Foundation Models for Reinforcement Learning

Chengyang Ying

31 March 2023

Foundation Models

· A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks, like BERT, GPT-3, CLIP.

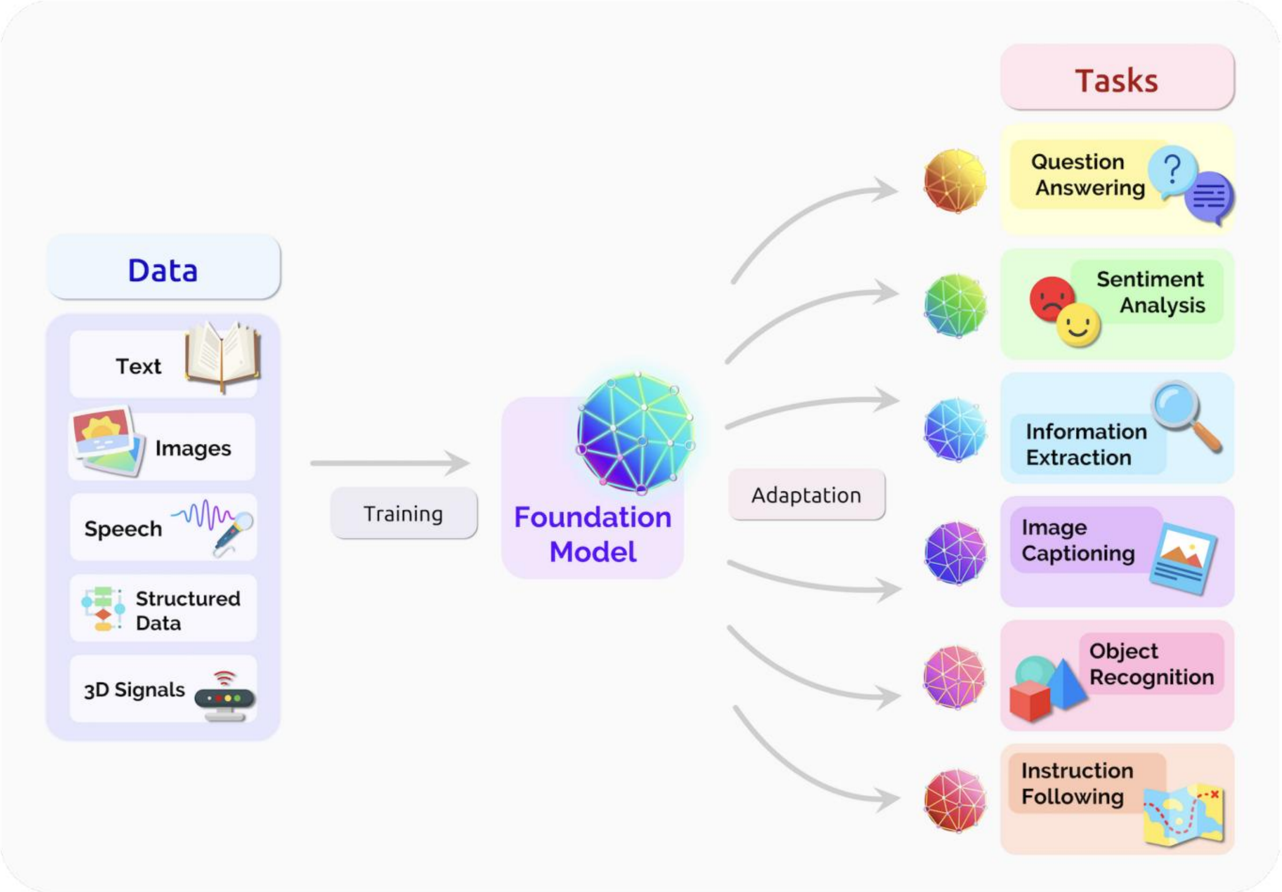
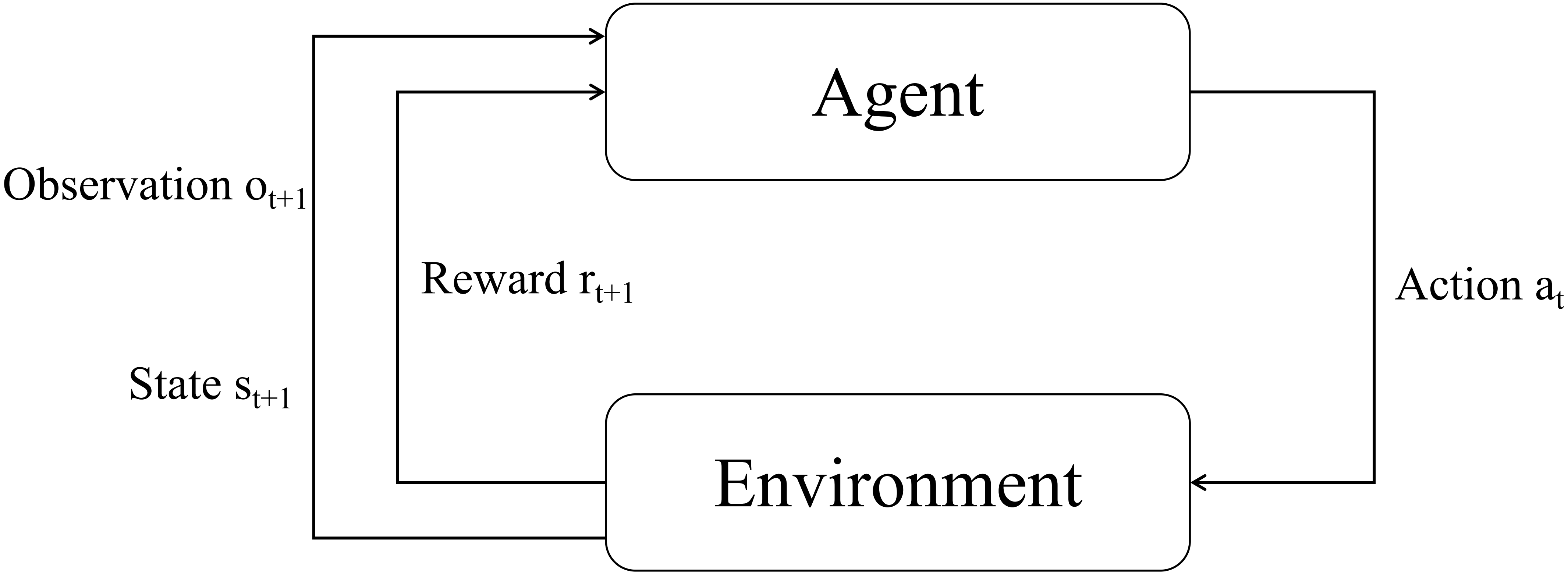


