Слайд 1

**Тема доклада:** «Действительно ли обучение с подкреплением развивает способность к рассуждению в больших языковых моделях сверх базовой модели?»

Слайд 2

**План доклада:**

* Что такое большие языковые модели (**Large Language Models** или **LLM**) и зачем они нужны?
* Что такое обучение с подкреплением с верифицируемыми наградами (**Reinforcement Learning with Verifiable Rewards** или **RLVR**) и его отличие от обычного обучения с подкреплением (**Reinforcement Learning** или **RL**)?
* Действительно ли RLVR расширяет способности LLM к рассуждениям или просто оптимизирует уже имеющиеся?
* Какие есть ограничения подхода RLVR при обучении языковых моделей?
* Результаты исследования и заключение.

Слайд 3

Что такое большие языковые модели (LLM) и зачем они нужны?

**Большие языковые модели**, или **LLM** (от английского **Large Language Models**) — это мощные алгоритмы искусственного интеллекта, обученные на огромных объемах текстовых данных. Они умеют генерировать осмысленные тексты, отвечать на вопросы, переводить языки и даже решать логические задачи.

**Примеры таких моделей:** GPT-4 (**OpenAI**), Gemini 1.5 (**Google**), DeepSeek-V3 (**DeepSeek**), Grok-1.5 (**xAI**), Qwen 72B (**Alibaba**) и другие.

Слайд 4

Что такое RLVR и чем оно отличается от обычного RL?

Обучение с подкреплением (**Reinforcement Learning**, **RL**) — это когда модель учится методом проб и ошибок, получая "награды" за правильные действия.

Но в случае с языковыми моделями появился особый вариант — **RLVR** (**Reinforcement Learning with Verifiable Rewards**). Его ключевые особенности:

* Автоматическая проверка
* Масштабируемость
* Специализация

**Автоматическая проверка (Как работает?)**

* **Обычное RL:** Человек вручную оценивает ответы (верно/неверно)
* **RLVR:** Проверка происходит алгоритмически через четкие правила

|  |  |
| --- | --- |
| Тип задачи | Критерий проверки |
| Математика | Сравнение с числовым ответом |
| Программирование | Запуск unit-тестов |
| Логика | Проверка шагов доказательства |

Плюсы:

* Нет субъективности (все проверяется кодом)
* Можно обрабатывать миллионы ответов без людей

**Масштабируемость (Почему это мощно?)**

* **Обычное RL:** Требует людей-оценщиков, а это дорого и медленно
* **RLVR:** Автономность (модель учится 24/7, получая мгновенную обратную связь) и скорость (за одно и то же время модель делает гораздо больше автоматических проверок)

**Специализация (Зачем это нужно?)**

* **Обычное RL:** "Универсал" — знает всё понемногу
* **RLVR:** "Профи" — заточена под конкретные задачи

Как достигается:

* Целевые награды (Только за нужные действия. Например, верный код → +1, ошибка → 0)
* Фильтрация шума (Игнорируются "креативные, но бесполезные" ответы)

Так же важно понимать, что у RLVR-моделей есть свои проблемы:

* **"Туннельное зрение".** Модель забывает нестандартные, но верные решения. Пример: В математике RLVR-модель пропускает красивое, но сложное доказательство.
* **Зависимость от тестов.** Если в тестах ошибка, то модель выучит неправильное.
* **Отсутствие креативности.** RLVR-поэт будет писать только в рамках шаблонов.

Слайд 5

Действительно ли RLVR расширяет способности LLM?

На слайде можно увидеть графики зависимости метрики эффективности (**pass@k**) от количества сэмплов (**k**) для базовых моделей и их версий, обученных на различных математических тестах (**AIME24** – конкурс по математике для старшеклассников, **AMC23** – конкурс по математике для старшеклассников, **MATH500** – кастомный датасет из 500 сложных математических задач, включая: олимпиадные задачи, университетские задачи, логические головоломки, **Minerva** – датасет созданный Google с математическими и научными задачами).

Синяя линия (**Base**) показывает производительность базовой модели. Оранжевая линия (**RL**) показывает производительность модели после RL-оптимизации.

Слайд 6

Когда лучше использовать базовую модель, а когда стоит применять RLVR?

На слайде вы можете увидеть сравнение базовой и RLVR моделей для двух задач.

Визуализация в виде дерева:

* **Корень дерева** — это начальная постановка задачи
* **Ветви** — возможные цепочки рассуждений модели (серые = невыбранные, черные = выбранные)
* **Листья** — конечные ответы (зелёные = правильные, красные = ошибочные)

Слева показаны деревья для базовой модели, а справа деревья для RLVR-модели. Сверху решения задачи A, снизу решения задачи B.

**Задача A (простая):**

* RLVR-модель быстро находит оптимальный путь
* Базовая модель требует больше итераций

**Задача B (сложная):**

* Базовая модель находит неочевидное решение через редкую ветвь
* RLVR-модель "зацикливается" на оптимизированных, но ошибочных путях

Слайд 7

Результаты, выводы и заключение.

Как показала данная работа, обучение с подкреплением не развивает способность к рассуждению в больших языковых моделях сверх базовой модели.

Мы можем использовать RLVR-модели, если нужны быстрые, точные ответы (чат-боты, автотесты). Но базовые модели оказываются эффективней для задач, где важны креативность или новые подходы.

Дальнейшее развитие этой темы и возможное нахождение альтернативы это создание гибрида (RLVR + базовая модель) — баланс скорости и гибкости.

Слайд 8

Спасибо за внимание!