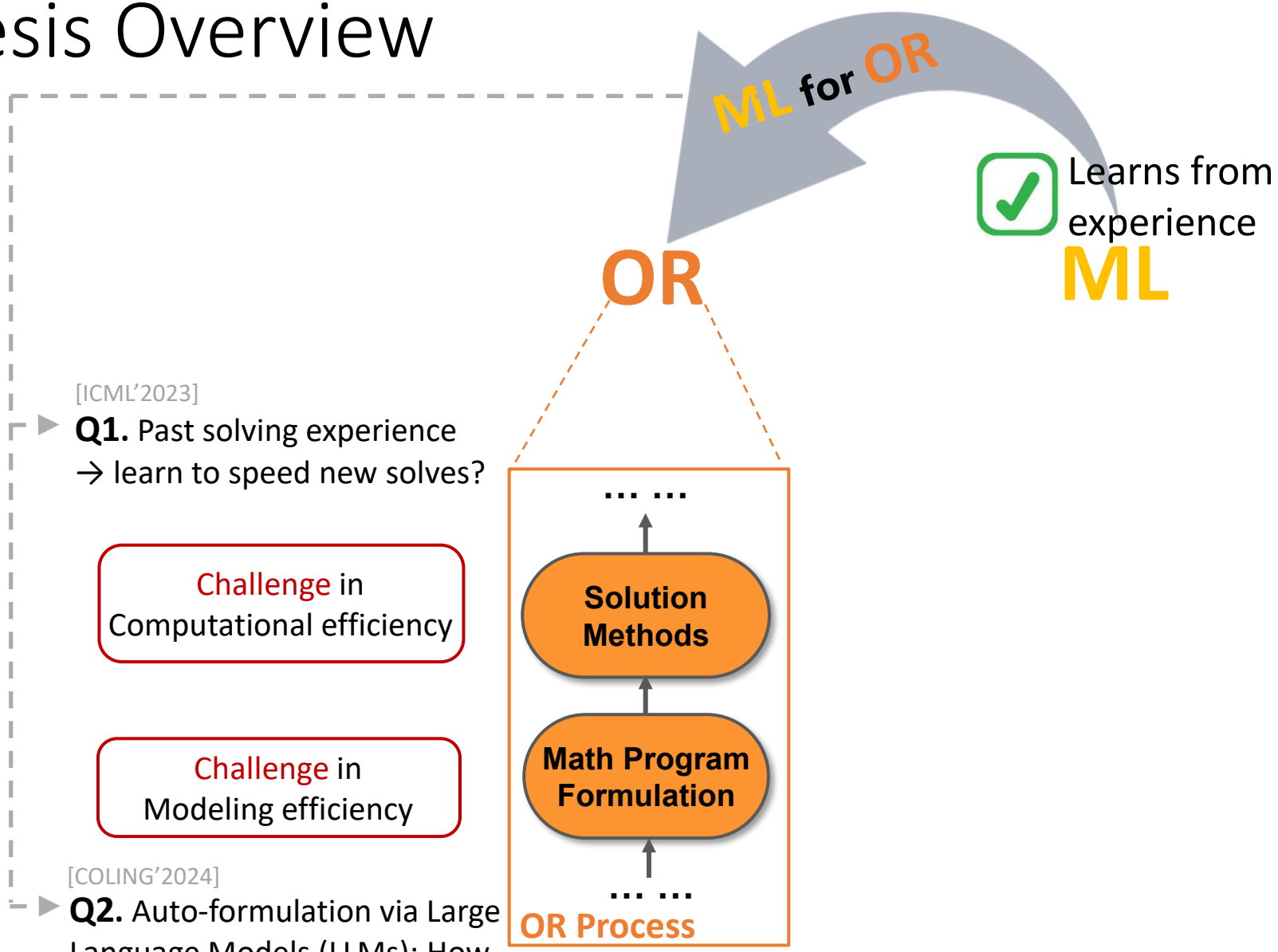
When $P = 1k,\ T =$ 30 days, the solving time ***exceeds*** 10 minutes.

Challenge in computational efficiency (solving)

Opportunity: Solving experience

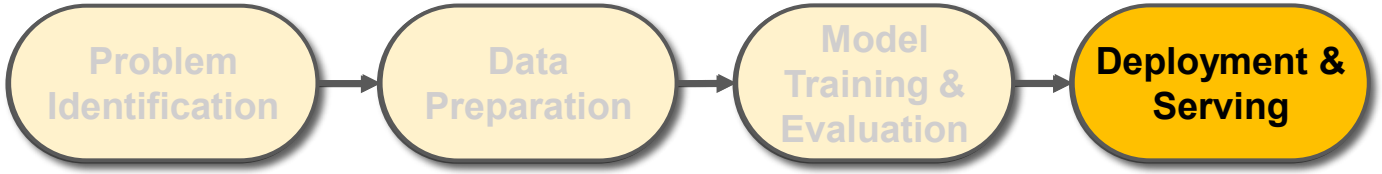


Thesis Overview

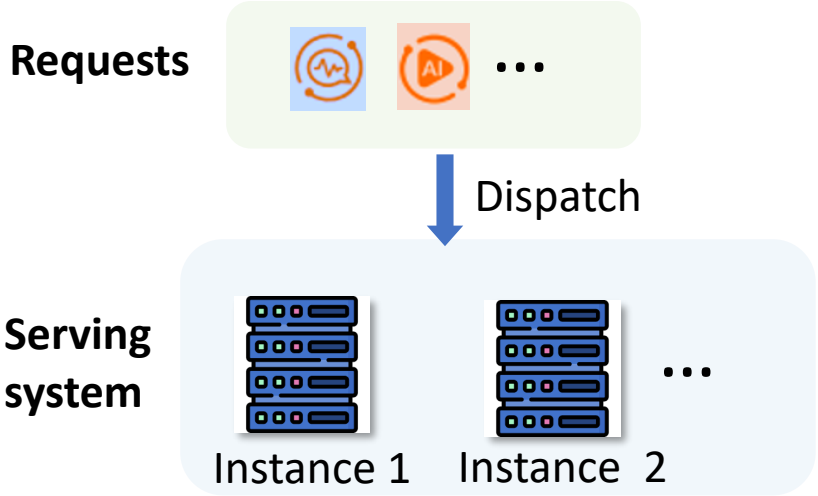


Bottlenecks in the ML lifecycle?

