
Адаптация методов обучения с подкреплением в задаче обучения LLM

A Preprint

Денисов Егор Александрович
МГУ им. М.В.Ломоносова
Факультет ВМК, кафедра ММП
Москва, Россия
s02220081@gse.cs.msu.ru

Сердюк Юлиан Анатольевич
МГУ им. М.В.Ломоносова
Факультет ВМК, кафедра ММП
Москва, Россия

Abstract

Современные большие языковые модели (БЯМ) демонстрируют впечатляющие способности к рассуждению, обобщению и генерации текста, однако их эффективность во многом зависит от качества этапов дообучения — дообучения с учителем (англ. SFT) и обучения с подкреплением (англ. RL). Несмотря на то, что данные подходы обширно изучаются, остаются открытыми вопросы о том, как корректно их совмещать для достижения наилучшего качества. В данной работе рассматриваются подходы к адаптации методов обучения с подкреплением для задач дообучения БЯМ с особым вниманием к двум аспектам - комбинации функции потерь и оптимальному чередованию этапов обучения. Эксперименты выполняются на стандартных бенчмарках для тестирования математического и логического рассуждения модели. Анализ результатов показывает, что гибридные схемы, предложенные нами, обеспечивают лучшее соотношение между устойчивостью, скоростью сходимости и способностью к обобщению.

1 Introduction

Современные большие языковые модели (БЯМ) становятся основным инструментом в задачах обработки естественного языка и генерации текста. Однако для достижения высокой точности и способности к обобщению им необходимо сложное многоэтапное обучение, включающее в себя различные этапы. Актуальность исследований в этой области обусловлена необходимостью повышения устойчивости, управляемости и эффективности БЯМ при решении задач, требующих рассуждений и логических выводов.

На сегодняшний день одним из ключевых направлений развития БЯМ является совершенствование этапов дообучения, в частности SFT и RL. Основной подход на стадии RL - обучение с подкреплением на основе человеческих ответов (англ. RLHF) ([Christiano et al., 2023, Schulman et al., 2017, Rafailov et al., 2024, Shao et al., 2024]), а также различные его улучшения. В качестве последних выступают различные модификации функции вознаграждения, стратегий обновления модели и схем совмещения RL с традиционным SFT.

Тем не менее, существующие решения имеют ряд ограничений. Во-первых, большинство подходов не учитывают сложное взаимодействие между этапами SFT и RL, что может приводить к забыванию уже приобретённых моделью знаний ([Yuan et al., 2025]). Во-вторых, функции вознаграждения зачастую плохо коррелируют с реальными показателями качества текста, особенно в задачах рассуждения. В-третьих, в RLHF используются данные собранные в онлайн-режиме. Зачастую такие данные ограничены, и используемые методы оптимизации нередко страдают от нестабильности и переобучения.

В данной работе предлагается подход к адаптации методов обучения с подкреплением для дообучения БЯМ, основанный на гибридной схеме чередования этапов SFT и RL с комбинированной функцией

потерь. Такой подход позволяет учитывать как сигналы от данных с учителем, так и награды, поступающие из среды. Особое внимание уделяется исследованию стратегий управления частотой переходов между режимами обучения и влиянию этих переходов на стабильность оптимизации. Кроме того, рассматриваются модификации функции вознаграждения, направленные на повышение чувствительности к ошибкам логического вывода.

Модели, обученные с использованием предложенного подхода, демонстрируют более быструю сходимость и лучшую способность к обобщению по сравнению с классическим RLHF. Таким образом, представленные результаты вносят вклад в развитие методологий дообучения БЯМ, предлагая более гибкий и устойчивый механизм совмещения SFT и RL, что открывает новые перспективы для дальнейшего повышения качества языковых моделей.

2 Related works

2.1 Парадигмы обучения

Дообучение с учителем является основополагающей техникой для адаптации больших языковых моделей под специфичные задачи. Благодаря своей эффективности, данный подход стал широко применим в различных областях, в том числе и в решении математических задач ([Cobbe et al., 2021, Hendrycks et al., 2021, Yuan et al., 2023]). Тем не менее, чистый SFT позволяет хорошо выучить шаблоны решений, но не помогает улучшить способность модели к рассуждению. Причиной этому служат ограниченные и несбалансированные данные, неоптимальная для данной задачи функция потерь.

Для решения существующих проблем после стадии SFT стало применяться обучение с подкреплением. Таким образом, SFT позволяет достичь высокого базового качества, а RL улучшает эффективность и подстраивает поведение модели под конкретную задачу. Ярким примером совмещения двух парадигм является DeepSeekMath ([Shao et al., 2024])

2.2 Совмещение SFT и RL

Последовательное выполнение стадий SFT и RL также имеет свои ограничения и недостатки. Вследствие этого исследуются способы динамически совмещать данные подходы. Так, в работах [Ma et al., 2025, Liu et al., 2025] предлагаются схожие стратегии чередования RL и SFT, где RL служит основой обучения, а SFT применяется для наиболее сложных примеров в выборке.

