From Algorithmic and RL-based to LLM-powered Agents

Bo AN

Nanyang Technological University
Skywork AI

August 7, 2024 @IJCAI'24

Agent Mediated Intelligence Research Group

Autonomous Agents: Problems and Approaches

➢ **Autonomous Agents** can address many critical and complex problems

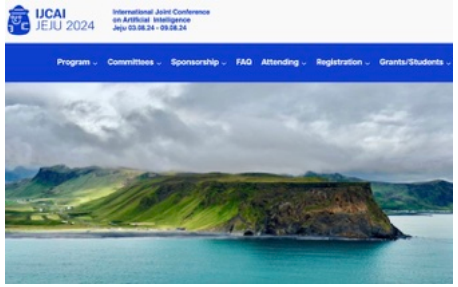
- Stochastic, open environment
- Multiple players
- Sequential decision, online
- Strategic (selfish) behavior
- Distributed optimization

Public Security

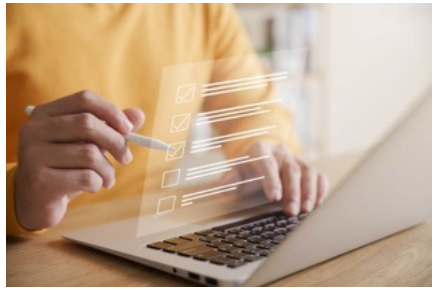
Pursuit Evasion

Recommendation

Financial Trading



Web Control



Computer Control

➢ Approaches

Algorithmic (since 2006)	**RL-based** (since 2017)	**LLM-powered** (since 2023)
Leveraging **optimization techniques** to solve **small-scale** problems	Leveraging **Reinforcement Learning (RL)** to solve large-scale **specific** problems	Leveraging **Large Language Models (LLMs)** for solving **general** problems

Outline

Algorithmic
(since 2006)
Leveraging **optimization techniques** to solve **small-scale** problems

RL-based
(since 2017)
Leveraging **Reinforcement Learning (RL)** to solve large-scale **specific** problems

LLM-powered
(since 2023)
Leveraging **Large Language Models (LLMs)** for solving **general** problems

- Security Games
- Equilibrium Finding
- Pursuit-Evasion Games

Security Games

- Real-world security problems

Terrorist attack

Robbery

Oil Tanker Attack Incident

- Security Games (pioneered by Milind Tambe)
  - Limited security resources
  - Adversary monitors defenses, exploits patterns
- Game theory for security resource scheduling **[60+@ AAAI, AAMAS, IJCAI, NeurIPS, ICML]**
  - INFORMS Daniel H. Wagner Prize for Excellence in Operations Research Practice (2012), etc
  - Operational Excellence Award from US Coast Guard (2012), etc
  - United States congressional hearing

Computing Quantal Correlated Equilibrium in NFGs [EC'22]

- Fighting against terrorism requires **coordinating** agents
  - Coordination → Correlated equilibrium
  - Signaler sends signal to other agents for coordination

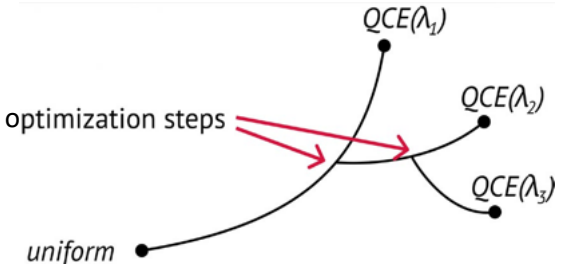
- Real-world players are **subrational**!
  - Subrational → Quantal response equilirium
  - Quantal correlated equilibrium (QCE) given signal λ

$$u_i(a_i|s_i) = \sum_{a_{-i}\in A_{-i}} \sum_{s_{-i}\in S_{-i}} \lambda(s_i, s_{-i})\delta_{-i}(a_{-i}|s_{-i})u_i(a_i, a_{-i}), \qquad \delta_i(a_i|s_i) = \frac{q_i(u_i(a_i|s_i))}{\sum_{a_i'\in A_i} q_i(u_i(a_i'|s_i))}$$

  - Existing methods cannot be applied as none supports both **coordination** and **subrationality**

- Our method
  - **PPAD-hard** to compute a QCE given the signal distribution
  - Computational methods for QCE
    - Trace QCE using a homotopic system
    - Optimize the signal structure

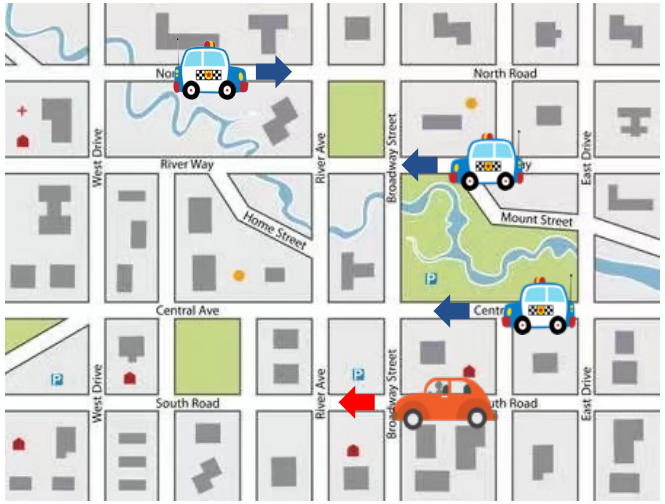
$$\lambda^{\iota+1} \leftarrow P_\Lambda\left(\lambda^\iota + \eta f'(\lambda, \delta_p(\lambda))\right)$$



Pursuit-Evasion Game (PEG)

- Chase scenes in the real world

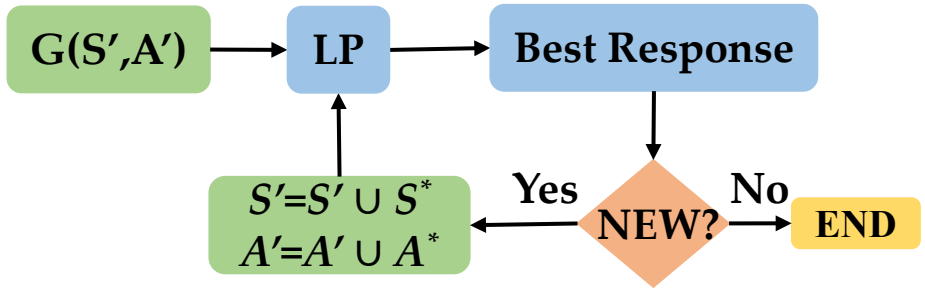
- Game model
  - Players: A team of Pursuers vs. Evader
  - Actions: Move on a graph
  - End conditions
    - Exit nodes: evader can escape from exit nodes
    - Limited time: play limited # of steps
    - Capture: one pursuer and the evader reach the same node around the same time



  - Zero-sum **imperfect-information extensive-form** game
  - **Scalability** challenge: action space **exponentially** increases with the size of the map, time horizon, # of pursuers

Algorithmic Approaches for PEGs

- Incremental Strategy Generation-based Approaches



Escape Interdiction Game Solver (EIGS) [IJCAI 2017]

- **Mixed-integer linear programming** for best response oracle
- **EIGS: 8*8** sparse grid map

Iteratively Generated Reachable States (IGRS++) [AAAI 2019]

- **Novel pruning techniques** for best response oracle
- **Manhattan** map with **702** nodes, **1216** edges, **4** pursuers, #game step=**26**

ISG variant for TME(ISGT,CISGT) [ICML 2020]

- First ISG **guaranteeing** to converging to **team-maxmin equilibria**
- Every pursuer takes actions **independently**
- **10*10** grid map

- Limitations
  - **Scalability**: Cannot scale up to **realistic-sized** games
  - **Transferability**: Cannot quickly adapt to other **similar** games

Reinforcement Learning is prominent in solving long-horizon and large-scale decision making problems

Outline

Algorithmic
(since 2006)
Leveraging optimization techniques to solve small-scale problems

RL-based
(since 2017)
Leveraging **Reinforcement Learning (RL)** to solve large-scale **specific** problems

LLM-powered
(since 2023)
Leveraging Large Language Models (LLMs) for solving general problems

- DRL for Pursuit-Evasion Games
- DRL for Industry, esp recommender systems
- DRL for Financial Trading

Deep (R)L-based Approaches for PEGs

- 15*15 grid map, 8 pursuers

- Singapore map, 8 pursuers
- Manhattan map, 6 pursuers

- Singapore map, no need to train NE from scratch for similar games
- Quickly adapt to PEGs with different initial conditions

CFR-MIX [IJCAI 2021]

NSGZero [AAAI 2022]

Grasper [AAMAS 2024]

2021 → 2022 → 2023 → 2024

NSG-NFSP [IJCAI 2021]

PSRO with Representation Learning [AAAI 2023]

- Singapore map with 372 nodes, 1470 edges, 4 pursuers, #game step = 300

- 100*100 grid map, no need to train the best response from scratch in PSRO

From Specialist to More Scenarios

➢From the **single scenario** to more **different** decision-making scenarios

Pursuit-Evasion Games (PEGs)

Zero-Sum

[IJCAI 2024]

RENES

Normal-Form

- Use **RL to modify game payoffs** to solve more effectively
- **GNN** to handle games with **different sizes** of action sets

SPSRO

Extensive-Form

- **Self-adaptive PSRO**: A **unification** of different PSRO variants
- **Transformer-based** HPO policy capable of transfe-ring to diff. games

Zero-Sum
General-Sum

[ICML 2024] **CMD**

- **Configurable Mirror Descent (CMD)**: A **single algorithm** to handle **more different** decision-making scenarios
- Different nunmbers of agents, relationships, solution concepts, and evaluation measures

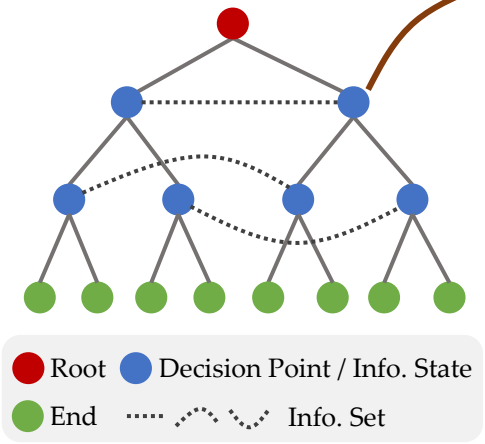
Single-Agent
Cooperative
Zero-Sum
General-Sum
Mixed Coop. & Comp.

Configurable Mirror Descent (CMD) [ICML'24]

- Motivating scenario: develop a skilled robot
  - The user may ask the robot to learn to complete **different tasks**

Single-Agent, Cooperative, Zero-Sum, General-Sum, Mixed Coop.-Comp. (MCC)

Can we develop A Single Algorithm to handle these different tasks?