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

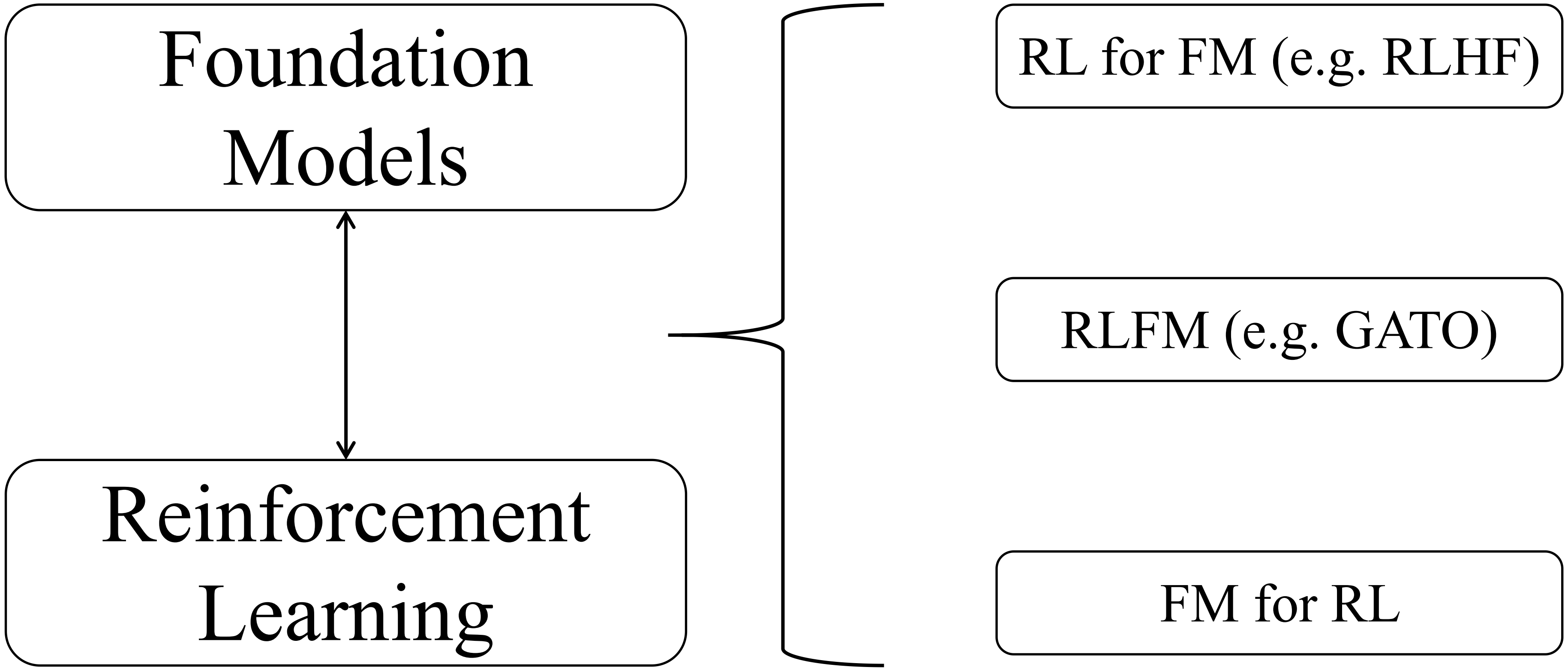
Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models[J]. arXiv preprint arXiv:2108.07258, 2021.