Другой класс работ предлагает пошаговую адаптацию и мониторинг тренировочной динамики: SASR ([Chen et al., 2025a]) и близкие методы отслеживают градиентные нормы и распределение выходов модели, чтобы динамически корректировать вклад SFT и RL на каждом шаге обучения и тем самым уменьшить эффект катастрофического забывания и переобучения.

Наконец, в сторону более формализованных схем кооперации шагнули BRIDGE ([Chen et al., 2025b]), который использует двухуровневую оптимизацию для одновременного обучения «нижнего» уровня (RL-обновления) и «верхнего» уровня (SFT). Такая кооперативная постановка направлена на то, чтобы SFT на каждой итерации предоставлял улучшенную инициализацию для стадии обучения с подкреплением, что даёт прирост итоговой точности.

3 Постановка задачи

3.1 Данные

Рассматривается датасет $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, где x_i - текстовое задание, представляющее собой условие математической задачи; y_i - эталонное текстовое решение или ответ. Каждое задание и решение может рассматриваться как последовательность токенов $x_i = (x_{i,1}, \dots, x_{i,T_{x_i}})$, $y_i = (y_{i,1}, \dots, y_{i,T_{y_i}})$, где T_{x_i}, T_{y_i} - длины последовательностей x_i, y_i соответственно.

3.2 Отображение

Модель $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ отображает входной текст x в распределение над возможными выходными последовательностями y :

$$f_\theta(x) = \arg \max_{y \in \mathcal{Y}} \pi_\theta(y|x)$$

Здесь $\pi_\theta(y|x) = \prod_{t=1}^T \pi_\theta(y_t|x, y_{<t})$.

3.3 Внешний критерий качества

Для задач математического рассуждения естественным внешним критерием качества служит точность рассуждения, оцениваемая на уровне финального ответа. В данном случае таким критерием выступает метрика Pass@k.

Pass@k - это вероятность того, что среди первых k решений модели есть хотя бы одно правильное. Математически, если x - задача, $\{y_i\}_{i=1}^m$ - ответы модели, $c = \sum_{i=1}^m \mathbb{I}[y_i -]$ - количество правильных ответов модели. Тогда $\text{Pass@k} = 1 - \frac{C_{m-c}^k}{C_m^k}$ при условии того, что $C_{m-c}^k = 0$, если $k > m - c$

3.4 Оптимизационная задача

Оптимизационная задача выглядит следующим образом:

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{total}(\theta)$$

Здесь совокупная функция потерь определяется как

$$\mathcal{L}_{total}(\theta) = \lambda_{SFT} \mathcal{L}_{SFT}(\theta) + \lambda_{RL} \mathcal{L}_{RL}(\theta)$$

Коэффициенты $\lambda_{SFT}, \lambda_{RL} \geq 0$ регулируют вклад каждого из этапов в данный момент времени. Компонента \mathcal{L}_{SFT} соответствует стандартной кросс-энтропийной функции потерь:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \sum_{t=1}^T \log \pi_\theta(y_t|x, y_{<t})$$

В качестве компоненты \mathcal{L}_{RL} используется функция потерь метода *GRPO* ([Shao et al., 2024]):

$$\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_\theta(\cdot|x)}$$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{T_{y_i}} \sum_{t=1}^{T_{y_i}} \left\{ \min \left[\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL}(\pi_\theta || \pi_{\theta_{old}}) \right\}$$

β, ε - гиперпараметры, $\hat{A}_{i,t}$ - выигрыш относительно средней ожидаемой награды.

4 Предлагаемое решение

4.1 Совмещение SFT и RL

В данной работе рассматривается подбор оптимальной стратегии чередования этапов SFT и RL.

SFT На данном этапе модель обучается на стандартной задаче обучения с учителем, минимизируя кросс-энтропийную функцию потерь:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \sum_{t=1}^T \log \pi_\theta(y_t|x, y_{<t})$$

Это способствует запоминанию общих подходов к решению и базовых знаний.

RL В ходе стадии RL модель обучается на следующую функцию потерь:

$$\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_\theta(\cdot|x)}$$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{T_{y_i}} \sum_{t=1}^{T_{y_i}} \left\{ \min \left[\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{old}}(y_{i,t}|x, y_{i,<t})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL}(\pi_\theta || \pi_{\theta_{old}}) \right\}$$

Этот этап используется для улучшения обобщающей способности модели, помогая ей применять свои знания для решения новых задач.

