

Оптимизация агентных способностей Qwen 2.5 (1.5B) с помощью GRPO и Curriculum Learning

1 Постановка задачи

Цель проекта - обучить компактную языковую модель (1.5B параметров) решать задачу поиска пути в графе.

Агент должен проанализировать условия, построить цепочку рассуждений и выдать итоговый маршрут в строгом XML-формате: сначала процесс принятия решений в тегах `<reasoning>`, затем итоговая последовательность узлов (чисел) в тегах `<answer>`.

2 Проводимые эксперименты

Для стабилизации обучения и улучшения агентных метрик был протестирован метод **Curriculum Learning** (обучение по учебному плану с постепенным усложнением задач). Базовая модель: `unslloth/Qwen2.5-1.5B-Instruct`.

Оценка проводилась на трех отложенных датасетах разной сложности (`test_easy`, `test_medium`, `test_hard`)

Сравнивались три конфигурации:

1. **Базовая модель (Baseline):** Исходный Qwen 2.5 1.5B без дообучения (zero-shot).
2. **Base GRPO:** Обучение алгоритмом GRPO (Group Relative Policy Optimization) на смешанных данных без градации сложности.
3. **Curriculum GRPO:** Обучение алгоритмом GRPO с постепенным повышением сложности задач (от простых путей к сложным).

3 Результаты

Внедрение Curriculum Learning оказалось критически важным для успешного применения GRPO.

Стандартный подход (Base GRPO) привел к деградации модели, в то время как Curriculum обеспечил кратный рост целевых метрик.

Сводная таблица результатов (на примере датасета Easy):

Метрика	1. Baseline (Qwen)	2. Base GRPO	3. Curriculum GRPO	Динамика (K базе)
Accuracy (Доля игр, когда был найден путь из A в Б)	9.0%	1.0%	14.0%	🚀 Рост в 1.5 раза
Optimal Rate (Доля игр, когда был найден оптимальный путь из A в Б)	6.0%	0.0%	12.5%	↗ Рост в 2 раза
Optimality Gap ('На сколько в среднем стоимость путей агента больше идеальных)	3.67	7.50	0.82	⬇ Снижение в 4.5 раза
Hallucination Rate (Доля игр, где модель придумала несуществующий путь)	7.5%	20.5%	6.0%	🌐 Минимальный уровень
Format Compliance (Доля игр, в которых модель соблюдала правильный формат \n (сначала reasoning, потом answer))	18.5%	23.0%	28.5%	📝 Улучшение на 10 п.п.
Reasoning Len (Средняя длина reasoning части)	86	255	122	⚖️ Оптимальный баланс

Ключевые выводы:

- Крах базового GRPO:** Запуск RL-алгоритма без учебного плана разрушил полезные веса претрейна. Модель провалилась в локальный минимум: точность упала до 1%, длина рассуждений выросла в 3 раза (модель начала «лить воду»), а галлюцинации взлетели до 20.5%. Модель выучила только синтаксис ответа, потеряв логику.
- Эффективность Curriculum Learning:** Постепенное усложнение задач не только спасло политику от катастрофического забывания, но и позволило превзойти базовую модель. Способность находить абсолютно оптимальный путь выросла в 2 раза (до 12.5%).
- Снижение цены ошибки:** Метрика Optimality Gap у Curriculum-модели упала до 0.82. Это означает, что даже в случае неверного ответа агент ошибается незначительно, выстраивая маршрут, максимально близкий к математическому идеалу.

Итог: Для малогабаритных моделей (1.5B), решающих сложные агентные задачи, применение GRPO в чистом виде деструктивно. Использование Curriculum Learning является необходимым условием для успешного Alignment-процесса и защиты от Reward Hacking.

