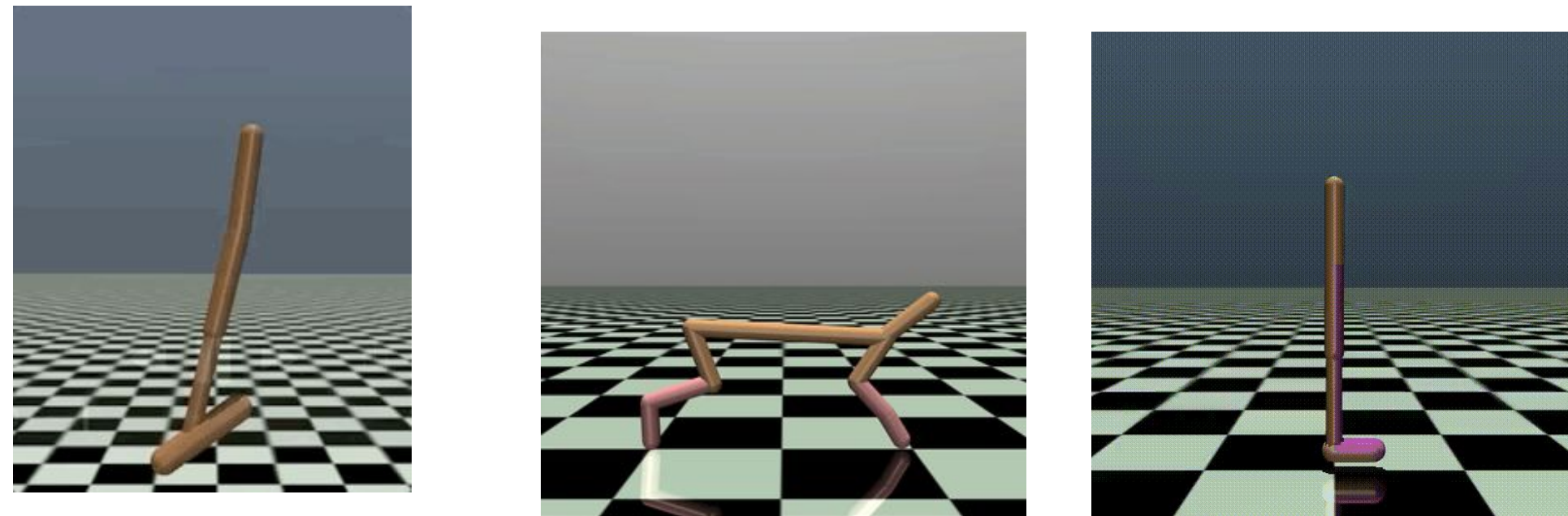# Prompts and Pre-Trained Language Models for Offline Reinforcement Learning
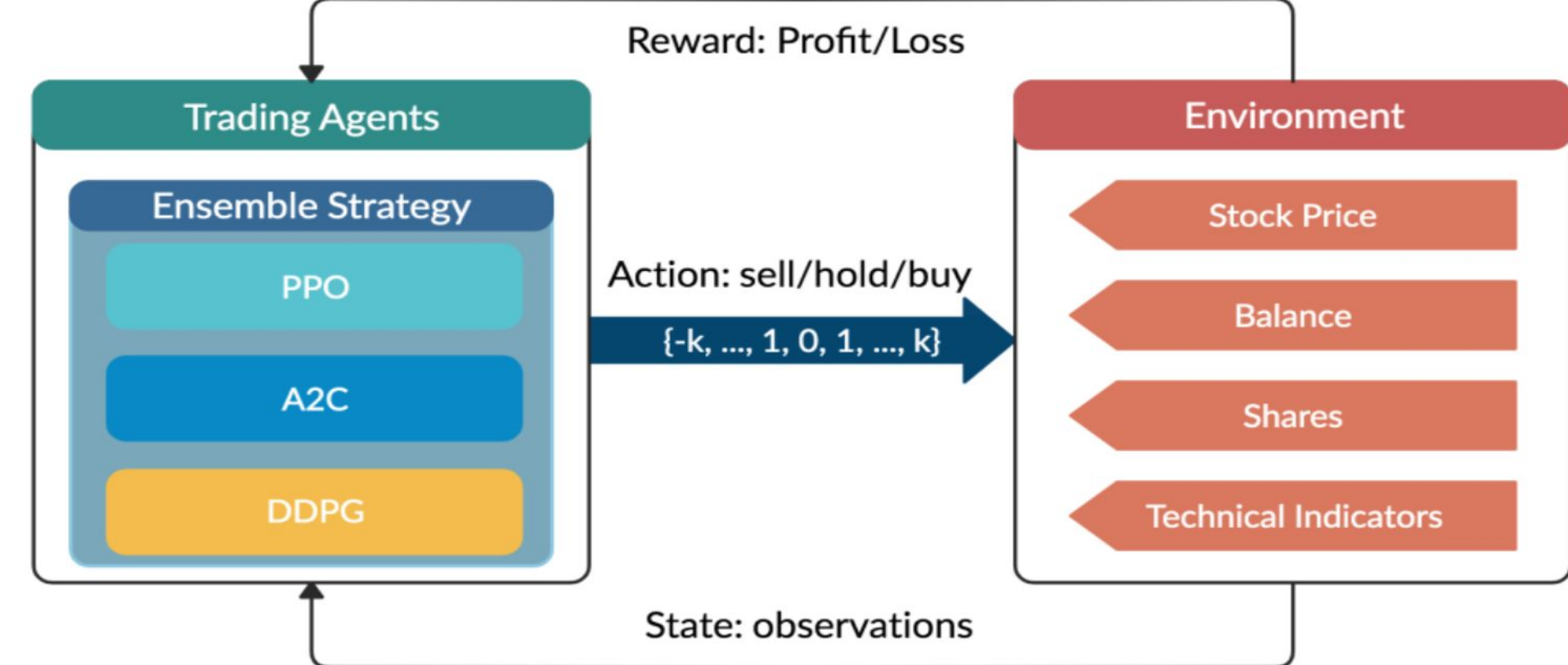
Denis Tarasov, Vladislav Kurenkov, Sergey Kolesnikov @ Tinkoff

