KDD2026
August 9–13, 2026
Jeju, Korea

AlphaSearch : Agentic AI for Price Arbitrage in Energy Systems

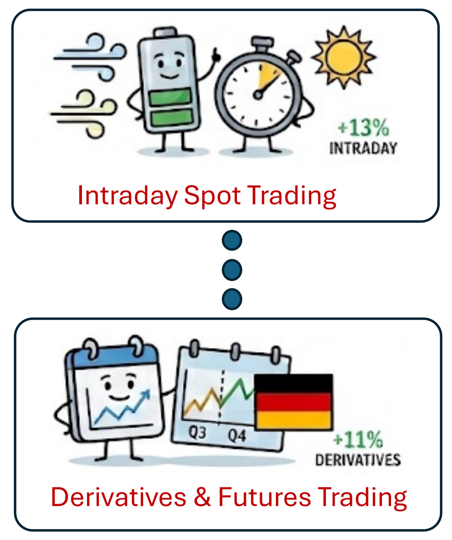
Millend Roy*, Vladimir Pyltsov*

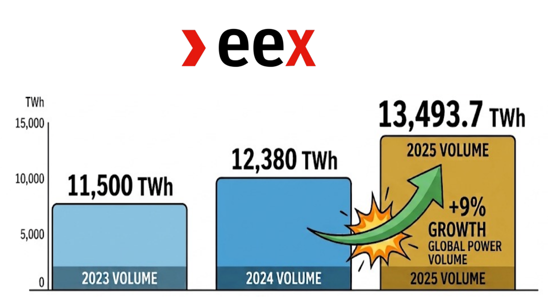
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science
IEOR Industrial Engineering & Operations Research COLUMBIA | ENGINEERING
Quadracci SUSTAINABLE ENGINEERING LAB.

1 Motivation

- Energy markets present lucrative trading opportunities, contributing to the steady rise in trading volume.

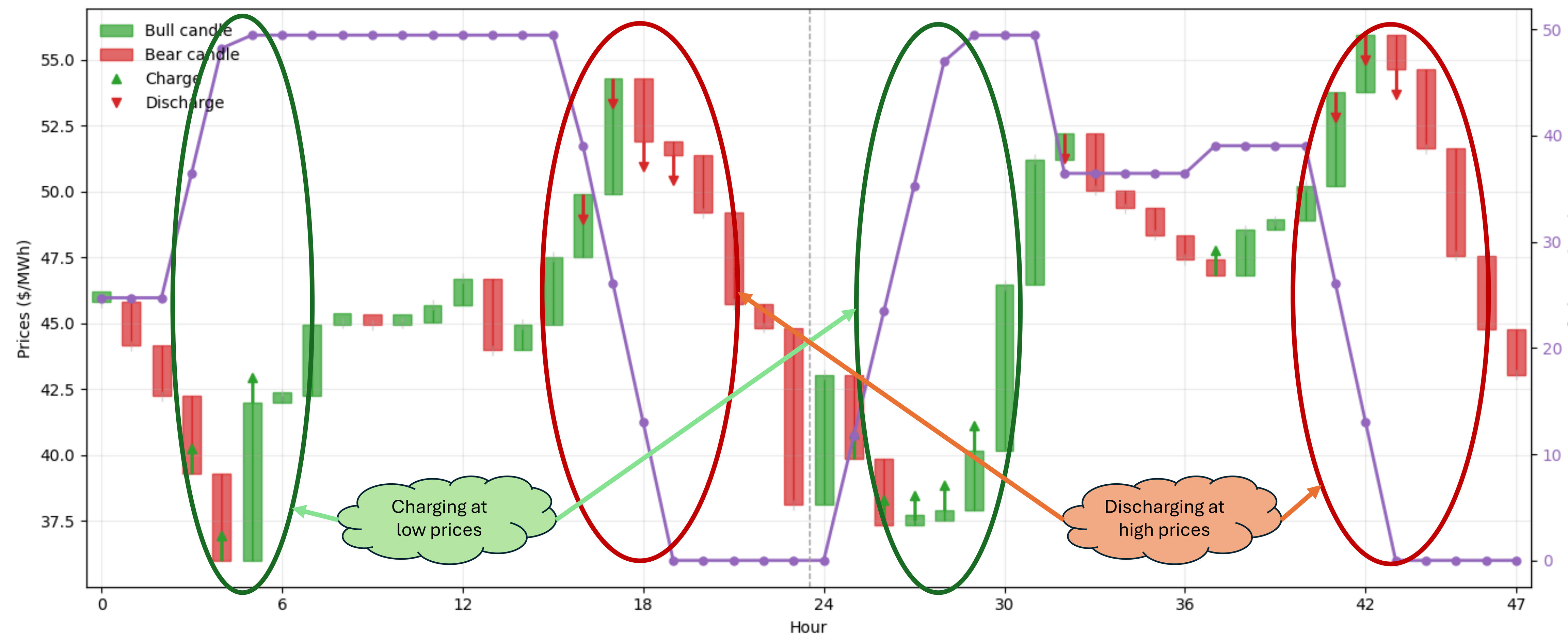
Strategies:





- Pre-trained Foundation Models show strong generalization with zero-shot capabilities
  - How to employ cross-modality adaptation to extract trading opportunities from Time-Series via Text-Prompts.
  - Reasoning Capabilities to provide more interpretation into the decision-making?

2 Introduction: "Temporal Arbitrage" as Intraday Trading Strategy!



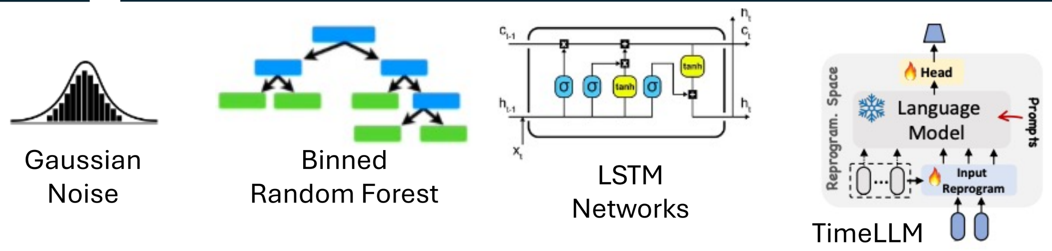
Challenges:

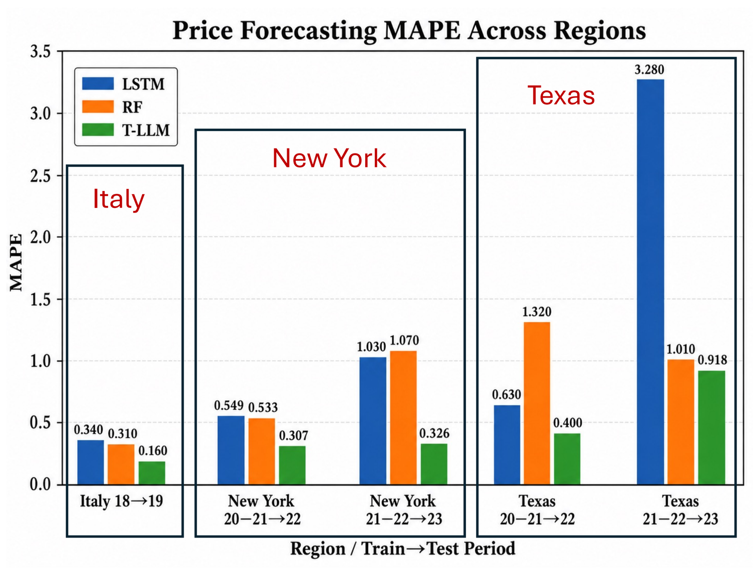
- Not just a buy-low/sell-high problem
- Practical Deployment requires more than a single optimization model:
  - Validate data
  - Forecast future price candles
  - Optimization based on trading strategy
  - Feasibility Checking
  - Benchmarking Trading PnL
  - Stress Testing on scenarios
  - Reasoning Explanation behind the decision making.

Data:

Zenodo

3 Forecast Layer Benchmarks





- Italy: stable market
  NY: moderately volatile market
  TX: highly volatile market
- Across Different Classes: TimeLLM emerges as winner.
- TimeLLM forms the building block for the forecasting layer in the architecture.

4 Optimization Problem Setup

Battery Parameters $\theta^{\text{bat}} = \begin{pmatrix} C, \mathrm{SoC}_{init}, \mathrm{SoC}_{\min}, \mathrm{SoC}_{\max}, \\ P_c^{\max}, P_d^{\max}, \eta_c, \eta_d, \mathrm{SoC}_{target} \end{pmatrix}$

Market ISO Daily Data $\xi_i = \left(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_i, \mathbf{D}_i, \hat{\mathbf{D}}_i, \mathbf{X}_i, \theta_i^{\text{market}}\right)$