A simplified ML lifecycle (De Silva et al., 2022)



Large language model (LLM) serving

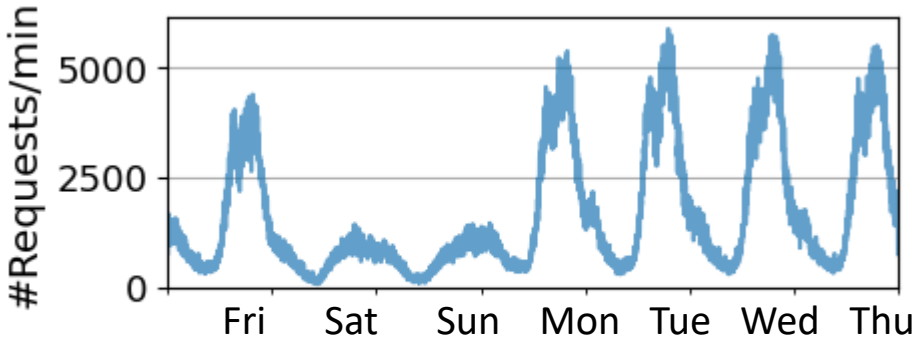


Heterogeneous Workloads

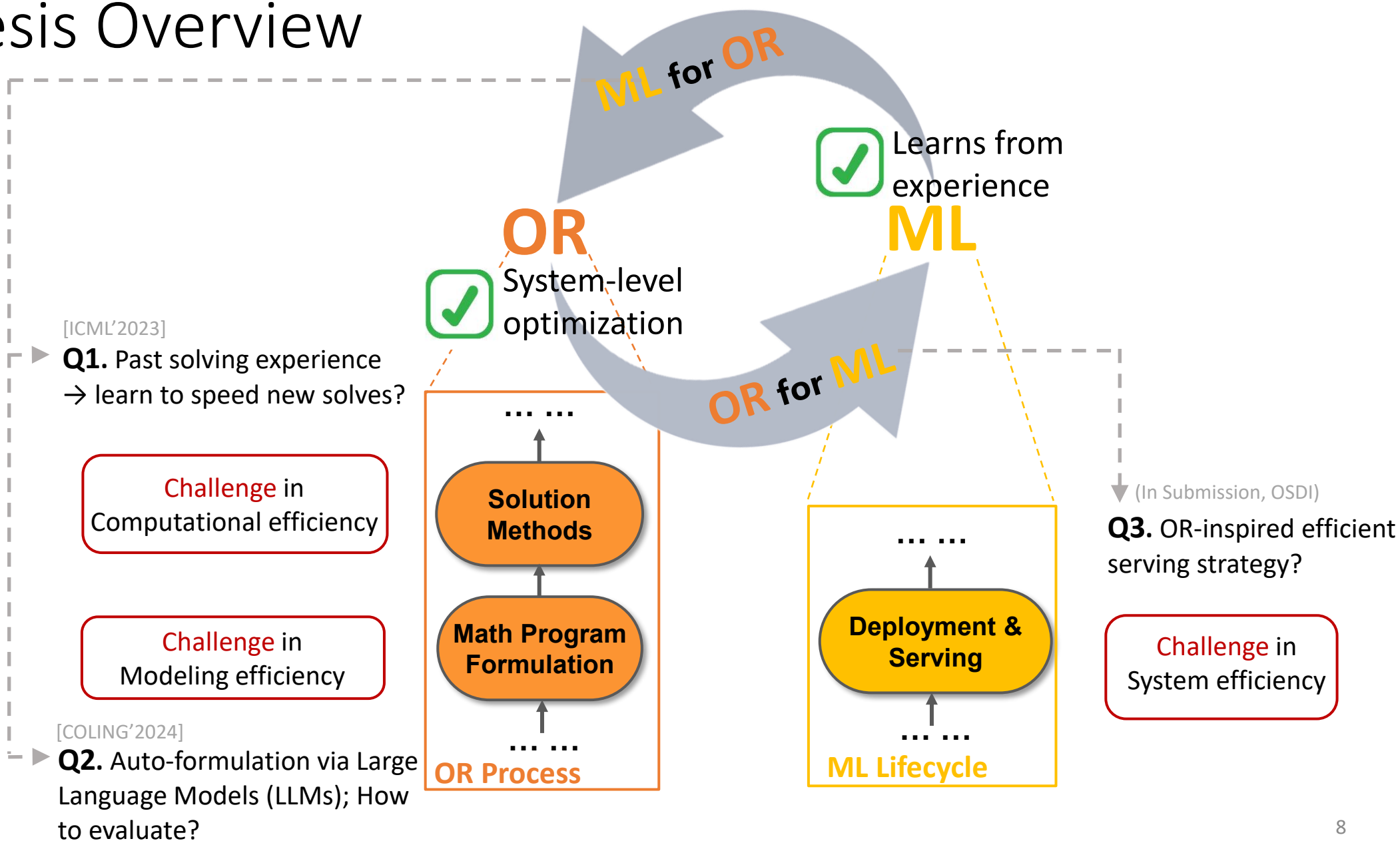
Application	Workload	Requirement
Summarizer	Heavy	Offline, response in 24 hours
ChatBot	Light	Response 1 token every 50 ms

Time-Varying Workloads

Request rate of a Chatbot app in a week

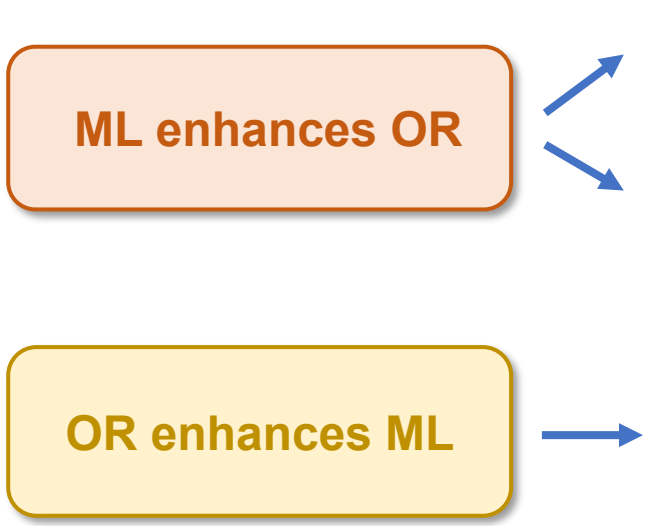


Thesis Overview



Roadmap

- Introduction
- To answer the three research questions:



Smart initial basis selection for linear programs [ICML' 23]

Towards human-aligned evaluation for linear programming word problems [COLING' 24]

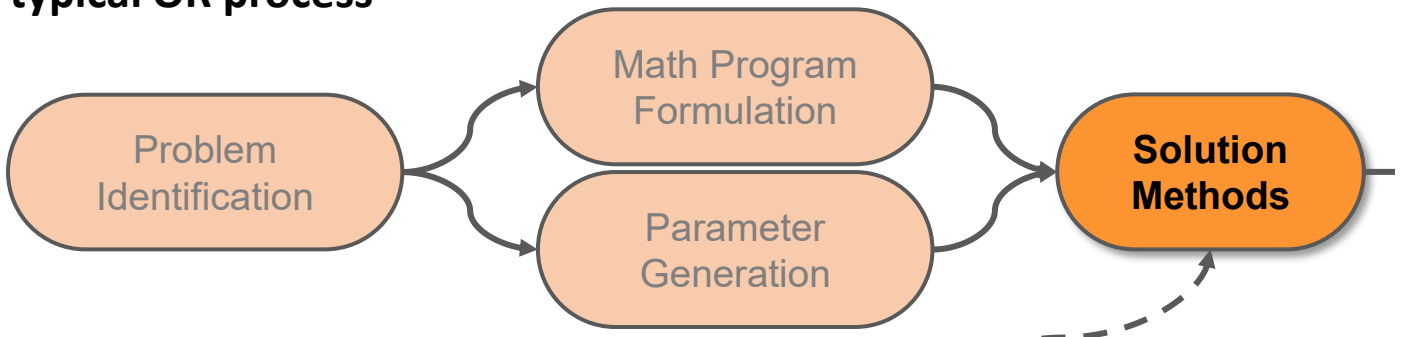
HFX: Joint Design of Algorithms and Systems for Multi-SLO Serving and Fast Scaling (In Submission, OSDI)

- Conclusion and future directions

Scope and background

Q1 Past solving experience → learn to speed new solves?