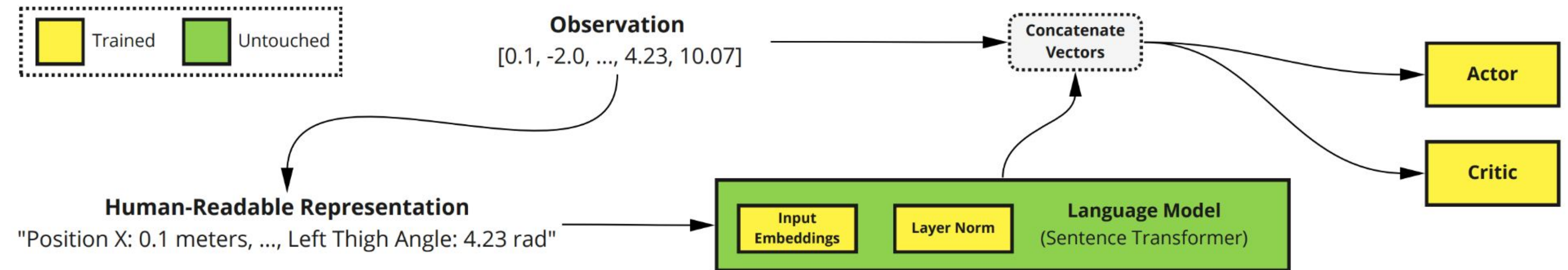## Can We Use Language Models for Text-free Environments?

MuJoCo

FinRL

Reward: Profit/Loss

Trading Agents
Ensemble Strategy
PPO
A2C
DDPG

Action: sell/hold/buy
{-k, ..., 1, 0, 1, ..., k}

Environment
Stock Price
Balance
Shares
Technical Indicators

State: observations

Yes! Prompt Tuning for Offline RL

Trained   Untouched

Observation
[0.1, -2.0, ..., 4.23, 10.07] → Concatenate Vectors → Actor / Critic

Human-Readable Representation
"Position X: 0.1 meters, ..., Left Thigh Angle: 4.23 rad" → Input Embeddings | Layer Norm | Language Model (Sentence Transformer) → Concatenate Vectors

Inspired by the recent success in informing deep RL algorithms with pre-trained language models, we propose an alternative approach to leverage them in environments that are not obviously suited for the textual representation, similar to Reid et al. (2022); Li et al. (2022). Here, we suggest a method based on prompt-tuning (Liu et al., 2021a) depicted in figure above: (1) re-write an observation into a human-readable textual representation (2) encode it using a pre-trained language model, and (3) concatenate the resulting embedding with the original observation.

Benchmark Results

FinRL

Task	TD3 + BC	TD3 + BC + LM-No-Train
FinRL-L-99	742	1043
FinRL-M-99	544	627
FinRL-H-99	808	859
FinRL-L-999	711	804
FinRL-M-999	1024	854
FinRL-H-999	879	883

MuJoCo

Task	TD3 + BC	TD3 + BC + LM	TD3 + BC LM-No-Train
Hopper-L-99	660	762	654
Walker2d-L-99	2480	2669	2564
HalfCheetah-L-99	4171	4084	4171

Examples of Prompts

Hopper

-0.00287, 1.24741, -0.00328,
-0.00231, -0.00302, -0.00395,
-0.00037, 0.00124, -0.00342,
0.00295, 0.00157, -0.00121

→

x position: -0.00287 m, z position: 1.24741 m, y angle: -0.00328 rad, thigh angle: -0.00231 rad, leg angle: -0.00302 rad, foot angle: -0.00395 rad, x velocity: -0.00037 m/s, z velocity: 0.00124 m/s, y angular velocity: -0.00342 rad/s, thigh angular velocity: 0.00295 rad/s, leg angular velocity: 0.00157 rad/s, foot angular velocity: -0.00121 rad/s

FinRL

1000000.0, 2.79591, 0.0, 0.00373, 100.0,
66.66666, 100.0, 15.80062, 0.0, 0.0143, 100.0,
66.66666, 100.0,
...
42.72148, 0.0, 0.01877, 100.0, 66.66666, 100.0,
53.48384, 0.0, 0.0266, 100.0, 66.66666, 100.0

→

Current balance: 1000000.0, AAPL: 2.79591 0.0 0.00373 100.0 66.66666 100.0, MSFT: 15.80062 0.0 0.0143 100.0 66.66666 100.0,
...
WBA: 42.72148 0.0 0.01877 100.0 66.66666 100.0, DD: 53.48384 0.0 0.0266 100.0 66.66666 100.0