Policy Optimization at Each Decision Point / Info. State

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}_{a\sim\pi} Q(a)$$

SOTA: Magnetic Mirror Descent (MMD)

$$\pi_{k+1} = \arg\max_{\pi\in\Pi} \langle Q(\pi_k), \pi\rangle - \alpha \mathrm{KL}(\pi, \pi_1) - \frac{1}{\eta}\mathrm{KL}(\pi, \pi_k)$$

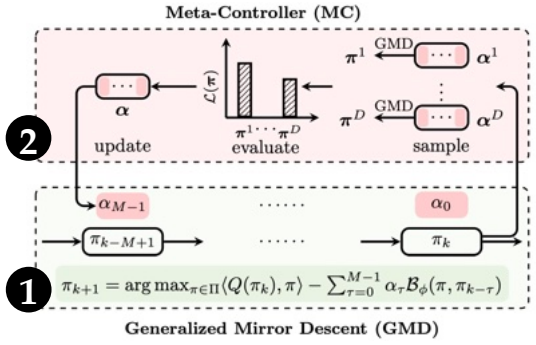
- MMD only focuses on **single-agent** and **two-player zero-sum** games/tasks

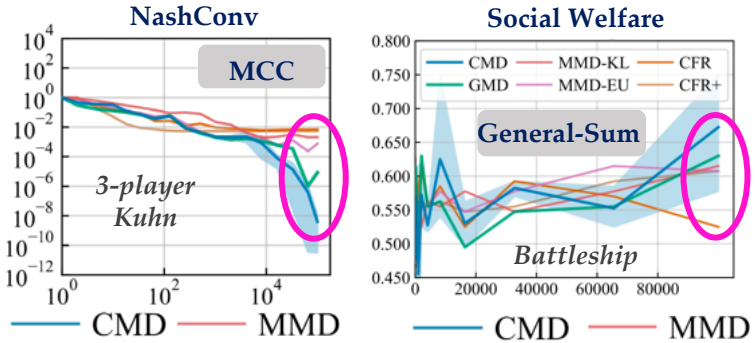
Configurable Mirror Descent (CMD)

$$\pi_{k+1} = \arg\max_{\pi\in\Pi} \langle Q(\pi_k), \pi\rangle - \sum_{\tau=0}^{M-1} \alpha_\tau \mathcal{B}_\phi(\pi, \pi_{k-\tau})$$

(2) α_τ (1) $\mathcal{B}_\phi(\pi, \pi_{k-\tau})$

1. No one regularizer is **panacea** for **all** tasks → Consider **other convex functions** (***i.e.*, more different types** of Bregman div.)
2. **Meta-Controller** to adjust hyper-para. **conditional on solution concepts and evaluation measures (configurable)**





- CMD achieves consistent **competitive** or **better** performance in diff. tasks compared to SOTA
- CMD can be easily **configured** for **different solution concepts** and **evaluation measures**

RL for Industry: Some of Our Deployed Applications

Fraud Detection

Impression allocation [IJCAI'18]

Robust fraud detection [WWW'19]

Ride-hailing

Joint pricing and dispatch [ICDM'19]

Recommendation

Online mobile [RecSys'20]

Multi-module [RecSys'20]

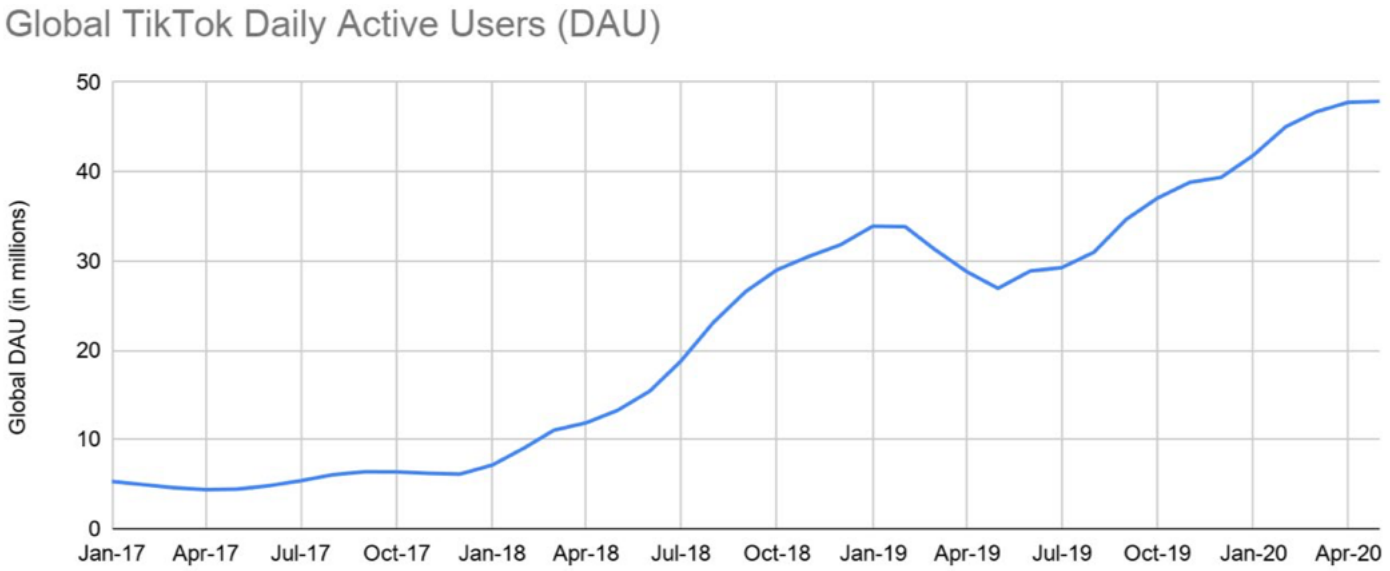
Social media platform[ICLR'23, KDD'23, NeurIPS'23]

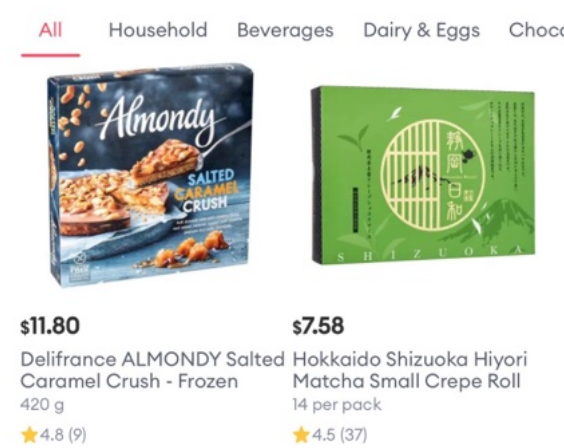
Financial Trading

Portfolio management [AAAI'21, WWW24]

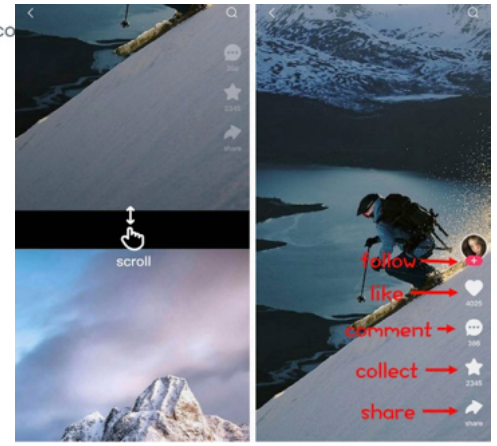
Algorithmic trading [CIKM'22, AAAI'24, KDD'24]

RL for Optimizing User Retention in Recommender Systems



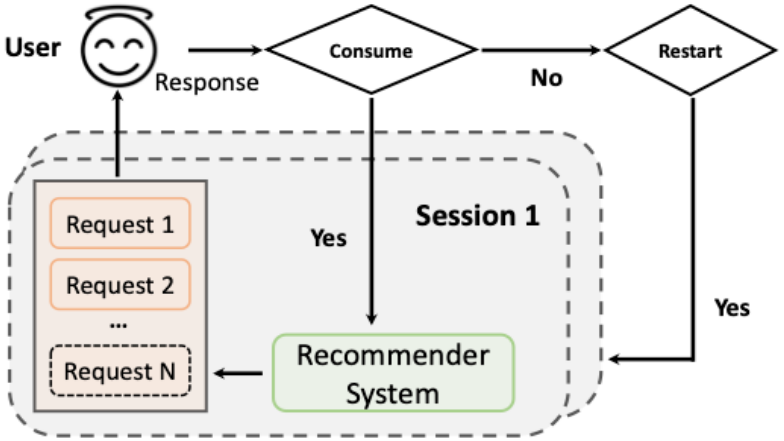


Goods



Short Video

- RL for Recommendation
  - User as environment, item as policy
- User Retention
  - Optimize user return time & session length
  - Directly related to the key metric: DAU



- Challenges: **Delayed feedback, high training cost, non-stationary environment**

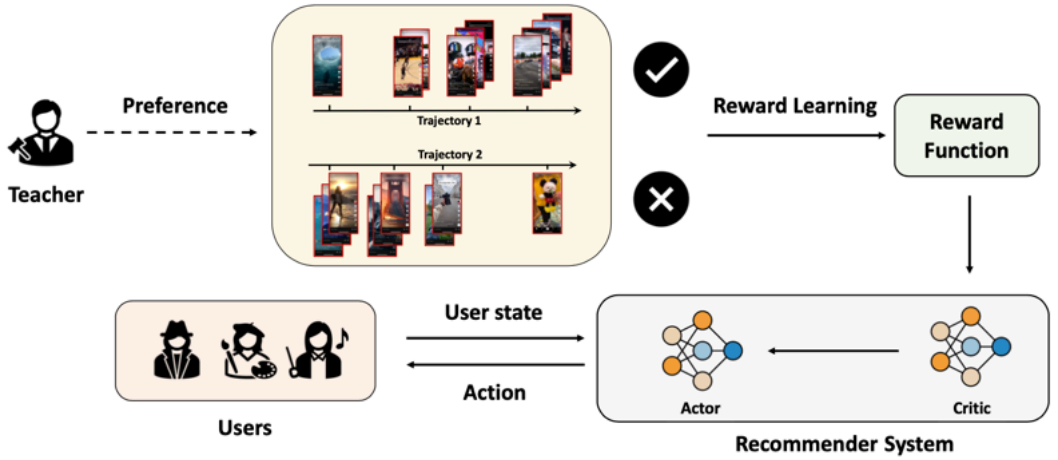
PrefRec: Recommender Systems with Human Preferences for Reinforcing Long-term User Engagement [KDD' 23]