A typical OR process



Consider following LP form:

$$\min_{x\in\mathbb{R}^n, s\in\mathbb{R}^m} \quad c^T x$$

$$\text{s.t.} \quad Ax = s$$

$$\ell^x \le x \le u^x$$

$$\ell^s \le s \le u^s$$

Solving linear programs --- Simplex algorithm

Rule-based initial basis selection strategy → ML-enhanced strategy

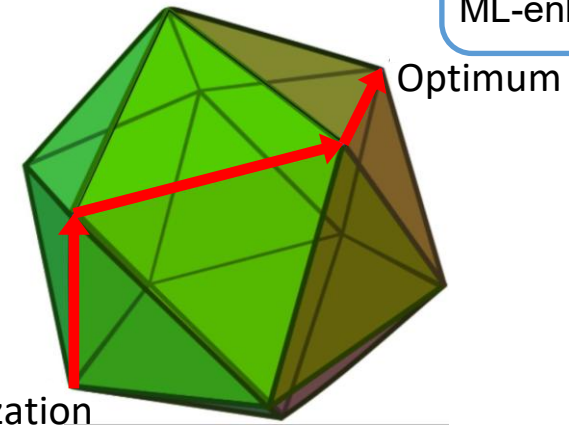


Figure 1 Illustration of Simplex algorithm

- Background Concepts:
  - Vertex on Simplex ⟺ Basic feasible solution ⟺* Valid basis
  - Basic: an index set of *size m* from $[x, s]$
  - Valid = *2 other properties*
    - The m basic variables from a ***non-singular*** linear system
    - Value of remaining non-basic variables are ***consistent with bounds***

* Comment: Refer to Link for details about Valid basis → Basic **in**feasible solution → Find a basic feasible solution

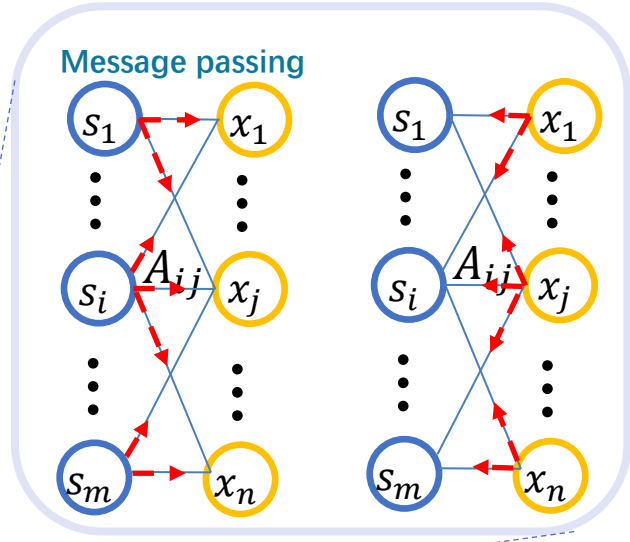
Problem definition

- Given historical solving experience (LP problem *v.s.* its optimal basis), learn a mapping: a new LP instance → a close-to-optimal basis.

- Challenge: This mapping should have following properties
  1. Handle LPs of varying sizes (#constraints and #decision variables)
  2. Permutation equivariance.
  3. Output a valid basis.

Method design

- Mapping ≈ Graph neural network (GNN) + PostProcess
  - Benefits of GNN (Scarselli et al., 2009):
    - 1. Accept variable-sized graphs as input
    - 2. node-level predictions are permutation equivariant
  - Post processing ensures a valid basis



GNN

basis status probability vectors

LP form

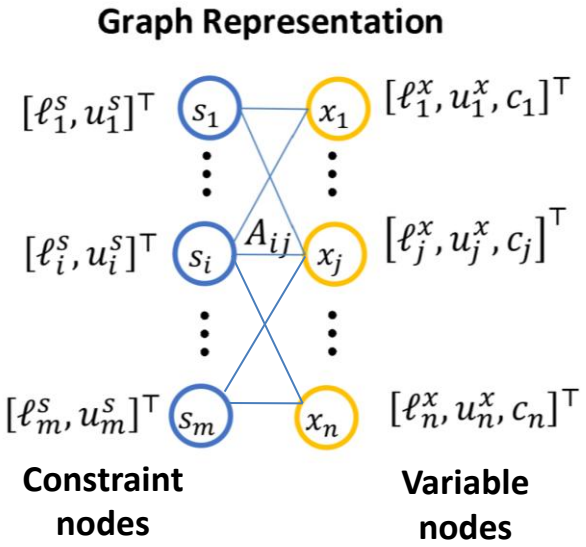
$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{s}\in\mathbb{R}^m} \quad \boldsymbol{c}^T\boldsymbol{x}$$

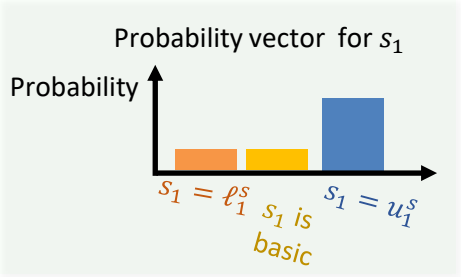
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{s}$$

$$\boldsymbol{\ell}^x \le \boldsymbol{x} \le \boldsymbol{u}^x$$

$$\boldsymbol{\ell}^s \le \boldsymbol{s} \le \boldsymbol{u}^s$$

Convert





Method: Post process: output a *valid* basis

- Recall: GNN prediction is in format of $m + n$ probability vectors
  - Each variable's status is represented in its independent probability vector
  - But this is still an intermediate output, not a *valid basis*.
- *Valid basis* means:
  - The status of non-basic variables must be consistent with their bounds.
    - Knowledge-based masking: adding large penalty to the logits of unreachable bounds
  - Size of basis must be m
    - Basis generation: select top-m constraint and variable indices as basis $(\mathcal{B}_x, \mathcal{B}_s)$
  - The basis matrix corresponding to the selected initial basis must be non-singular.
    - Basis adjustment: make sure the basis is valid by trying to *factorize* [4] the corresponding constraint matrix $[A_{\mathcal{B}_x} \quad -I^m_{\mathcal{B}_s}]$

"Valid basis" means:	**Steps**
Non-basic variables must respect their bounds	Knowledge-based masking: Penalize logits of unreachable bounds
Basis size = m	Basis selection: Pick top-m probability
Basis matrix must be non-singular	Adjustment: Factorize to ensure non-singularity

[4] Bixby, R. E. and Saltzman, M. J. Recovering an optimal LPbasis from an interior point solution. Oper. Res. Lett., 15(4):169–178, 1994.

Experiment: Does GNN predicted basis provide effective warm-start?

