



Московский государственный университет имени М.В.Ломоносова  
Факультет вычислительной математики и кибернетики  
Кафедра суперкомпьютеров и квантовой информатики

Алексеев Никита Константинович,  
Попова Нина Николаевна

Улучшение малых языковых моделей с помощью  
итеративного многошагового рассуждения

Москва, 2026

# Введение

- Благодаря методу Chain-of-Thought LLM показывают себя крайне хорошо в решении сложных задач, требующих нескольких логических шагов.
- Однако во многих практических сценариях применение больших моделей может быть невыгодным или вовсе невозможным.
- Альтернатива – SLM. В плане эффективности они выигрывают у LLM, однако перенос подхода CoT сопряжён с рядом трудностей.
- Соответственно необходимо совершенствовать методы обучения рассуждениям малых моделей.

# Основные подходы к обучению рассуждениям

- Обучение на предпочтениях – на основе положительных и отрицательных примеров.
- Обучение с подкреплением – на основе верификатора, оценивающего ответы обучаемой модели.
- Ключевые проблемы и ограничения:
  - Малая ёмкость моделей.
  - Деградация при обучении.
  - Стандартные методологии обучения разработаны под LLM.

# Применяемые стратегии

- Дистилляция знаний из более крупных моделей (как полноценная, так и через создание большой синтетической выборки).
- Добавление обучения на предпочтениях, так как при генерации синтетики LLM отвечает не всегда корректно.
- Использование верификаторов для проверки данных генерируемых обучаемой моделью, которые потом будут подаваться на вход при обучении.

# Идея предложенного метода

- Комбинирование методов обучения с алгоритмическим отбором данных между этапами.
- Повышение вероятности правильных ответов в противовес ложным за счёт обучения на предпочтениях – парах положительный/отрицательный пример.
- Улучшение рассуждающей способности модели (расширения пространства возможных решений) с помощью обучения с подкреплением.
- Итеративное применение выделенных методов при условии повышения качества модели.

# Рассматриваемая проблемная область

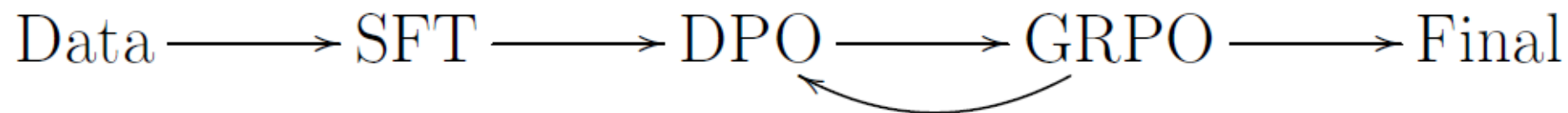
- Математический датасет GSM8K, состоящий из математико-логических задач.
- Малые языковые модели: Llama-3.2-1B и Qwen2.5-1.5B.

Таб. 1: Характеристики выбранных моделей

| Характеристики       | Llama-3.2-1B | Qwen2.5-1.5B |
|----------------------|--------------|--------------|
| Число параметров     | 1.23 B       | 1.54 B       |
| Число слоёв          | 16           | 28           |
| Число голов внимания | 20           | 14           |
| Размер скрытого слоя | 2048         | 1536         |

# Предлагаемое решение

- Итеративное обучение SLM на основе SFT, DPO, GRPO.



- SFT – обучение модели базовым рассуждениям, умению воспроизводить простые логические цепочки.
- DPO – оптимизация моделей с помощью предпочтений задаваемых вручную.
- GRPO – групповая относительная политика оптимизации.
- Выход из цикла, в случае выполнения условия после очередной итерации DPO+GRPO не выросло.