- User retention signals are **sparse** and **delayed**
  - How to design immediate reward value to evaluate each recommending action?
- Reinforcing retention from preferences rather than rewards
  - Predicting the preference via Bradley-Terry model

$$P_\psi\left[\sigma^1 \succ \sigma^0\right] = \frac{\exp\left(\sum_t \hat{r}\left(\mathbf{s}_t^1, \mathbf{a}_t^1; \psi\right)\right)}{\sum_{i\in\{0,1\}} \exp\left(\sum_t \hat{r}\left(\mathbf{s}_t^i, \mathbf{a}_t^i; \psi\right)\right)}.$$

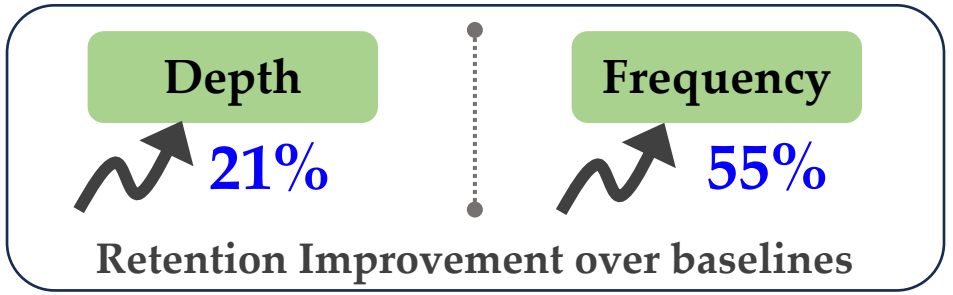
  - Learn the reward model automatically

$$\mathcal{L}_\psi = -\mathop{\mathbb{E}}_{(\sigma^0,\sigma^1,y)\sim\mathcal{D}_p}\left[y(0)\log P_\psi\left[\sigma^0 \succ \sigma^1\right] + y(1)\log P_\psi\left[\sigma^1 \succ \sigma^0\right]\right]$$



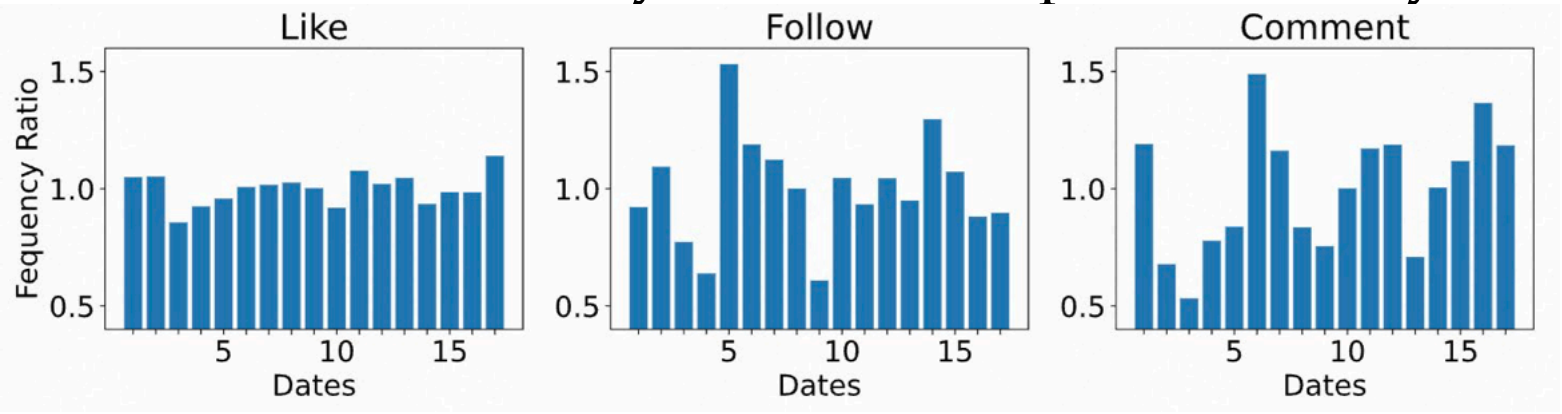
First study of RLHF in recommender systems

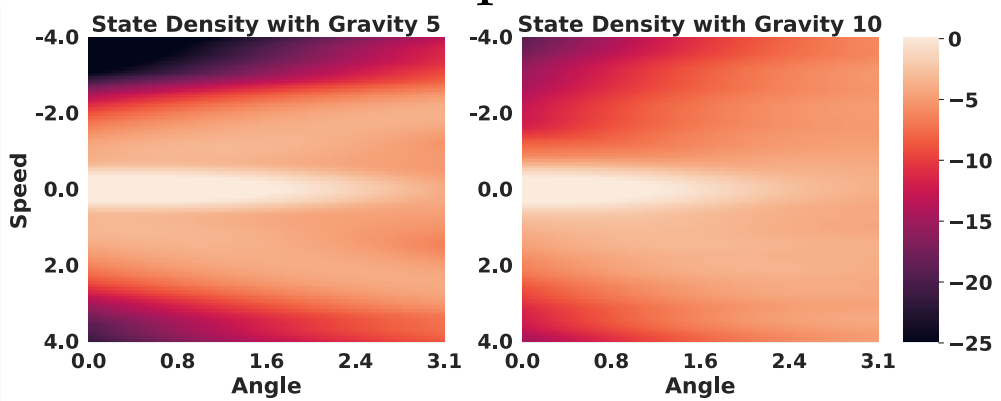
- Robust training under learned reward functic
  - Learn both V and Q for stable updates
  - Expectile regression for the V-function



State Regularized Policy Optimization with Dynamics Shift [NeurIPS'23]

- How to effectively train **online** policies for dynamic user behavior patterns?



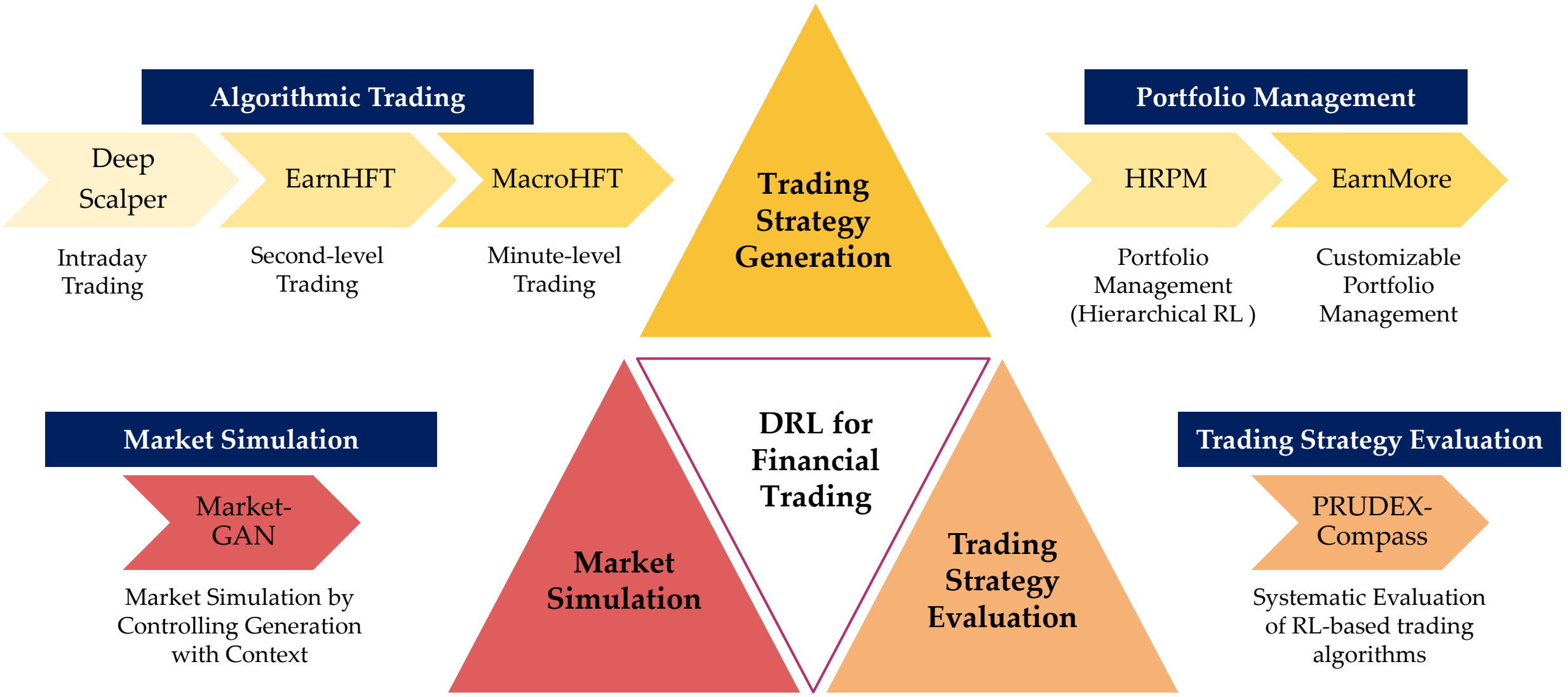


- Motivation: Different MDPs with similar structures can have similar optimal stationary state distribution
- Regularize policy training with the invariant and optimal state distribution.

$$\max_{\pi} \mathbb{E}_{s_t, a_t \sim \tau_\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad \text{s.t.} \quad D_{\mathrm{KL}}\left(d_\pi(\cdot) \| \zeta(\cdot)\right) < \varepsilon$$

- Live A/B tests
  - On a short-video recommendation platform with **millions** of users
  - Involve **1%** of the data flow, last **15** days
  - Reduce hate rate by **1.87%**, improve return time by **0.138‰**, session length by **0.263‰**

DRL for Financial Trading



DRL for Algorithmic Trading [CIKM'22, AAAI'24, KDD'24]

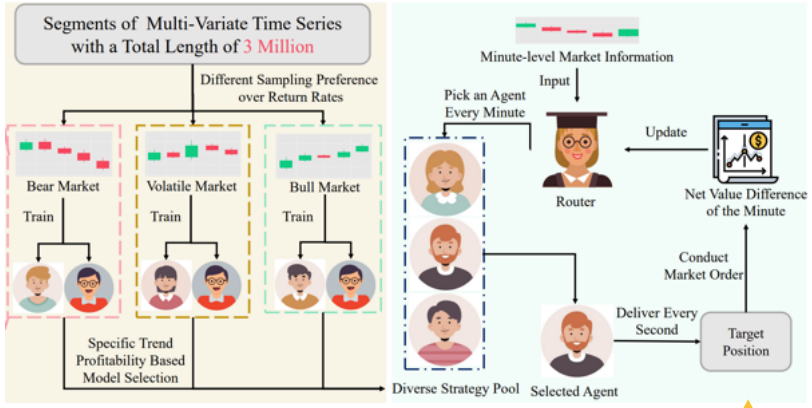
➢Challenges

- ❑Non-stationary market
- ❑Temporal Distribution Shift

Bull **Volatile** **Bear**

➢Approaches

- ❑Low-level RL agents with optimal supervisor for various market conditions
- ❑High-level RL agent to pick agents dynamically