- Setup
  - Dataset: open-sourced and private large-scale ones
  - Optimization Solvers: open-source HIGHS and commercial OptVerse
  - Baselines: Default and traditional (CA / CA-MPC / CA-ANG) initial-basis strategies

Dataset	#LPs	m=#constraints	n=#variables
LIBSVM	100	$20.0K$	$20.0K$
MIRP	28	$28.2K_{\pm 25.2K}$	$28.7K_{\pm 25.0K}$
STOCH	100	$52.3K_{\pm 1.9K}$	$107.0K_{\pm 3.8K}$
GEN	100	$1.0K$	$1.0K$
SC-1	525	$312.9K_{\pm 177.4K}$	$659.1K_{\pm 386.4K}$
SC-2	190	$1.4M_{\pm 199.1K}$	$2.9M_{\pm 450.6K}$

Table 1 Dataset statistic

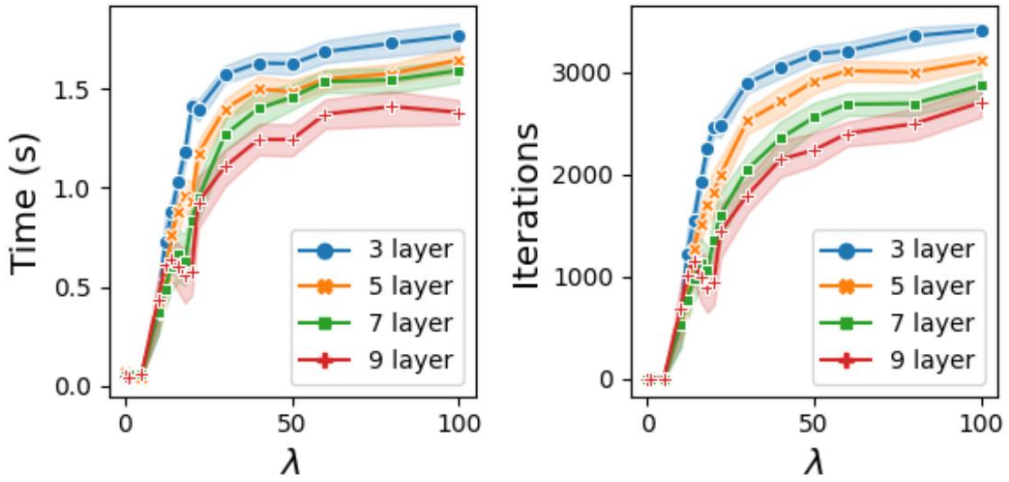
Dataset	DEFAULT	CA	CA-MPC	CA-ANG	GNN(Ours)
LIBSVM	$14.9K_{\pm 9.5K}$	$14.9K_{\pm 9.5K}$	$21.0K_{\pm 4.8K}$	$15.2K_{\pm 1.1K}$	$\mathbf{9.1K}_{\pm \mathbf{3.1K}}$
MIRP	$40.3K_{\pm 23.3K}$	$34.8K_{\pm 20.2K}$	$36.7K_{\pm 20.8K}$	$39.6K_{\pm 22.7K}$	$\mathbf{25.9K}_{\pm \mathbf{16.9K}}$
STOCH	$75.3K_{\pm 4.3K}$	$52.5K_{\pm 4.8K}$	$48.7K_{\pm 5.2K}$	$53.3K_{\pm 1.7K}$	$\mathbf{31.8K}_{\pm \mathbf{14.3K}}$
GEN	$2.4K_{\pm 225.0}$	$2.4K_{\pm 225.0}$	$2.4K_{\pm 225.0}$	$2.4K_{\pm 225.0}$	$\mathbf{552.8}_{\pm \mathbf{642.9}}$
SC-1	$272.3K_{\pm 151.9K}$	$158.9K_{\pm 89.1K}$	$266.9K_{\pm 148.5K}$	$269.2K_{\pm 151.5K}$	$\mathbf{26.6K}_{\pm \mathbf{15.4K}}$
SC-2	$1.2M_{\pm 170.7K}$	$1.1M_{\pm 172.2K}$	$1.2M_{\pm 163.5K}$	$431.9K_{\pm 99.0K}$	$\mathbf{169.1K}_{\pm \mathbf{34.3K}}$

Table 2 Performance comparison between the GNN and baselines strategies, measured by the Number of Solver Iterations.

Experiment: limitations

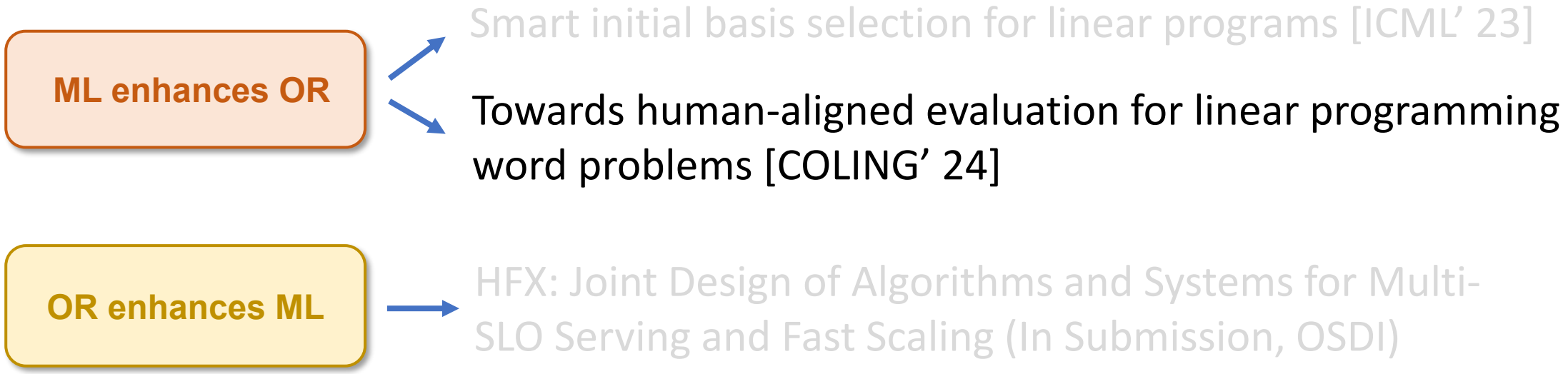
- How does dataset diversity affect performance of GNN-strategy?
  - Dataset diversity is the diversity of the *mappings from LP instances to their corresponding optimal bases* within a dataset.
  - We customized a LP-generating strategy and designed a parameter λ (Larger λ, more diverse).
- Does the GNN-strategy generalize across datasets from different sources?