Reinforcement Learning



In fully-observed MDP, o=s

Foundation Models and RL



Foundation Models for RL

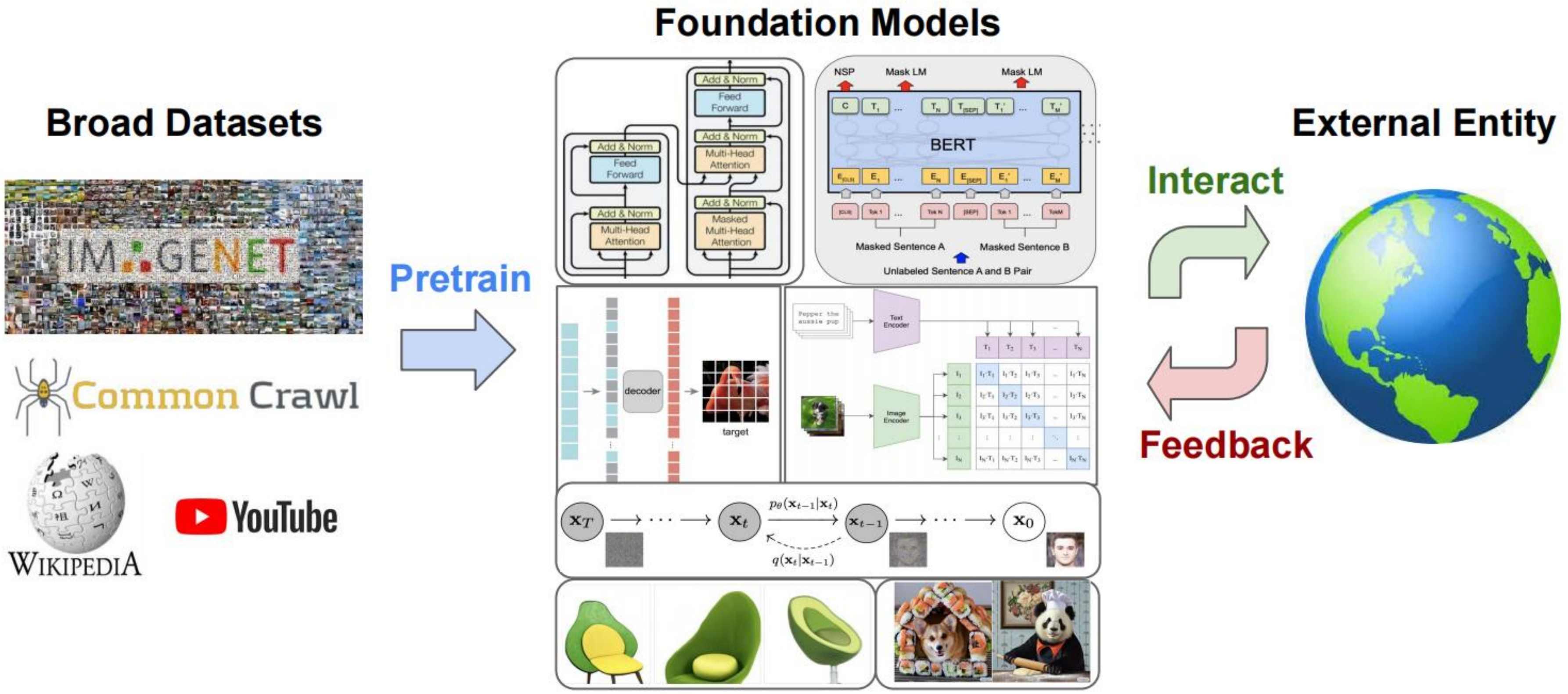


Fig. 1. Overview of foundation models for decision making. Foundation models pretrained on broad data are adapted to accomplish specific tasks by interacting with external entities and receiving feedback.

Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

Foundation Models for RL

Two major directions:

- Foundation models serve as generative models
  - like behavior, models
- Foundation models serve as representation learners
  - plug-and-play (use pretrained models to capture features of images or task-discription)
  - learning sequential representations

Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

Foundation Models for RL: as Conditional Generative Models

A key advantage to generative modeling of behaviors lies in scaling up

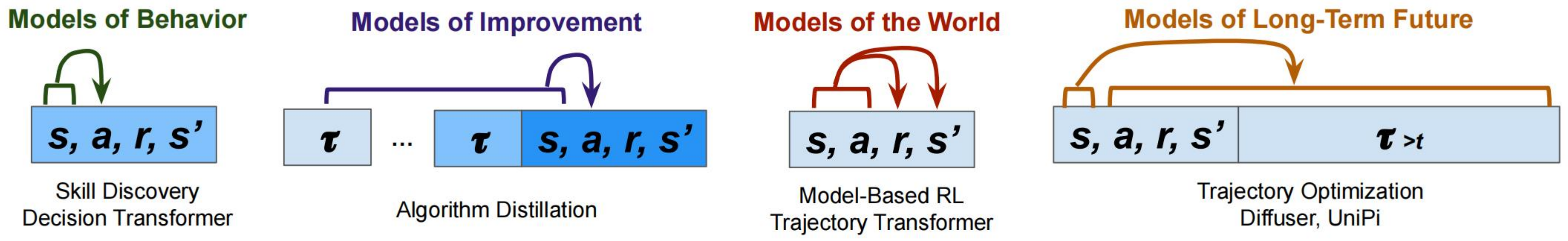


Fig. 3. Illustrations of how conditional generative models can model behaviors, improvements, environments, and long-term futures given a trajectory $\tau \sim \mathcal{D}_{\text{RL}}$. Dark blue indicates transitions with higher rewards. Models of behavior (Decision Transformers [Lee et al. 2022]) and self-improvement (Algorithm Distillation [Laskin et al. 2022]) require near-expert data. Models of the world (Trajectory Transformer [Janner et al. 2021]) and long-term future (UniPi [Du et al. 2023b]) generally require data with good coverage.

Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

Foundation Models for RL: as Representation Learners

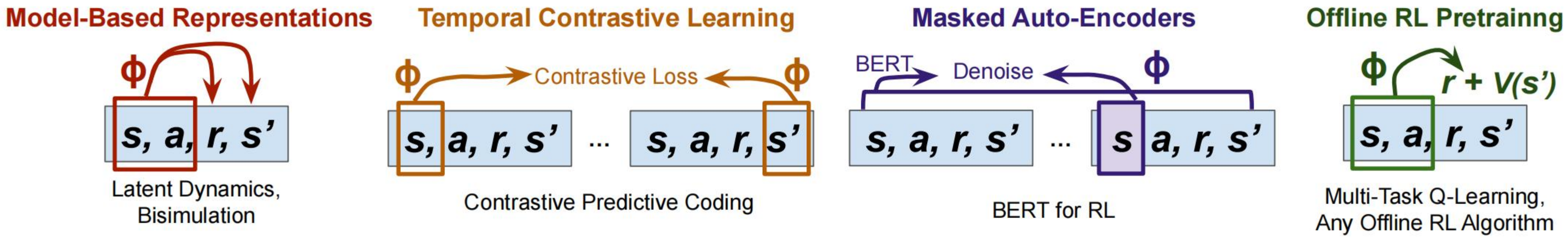


Fig. 4. Illustrations of different representation learning objectives such as model-based representations [Nachum and Yang 2021], temporal contrastive learning [Oord et al. 2018], masked autoencoders [Devlin et al. 2018], and offline RL [Kumar et al. 2022], on a trajectory $\tau \sim \mathcal{D}_{\mathrm{RL}}$ specifically devised for sequential decision making.

Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

Foundation Models for RL

main Challenges:

· How to bridge languange/vision dataset and decision making datasets

· How to structure Environments and Tasks

· How to improve large foundation models / decision making

Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

Reward Design with Language Models (ICLR23)

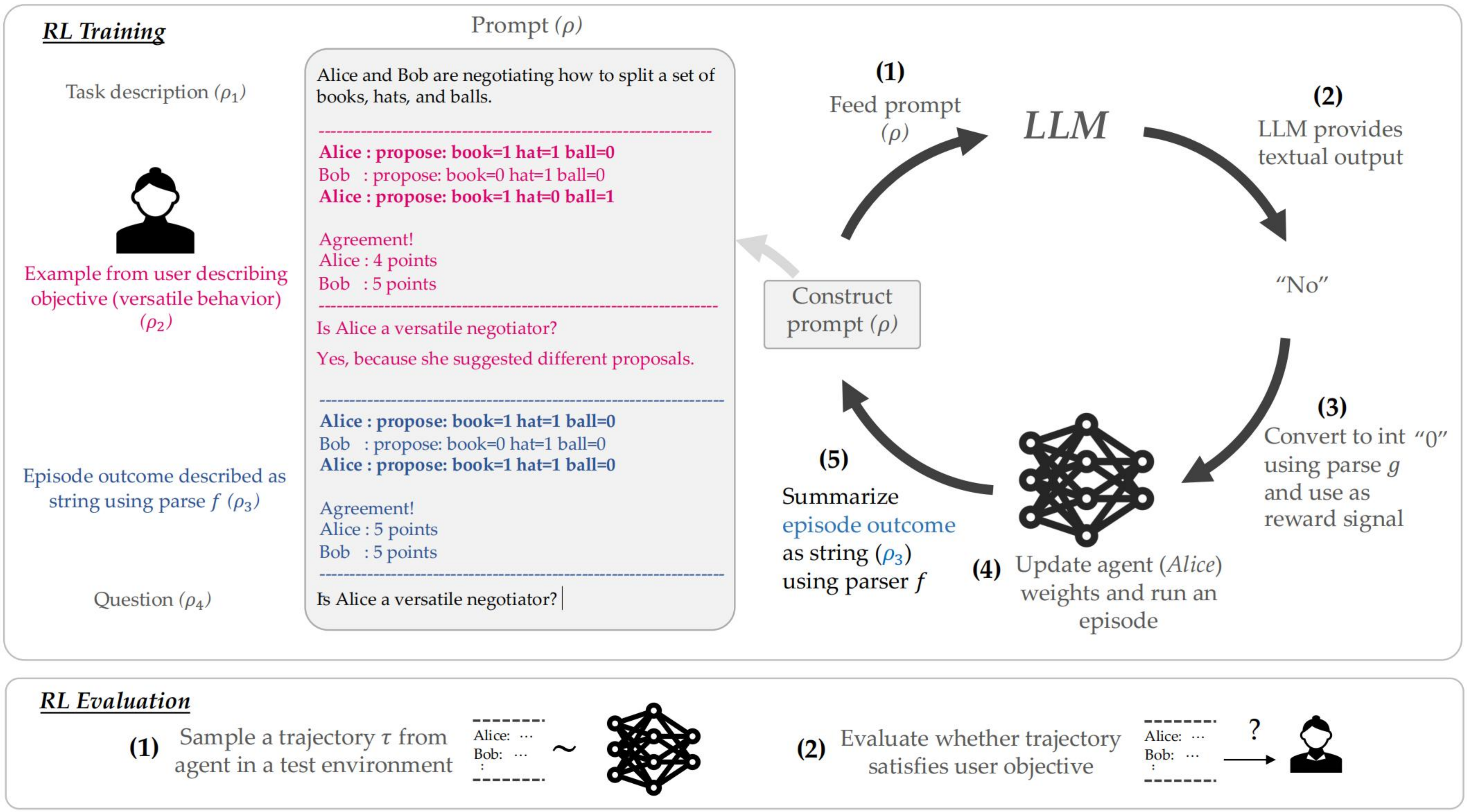
· Insights: LLMs are great in-context learners and can capture meaningful commonsense priors about human behavior.

· Goal: use LLMs (like GPT-3) as a proxy reward function by providing a textual prompt containing a few examples (few-shot) or a description (zero-shot) of the desired behavior

· *Generate the input of LLM from states:* the task description, the user-specified examples/description, a question asking if the outcome satisfies the objective

· *Generate rewards from the output of LLM:* use a handcrafted, task-specific parser e.g. "No"→0, "Yes"→1

Kwon M, Xie S M, Bullard K, et al. Reward Design with Language Models[C]//The Eleventh International Conference on Learning Representations.



Kwon M, Xie S M, Bullard K, et al. Reward Design with Language Models[C]//The Eleventh International Conference on Learning Representations.