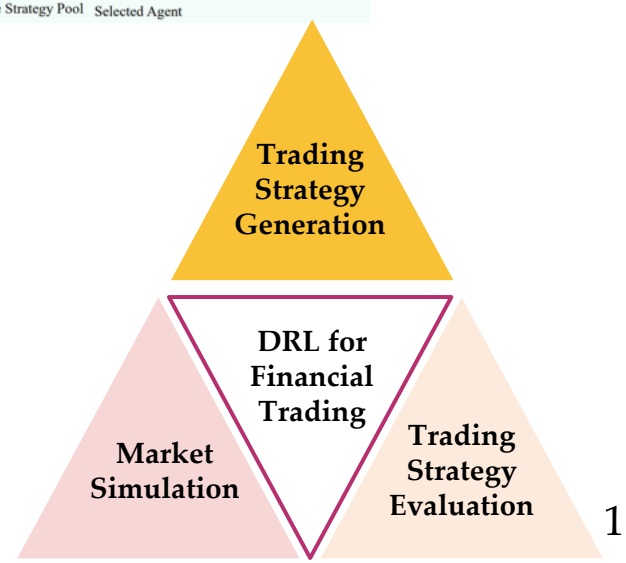
➢Results

4 cryptocurrencies

80%

Improvement over baselines

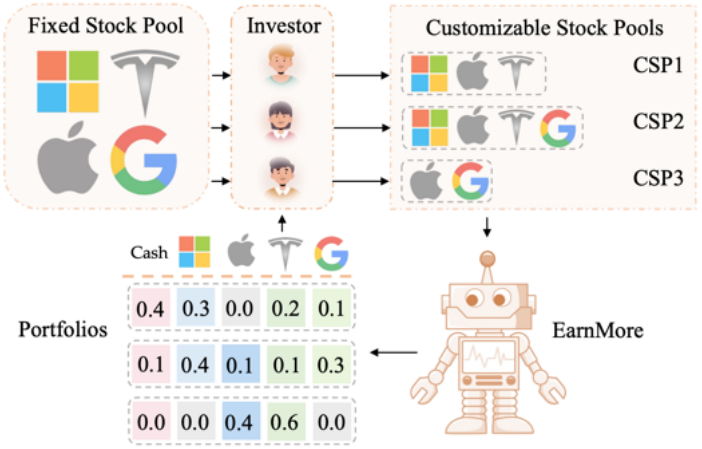
Trading Example on BTCUSDT



DRL for Portfolio Management [AAAI'21, WWW'24]

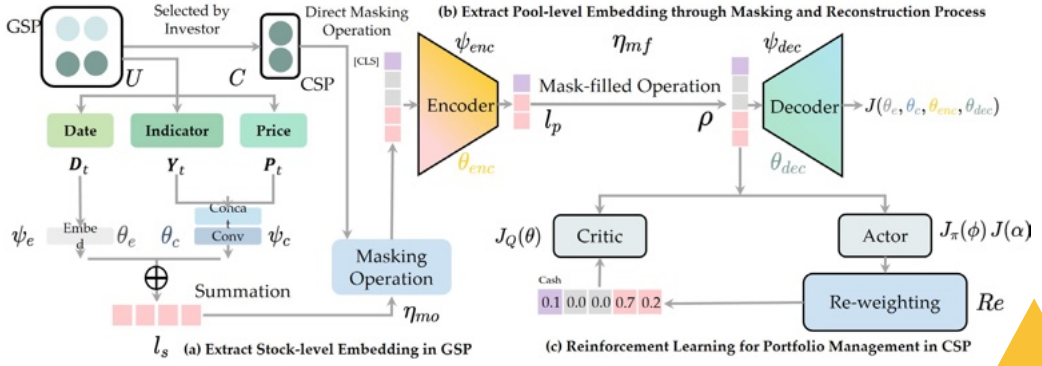
➢Motivation

- ❑Customizable stock pool



➢Approaches

- ❑Learnable masked token for customizable stock pools
- ❑Meaningful embeddings from a self-supervised masking and reconstruction process
- ❑Reweighting mechanism for favorable stocks

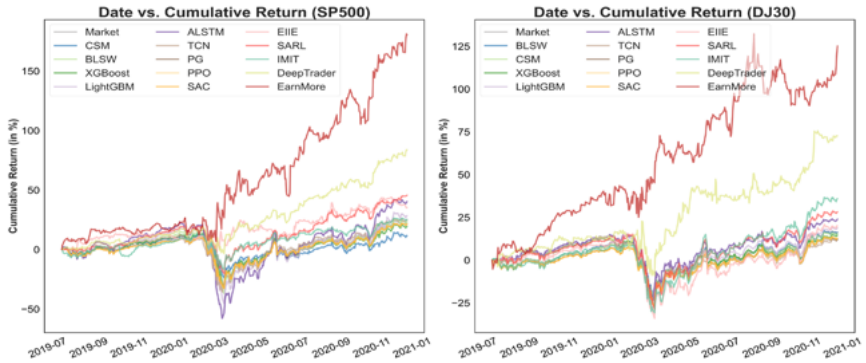


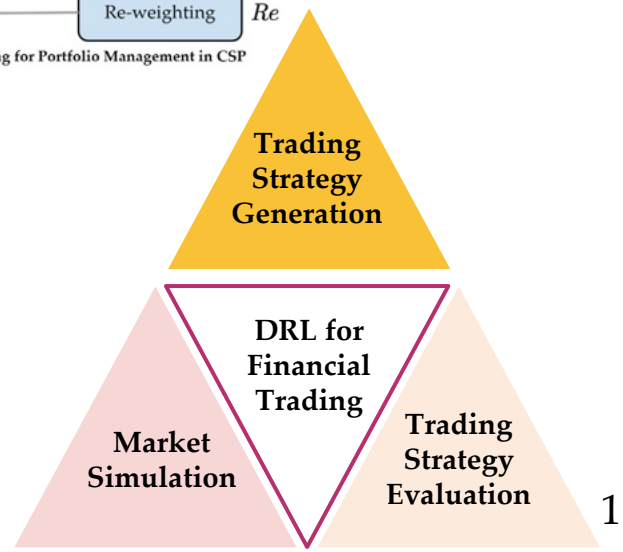
➢Results

SP500 61%

DJ30 47%

Improvement over baselines





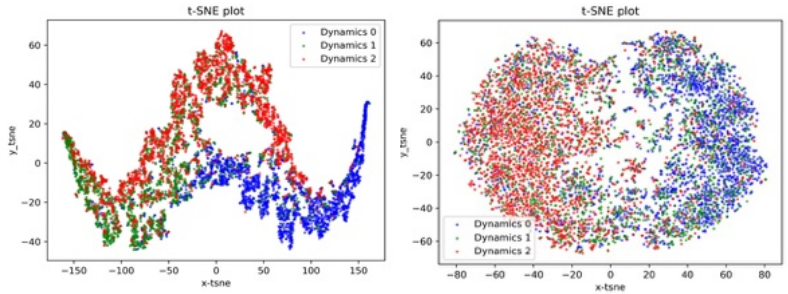
Market Simulation and Strategy Evaluation [AAAI'24, TMLR'23]

- Market-GAN: Towards Context-aligned Market Data Generator
  - Supervisor to transfer context knowledge to the generator
  - Two-stage training scheme to capture intrinsic market distribution

Alignment 89%

Prediction 36%

Improvement over baselines

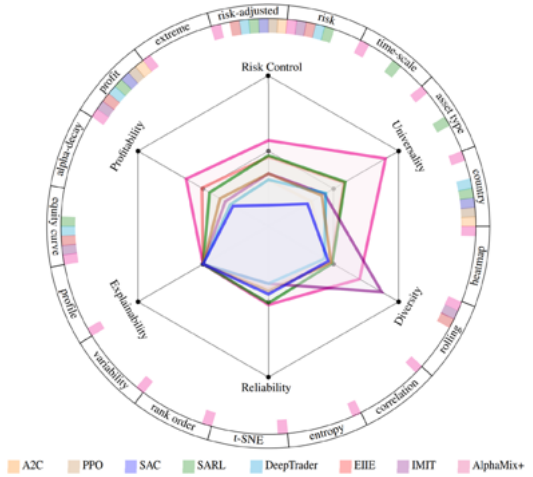


Market-GAN Real Data

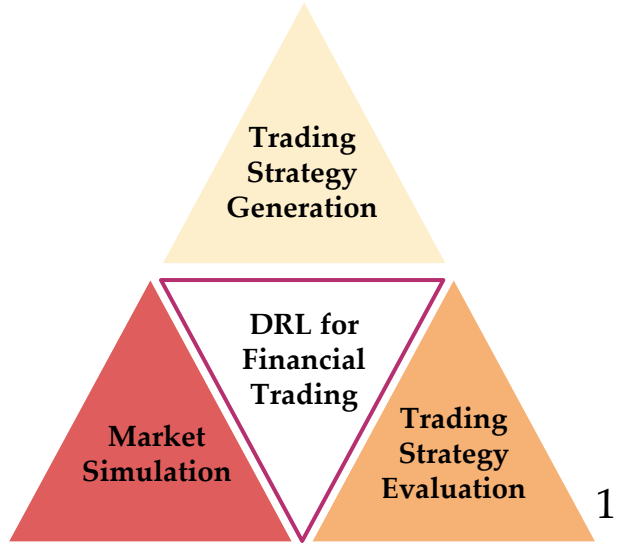
- PRUDEX-Compass: Towards Systematic Evaluation