SFT + RL В качестве стратегии совмещения двух стадий рассматриваются следующие варианты:

1. Использование комбинированной функции потерь:

$$\mathcal{L}_{total}(\theta) = \lambda \mathcal{L}_{SFT}(\theta) + (1 - \lambda) \mathcal{L}_{RL}(\theta), \lambda \in [0, 1]$$

2. Переключение между стадиями на основе анализа метрик валидации - темпа изменения функции потерь, стабильности KL-дивергенции, уровня разнообразия ответов. В данном случае выбирается, какая продолжительность той или иной стадии является оптимальной. Является частным случаем первой стратегии с $\lambda \in \{0, 1\}$

4.2 Энтروпийная регуляризация

Стадия RL является достаточно стохастичной, что может привести к "коллапсу энтропии когда энтропия распределения становится близка к нулю, и политика становится близка к детерминированной. Чтобы этого избежать, рассматривается влияние регуляризационного члена, имеющего вид:

$$R = \beta H(\pi_\theta(y|x)), \quad H(\pi_\theta(y|x)) = - \sum_{t=1}^T \pi_\theta(y_t|x, y_{<t}) \log \pi_\theta(y_t|x, y_{<t})$$

Здесь β - гиперпараметр регуляризации, H - энтропия распределения.

5 Эксперименты

Основной целью экспериментов из данной секции является получение качества, сравнимого с имеющимися результатами на известных бенчмарках, оценивающих математическое рассуждение языковых моделей.

5.1 Бенчмарки

Так как в данной работе используются небольшие модели (0.5B), для оценки качества были выбраны популярные бенчмарки, содержащие математические задачи уровня средней и старшей школы - GSM8K, MATH-500 ([Cobbe et al., 2021, Hendrycks et al., 2021])

GSM8K - это датасет, содержащий около 8 500 задач по арифметике и элементарной математике, сформулированных в виде текстовых описаний на английском языке. Каждое задание сопровождается подробным решением в пошаговом виде и корректным числовым ответом.

MATH-500 - это сокращённая версия набора MATH, включающая 500 задач повышенной сложности, охватывающих широкий спектр разделов школьной и олимпиадной математики: алгебру, геометрию, комбинаторику, теорию чисел и вероятности. Каждая задача содержит подробное текстовое описание и ожидаемый итоговый ответ в числовой или символьной форме.

5.2 Бейзлайны

Для объективной оценки эффективности предложенного подхода проводилось сравнение с рядом современных методов дообучения больших языковых моделей с использованием обучения с подкреплением. В качестве бейзлайнов были выбраны три наиболее релевантных решения: ReLIFT, SuperRL и DeepSeekMath.

ReLIFT - гибридный подход, в котором основное обучение происходит при помощи RL, а для наиболее сложных задач используется SFT

SuperRL - метод, схожий по своей структуре с предыдущим.

DeepSeekMath - классический подход, сочетающий SFT и GRPO

Результаты данных методов на рассматриваемых бенчмарках представлены в таблице ():

Модель	GSM8K	MATH-500
ReLIFT	-	87.4
SuperRL	78.9	-
DeepSeekMath	64.2	-

Таблица 1: Результаты бейзлайнов на GSM8K и MATH-500 (Exact Match, %).



Рис. 1: Sample figure caption.

6 Examples of citations, figures, tables, references

6.1 Citations

Citations use **natbib**. The documentation may be found at

<http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf>

Here is an example usage of the two main commands (**citet** and **citep**): Some people thought a thing [??] but other people thought something else [?]. Many people have speculated that if we knew exactly why ? thought this...

6.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi. See Figure 1. Here is how you add footnotes.¹ Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetur eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

¹Sample of the first footnote.

Таблица 2: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	~ 100
Axon	Output terminal	~ 10
Soma	Cell body	up to 10^6

6.3 Tables

See awesome Table 2.

The documentation for `booktabs` (‘Publication quality tables in LaTeX’) is available from:

<https://www.ctan.org/pkg/booktabs>

6.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

Список литературы

- Paul F Christiano et al. Deep reinforcement learning from human preferences. arXiv:1706.03741v4, 2023.
- John Schulman et al. Proximal policy optimization algorithms. arXiv:1707.06347v2, 2017.
- Rafael Rafailov et al. Direct preference optimization: Your language model is secretly reward model. arXiv:2305.18290v3, 2024.
- Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300v3, 2024.
- Xiangchi Yuan et al. Mitigating forgetting between supervised and reinforcement learning yields stronger reasoners. arXiv:2510.04454v1, 2025.
- Karl Cobbe et al. Training verifiers to solve math word problems. arXiv:2110.14168v2, 2021.
- Dan Hendrycks et al. Measuring mathematical problem solving with the math dataset. arXiv:2103.03874v2, 2021.
- Zheng Yuan et al. Scaling relationship on learning mathematical reasoning with large language models. arXiv:2308.01825v2, 2023.
- Lu Ma et al. Learning what reinforcement learning can’t: Interleaved online fine-tuning for hardest questions. arXiv:2506.07527v2, 2025.
- Yihao Liu et al. Superrl: Reinforcement learning with supervision to boost language model reasoning. arXiv:2506.01096v2, 2025.
- Jack Chen et al. Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. arXiv:2505.13026v3, 2025a.
- Liang Chen et al. Beyond two-stage training: Cooperative sft and rl for llm reasoning. arXiv:2509.06948v2, 2025b.