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), 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. 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, поступа-

ющие 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. 2021a; 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. 2025) 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 шагнул BRIDGE (Chen et al. 2025), 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 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 3.

3.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis

cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

3.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

4 Examples of citations, figures, tables, references

4.1 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...

4.2 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 consectetuer 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, consectetuer 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.

4.3 Tables

See awesome Table 1.

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

`https://www.ctan.org/pkg/booktabs`

4.4 Lists

- Lorem ipsum dolor sit amet

[1] Sample of the first footnote.

Рис. 1: Sample figure caption.

Таблица 1: Sample table title

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

- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

Список литературы