УДК 004.4/.8 EDN: FIHEJC

ТРАНСФЕР ЗНАНИЙ ДЛЯ LLM-ОРИЕНТИРОВАННЫХ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В МУЛЬТИАГЕНТНЫХ СИСТЕМАХ

К.А. Морозов kmorozov@bmstu.ru
А.Н. Алфимцев alfim@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Способность больших языковых моделей справляться с интеллектуальными задачами является крайне важным навыком в различных средах, для которых требуется принимать решения на основе общедоступной информации. В обучении с подкреплением, особенно в мультиагентном, независимо от сложности среды, крайне важно на основе простых действий достигать значимых результатов, которые с точки зрения ретроспективы могли казаться невыполнимыми. Рассмотрена возможность использования большой языковой модели Mistral-7B Instruct-v0.3 для применения в задаче мультиагентного обучения с подкреплением. Разработан метод взаимодействия с Large Language Model (LLM) для использования рассуждения большой языковой модели для задачи планирования и распределения действий. Проведена оценка рефлексии большой языковой модели в результате действий, которые обозначены как необходимые для достижения поставленной в среде цели. Реализуемый трансфер знаний из LLM позволяет использовать успешные подходы для задач мультиагентного обучения с подкреплением в среде мира-сетки. Выполнено экспериментальное сравнение алгоритмов машинного обучения, которые могут эффективно взаимодействовать с предоставляемой им информацией, полученной в результате взаимодействия с большой языковой моделью. Предлагаемый метод позволяет встроить в обучение мультиагентной системы структуру рассуждения LLM

Ключевые слова

Мультиагентное обучение с подкреплением, большие языковые модели, машинное обучение, рассуждение

Поступила 11.06.2025
Принята 05.12.2025

Работа выполнена в рамках Государственного задания (№ FSFN-2024-0059)

Введение. Современные большие языковые модели (Large Language Models, LLM) демонстрируют возможности при решении различных задач, требующих анализа информации, логического вывода и принятия решений. Способности, позволяющие обобщать знания и выстраивать цепочки последовательных действий, открывают новые перспективы для областей, где критически важно эффективно использовать общедоступные данные. Вместе с этим изменяются и принципы разработки программного обеспечения. Большие языковые модели способны заменить многие существующие компьютерные приложения и настольные элементы программных систем. Для развития программного проектирования требуются инструменты дополнительного планирования, позволяющие использовать способности LLM так, чтобы была возможность создания новых конкурентно способных цифровых продуктов. Один из таких инструментов — трансфер знаний. Подобные инновации могут оказаться эффективными в задачах автоматизации, управления динамическими средами и координации различных частей распределенных систем. Особый интерес представляет интеграция LLM в обучение с подкреплением (Reinforcement Learning, RL) [1], где одной из ключевых задач является формирование выверенной точной стратегии, позволяющей достигать долгосрочных целей с использованием цепочки доступных действий. В мультиагентных средах эта задача усложняется необходимостью координации агентов, чьи решения должны быть согласованными и адаптивными. В свою очередь в соревновательных средах агенты должны совершать действия для того, чтобы оказаться успешнее оппонентов.

В настоящей работе исследована возможность применения большой языковой модели Mistral-7B Instruct-v0.3 [2] в рамках мультиагентного обучения с подкреплением (Multi-Agent Reinforcement Learning, MARL) [3–5]. Основной фокус направлен на использование способности LLM к планированию, рефлексии и распределению ролей между агентами. В отличие от традиционных подходов [6], где алгоритмы RL очень часто опираются на жестко заданные правила, предлагаемый метод предоставляет трансфер знаний из LLM для агентов в среде мира-сетки, упрощенной в целях первичного тестирования возможностей больших языковых моделей. Однако в подобной среде требуется четкая координация действия и выстраивание стратегии поведения.

Трансфер знаний (трансферное обучение, Transfer Learning, Knowledge Transfer) представляет собой процесс, при котором информация, полученная при выполнении одной задачи или набора данных, используется

для улучшения эффективности при реализации другой связанной задачи или набора данных [7].

Проведена оценка того, как LLM формирует действия агентов, анализирует последствия решений, адаптирует стратегию на основе предоставляемой ей изначальной информации. Особое внимание уделено рефлексии: действительно ли успешно LLM смогла справится с поставленной задачей или только проимитировала решение, когда можно принимать сценарий действий, сформированный LLM, как успешное выполнение заданной задачи? Рассмотрена и протестирована совместимость рассуждения LLM с классическими алгоритмами RL, в частности IQL (Independent Q-Learning) [8], CDQN (Central Deep Q-Network), представляющего собой модификацию DQN (Deep Q-Network) [9], MADDPG (Multi-Agent Deep Deterministic Policy Gradient) [10].

В результате предлагаемый метод позволяет создать гибридную систему, сочетающую значительные общедоступные познания большой языковой модели с эффективностью традиционных подходов в мультиагентном обучении с подкреплением. Полученные в процессе исследований результаты тестирования и анализа могут стать основой для разработки более гибких, точных и эффективных систем, способных действовать не только в среде мира-сетки, но и в условиях частой неопределенности, характерной для реального мира с учетом особенностей больших языковых моделей.