Reward Design with Language Models

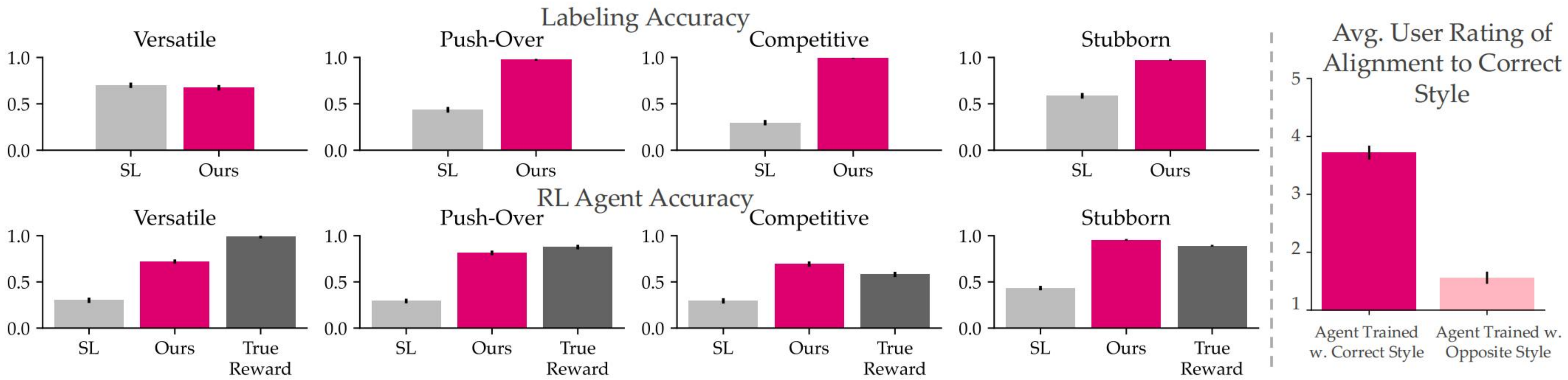


Figure 4: **DEALORNODEAL, Few-shot**. (Top) Accuracy of reward signals provided by LLM and SL during RL training. (Bottom) Accuracy of RL agents after training. (Right) Pilot study results. Agents trained with the user's preferred style were rated as significantly more aligned than an agent trained with the opposite style $p < 0.001$.

Kwon M, Xie S M, Bullard K, et al. Reward Design with Language Models[C]//The Eleventh International Conference on Learning Representations.

Reward Design with Language Models

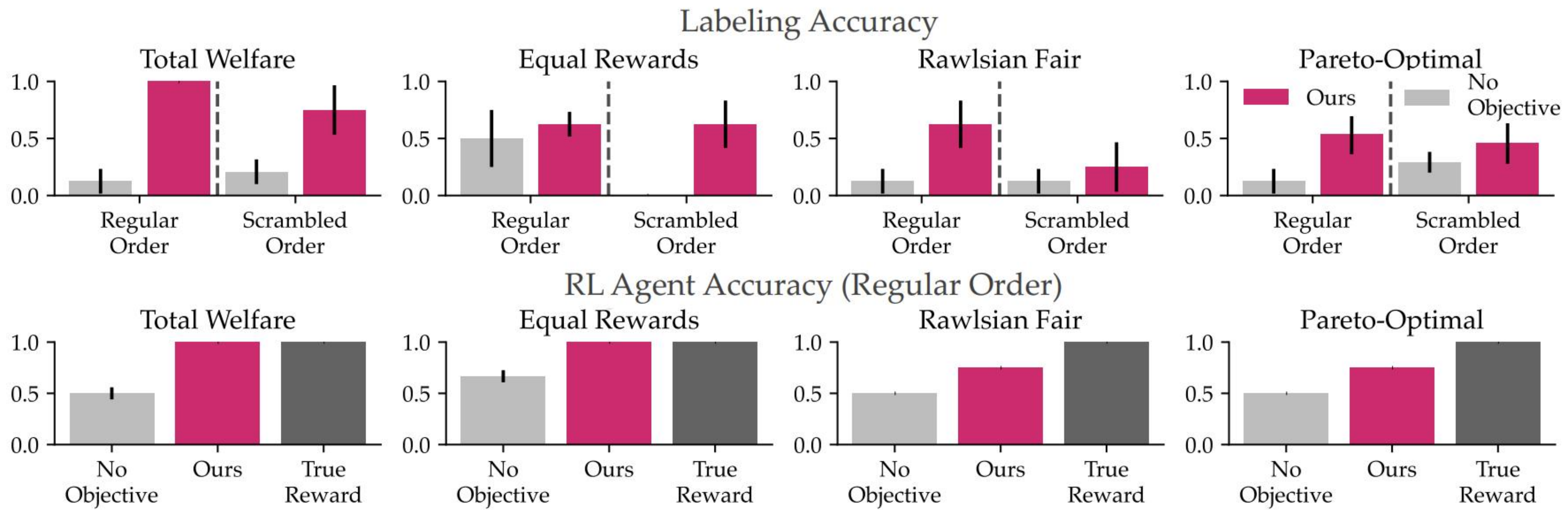


Figure 3: **Matrix Games, Zero-shot**. (Top) Accuracy of reward signals provided by LLM and a *No Objective* baseline during RL training. We report results for both regular and scrambled versions of matrix games. (Bottom) Accuracy of RL agents after training.

- Evaluate in zero shot via prompting well-known concepts such as Pareto-optimality

Kwon M, Xie S M, Bullard K, et al. Reward Design with Language Models[C]//The Eleventh International Conference on Learning Representations.

Can wikipedia help offline reinforcement learning?

Figure 1. Adapting pre-trained language models (e.g. from Wikipedia) to offline RL (e.g. in continuous control and games).

Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

Can wikipedia help offline reinforcement learning?

· Input sequence follows DT: $\mathbf{t} = (\hat{R}_1, s_1, a_1, \hat{R}_2, s_2, a_2, \ldots, \hat{R}_N, s_N, a_N)$

use linear projections to get input representations

· align state/action/reward and language representations:

minimize the negative the sum of the maximum similarity value for each embedding E and each input representation I via cosine similarity function as

$$\mathcal{L}_{\cos} = -\sum_{i=0}^{3N} \max_j \mathcal{C}(I_i, E_j)$$

· use K-means clustering over the embeddings to reduce the size of V

Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

Experimental Results

LM: trained in100 million tokens from full Wikipedia articles.