# Модификация методов DPO и GRPO

- Изменение функции потерь: добавлено взвешивание логитов на уровне токенов.
- Функция потерь для DPO:

$$L_{WDPO}(\theta) = -\mathbb{E}_{(x, y^+, y^-)} \left[ \log \sigma \left( \beta \left( \sum_{t=1}^{T^+} \mathbf{w}_t^+ * \log \frac{\pi_{\theta}(y_t^+ | x, y_{<t}^+)}{\pi_{ref}(y_t^+ | x, y_{<t}^+)} - \sum_{t=1}^{T^-} \mathbf{w}_t^- * \log \frac{\pi_{\theta}(y_t^- | x, y_{<t}^-)}{\pi_{ref}(y_t^- | x, y_{<t}^-)} \right) \right) \right]$$

- Функция потерь для GRPO:

$$L_{WGRPO}(\theta) = -\mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \sum_{t=1}^T \mathbf{w}_t * \log \frac{\pi_{\theta}(y_t | x, y_{<t})}{\pi_{ref}(y_t | x, y_{<t})} - R(x, y) \right) \right]$$

# Специфика метода DPO для SLM

- Необходимо более строго контролировать данные предпочтений.
- Слишком большая разница в их относительных вероятностях приводит к нестабильному обучению.
- Необходим отбор наиболее информативных пар, на основе разности логарифмов правдоподобия.
- Это позволяет сократить объём обучающей выборки, одновременно повышая стабильность обучения и скорость сходимости без ухудшения итогового качества модели.

# Специфика метода GRPO для SLM

- Повышенная чувствительность к сигналу наград.
- Для стабильности обучения необходимо использовать относительно простые, но строгие функции.
- Слабый положительный сигнал для частично правильных ответов снижает дискриминативность функции награды.
- Дополнительно важную роль играет явный контроль длины генераций, позволяющий ограничить рост нерелевантных токенов и повысить информативность сигнала награды.

# Связь DPO и GRPO

- DPO — перераспределяет вероятности внутри существующего распределения. Ранжирует, но не создаёт новых стратегий вывода.
- GRPO — использует абсолютный сигнал награды и оценивает каждую последовательность независимо. Стимулирует исследование пространства решений и появление новых корректных рассуждений.
- Соответственно последовательное применение (сначала DPO, затем GRPO) стабилизирует распределение, а затем его расширяет.

# Методика проведения эксперимента

- Фреймворк: pytorch
- Библиотеки: TRL, vLLM.
- Вычислительная платформа: узел суперкомпьютера «270 лет МГУ».
- Минимальное требуемое время для обучения (две итерации) – 15 гри-часов (для набора данных из ~7500 примеров). Всего же необходимо проводить десятки таких запусков, в связи с большим числом комбинаций возможных модификаций методов обучения и отбора данных.