Материалы и методы решения задач, принятые допущения. Использована большая языковая модель Mistral-7B Instruct-v0.3. Взаимодействие с ней выполнено через API (Application Programming Interface). Выбор модели обусловлен способностью решать задачи в среде мира-сетки и возможностью ее открытого использования (в отличие от моделей компании OpenAI). Так на основе полученных при тестировании результатов, аналогичных представленным в табл. 1 (демонстрируют способность справляться с указанными задачами) [1], выбрана Mistral-7B Instruct-v0.3.

Таблица 1

Результаты работы LLM для RL-сред

Большая языковая модель	Размер среды мира-сетки							
	3 × 3		4 × 4		5 × 5	6 × 6	7 × 7	8 × 8
	Цель	Ловушка	Цель	Ловушка	Цель			
GPT-2	+	–	–	–	–	–	–	–
GPT-3.5-turbo	+	+	+	+	+	+	+	+

Окончание табл. 1

Большая языковая модель	Размер среды мира-сетки							
	3 × 3		4 × 4		5 × 5	6 × 6	7 × 7	8 × 8
	Цель	Ловушка	Цель	Ловушка	Цель			
GPT-4	+	+	+	+	+	+	+	+
Mistral-Medium	+	+	+	–	+	+	+	+
Mistral-Large	+	+	+	+	+	+	+	+
LLaMa-2	+	–	+	–	+	+	+	+
Vicuna-7b	+	+	+	–	+	+	+	+

В качестве основных алгоритмов машинного обучения для задачи получения и использования информации в результате трансфера знаний применено независимое глубокое обучение (IQL), централизованное обучение с использованием сверточной нейронной сети (CDQN) (включая мультиагентный случай) и алгоритм мультиагентного глубокого детерминированного градиента стратегий (MADDPG). В качестве мультиагентной среды, в которой проведены тестирования, реализован лабиринт в виде мира-сетки. Реализация алгоритмов взаимодействия с LLM и среды выполнена на языке Python. Для обучения использован графический процессор Nvidia GeForce RTX 3060 Ti. В качестве основных допущений приняты следующие: LLM способна воспринимать информацию о среде мире-сетки если не графически, то в матричном виде, и на основе допустимых (заранее обозначенных) действий выстраивать стратегию поведения.

Результаты. В процессе исследования протестированы и получены результаты, подтверждающие способность LLM справляться с задачами в среде мире-сетки, однако не без недостатков. Тестирование проведено на 200 сгенерированных задачах в лабиринте мире-сетки [11], где от LLM требовалось сформировать путь для каждого агента так, чтобы один из них достиг поля с выходом, при этом не упираясь в стены (ограничивающие лабиринт) и не переходя на клетки с ловушками (которые означали поражение в эпизоде и автоматически блокировали возможность успешного достижения выхода). На число действий для каждого агента в рамках одного эпизода накладывалось ограничение размером 50. Поля с выходом для каждого агента свои и могут изменяться в зависимости от начального расположения агентов. Экземпляр среды мира-сетки показан на рис. 1. Для агента, обозначенного зеленым цветом, выходом из лабиринта (соответственно целью) является клетка желтого цвета, а для агента синего цвета — клетка красного цвета.

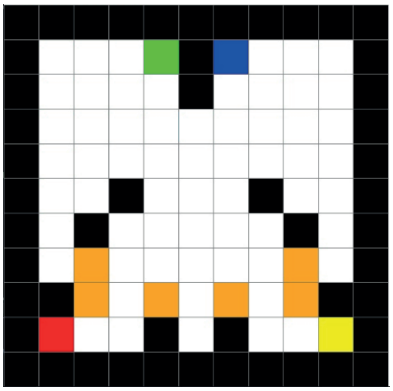

Рис. 1. Пример среды мира-сетки с доступной областью перемещения размером 9 × 9 (синим и зеленым цветом обозначены агенты, красным и желтым — цели, оранжевым — ловушки, черным — стены, ограничивающие переход)

Допустимыми действиями в представленной среде являются повороты направо и налево, движение вперед. Информация о наградах и штрафах (отрицательных вознаграждениях) в условных единицах для агентов представлена в табл. 2.

Таблица 2

Распределение наград за действия (шаги) агентов

Шаг	Награда, усл. ед.	
	Агент 1	Агент 2
В ловушку (конец эпизода)	– 100	– 100
В стену	– 3	– 3
На пройденный путь	– 1	– 1
На свободную клетку	– 0,5	– 0,5
На клетку с выходом (конец эпизода)	100 / 10	10 / 100

В результате предоставления LLM информации о лабиринте, необходимой цели, допустимых действиях (включая их максимальное число за эпизод) и награде за них формируется вывод в виде рассуждений большой языковой модели о том, как достичь выхода в этой среде, совершая каждое действие из цепочки рассуждений шаг за шагом [12, 13]. В результате сформирована информация (табл. 3), отображающая возможность за счет представленных LLM цепочек рассуждений достигнуть выхода, включая ее собственное убеждение о правильности действий, которое детектировалось за счет того, что LLM завершала формирование необходимых действий (шагов) до достижения лимита, а также путем добавления ключевой фразы *The agents successfully achieved their goals!*

Таблица 3

Информация о достижении полученной от LLM цели в мире-сетки

Размер мира-сетки	Вероятность достижения цели с агентами		
	двумя	одним	одним при наличии наград
3 × 4	0,33	0,32 (0,98)	0,00 (0,97)
4 × 4	0,26	0,26 (1,00)	0,10 (1,00)
5 × 5	0,02	0,00 (0,75)	0,00 (0,78)
6 × 6	0,04	0,00 (0,51)	0,00 (0,67)
9 × 9	0,00	0,00 (0,00)	0,00 (0,25)

Примечание. В скобках указаны значения, полученные LLM.