LM

L-VM: discard the image encoder

VM

Baselines

Dataset	Environment	ChibiT	GPT2	CLIP	iGPT	DT	CQL	TD3+BC	BRAC-v	AWR	BC
Medium Expert	HalfCheetah	**91.7 ± 1.1**	**91.8 ± 0.5**	91.3 ± 0.4	1.9 ± 0.1	86.8	62.4	90.7	41.9	52.7	59.9
	Hopper	**110.0 ± 1.2**	**110.9 ± 1.6**	110.2 ± 0.1	6.9 ± 3.7	107.6	**111.0**	98.0	0.8	27.1	79.6
	Walker	108.4 ± 0.2	108.9 ± 0.3	108.5 ± 0.6	0.5 ± 0.7	108.1	98.7	**110.1**	81.6	53.8	36.6
Medium	HalfCheetah	43.3 ± 0.1	42.8 ± 0.1	42.3 ± 0.2	1.5 ± 0.1	42.6	44.4	**48.3**	46.3	37.4	43.1
	Hopper	**82.1 ± 4.6**	79.1 ± 1.1	66.9 ± 0.9	5.7 ± 1.5	67.6	58.0	59.3	31.1	35.9	63.9
	Walker	77.8 ± 0.1	78.3 ± 1.5	74.1 ± 0.9	0.4 ± 0.4	74.0	79.2	**83.7**	81.1	17.4	77.3
Medium Replay	HalfCheetah	39.7 ± 0.5	40.3 ± 2.3	37.9 ± 0.2	1.6 ± 0.1	36.6	46.2	44.6	**47.7**	40.3	4.3
	Hopper	81.3 ± 5.0	**94.4 ± 2.5**	85.8 ± 0.3	5.7 ± 0.9	82.7	48.6	60.9	0.6	28.4	27.6
	Walker	71.3 ± 2.0	72.7 ± 1.2	69.9 ± 0.3	9.1 ± 7.7	66.6	26.7	**81.8**	0.9	15.5	36.9
Average (All Settings)		**78.3**	**80.1**	**76.3**	3.7	74.7	63.9	75.3	36.9	34.3	46.4

Table 2. Results for D4RL datasets[4]. We report the mean and variance for three seeds. Language model pre-trained models are consistently better than the Decision Transformer, and outperform/are competitive other baselines.

Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

Ablation Study

Model	Walker2d	HalfCheetah	Hopper
DT (GitHub)	3h14m	3h23m	2h47m
ChibiT (ours)	43m	48m	36m
GPT2 (ours)	1h27m	1h32m	1h2m

Table 3. Training time comparison (measured in hours and minutes on a single V100 GPU on the medium-expert setting) between the Decision Transformer and two pre-trained models: ChibiT and GPT2 on OpenAI gym tasks. Note that GPT2 is 144x larger than the other models with 84M model parameters.

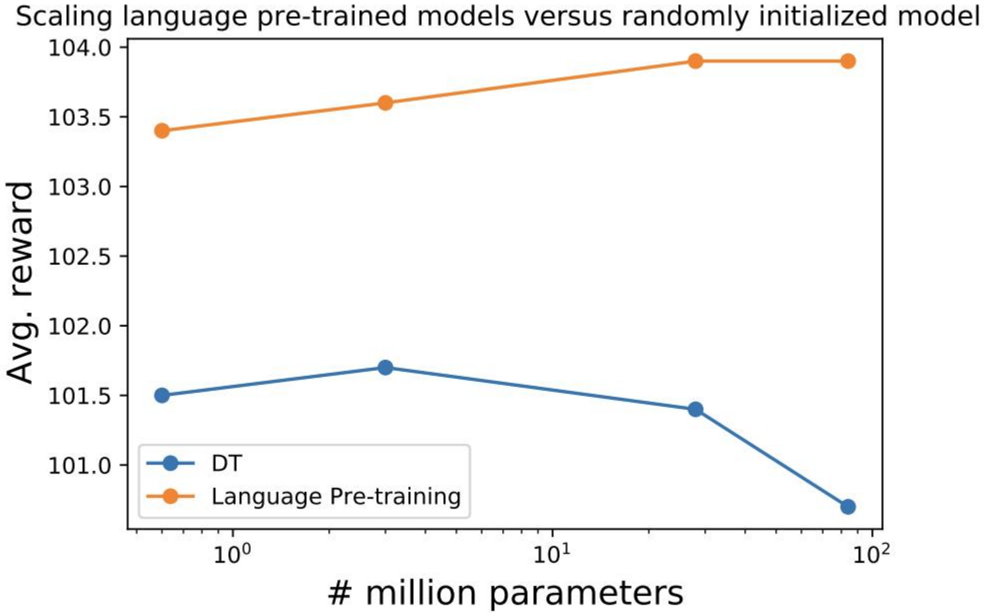


Figure 3. Comparison of Average *Medium-Expert* reward for various model sizes on OpenAI Gym.

Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

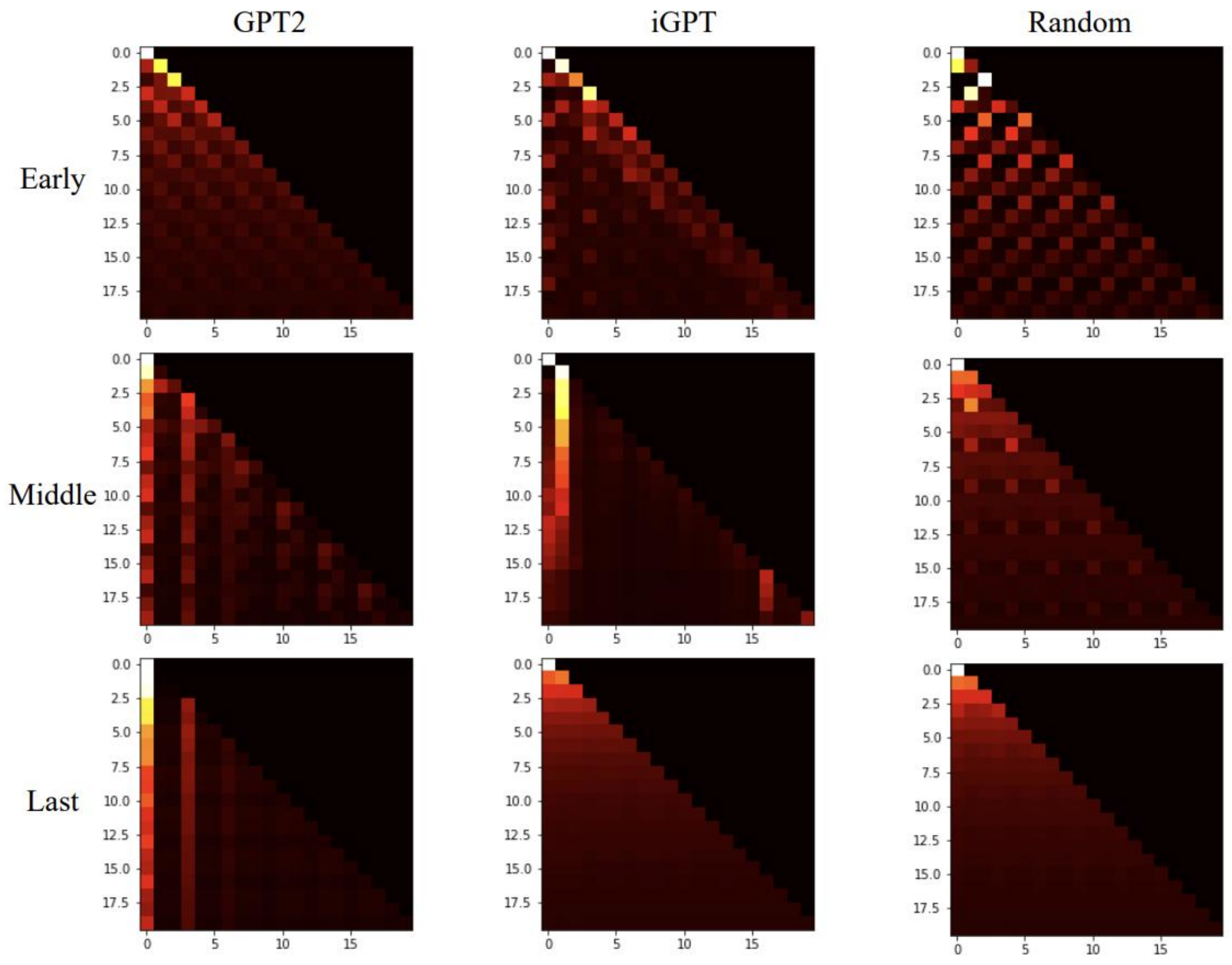


Figure 2. **Attention analysis.** We visualize early, middle and last attention weights computed by GPT-2, iGPT, and randomly initialized DT models on Hopper-medium to study how pre-training on different modalities affects how the model attends to previous timesteps. The x-axis represents keys (representations that are being "looked at") while the y-axis represents queries (i.e. representations that are "looking at" other representations) for a given timestep. Ligher colors represent higher attention weights, while darker colors represent lower weights.

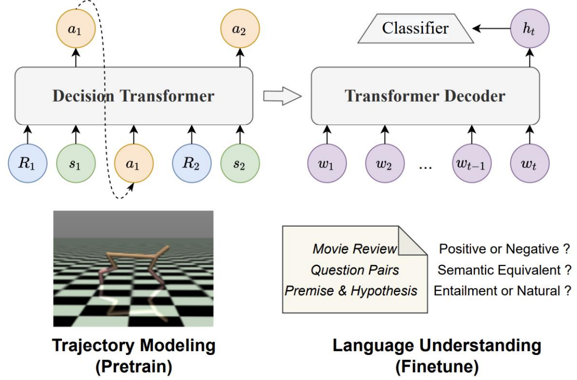
· GPT2/Random tend to attend to positions with multiples of 3 timesteps behind the current position

· GPT2 starts showing a stronger preference for previous returns-to-go

· Random/iGPT exhibit a behaviour closer to mean pooling. GPT2 is reliant on the initial returns-to-go

Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

Offline RL → NLP



random init

Model	SST2	MRPC	QQP	QNLI	WNLI	RTE	Avg.
Random-GPT	79.91±0.85	69.9±0.39	76.95±0.15	61.33±0.31	47.61±5.8	48.88±1.13	64.10
hopper-expert	79.06±0.26	69.46±0.78	78.41±0.1	61.35±0.18	53.8±1.38	51.12±2.99	65.53
hopper-medium-replay	78.99±0.78	69.22±0.43	78.44±0.12	61.15±0.28	49.3±2.95	51.26±0.6	64.73
hopper-random	79.7±1.14	69.07±0.42	77.35±0.05	60.98±0.27	51.55±3.52	48.45±0.53	64.51
walker2d-expert	77.78±0.61	69.85±0.76	77.74±0.14	60.52±0.27	52.11±5.12	49.68±2.53	64.61
walker2d-medium-replay	79.08±0.41	70.05±0.68	78.33±0.1	61.04±0.2	52.96±6.34	50.76±2.16	65.37
walker2d-random	73.51±0.95	68.68±0.18	76.93±0.07	59.34±0.57	51.83±3.01	52.27±4.22	63.76
halfcheetah-expert	77.98±0.58	69.61±0.51	77.5±0.09	61±0.39	49.86±1.69	52.27±2.15	64.70
halfcheetah-medium-replay	79.59±0.63	70.69±0.95	77.73±0.22	60.54±0.38	50.7±4.23	51.77±2.1	65.17
halfcheetah-random	74.04±0.95	68.92±0.33	76.92±0.08	59±0.47	50.99±4.66	52.42±1.77	63.72
antmaze-large-diverse	76.86±0.59	70±0.74	77.37±0.16	60.37±0.33	54.65±5.94	53.29±1.42	65.42
maze2d-large	77.5±0.6	67.91±0.37	77.67±0.18	60.42±0.33	**54.93±3.98**	52.78±2.77	65.20
LM-GPT	**82.43±0.64**	**71.81±0.78**	**78.62±0.05**	**63.22±0.38**	51.55±4.85	**56.97±0.42**	67.43