Source \ Target	LIBSVM	MIRP	STOCH	GEN	SC-1	SC-2
LIBSVM	**0.8**	1.1	2.1	1.8	1.8	2.5
MIRP	1.1	**0.7**	1	3.5	1.2	1.2
STOCH	1	1.6	**0.9**	0.7	1.7	3.9
GEN	2	1.1	2.1	**0.1**	3.1	7.4
SC-1	1.2	0.8	3.1	1.1	**0.3**	**0.3**
SC-2	1	1	1.2	1	0.6	**0.3**



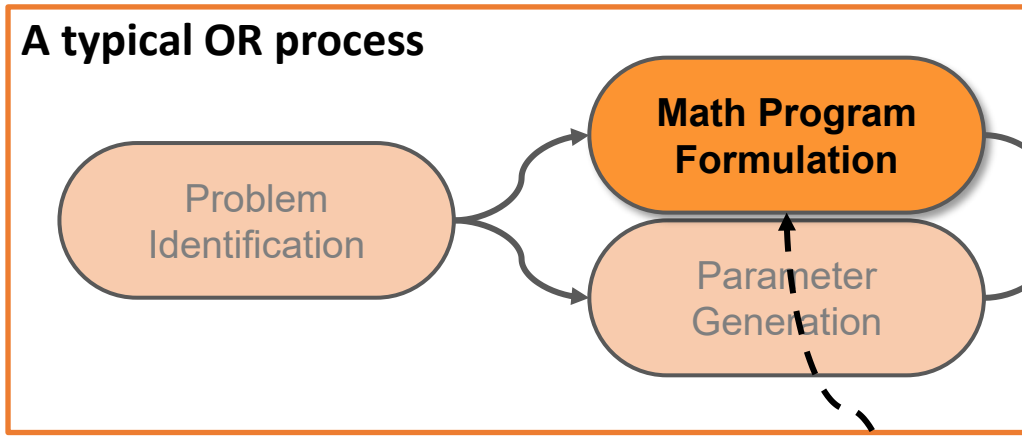
Roadmap

- Introduction
- To answer the three research questions:



- Conclusion and future directions

Scope



Problem description

Ben is growing apples and pears on his orchard. He has 50 acres available on which he must grow a minimum of 5 acres of apples and a minimum of 10 acres of pears to meet demands. The profit per apple is $2 and the profit per pear is $4. He prefers to grow more pears than apples but limitations in his workforce allow him to grow at most twice the amount of pears as apples. How many of each fruit should Ben grow in order to maximize his profit?

Math program (in variables-objective-constraint format)

```
Variables:
    x (acres of apples)                    Variables
    y (acres of pears)
Maximize: 2x + 4y (profit)                 Objective
Subject to:
    x + y <= 50 (land constraint)
    x >= 5 (apple minimum)                 Constraints
    y >= 10 (pear minimum)
    x <= y <= 2x (pear to apple ratio)
```

Expert or LLM

Formulation often requires expert experience. With the emerge of LLMs, non-experts can utilize it to generate formulations now. (Fan et al., 2025; Tasnim et al., 2024)

LLM

A new evaluation metric

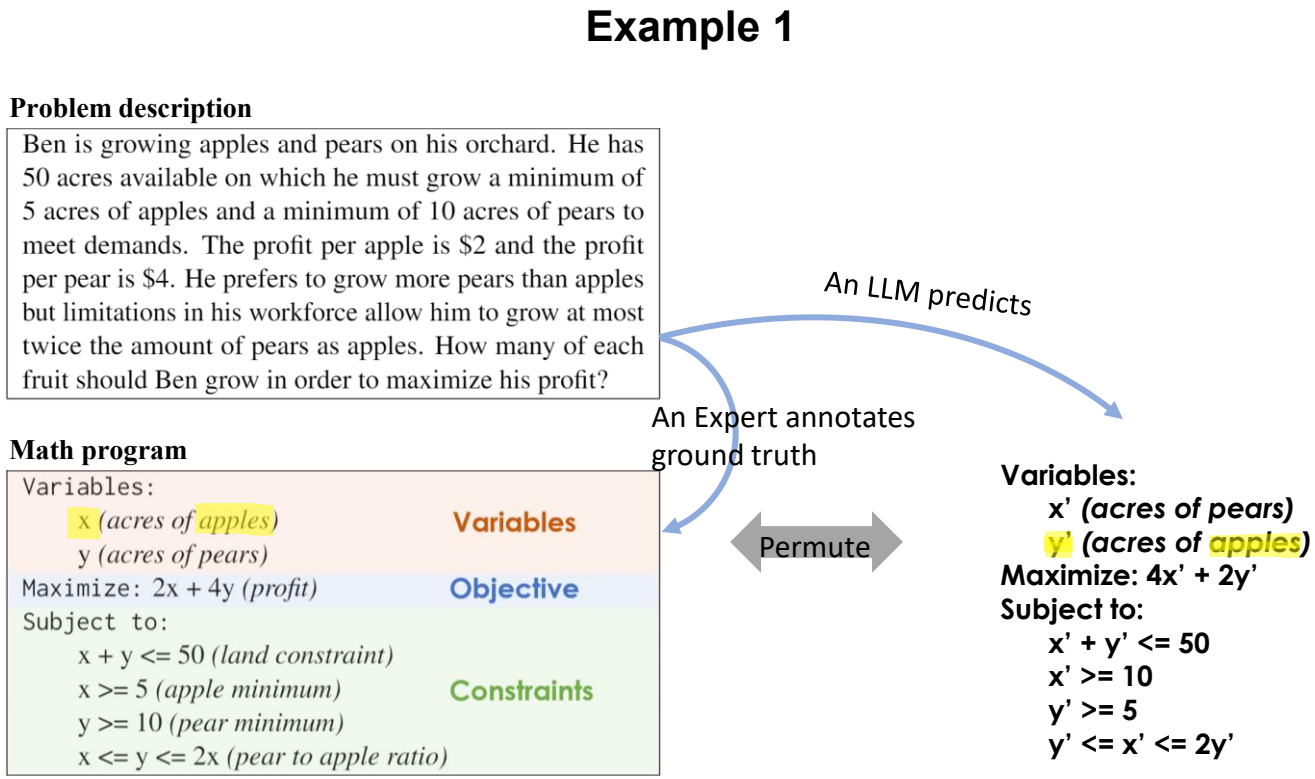
Existing metrics:

- Declaration-based evaluation metric
- Execution-based evaluation metric

Q2 Auto-formulation via Large Language Models (LLMs); How to evaluate?