На основе информации можно сделать следующий вывод: LLM считает, что достигает успеха и поставленных целей в большинстве случаев (с одним агентом). Однако реальные значения показывают, что выводы LLM расходятся с действительностью [14], т. е. LLM выстраивает цепочки рассуждений, в которых на самом деле не может достигнуть поставленных целей, однако утверждает обратное. Кроме того, информация о распределении наград (как и о их наличии) влияет на внутреннюю уверенность в правильности действий, выстраиваемых большой языковой моделью [15]. В рассматриваемом случае это не улучшает реальные результаты, которые можно использовать на практике для трансфера знаний в мультиагентную систему, а ухудшает их. Данные, представленные в табл. 3, свидетельствуют о том, что LLM справляется достаточно успешно в среде мира-сетки размерами 3 × 4, 4 × 4, при этом формируя действия сразу для двух агентов в среде. Однако, ввиду того, что среда размером 3 × 4 не пригодна для размещения ловушек [16] и является начальной точкой тестирования возможности LLM справляться с подобными задачами, основным размером среды для дальнейшего тестирования является 4 × 4.

С учетом изложенного сформирована обобщенная структура метода взаимодействия с LLM для мультиагентного обучения с подкреплением (рис. 2).

Согласно представленной структуре, следующей значимой задачей обозначено определение алгоритмов, которые могут наиболее успешно справляться с необходимостью предоставления агентам информации, полученной в результате успешных цепочек рассуждения от LLM.

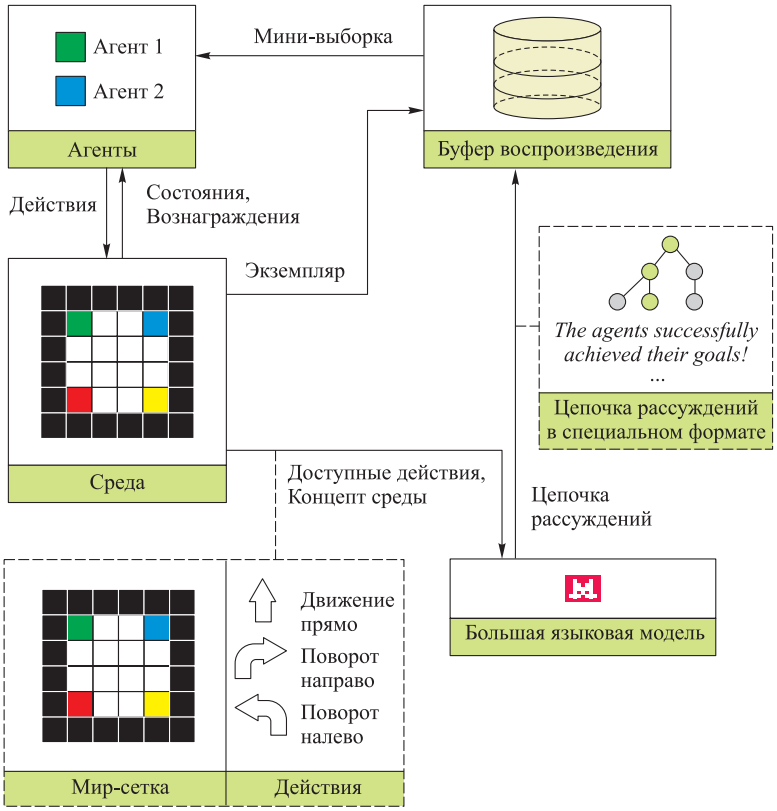

Рис. 2. Обобщенная структура метода трансфера знаний для LLM-ориентированных алгоритмов в мультиагентных системах (штриховые линии описывают содержимое данных, передаваемых к LLM и от LLM, сплошные линии — взаимодействие различных элементов структуры)

Пошаговая структура взаимодействия с алгоритмом CDQN представлена в алгоритме в качестве демонстрации общего метода трансфера знаний.

Алгоритм. CDQN с трансфером знаний

1. **Инициализировать:**

$\alpha \in (0,1]$; $\gamma \in [0,1]$; $\varepsilon \in [0.1,1]$; $Ne \in [1,N]$;

$Eb \leftarrow 0$; $Bl \leftarrow 10000$; $Bs \leftarrow 32$; $Cs \leftarrow 100$;

$i \leftarrow 1,\ldots,n$; $a_i \in A_i$; $j \leftarrow 1,\ldots,Bs$; $K_i \leftarrow \mid A_i \mid$;

$\theta \leftarrow \xi$; $\theta' \leftarrow \xi$, $\forall o \in \mathbb{O}$.