pretrained on wikitext-103 dataset

Table 2: Comparison for results on six NLU tasks. The second best data has been underlined

Zhang Z, Wang Y, Zhang Y, et al. Can Offline Reinforcement Learning Help Natural Language Understanding?[J]. arXiv preprint arXiv:2212.03864, 2022.

Prompts and Pre-Trained Language Models for Offline Reinforcement Learning

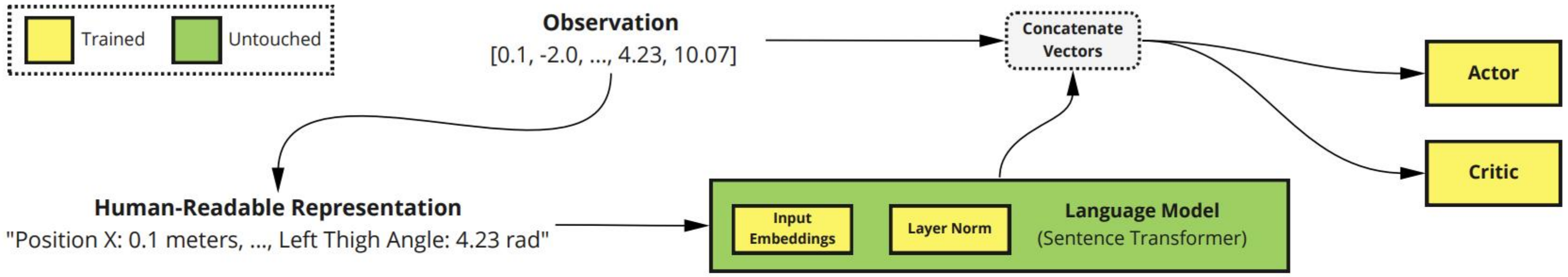


Figure 1: The proposed scheme for applying pre-trained language models in deep offline reinforcement learning. Note that the practitioner should only provide human-readable descriptions of the state dimensions. This scheme can be applied to most deep offline RL algorithms.

· mapping each dimension in observations into **human-readable labels** like "Position X: 10.0 meters"/ "Left Thigh Angle: 0.4 rad"

Tarasov D, Kurenkov V, Kolesnikov S. Prompts and Pre-Trained Language Models for Offline Reinforcement Learning[C]//ICLR 2022 Workshop on Generalizable Policy Learning in Physical World.

Prompts and Pre-Trained Language Models for Offline Reinforcement Learning

Table 1: Results for a subset of NeoRL dataset. The proposed approach outperforms the base model in two out of three settings. LM-No-Tune corresponds to a version of the method, where the language model is fixed throughout the training.

Task	BC	CQL	PLAS	BCQ	TD3+BC	TD3+BC LM-Finetune	TD3+BC LM-No-Tune
Hopper-v3-L-99	515	527	527	545	660	**762**	654
Walker2d-v3-L-99	1749	2370	42	1407	2480	**2669**	2564
HalfCheetah-v3-L-99	3260	3792	3451	3363	**4171**	4084	4171

only finetune
the input
embeddings
and layer norm
parameters

Tarasov D, Kurenkov V, Kolesnikov S. Prompts and Pre-Trained Language Models for Offline Reinforcement Learning[C]//ICLR 2022 Workshop on Generalizable Policy Learning in Physical World.

Prompts and Pre-Trained Language Models for Offline Reinforcement Learning

Table 2: Results for the FinRL subset of NeoRL benchmark. Tuning the language model may not be required to improve over the base model, where vector observations correspond to financial indicators.

Task	BC	CQL	PLAS	BCQ	TD3+BC	TD3+BC LM-No-Tune
FinRL-L-99	136	487	447	330	742	**1043**
FinRL-L-999	137	416	396	323	544	**627**
FinRL-M-99	355	700	388	376	808	**859**
FinRL-M-999	504	621	470	356	711	**804**
FinRL-H-99	252	671	464	426	**1024**	854
FinRL-H-999	270	444	495	330	879	**883**

The observation space here represent different financial indicators (e.g., price of a certain stock, an amount of a specific stock, etc.) and are more common than the proprioceptive inputs when training the LMs, and therefore a non-tuned LM should result in consistent improvements.

Tarasov D, Kurenkov V, Kolesnikov S. Prompts and Pre-Trained Language Models for Offline Reinforcement Learning[C]//ICLR 2022 Workshop on Generalizable Policy Learning in Physical World.

Reference

· Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models[J]. arXiv preprint arXiv:2108.07258, 2021.

· Yang S, Nachum O, Du Y, et al. Foundation Models for Decision Making: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2303.04129, 2023.

· Reid M, Yamada Y, Gu S S. Can wikipedia help offline reinforcement learning?[J]. arXiv preprint arXiv:2201.12122, 2022.

· Kwon M, Xie S M, Bullard K, et al. Reward Design with Language Models[C]//The Eleventh International Conference on Learning Representations.

Thanks for Listening

Questions?