Comprehensive Evaluation

6	17	6
Main Metrics	Specific Measures	Visualization Toolkits



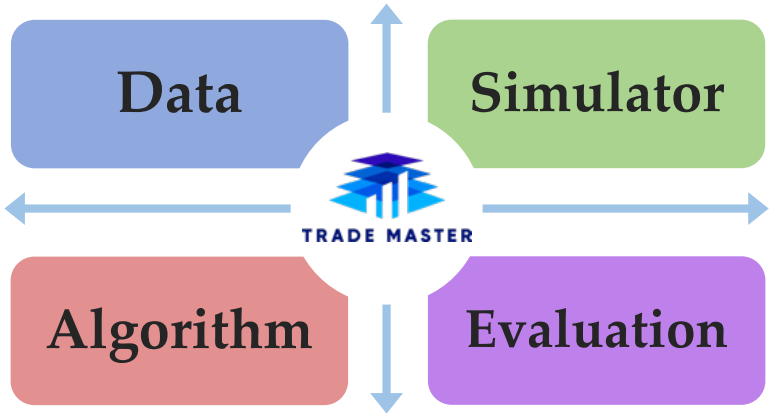
Metrics Hexagon



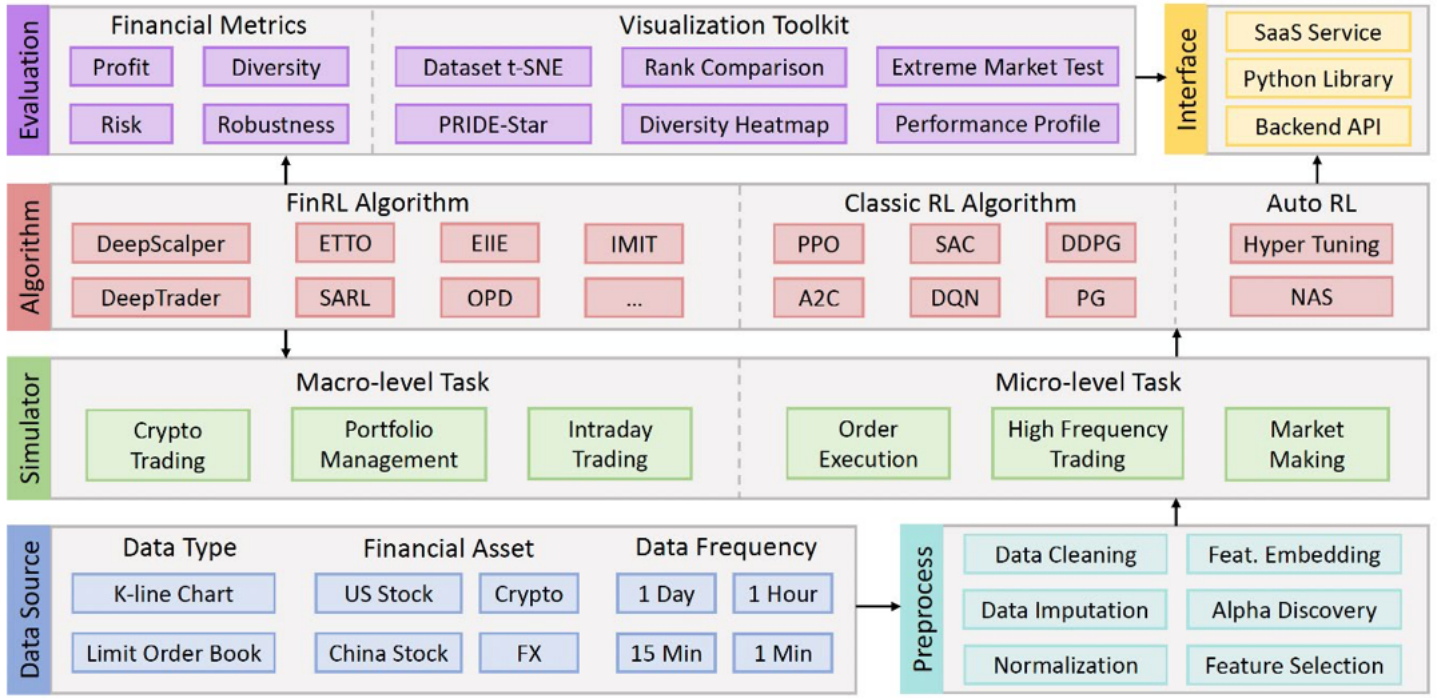
TradeMaster Platform [NeurIPS'23]

GitHub Over 1.3K stars

- A first-of-its kind, best-in-class RL platform for trading



Four Core Components



Data Source		Trading Algorithms		Market Simulator		Trading Evaluation	
6	13	6	16	18	29	6	10
Markets	Datasets	Trading Scenarios	RL-based Trading Algorithms	Controllable Parameters	Stock Tickers from DJIA index	Visualization Toolkits	Evaluation Measures

Github link: https://github.com/TradeMaster-NTU/TradeMaster.git

Outline

Algorithmic
(since 2006)
Leveraging **optimization techniques** to solve **small-scale** problems

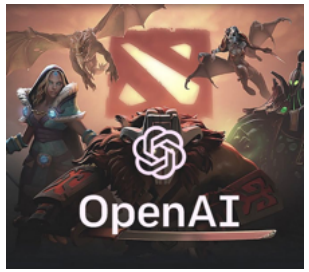
RL-based
(since 2017)
Leveraging **Reinforcement Learning (RL)** to solve large-scale **specific** problems

LLM-powered
(since 2023)
Leveraging **Large Language Models (LLMs)** for solving **general** problems

- LLM+RL for Decision and Reasoning
- LLM-powered Agent for Computer Control
- LLM-powered Trading Agent

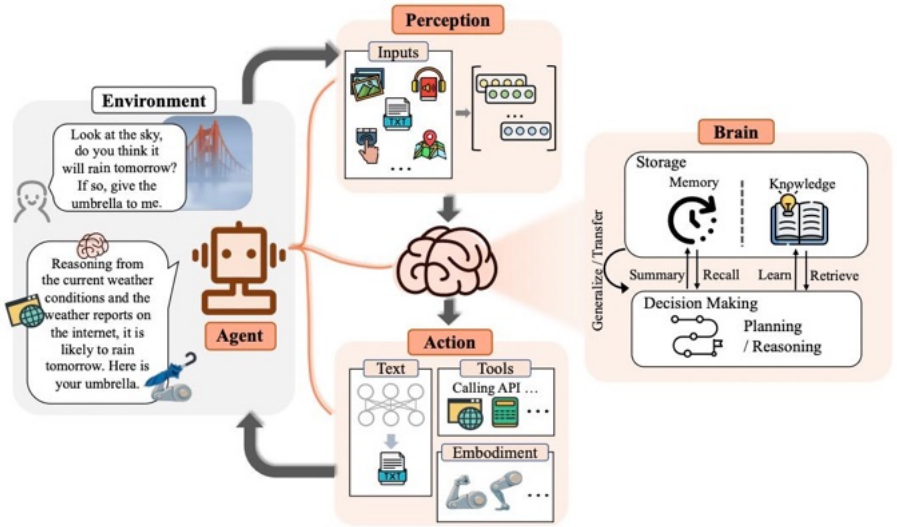
From RL-based Agents to LLM-powered Agents

- Limitations/Challenges of RL-based agents
  - Sample-inefficient
  - Task-specific
  - Long-horizon tasks
  - Sparse reward



- LLM-powered agents
  - World knowledge
  - Reasoning/planning
  - Tool use (especially code)
  - In-context learning

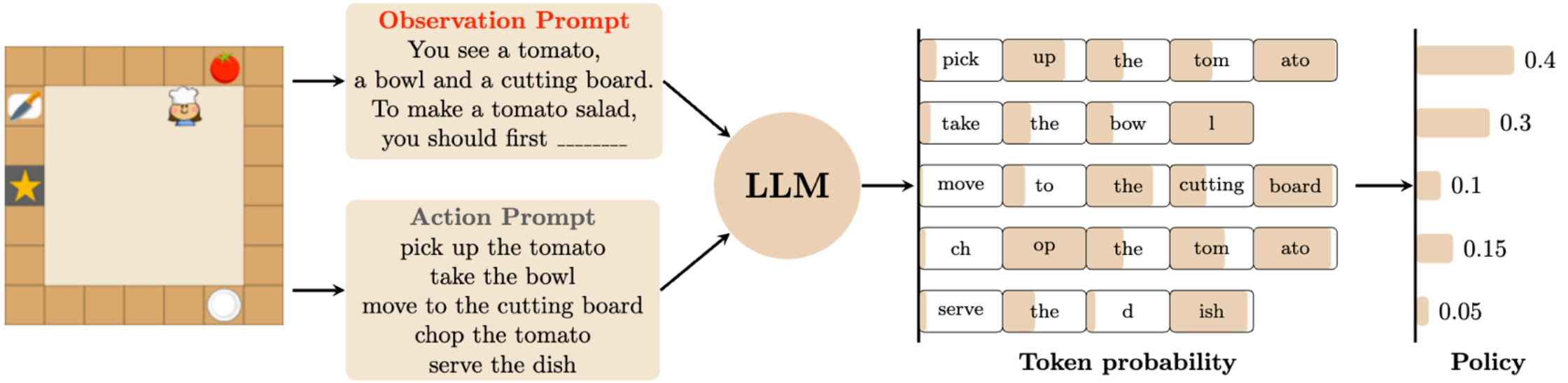
In the next few years, **AI agents** will utterly change how we live our lives, online and off.



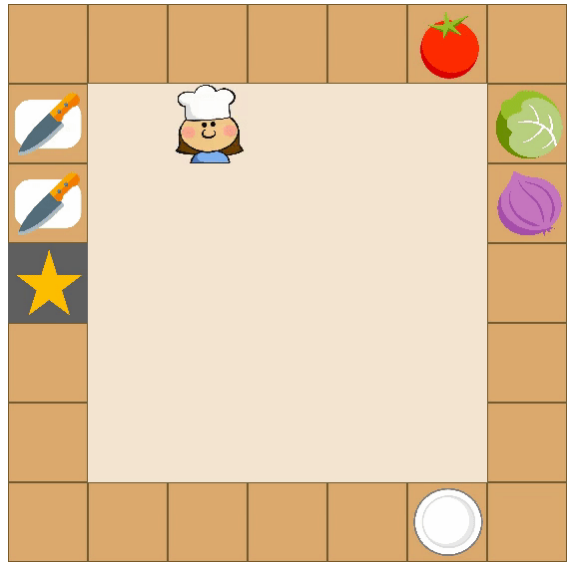
- Exciting new area with explosion of concepts, techniques, and components

Image credit: OpenAI Five, DeepMind AlphaStar, AlphaGo, and ChatGPT

True Knowledge Comes from Practice: Aligning LLMs with Embodied Environments via Reinforcement Learning [ICLR'24]



- Challenges
  - LLMs have rich prior knowledge but often fail in simple decision-making tasks due to the **misalignments** with environments
  - RL agents are always aligned well with environments but **struggle with leveraging prior knowledge**
- Solution
  - Use LLMs to form RL policy and optimized with PPO by interacting with embodied environments