2. Первично заполнить буфер воспроизведения Eb данными, полученными от LLM.

3. **Цикл:**

4. Получить состояние o.

5. Передать наблюдение o в основную нейронную сеть и получить оценку $Q(o, a_1, \ldots, a_n; \theta)$.

6. Выделить в $Q(o, a_1, \ldots, a_n; \theta)$ оценки действий $Q_{1\ldots K_i}(o, a_1; \theta), \ldots, Q_{1\ldots K_i}(o, a_n; \theta)$ для каждого агента i.

7. Выбрать возможное действие a_i в соответствии с максимальным значением $Q_1(o, a_i; \theta), Q_2(o, a_i; \theta), \ldots, Q_{K_i}(o, a_i; \theta)$ и с учетом параметра ε.

8. Выполнить возможное действие a_i.

9. Получить вознаграждение r и следующее наблюдение o'.

10. Сохранить кортеж $(o, a_1, \ldots, a_n; r, o')$ в буфере воспроизведения Eb с учетом соблюдения объема буфера Bl.

11. Получить случайную мини-выборку $(o, a_1, \ldots, a_n; r, o')$ объемом Bs из буфера воспроизведения Eb.

12. Вычислить для каждого j-го кортежа $(o, a_1, \ldots, a_n; r, o')$ из мини-выборки значение $Q^j(o', a'_1, \ldots, a'_n; \theta)$, предварительно передав в целевую нейронную сеть o'.

13. Вычислить

$$y^j \leftarrow \begin{cases} r, & \text{если } t = T, \\ r + \gamma \max\limits_{a'_1, \ldots, a'_n} Q^j(o', a'_1, \ldots, a'_n; \theta'), & \text{если } t \neq T. \end{cases}$$

14. Вычислить для каждого j-го кортежа $(o, a_1, \ldots, a_n; r, o')$ из мини-выборки значение $Q^j(o, a_1, \ldots, a_n; \theta)$, предварительно передав в целевую нейронную сеть o.

15. Вычислить функцию потерь

$$L^j(\theta) \leftarrow \frac{1}{K} \sum_{p=1}^{K} (y_p^j - Q_p^j(o, a_1, \ldots, a_n; \theta))^2.$$

16. Обновить веса θ основной нейронной сети с использованием $L^j(\theta)$ и скорости обучения α.

17. Заменить веса θ' целевой нейронной сети весами θ основной нейронной сети, если закончился очередной период Cs.

Использованы следующие обозначения: α — скорость обучения; γ — шаг дисконтирования; ε — параметр эпсилон; Ne — число эпизодов обучения; Eb — буфер воспроизведения; Bl — объем буфера воспроизведения; Bs — объем мини-выборки из буфера воспроизведения; Cs — период копирования весов из основной нейронной сети в целевую сеть (в эпизодах); i — номер агента; a_i — действие агента i, K_i — число действий, доступных агенту i; A_i — множество действий, доступных агенту i; j — кортеж из мини-выборки; θ, θ' — веса основной и целевой нейрон-

ных сетей; o — совместное наблюдение (каждый агент i получает одинаковое наблюдение); o' — следующее наблюдение; $\mathbb{O}$ — множество совместных наблюдений; Q — нейронная Q-сеть; r — вознаграждение; t — момент времени; T — длительность временного интервала.

Реализована возможность трансфера знаний для агентов, обучающихся в среде мира-сетки на основе алгоритмов IQL, CDQN, MADDPG. В рамках взаимодействия со средой каждый агент мог совершить в общей сложности до 100 тыс. действий в процессе обучения. Полученные результаты представлены на рис. 3.

Пример для алгоритма CDQN достижения цели агентами в мультиагентной среде размером 4×4 на основе обучения с подкреплением с учетом использования трансфера знаний из LLM показан на рис. 4.

Обсуждение полученных результатов. Представлены возможности трансфера знаний для мультиагентной системы, определены особенности передачи данных, перечислены факторы, которые привели к улучшению или ухудшению восприятия задачи с помощью LLM. Продемонстрированы результаты обучения агентов в мультиагентной среде на основе трансфера знаний и протестированы различные алгоритмы.

Согласно результатам, представленным на рис. 3, следует выделить алгоритмы, которые наиболее успешно подходят для рассматриваемой задачи. Наибольшую эффективность в обучении показал алгоритм CDQN, в то время как MADDPG не смог обучить агентов столь эффективно с учетом предоставляемой информации от LLM. Возможно, это связано с достаточно высокими штрафами, которые накладываются на агентов, а с учетом информации в результате трансфера знаний алгоритм обучения может оказаться недостаточно успешным для рассматриваемой задачи. Это привело к тому, что агенты не смогли научиться координировать действия для достижения общей цели.

Следует отметить, как теперь обучены агенты в случае с двумя другими алгоритмами (IQL, CDQN). LLM выбрала стратегию формирования цепочки рассуждения так, чтобы агент 1 совершал действия для достижения цели, а агент 2 не мешал ему [17–23], что позволило минимизировать число шагов. В процессе обучения эта информация усвоена агентами и преобразована в эффективную стратегию, в которой агент 1 выполняет действия для достижения цели (выхода из лабиринта) и приводит к наибольшей награде и завершению игрового эпизода. Агент 2, чтобы не перемещаться по пройденному пути и не мешать агенту 2, в большинстве случаев выбирает тактику умышленного бездействия (совершает действия, которые не приведут к его перемещению на другую клетку).