Limitation of existing metrics

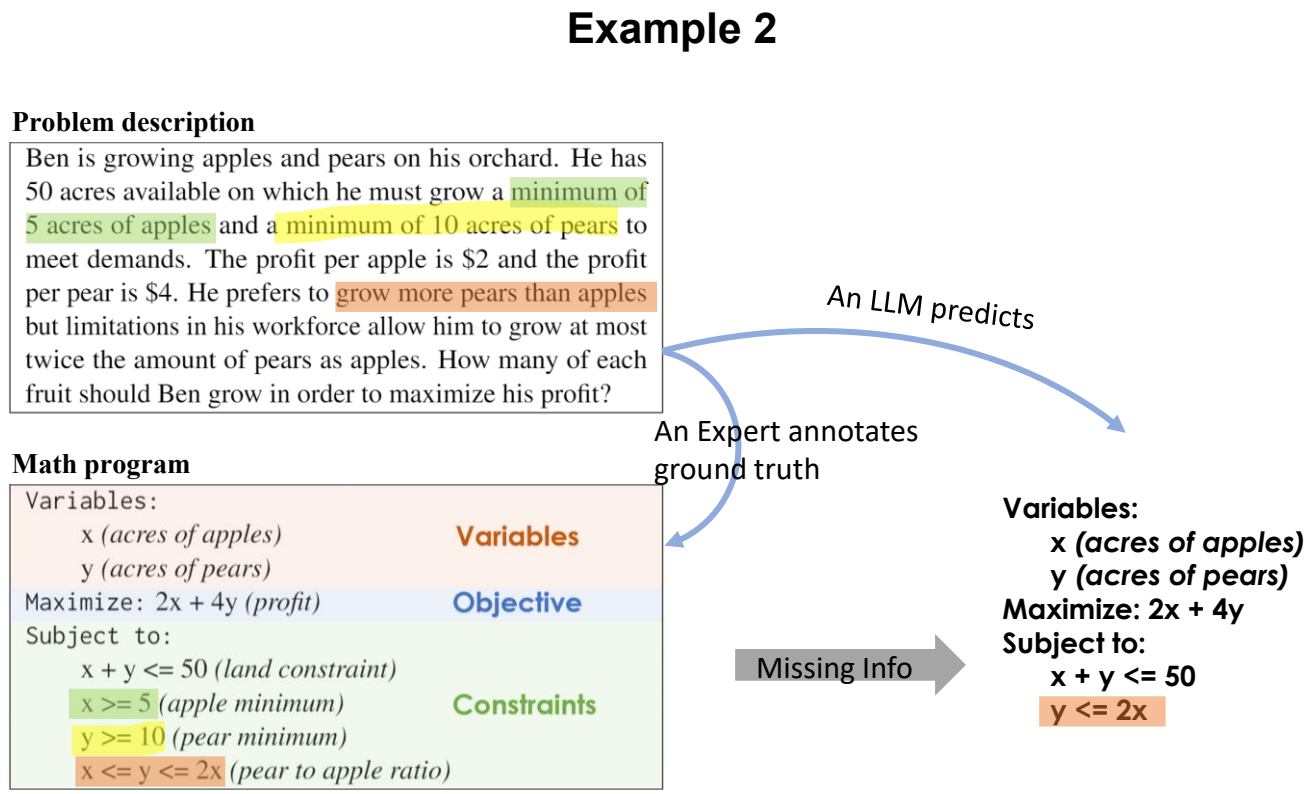
- Declaration-based evaluation metric (Ramamonjison et al., 2022)

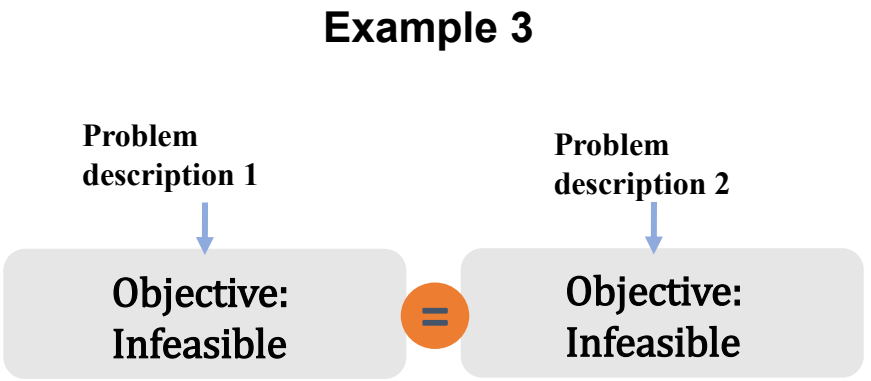


Permutation invariance

Limitations of existing metrics

- Execution-based evaluation metric (Prasath and Karande, 2023)





Variables:
$x = 16.\dot{6}, y = 33.\dot{3}$
Objective: $2x + 4y = 50$

=

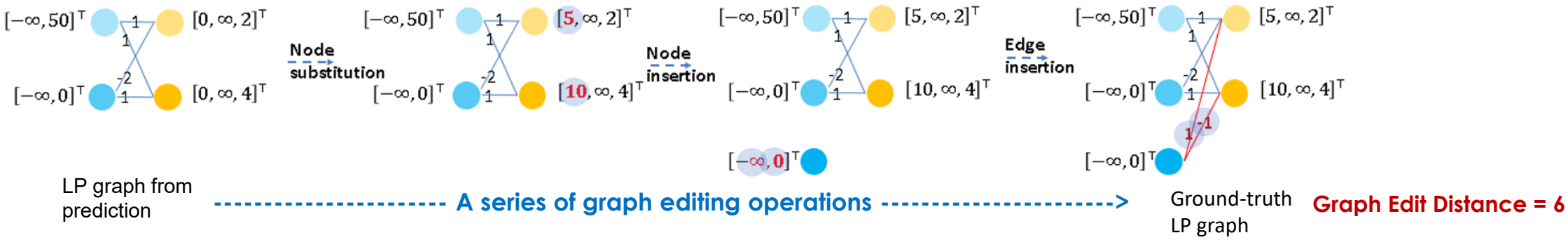
Variables:
$x = 16.\dot{6}, y = 33.\dot{3}$
Objective: $2x + 4y = 50$

✗ Measure consistency → original info

Proposed Evaluation Metric

- Design a metric that is better aligned with human judgment. Desired properties:
  1. Permutation invariance ⟵ Graph representation
  2. Measure consistency with the original problem's information. ⟵ #Attribute-level edits

- Propose a metric based on **graph edit distance** (GED-based metric)**.**
  - > Step 1&2: Format conversion. LLM prediction → LP general form → LP graph
  - > Step 3: Compute the GED between the LP graph from the LLM *prediction* and the *ground-truth* LP graph.



Experiment: which metric aligns better with human judgement?

- Setup
  - Dataset: NL4OPT test set (289 samples)
  - For each sample:
    - 4 predicted formulations from 4 LLMs
    - 1 human-written reference formulation
  - Human Evaluation → Human Ranking List
  - Automatic Metrics → Metric-based Ranking List
  - **Metrics**:
    - ***Exact match rate***:
      Are the two ranking lists identical?
    - ***Pairwise match rate***:
      Breaks down the ranking lists into pairs.

Metrics	Exact match rate	Pairwise match rate
Execution-based	9 / 289	716 / 1734
Declaration-based	64 / 289	1336 / 1734
GED-based	**178 / 289**	**1641 / 1734**

Table 3: Ranking match rate between automatic evaluation metrics and human judgements.

Roadmap

- Introduction
- To answer the four research questions:

ML enhances OR

- Smart initial basis selection for linear programs [ICML' 23]
- Towards human-aligned evaluation for linear programming word problems [COLING' 24]

OR enhances ML

- HFX: Joint Design of Algorithms and Systems for Multi-SLO Serving and Fast Scaling (In Submission, OSDI)

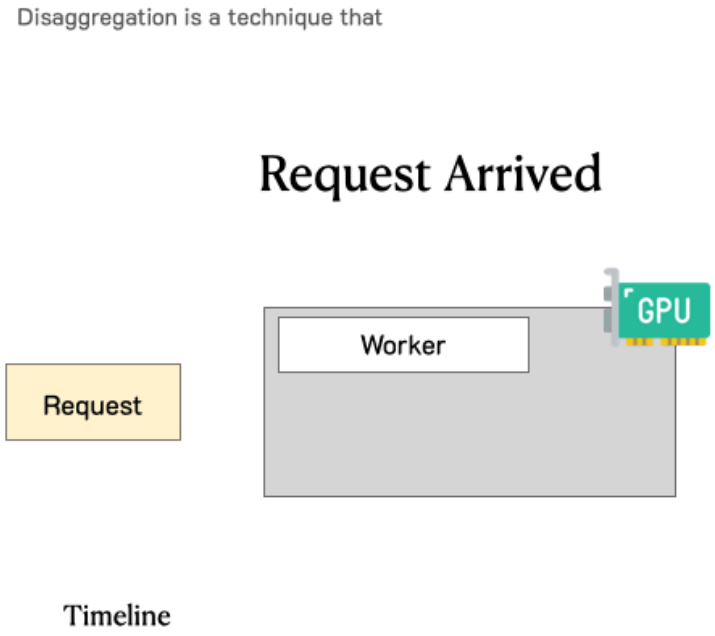
- Conclusion and future directions

Background

Challenge:
Heterogeneity & Variability

Q3 OR-inspired efficient serving strategy?