Q*: Multi-step Reasoning as Deliberative Planning

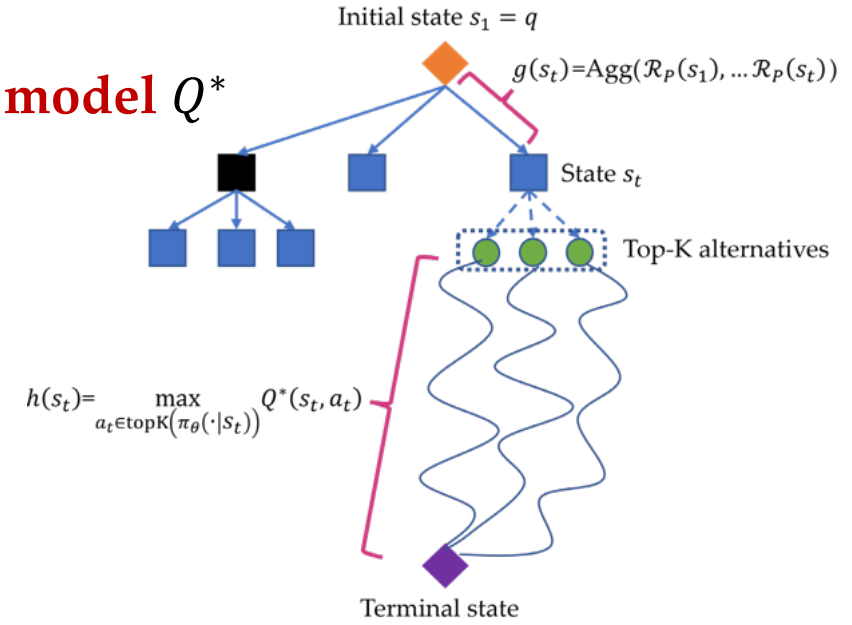
- Multi-step reasoning is challenging for LLMs
  - **SFT/RLHF**: computational burden, extensive human annotation cost
  - **Prompt engineering**: laborious expertise to craft, difficult to generalize
  - **Training verifiers**: no guidance for intermediate reasoning steps

- In-depth deliberation with A* search
  - Intermediate reasoning steps as states
  - Evaluate each state with process-based reward $\mathcal{R}_P$ & **Q-value model** Q^*

$$f(s_t) = g(s_t) + \lambda h(s_t)$$

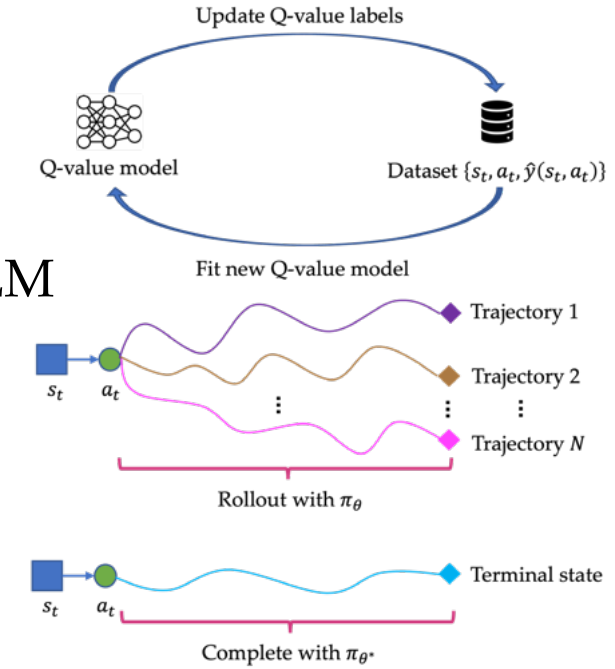
    - Aggregated utility $g(s_t) = \text{Agg}\big(\mathcal{R}_P(s_1), \dots, \mathcal{R}_P(s_t)\big)$
    - Heuristic value $h(s_t) = \max_{a_t \in \text{topK}\left(\pi_\theta(\cdot|s_t)\right)} Q^*(s_t, a_t)$

  - Best-first, informed search over reasoning steps
  - General, versatile & agile reasoning framework for LLMs

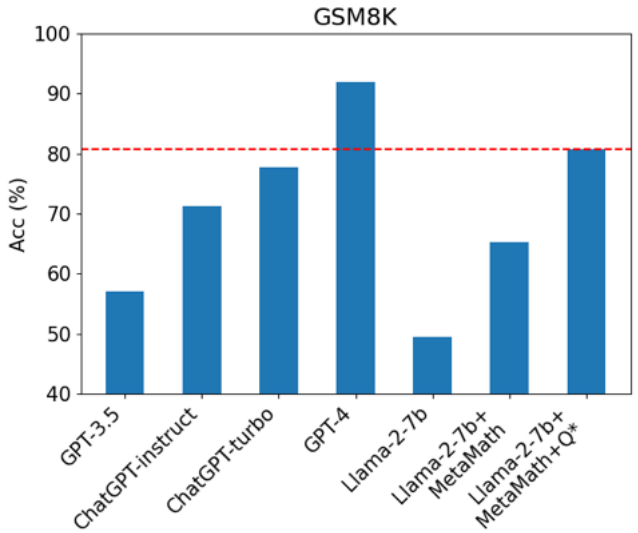


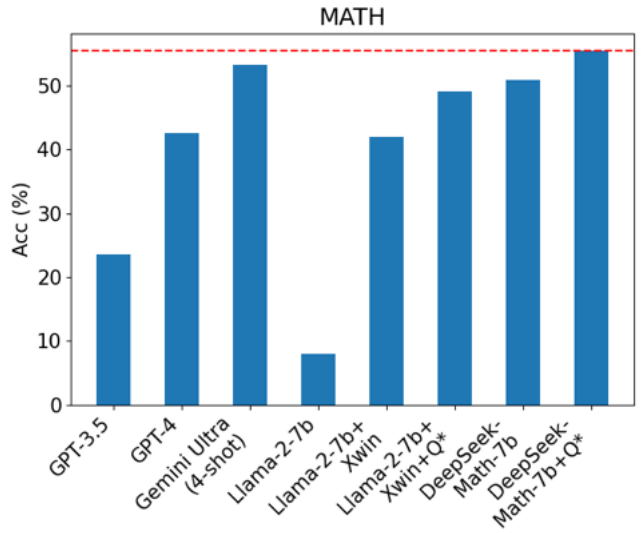
Q*: Multi-step Reasoning as Deliberative Planning

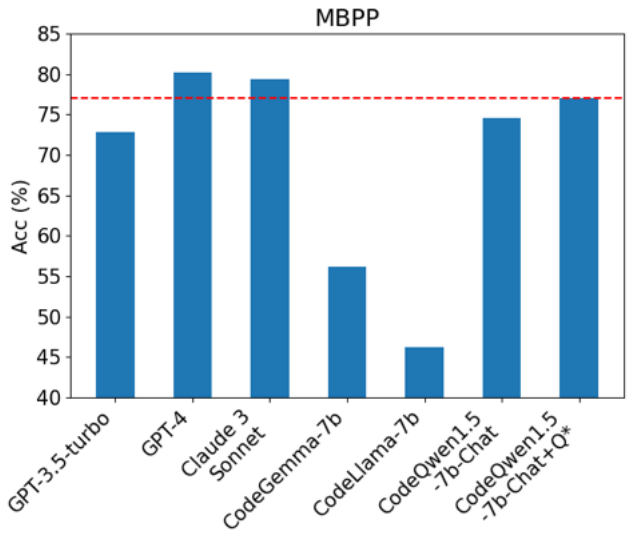
- Learning Q-value model (QVM)
  - **Offline RL**: alternatively update Q-value labels and fit QVM
  - **Rollout**: use the reward of best-performed rollout trajectory
  - **Completion**: use the reward of trajectory completed with stronger LLM



- Experimental results
  - GSM8K (on Llama2-7b): 65.2% → 80.8% (**+15.6%**)
  - MATH (on DeepSeek-Math-7b): 50.8% → 55.4% (**+4.6%**)
  - MBPP (on CodeQwen1.5-7b): 74.6% → 77% (**+2.4%**)

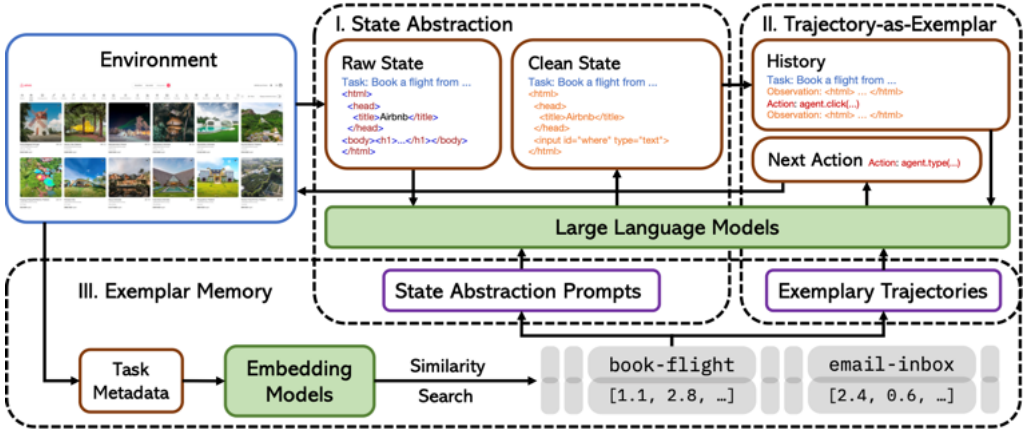


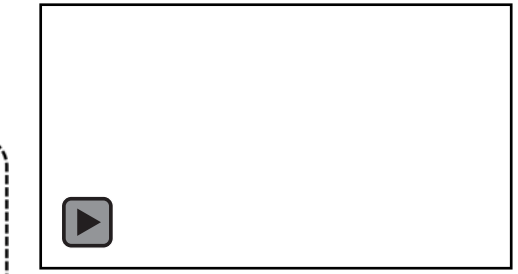


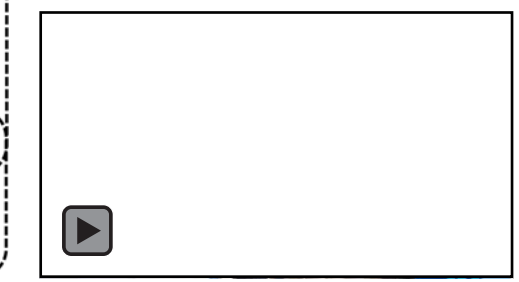


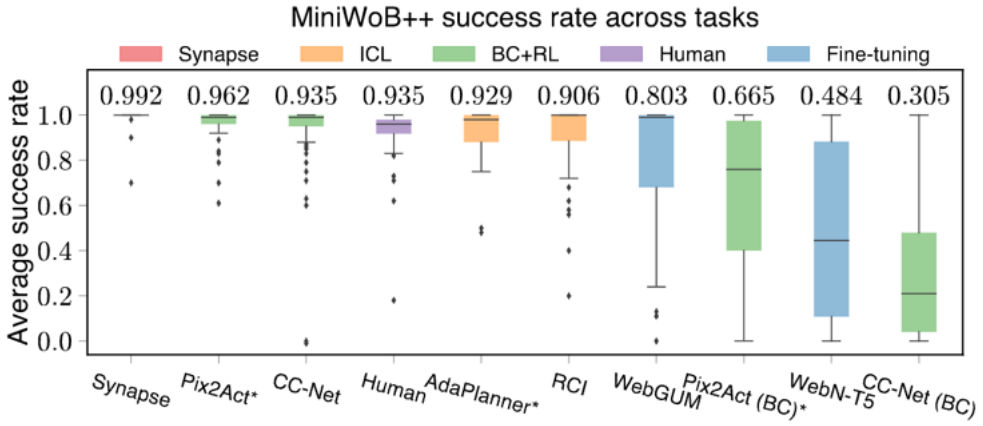
Synapse: Trajectory-as-Exemplar Prompting with Memory for Computer Control [ICLR'24]