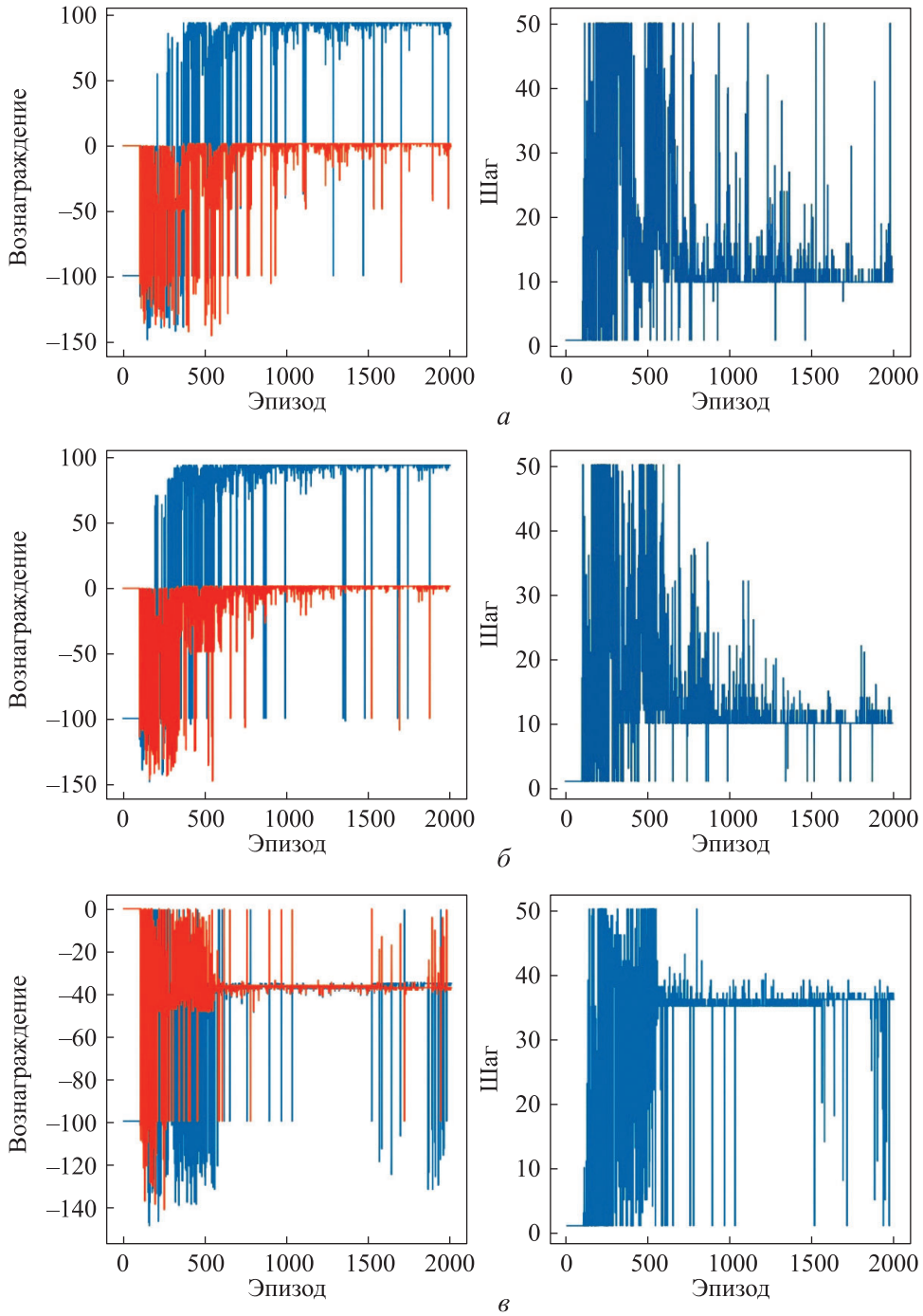

Рис. 3. Результаты обучения агентов 1 (—) и 2 (—)
на основе трансфера знаний с использованием алгоритмов IQL (*а*), CDQN (*б*), MADDPG (*в*)

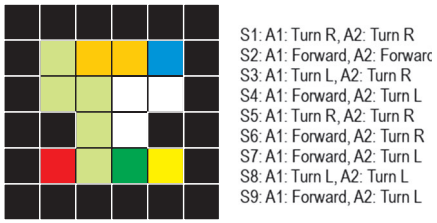

Рис. 4. Пример визуального отображения игрового эпизода в среде мира-сетки размером 4 × 4

Заключение. Использована большая языковая модель Mistral-7B Instruct-v0.3, с помощью которой формировались рассуждения о возможных последовательностях действий, приводящих к достижению конечных целей в мультиагентных средах. Определено, каким образом следует взаимодействовать с LLM для достижения лучших результатов относительно успешности прохождения задач в среде мира-сетки. Реализован метод трансфера знаний для LLM-ориентированных алгоритмов машинного обучения в мультиагентной системе. Установлено, что алгоритм CDQN оказался наиболее подходящим для восприятия информации от большой языковой модели в мультиагентной системе.