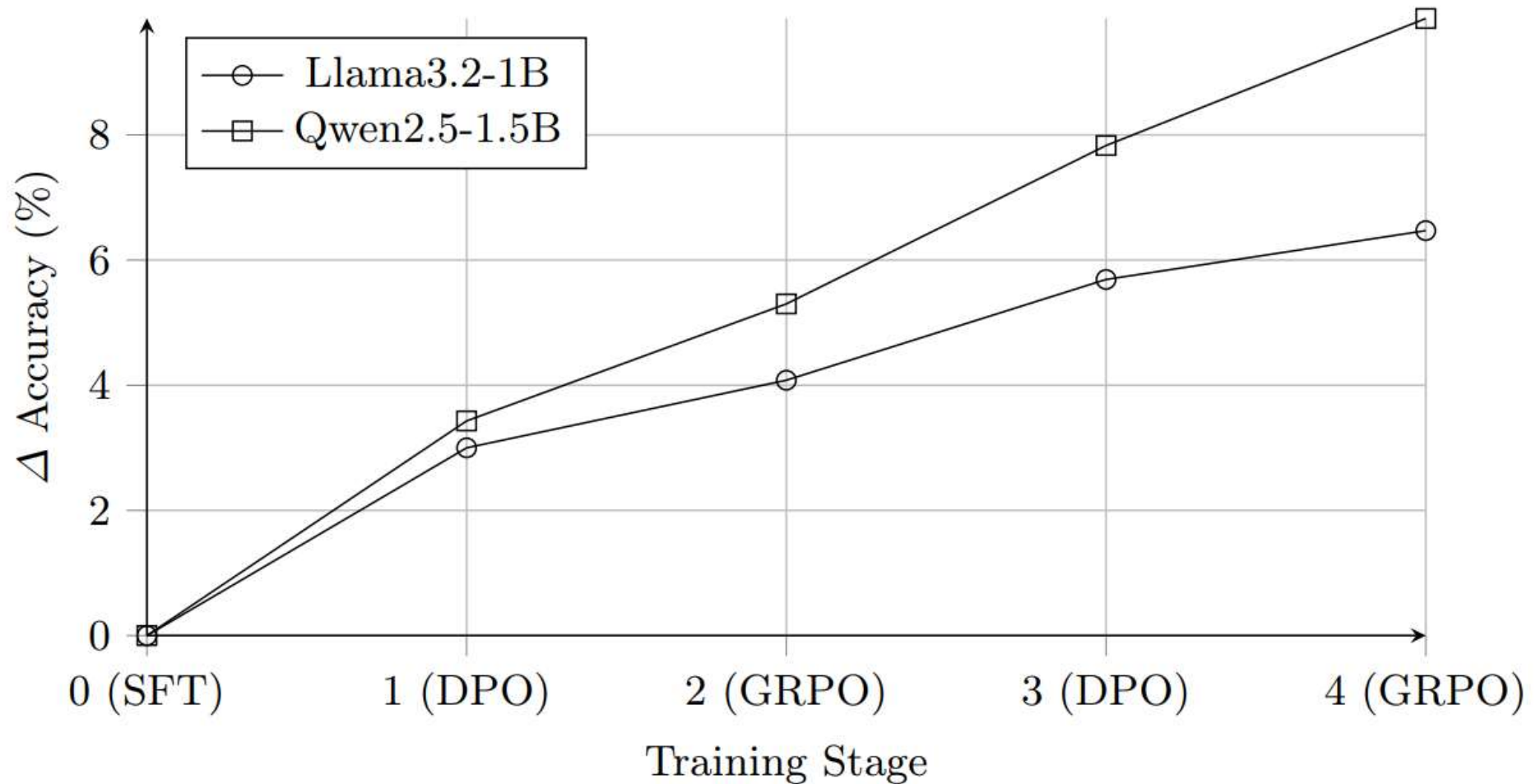
# Результаты экспериментов

Таб. 2: Процент правильных ответов на датасете gsm8k

| Итерация | Модели  |          |
|----------|---------|----------|
|          | Llama1B | Qwen1.5B |
| SFT      | 45.56   | 54.13    |
| DPO-1    | 48.46   | 57.56    |
| GRPO-1   | 49.54   | 59.43    |
| DPO-2    | 51.15   | 61.96    |
| GRPO-2   | 51.93   | 63.99    |

- Обучение базовых (pretrain) моделей проводилось только на GSM8K, для исключения «загрязнения» данных.

## Результаты экспериментов (2)



- Прирост точности относительно SFT baseline

# Выводы по проведённым экспериментам

- Итерационное применение методов DPO и GRPO стабильно улучшает способность малых моделей решать математические задачи, требующие рассуждений.
- Улучшение достигается за счёт стабилизации (ускорения сходимости) процесса обучения, а не специфики домена, поэтому данная стратегия представляется переносимой на другие специализированные области при наличии доменной адаптации и явных метрик качества.

# Заключение

- Предложена итеративная стратегия обучения рассуждениям SLM основанная на последовательном применении методов SFT, DPO и GRPO.
- Предложен алгоритмический подход к формированию и отбору данных предпочтений, не требующий использования внешних языковых моделей.
- Применение метода привело к улучшению: для Llama1B с 45.46% до 51.93% и для Qwen1.5B с 54.13% до 63.99%.
- Дальнейшее развитие: изменения в архитектуре SLM, и как следствие доработка существующих методов обучения.

Спасибо за внимание

# Pass@k: метрика качества LLM

- Pass@k – вероятность решить задачу хотя бы раз из k попыток.
- Соответственно Pass@1 требует правильного ответа с первого раза, зачастую проводится при «жадном» семплинге.
- В Pass@k ( $k > 1$ ) же генерация новых токенов идёт при более высокой температуре, что увеличивает их разнообразие и сильно повышает вероятность появления правильного ответа. Именно поэтому Pass@k резко обгоняет Pass@1.
- Однако цена этому – увеличение стоимости генерации в k раз.