Goal : Generate a schedule by $\min \sum_{t=0}^{T-1} \pi_t \, P_{\text{net}}(t)$

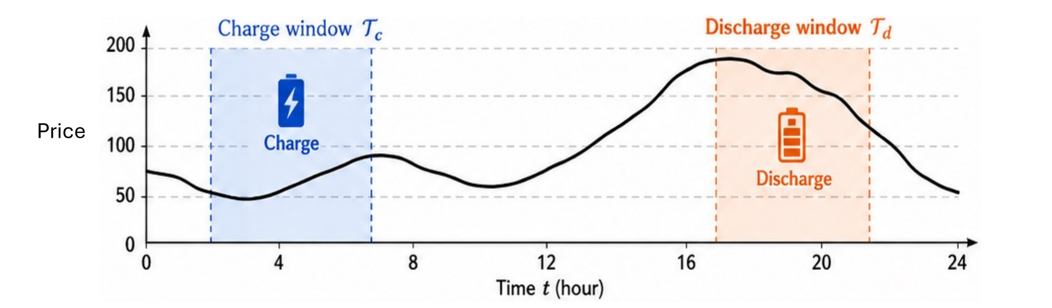
$$P_{\text{net}}(t) = D_t + P_c(t) - P_d(t)$$
$$0 \le P_c(t) \le P_c^{\max}$$
$$0 \le P_d(t) \le P_d^{\max}$$

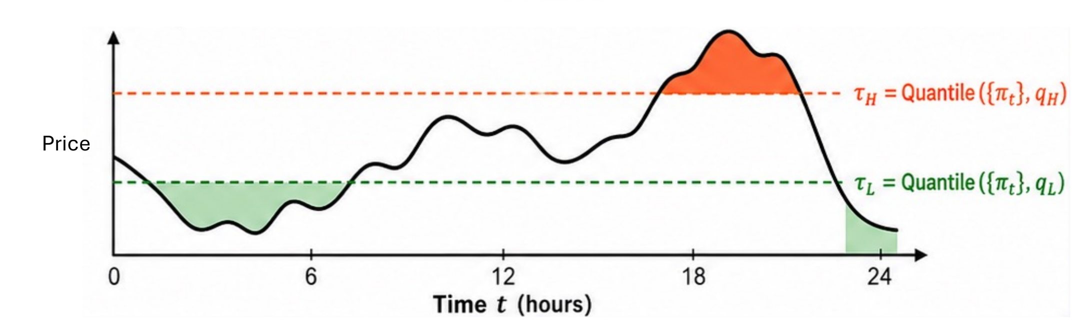
$$\mathrm{SoC}_{\min} \le \mathrm{SoC}_t \le \mathrm{SoC}_{\max}$$
$$\mathrm{SoC}_0 = \mathrm{SoC}_{init}$$
$$\mathrm{SoC}_T \ge \mathrm{SoC}_{target}$$

$$\mathrm{SoC}_{t+1} = \Pi_{[\mathrm{SoC}_{\min}, \mathrm{SoC}_{\max}]}\left(\mathrm{SoC}_t + \frac{\eta_c P_c(t)\Delta t}{C} - \frac{P_d(t)\Delta t}{\eta_d C}\right)$$

5 Optimization Benchmarks

1) Heuristic Baselines:





2) Mixed Integer Linear Programming Baselines:

$$0 \le P_c(t) \le P_c^{\max} y_c(t), \qquad t = 0, \dots, T-1,$$
$$0 \le P_d(t) \le P_d^{\max} y_d(t), \qquad t = 0, \dots, T-1.$$
$$y_c(t) + y_d(t) \le 1, \qquad t = 0, \dots, T-1$$

Bias-Corrected

bias $\hat{b}_t = \frac{1}{N}\sum_{j=1}^{N}(\pi_t^j - \hat{\pi}_t^j)$

$\tilde{\pi}_t = \hat{\pi}_t + \hat{b}_t$

$\min \sum_{t=0}^{T-1} \tilde{\pi}_t \, P_{\text{net}}(t)$

Stochastic

Construct scenarios $\left\{\hat{\pi}_t^{(s)}\right\}_{s=1}^{S}$

$\hat{\pi}_t^{(s)} = \tilde{\pi}_t + \epsilon_t^{(s)}$

$\min \frac{1}{S}\sum_{s=1}^{S}\sum_{t=0}^{T-1} \hat{\pi}_t^{(s)} P_{\text{net}}^{(s)}(t)$

Risk-Aware

Scenario Cost

$L_s = \sum_{t=0}^{T-1} \hat{\pi}_t^{(s)} P_{\text{net}}^{(s)}(t)$

$\min_{\alpha, z_s} \quad \alpha + \frac{1}{(1-\beta)S}\sum_{s=1}^{S} z_s$

$z_s \ge L_s - \alpha \ge 0, \forall s.$

α VaR & β quantile

6 Our Models

1) Reinforcement Learning (PPO)

State

$s_t = [\mathrm{SoC}_t, \hat{\pi}_{t:t+W}]$

Daily z score prices

PPO agent

Policy $\pi_\theta(a_t|s_t)$

$a_t \in [-1, 1]$

Near Horizon Mask.