Большие языковые модели демонстрируют потенциал для замены ряда существующих приложений и элементов программных систем, что создает новые возможности для адаптации задач и цифровых продуктов. Для успешного внедрения LLM в процесс создания подобного требуется переосмысление подходов к планированию и проектированию. Кроме того, использование таких инновационных инструментов, как трансфер знаний, позволит в перспективе не только повысить эффективность разработки, но и создавать конкурентоспособные решения, соответствующие современным требованиям рынка. Внедрение подобных практик станет важным шагом на пути к развитию качественно новых форм программного обеспечения.

Успешно реализован и протестирован метод, позволяющий LLM направлять действия агентов, предоставляя информацию на основе цепочки рассуждений о наиболее подходящих шагах для достижения целей. Однако возможность Mistral-7B Instruct-v0.3 справляться лишь с крайне ограниченными небольшим средами свидетельствует о необходимости дальнейшего рассмотрения других LLM с большим числом параметров, что приведет к увеличению требований к используемому оборудованию. В рамках последующих работ допустимо исследовать возможность более глубокого внедрения LLM для большей координации и взаимодействия с агентами и средой в рамках рекомендательной мультиагентной системы.

ЛИТЕРАТУРА

[1] Морозов К.А. Баланс между использованием большой языковой модели и обучением с подкреплением. *Наука, технологии и бизнес. VI Межвуз. конф. аспирантов, соискателей и молодых ученых*. М., Изд-во МГТУ им. Н.Э. Баумана, 2024, с. 328–334.

[2] Jiang A.Q., Sablayrolles A., Mensch A., et al. Mistral 7B. *arXiv:2310.06825*.
DOI: https://doi.org/10.48550/arXiv.2310.06825

[3] Морозов К.А. Особенности алгоритма обучения с подкреплением в мультиагентных средах на основе нейронных сетей трансформеров. *ИИАСУ'23. Сб. ст. II Всерос. науч. конф.* Т. 1. М., КДУ, Добросвет, 2023, с. 188–195.
DOI: https://doi.org/10.31453/kdu.ru.978-5-7913-1351-5-2023-435

[4] Velichko N.A. Distributed multi-agent reinforcement learning based on feudal networks. *6th REEPE*, 2024. DOI: https://doi.org/10.1109/REEPE60449.2024.10479775

[5] Morgunov E.F., Alfimtsev A.N. The "Stag Hunt" social dilemma in multi-agent reinforcement learning. *6th REEPE*, 2024.
DOI: https://doi.org/10.1109/REEPE60449.2024.10479770

[6] Morozov K.A. Models as a key factor of environments design in multi-agent reinforcement learning. *6th REEPE*, 2024.
DOI: https://doi.org/10.1109/REEPE60449.2024.10479882

[7] Zhu Z., Lin K., Jain A.K., et al. Transfer learning in deep reinforcement learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell*., 2023, vol. 45, pp. 13344–13362.
DOI: https://doi.org/10.1109/TPAMI.2023.3292075

[8] Kostrikov I., Nair A., Levine S. Offline reinforcement learning with implicit Q-learning. *arXiv:2110.06169*. DOI: https://doi.org/10.48550/arXiv.2110.06169

[9] Mnih V., Kavukcuoglu K., Silver D., et al. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, pp. 529–533.
DOI: https://doi.org/10.1038/nature14236

[10] Lowe R., Wu Y., Tamar A., et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv:1706.02275*.
DOI: https://doi.org/10.48550/arXiv.1706.02275

[11] Leike J., Martic M., Krakovna V., et al. AI safety gridworlds. *arXiv:1711.09883*.
DOI: https://doi.org/10.48550/arXiv.1711.09883

[12] Wei J., Wang X., Schuurmans D., et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*.
DOI: https://doi.org/10.48550/arXiv.2201.11903

[13] Lightman H., Kosaraju V., Burda Y., et al. Let's verify step by step. *arXiv 2305.20050*.
DOI: https://doi.org/10.48550/arXiv.2305.20050

[14] Huang L., Yu E., Ma W., et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 2025, no. 2, vol. 43, pp. 1–55. DOI: https://doi.org/10.1145/3703155

[15] Pawitan Y., Holmes C. Confidence in the reasoning of large language models. *arXiv:2412.15296*. DOI: https://doi.org/10.48550/arXiv.2412.15296

[16] Xue Y., Kudenko D., Khosla M. Graph learning-based generation of abstractions for reinforcement learning. *Neural Comput. & Applic.*, 2025, vol. 37, no. 19, pp. 13187–13207. DOI: https://doi.org/10.1007/s00521-023-08211-x

[17] Величко Н.А., Голубев Е.Ж., Моргунов Е.Ф. и др. Пешеходные ловушки как социальные дилеммы умного города и их решение алгоритмом WoLF-PHC. *ИИАСУ'22. Сб. ст. Всерос. науч. конф.* Т. 1. М., Изд-во МГТУ им. Н.Э. Баумана, 2022, с. 181–191. EDN: HWZUKR