For a single request, inference consists of 2 stages:
Prefill stage and **Decode** stage



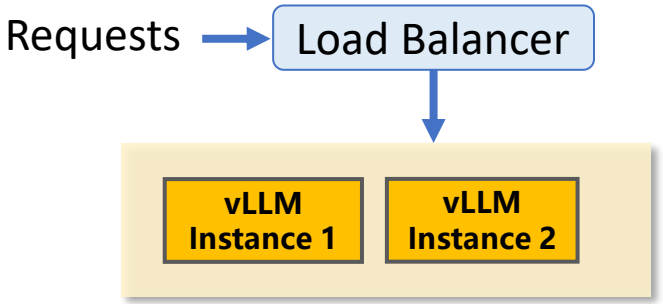
From *https://hao-ai-lab.github.io/blogs/distserve/*

Key service-level objectives (**SLO**s)

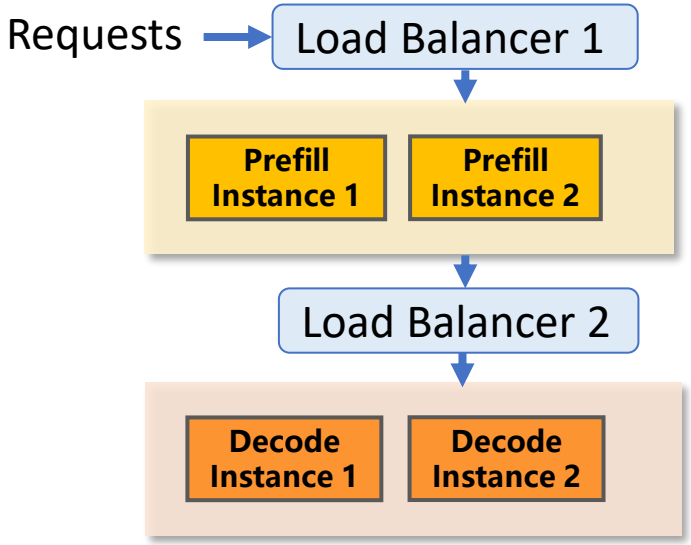
- Time to first token (TTFT) requirement.
- Time per output token (TPOT) requirement.

- Existing serving systems

vLLM (Kwon et al., 2023): colocated mode



DistServe (Zhong et al., 2024): disaggregated mode



Algorithms and Systems Designs

Challenges	Features	Designs
Heterogeneity	SLO-aware	Heuristic algo
Variability	Elasticity	
	Real-time decision	
	Fast scaling	System tech
2 modes	Compatibility	

Problem definition:

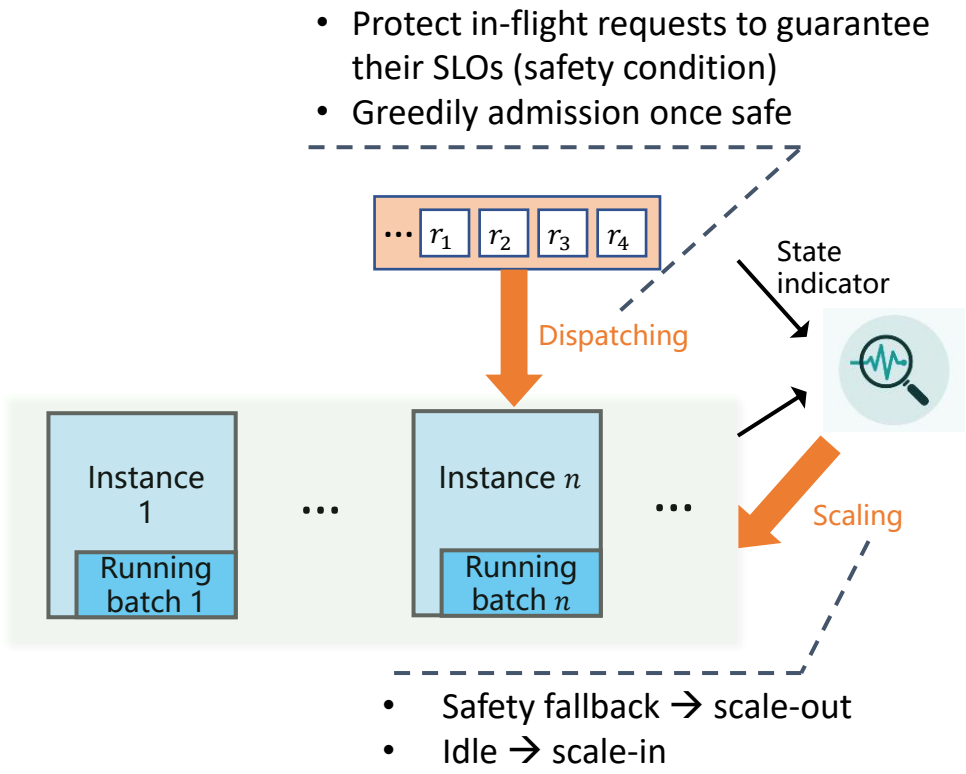
Decision Timing: On prefill/decode iteration completion and new request arrival

Decisions

- Dispatch: assign pending requests to workers
- Scaling: decide whether to scale and which workers to remove

Objectives: Minimize system cost & Maximize SLO attainment

Heuristic algo:

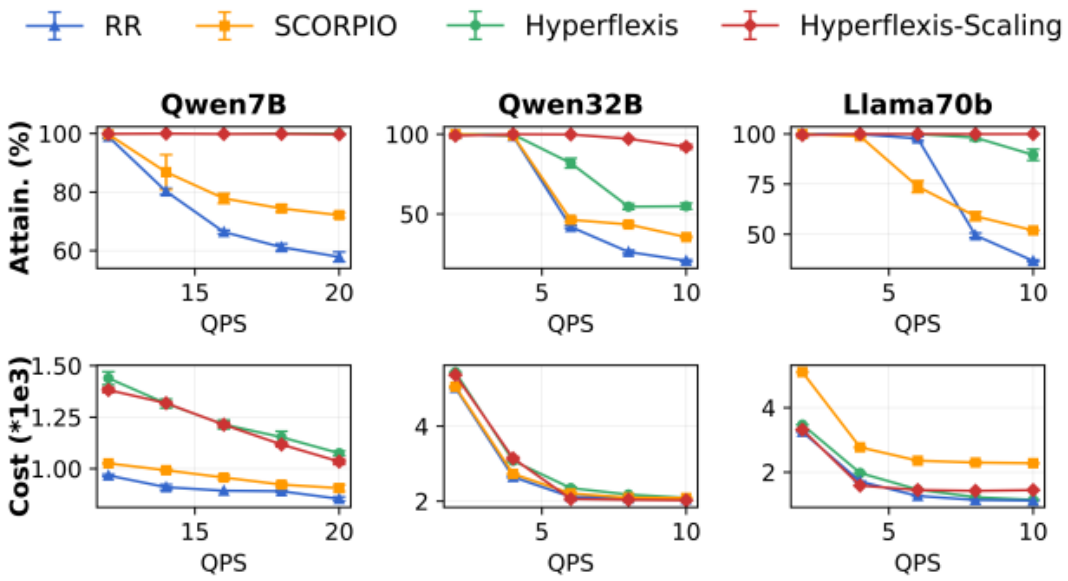


Experiment

- Setup:
  - ***Workloads***: 4-task multi-SLO benchmarks
  - ***Models:*** Qwen7B, Qwen32B, LLaMA70B (FP16, TP=1/2/8)
    - Deployment Mode: Colocated
  - ***#Instances:*** 2 by default, up to 4 with scaling
  - ***Baselines:***

Method	Multi-SLO aware	Scaling	Disagg.
RR	✗	✗	✓
SCORPIO	✓	✗	✗
HyperFlexis-Scaling	✓	✓	✓

  - ***Main metrics:*** SLO Attainment, Cost



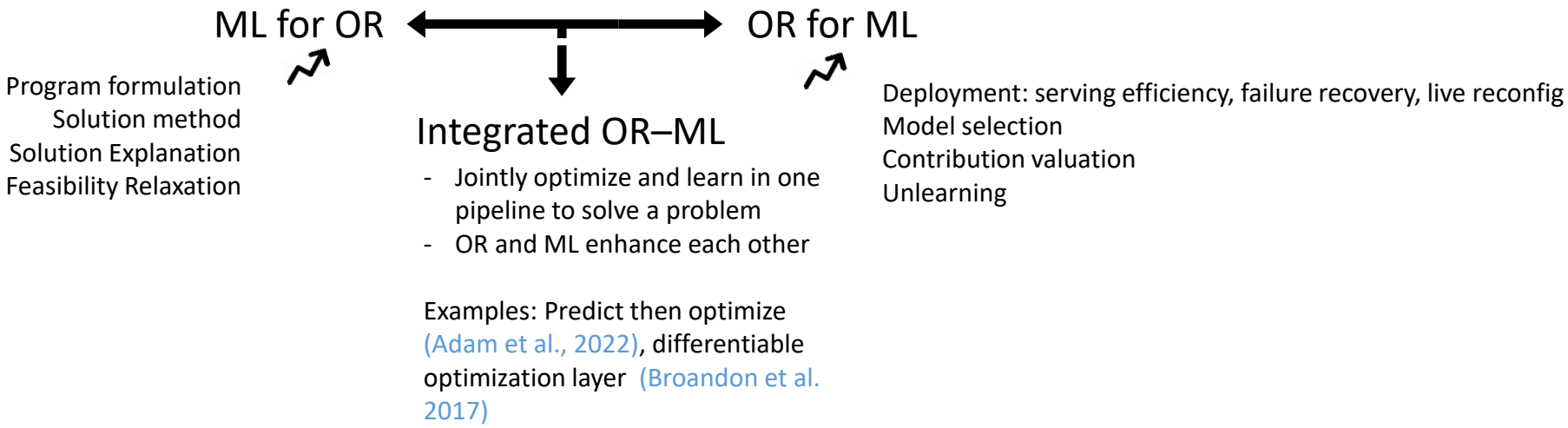
SLO attainment ↑↑
Cost modest ↑

SLO attainment ↑↑
Cost modest roughly same

Conclusion and future directions

- Demonstrated synergy between ML and OR.
  - **ML for OR:**
    - Proposed ML-guided LP initialization and accelerated correlated LPs' solving.
    - For LLM-based LP modeling, developed an improved evaluation metric.
  - **OR for ML:** Designed OR-inspired methods improving efficiency in LLM serving.

- Landscape of future directions:



Reference

(*co-first authors)

- [ICML' 23] Fan, Zhenan*, Xinglu Wang*, Oleksandr Yakovenko*, Abdullah Ali Sivas, Owen Ren, Yong Zhang, and Zirui Zhou. 2023. "Smart initial basis selection for linear programs." In *Proceedings of the 40th International Conference on Machine Learning*, 9650–9664.
- [COLING' 24] Xing, Linzi*, Xinglu Wang*, Yuxi Feng, Zhenan Fan, Jing Xiong, Zhijiang Guo, Xiaojin Fu, Rindra Ramamonjison, Mahdi Mostajabdaveh, Xiongwei Han, Zirui Zhou, and Yong Zhang. 2024. "Towards Human-aligned Evaluation for Linear Programming Word Problems." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 16550–16556.
- (In Submission, OSDI) Yousefijamarani, Zahra*, Xinglu Wang*, Qian Wang*, Morgan Lindsay Heisler, Taha Shabani, Niloofar Gholipour, Parham Yassini, Hong Chang, Kan Chen, Qiantao Zhang, Xiaolong Bai, Jiannan Wang, Ying Xiong, Yong Zhang, and Zhenan Fan. 2025. "HyperFlexis: Joint Design of Algorithms and Systems for Multi-SLO Serving and Fast Scaling." *arXiv preprint* arXiv:2508.15919.
- Other papers during PhD study:
  - Fan, Zhenan, Huang Fang, Xinglu Wang, Zirui Zhou, Jian Pei, Michael P. Friedlander, and Yong Zhang. 2024. "Fair and Efficient Contribution Valuation for Vertical Federated Learning." In *Proceedings of the Twelfth International Conference on Learning Representations.*
  - Gholami, Mohsen, Mohammad Akbari, Xinglu Wang, Behnam Kamranian, and Yong Zhang. 2023. "Etran: Energy-Based Transferability Estimation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18613–18622.
  - Fan, Zhenan, Bassem Ghaddar, Xinglu Wang, Lingzi Xing, Yong Zhang, and Zirui Zhou. 2025. "Artificial Intelligence for Optimization: Unleashing the Potential of Parameter Generation, Model Formulation, and Solution Methods." *European Journal of Operational Research*.
  - Singh, Gursimran, Xinglu Wang, Yifan Hu, Timothy Yu, Linzi Xing, Wei Jiang, Zhefeng Wang, Xiaolong Bai, Yi Li, Ying Xiong, Yong Zhang, Zhenan Fan. 2025. "Efficiently Serving Large Multimodal Models Using EPD Disaggregation." In *Proceedings of the 42nd International Conference on Machine Learning*.
  - Heisler, Morgan; Yousefijamarani, Zahra; Wang, Xinglu; Wang, Qian; *et al.* 2025. LLM Inference Scheduling: A Survey of Techniques, Frameworks, and Trade-offs. *TechRxiv Preprint.* DOI: 10.36227/techrxiv.176238087.79673350 (v1).

Thank you!