Avoid late charging

Action to Power

$P_c(t) = \max(0, a_t)\, P_c^{\max}$
$P_d(t) = \max(0, -a_t)\, P_d^{\max}$

Reward

$r_t = -\pi_t P_{\text{net}}(t)\Delta t - \lambda_{\text{smooth}}|a_t - a_{t-1}|$

$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$

2) LLM Solvers: API Call + QLoRA fine-tuned Qwen

QLoRA Fine-tuning (learn forecast to SoC mapping from MILP oracle trajectories)

- *Training Pairs* $x_i = (\xi_i, \theta^{\text{bat}})$, y_i *Prompt s*
- ChatML masking
- Supervised QLoRA adapters 4-bit base
- Fine-tuned Qwen p_θ

Inference

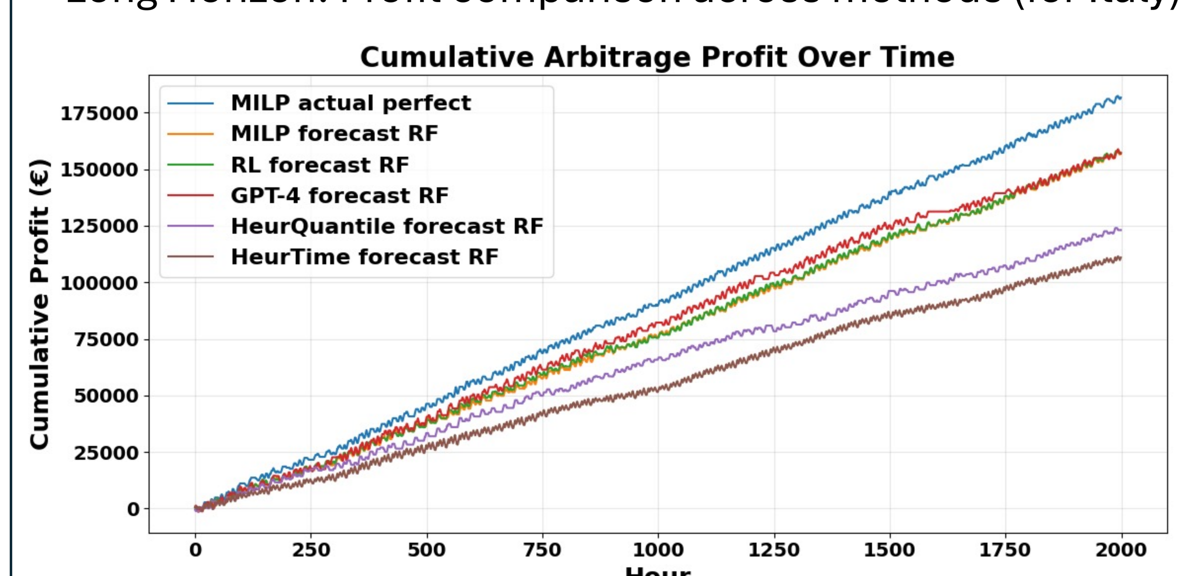
- *New instance* x_{new} *Prompt s*
- GPT-4 API-call: *CoT + greedy decode*
- Fine-tuned Qwen: *decode*
- Candidate schedule
- Validation & Repair Layer
- Feasible Schedule o/p

7 Experimental Results

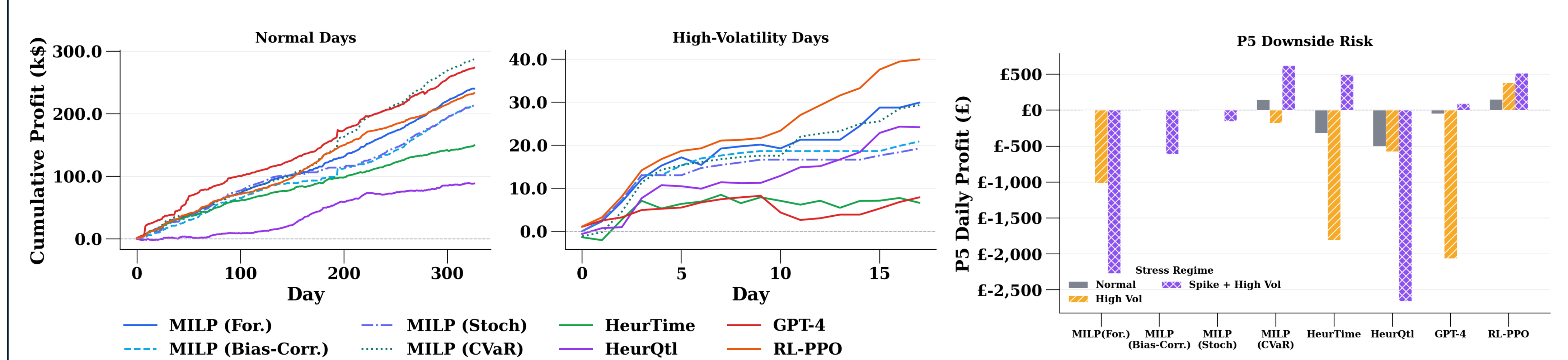
Profits across price forecasting methods (for Italy)