[18] Моргунов Е.Ф., Алфимцев А.Н. Распознавание и решение социальной дилеммы «охота на оленя» с помощью мультиагентного обучения с подкреплением. *ИИАСУ'23. Сб. ст. II Всерос. науч. конф.* Т. 1. М., КДУ, Добросвет, 2023, с. 182–187. DOI: https://doi.org/10.31453/kdu.ru.978-5-7913-1351-5-2023-435

[19] Zhang Y., Mao S., Ge T., et al. LLM as a mastermind: a survey of strategic reasoning with large language models. *arXiv:2404.01230*.
DOI: https://doi.org/10.48550/arXiv.2404.01230

[20] Liu I.J., Jain U., Yeh R.A., et al. Cooperative exploration for multi-agent deep reinforcement learning. *Proc. PMLR*, 2021, vol. 139, pp. 6826–6836.
URL: https://proceedings.mlr.press/v139/liu21j

[21] Алфимцев А.Н. Нечеткое агрегирование мультимодальной информации в интеллектуальном интерфейсе. *Программные продукты и системы*, 2011, № 3, с. 44–48. EDN: OWJLVH

[22] Vidmanov D.A., Alfimtsev A.N. MARLMUI: multi-agent reinforcement learning approach in mobile adaptive user interface. *5th REEPE*, 2023.
DOI: https://doi.org/10.1109/REEPE57272.2023.10086785

[23] Qiu W., Wang X., Yu R., et al. RMIX: learning risk-sensitive policies for cooperative reinforcement learning agents. *arXiv:2102.08159*.
DOI: https://doi.org/10.48550/arXiv.2102.08159

Морозов Кирилл Андреевич — аспирант, ассистент кафедры «Информационные системы и телекоммуникации» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Алфимцев Александр Николаевич — д-р техн. наук, заведующий кафедрой «Информационные системы и телекоммуникации» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Просьба ссылаться на эту статью следующим образом:

Морозов К.А., Алфимцев А.Н. Трансфер знаний для LLM-ориентированных алгоритмов машинного обучения в мультиагентных системах. *Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение*, 2026, № 1 (154), с. 80–95. EDN: FIHEJC

KNOWLEDGE TRANSFER FOR LLM-BASED MACHINE LEARNING ALGORITHMS IN MULTI-AGENT SYSTEMS

K.A. Morozov kmorozov@bmstu.ru
A.N. Alfimtsev alfim@bmstu.ru

BMSTU, Moscow, Russian Federation

Abstract

The ability of large language models to cope with intellectual tasks is an extremely important skill in a variety of environments that require making decisions based on publicly available information. In reinforcement learning, and especially in multi-agent learning, regardless of the overall complexity of the environment, it is extremely important to achieve significant results based on essentially simple actions that could seem impossible in retrospect. This article considers the possibility of using the large language model (LLM) *Mistral-7B Instruct-v0.3* for application in the problem of multi-agent reinforcement learning. A method for interaction with LLM is developed in order to use the reasoning of the large language model for the problem of planning and distributing actions. An assessment of the reflection of the large language model as a result of the actions it designated as necessary to achieve the goal set in the environment is carried out. The implemented knowledge transfer from LLM allows using successful approaches for multi-agent reinforcement learning problems in a grid-world environment. An experimental comparison of machine learning algorithms that can effectively interact with the information provided to them, obtained as a result of interaction with a large language model, is carried out. The proposed method allows embedding the LLM reasoning structure into the learning of a multi-agent system

Keywords

Multi-agent reinforcement learning, large language models, machine learning, reasoning

Received 11.06.2025
Accepted 05.12.2025

This work was carried out within the State Assignment (no. FSFN-2024-0059)

REFERENCES

[1] Morozov K.A. [Balance between using a large language model and reinforcement learning]. *Nauka, tekhnologii i biznes. VI Mezhvuz. konf. aspirantov, soiskateley i molodykh uchenykh* [Science, Engineering and Business, VI Interacademic Conf. for Graduate Students and Young Researchers]. Moscow, BMSTU Publ., 2024, pp. 328–334 (in Russ.).

[2] Jiang A.Q., Sablayrolles A., Mensch A., et al. Mistral 7B. *arXiv:2310.06825*.
DOI: https://doi.org/10.48550/arXiv.2310.06825

[3] Morozov K.A. Features of reinforcement learning algorithm in multi-agent environments based on neural networks of transformers. *IIASU'23. Sb. st. II Vseros. nauch. konf.* T. 1 [IIASU'23 – Artificial Intelligence in Management, Control, and Data Processing Systems. Proc. II All-Russian Sci. Conf. Vol. 1]. Moscow, KDU Publ., Dobrosvet Publ., 2023, pp. 188–195 (in Russ.).
DOI: https://doi.org/10.31453/kdu.ru.978-5-7913-1351-5-2023-435

[4] Velichko N.A. Distributed multi-agent reinforcement learning based on feudal networks. *6th REEPE*, 2024. DOI: https://doi.org/10.1109/REEPE60449.2024.10479775