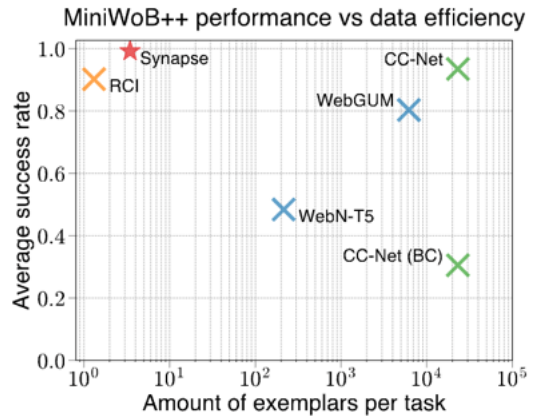
- In-context learning for computer control
- Challenges
  - Limited context length
  - Long-horizon error accumulation
  - Generalize to unseen tasks
- Solutions
  - State abstraction
  - Trajectory-as-exemplar
  - Exemplar memory
- SOTA in standard and real-world benchmarks







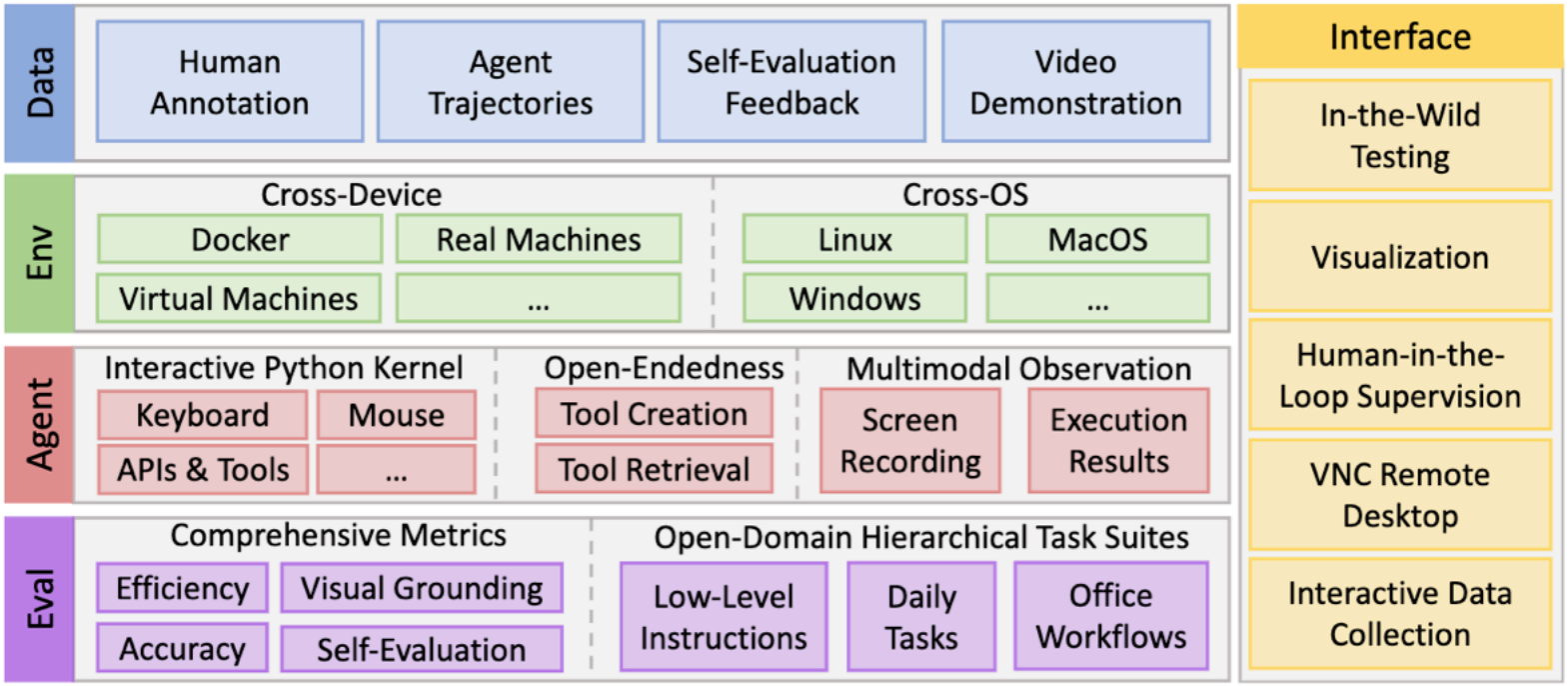




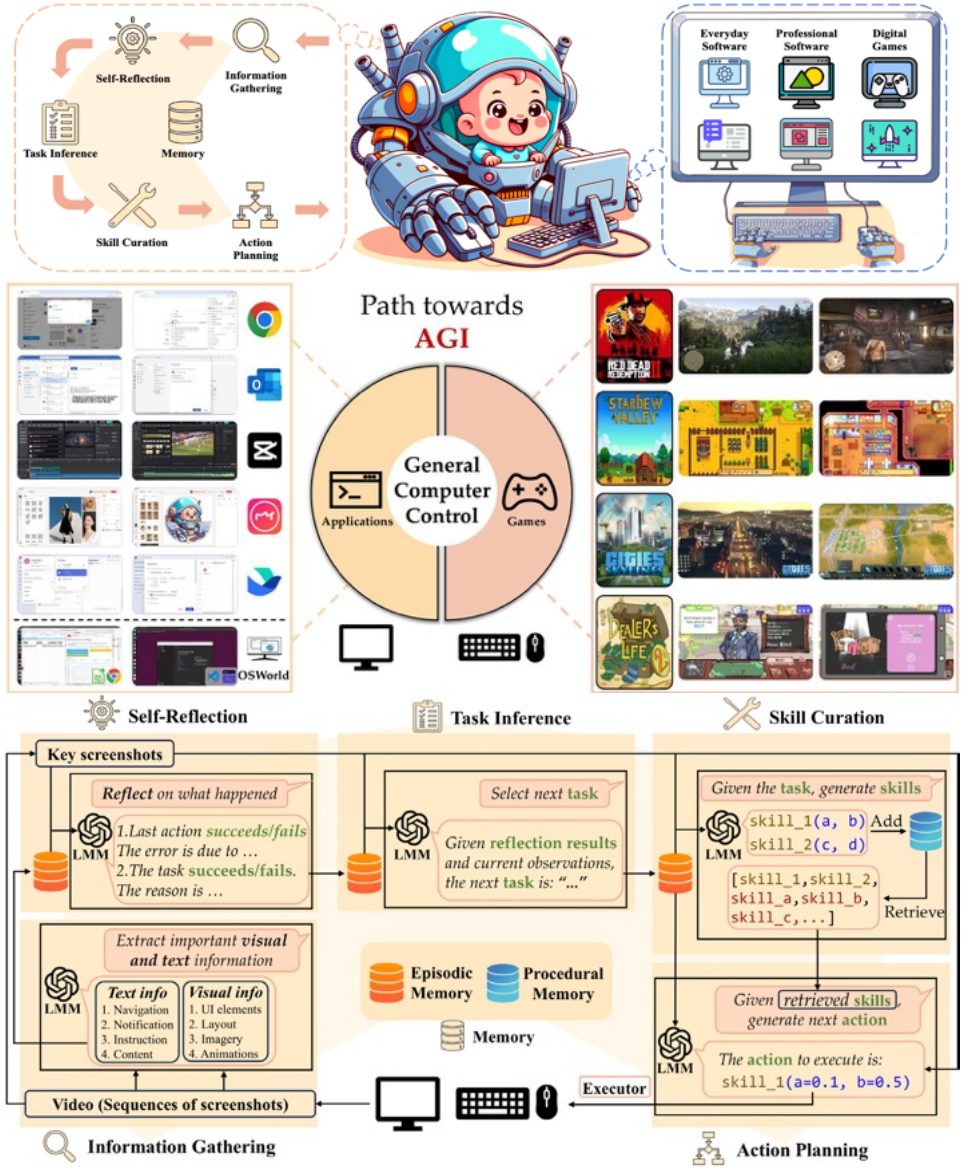
AgentStudio: A Toolkit for Building General Virtual Agents

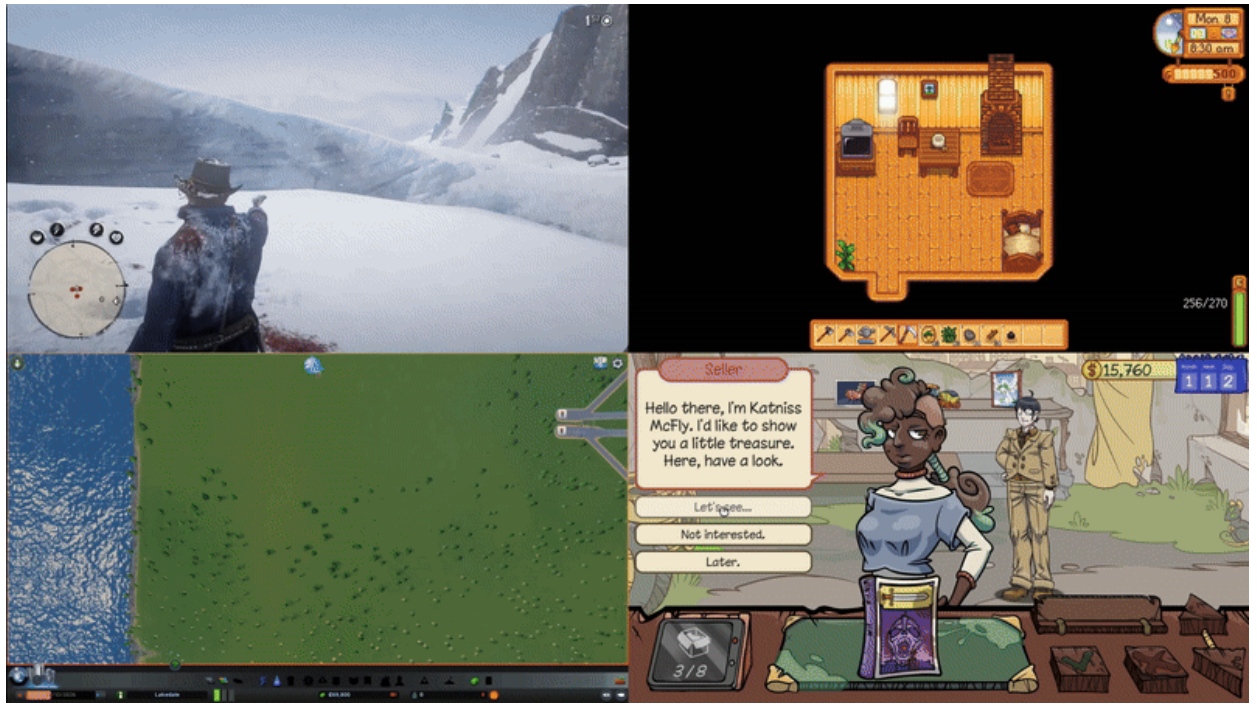
- Graphical Interface
  - Interactive Data Collection
  - Human-in-the-Loop Testing
  - Agent Visualization
- Agent Infra
  - General Obs/Act Spaces
  - Online & Diverse Envs
  - Native Tool Creation/Use
- Benchmark Suites
  - Open-Domain Tasks
  - Comprehensive Metrics
  - Expandable Eval Pipeline