Model	Perfect	RF	LSTM	T-LLM
MILP	2246.09	1954.99	1901.09	2050.31
MILP (Bias-Corr.)	**2246.17**	1955.13	1983.45	2110.62
MILP (Stochastic)	2246.17 ± 0	1958.88 ± 15.49	1942.35 ± 41.10	2111.43 ± 24.65
MILP ($CVaR_{\beta=0.95,\lambda=0.5}$)	2246.17 ± 0	1980.15 ± 44.89	1975.23 ± 25.99	2080.71 ± 64.88
Heuristic-1	1425.69	1425.69	1425.69	1425.69
Heuristic-2	1525.86	1910.13	**2048.62**	1009.23
Qwen (Fine-tuned)	2245.76	**2191.13**	2031.47	**2130.14**
GPT-4 (API-call)	2093.80	1547.55	1417.72	1703.67
RL (PPO)	2009.78 ± 169.57	1957.28 ± 118.51	1952.05 ± 83.54	2098.59 ± 42.46

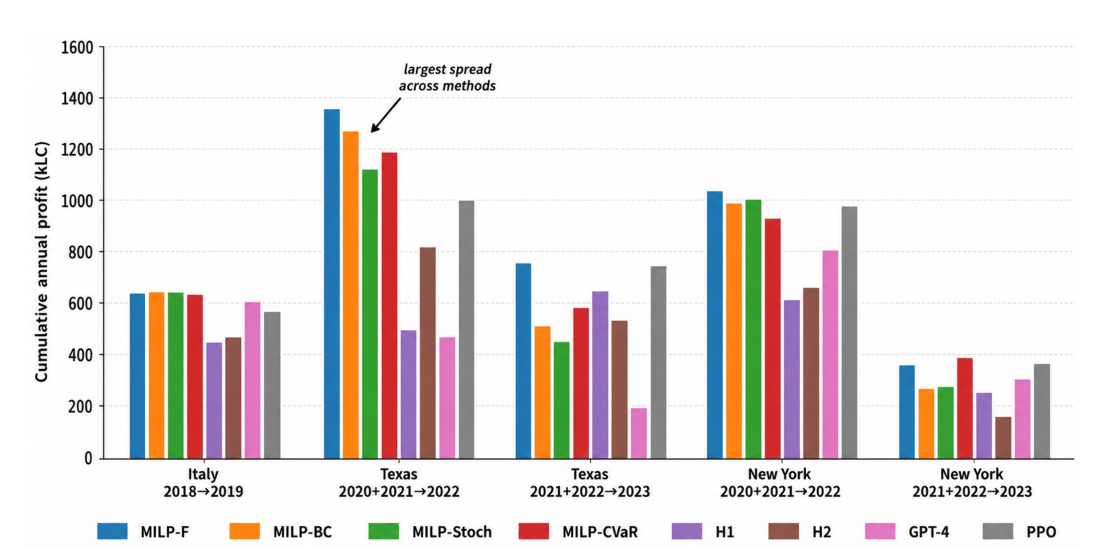
Long Horizon: Profit comparison across methods (for Italy)



8 Stress Test Results in Normal vs High Volatility Days in New York 2023



9 Cum. Profit Comparison across Regions



10 Limitations

- **No battery degradation model**
- **Forecast and RL sensitivity:** Profit depends on forecast quality, while PPO performance is sensitive to reward scaling, rollout length, learning rate, and exploration choices.
- **Limited long-horizon LLM performance:** Fine-tuned Qwen performs well on daily tasks but degrades over longer horizons due to possible overflowing numerical tokenization issues, and limited model capacity.
- **Post-hoc explanations only:** Generated rationales improve interpretability but are not formal guarantees and may omit solver-specific nuances.

Thank you, contact me at millend.roy@columbia.edu or my co-author at v.pyltsov@columbia.edu !