[5] Morgunov E.F., Alfimtsev A.N. The "Stag Hunt" social dilemma in multi-agent reinforcement learning. *6th REEPE*, 2024.
DOI: https://doi.org/10.1109/REEPE60449.2024.10479770

[6] Morozov K.A. Models as a key factor of environments design in multi-agent reinforcement learning. *6th REEPE*, 2024.
DOI: https://doi.org/10.1109/REEPE60449.2024.10479882

[7] Zhu Z., Lin K., Jain A.K., et al. Transfer learning in deep reinforcement learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023, vol. 45, pp. 13344–13362.
DOI: https://doi.org/10.1109/TPAMI.2023.3292075

[8] Kostrikov I., Nair A., Levine S. Offline reinforcement learning with implicit Q-learning. *arXiv:2110.06169*. DOI: https://doi.org/10.48550/arXiv.2110.06169

[9] Mnih V., Kavukcuoglu K., Silver D., et al. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, pp. 529–533.
DOI: https://doi.org/10.1038/nature14236

[10] Lowe R., Wu Y., Tamar A., et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv:1706.02275*.
DOI: https://doi.org/10.48550/arXiv.1706.02275

[11] Leike J., Martic M., Krakovna V., et al. AI safety gridworlds. *arXiv:1711.09883*.
DOI: https://doi.org/10.48550/arXiv.1711.09883

[12] Wei J., Wang X., Schuurmans D., et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*.
DOI: https://doi.org/10.48550/arXiv.2201.11903

[13] Lightman H., Kosaraju V., Burda Y., et al. Let's verify step by step. *arXiv 2305.20050*.
DOI: https://doi.org/10.48550/arXiv.2305.20050

[14] Huang L., Yu E., Ma W., et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 2025, no. 2, vol. 43, pp. 1–55. DOI: https://doi.org/10.1145/3703155

[15] Pawitan Y., Holmes C. Confidence in the reasoning of large language models. *arXiv:2412.15296*. DOI: https://doi.org/10.48550/arXiv.2412.15296

[16] Xue Y., Kudenko D., Khosla M. Graph learning-based generation of abstractions for reinforcement learning. *Neural Comput. & Applic.*, 2025, vol. 37, no. 19, pp. 13187–13207. DOI: https://doi.org/10.1007/s00521-023-08211-x

[17] Velichko N.A., Golubev E.Zh., Morgunov E.F., et al. [Pedestrian traps as social dilemmas of a smart city and their solution by the wolf-PHC algorithm]. *IIASU'22. Sb. st. Vseros. nauch. konf.* T. 1 [IIASU'22 – Artificial Intelligence in Management, Control, and Data Processing Systems. Proc. II All-Russian Sci. Conf. Vol. 1]. Moscow, BMSTU Publ., 2022, pp. 181–191 (in Russ.). EDN: HWZUKR

[18] Morgunov E.F., Alfimtsev A.N. Recognizing and solving the "Stag Hunt" social dilemma using multi-agent reinforcement learning. *IIASU'23. Sb. st. II Vseros. nauch. konf.* T. 1 [IIASU'23 – Artificial Intelligence in Management, Control, and Data Processing Systems. Proc. II All-Russian Sci. Conf. Vol. 1]. Moscow, KDU Publ., Dobrosvet Publ., 2023, pp. 182–187 (in Russ.).
DOI: https://doi.org/10.31453/kdu.ru.978-5-7913-1351-5-2023-435

[19] Zhang Y., Mao S., Ge T., et al. LLM as a mastermind: a survey of strategic reasoning with large language models. *arXiv:2404.01230*.
DOI: https://doi.org/10.48550/arXiv.2404.01230

[20] Liu I.J., Jain U., Yeh R.A., et al. Cooperative exploration for multi-agent deep reinforcement learning. *Proc. PMLR*, 2021, vol. 139, pp. 6826–6836.
Available at: https://proceedings.mlr.press/v139/liu21j

[21] Alfimtsev A.N. Fuzzy aggregation of multimodal information in an intelligent interface. *Programmnye produkty i sistemy* [Software & Systems], 2011, no. 3, pp. 44–48 (in Russ.). EDN: OWJLVH

[22] Vidmanov D.A., Alfimtsev A.N. MARLMUI: multi-agent reinforcement learning approach in mobile adaptive user interface. *5th REEPE*, 2023.
DOI: https://doi.org/10.1109/REEPE57272.2023.10086785

[23] Qiu W., Wang X., Yu R., et al. RMIX: learning risk-sensitive policies for cooperative reinforcement learning agents. *arXiv:2102.08159*.
DOI: https://doi.org/10.48550/arXiv.2102.08159

Morozov K.A. — Post-Graduate Student, Assistant, Department of Information Systems and Telecommunications, BMSTU (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Alfimtsev A.N. — Dr. Sc. (Eng.), Head of the Department of Information Systems and Telecommunications, BMSTU (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Please cite this article in English as:

Morozov K.A., Alfimtsev A.N. Knowledge transfer for LLM-based Machine Learning algorithms in multi-agent systems. *Herald of the Bauman Moscow State Technical University, Series Instrument Engineering*, 2026, no. 1 (154), pp. 80–95 (in Russ.). EDN: FIHEJC