Environment	Domain	Online	Action Space	Image	Video	Human Feedback	Tool Creation	Data Pipeline
World of Bits (Shi et al., 2017)	Simplified Web	✓	Keyboard & Mouse	✓	✓	✗	✗	✗
AndroidEnv (Toyama et al., 2021)	Android	✓	Touchscreen	✓	✓	✗	✗	✗
WebGPT (Nakano et al., 2021)	Web-assisted QA	✓	Web Operations	✗	✗	✗	✗	✗
WebShop (Yao et al., 2022)	Web Shopping	✓	Web Operations	✓	✗	✗	✗	✗
Mind2Web (Deng et al., 2023)	Real-World Web	✗	Web Operations	✓	✗	✗	✗	✗
AITW (Rawles et al., 2023)	Android	✗	Touchscreen	✓	✗	✗	✗	✗
GAIA (Mialon et al., 2023)	Tool-assisted QA	✗	APIs	✗	✗	✗	✗	✗
WebArena (Zhou et al., 2023)	Self-hosted Web	✓	Web Operations	✗	✗	✗	✗	✗
VisualWebArena (Koh et al., 2024)	Self-hosted Web	✓	Web Operations	✓	✗	✗	✗	✗
AgentStudio (Ours)	Real-World Devices	✓	Keyboard, Mouse and Code	✓	✓	✓	✓	✓



Cradle: Empowering Foundation Agents Towards General Computer Control

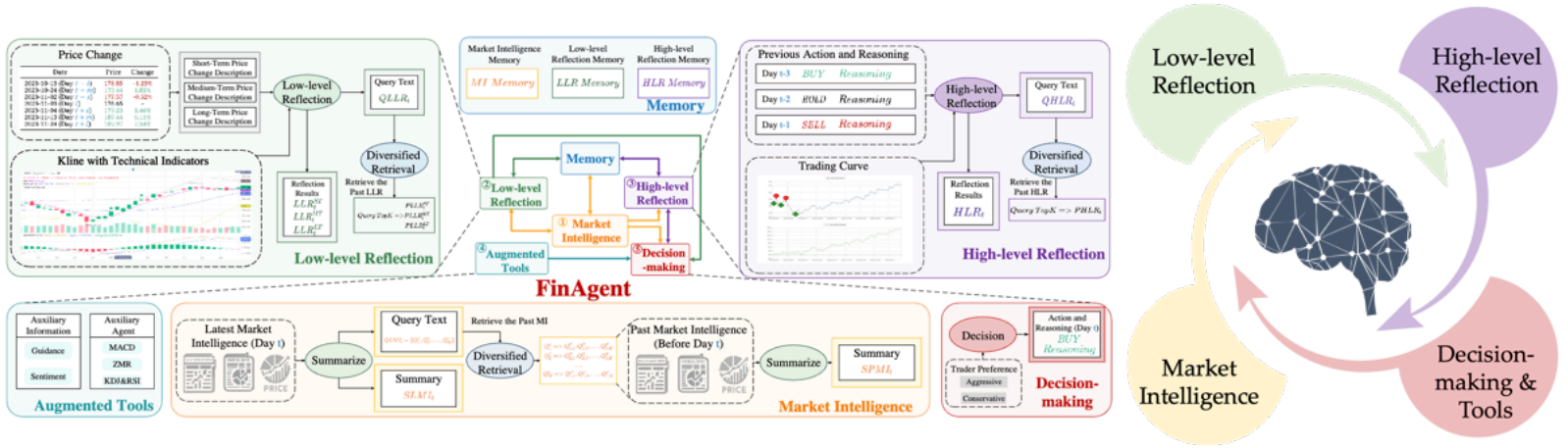




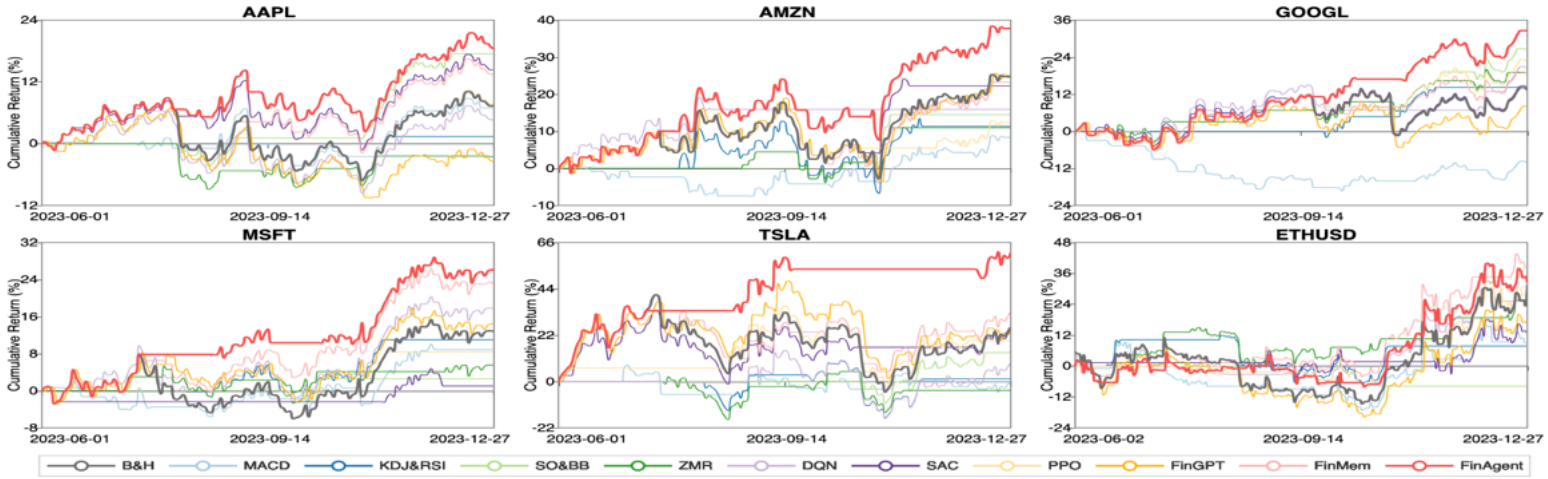
- Interact with **ANY** software via universal human-style interface
  - Receive input from screens and audio
  - Output keyboard and mouse actions
- **Six modules** enhancement
- Complete various **complex tasks** among **video games** (Red Dead Redemption 2, Stardew Valley, Dealer's Life 2, Cities: Skylines) and other **software** (Chrome, Outlook, CapCut, Meitu, Feishu)
- Greatly **extends** the **reach** of foundation agents

FinAgent: Foundation Multimodal Agent for Trading [KDD'24]

- **Multi-modal** foundational agent for financial trading
  - Process a **diverse range** of data to precisely understand market trend
  - **Five modules** to facilitate informed decision-making

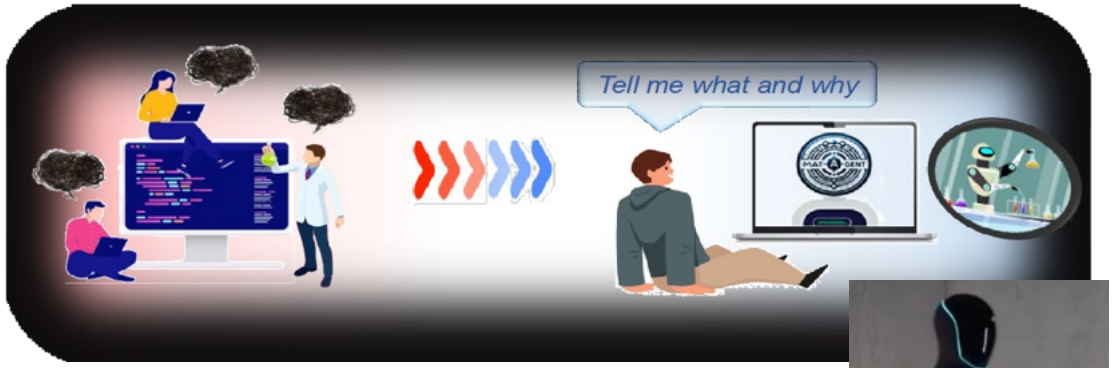


- Experimental results
  - Achieve the **highest** cumulative returns on 5 stocks
  - Offer at least **20%** improvement on TSLA dataset compared to the best-performing baseline

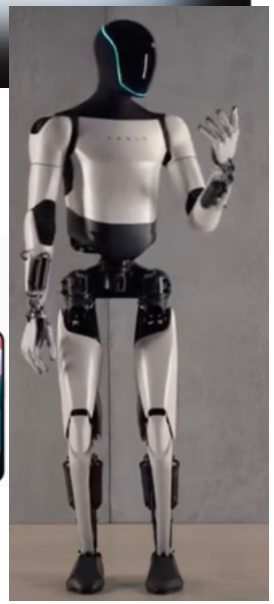


Future Directions

- LLM-powered agent for **science**
  - Drug and material discovery
  - Prediction, design, and experimentation
  - Autonomous, generalist, explainable



- LLM-powered agents in **digital and physical worlds**
  - Controlling all devices and robots
  - Apple Intelligence, Tesla Optimus



- **Safety and governance** of LLM-powered agents
  - Avoiding cartographic failures during deployment
  - Aligning agents with human preferences

- **Decision** foundation models
  - LMM, RL, ICL, …

Lessons Learned

- Algorithmic -> RL-based -> LLM-powered
  - In pursuit of scalability and generalizability
  - LLM (LMM) is preferred for generalizability

- No one is **panacea**

	Algorithmic	RL-based	LLM-power
(Appr.) Optimality	✓	-	✗
Scalability	✗	✓	-
Generalizability	✗	✗	✓

- Integrate the three paradigms to build **autonomous agents towards AGI**!

Thank You

Thanks to my mentors, collaborators, and students!

Email: boan@ntu.edu.sg