УДК 004.853:004.822

ОГЛЯД МЕХАНІЗМІВ САМОВДОСКОНАЛЕННЯ LLM

Іванов М.С.
e-mail: maksym.ivanov5@nure.ua
Науковий керівник – к.т.н., доц. Головянко М. В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

As LLMs increasingly serve as cognitive cores for agentic systems, autonomous refinement becomes crucial for ensuring accuracy. This work presents a conceptual overview of existing self-improvement mechanisms in large language models (LLMs) aiming to introduce current methodologies and establish a conceptual framework for understanding autonomous model alignment.

Широке застосування LLM як когнітивного ядра агентних систем вимагає від моделей здатності виходити за межі шаблонів, закладених під час первинного навчання. Для задач з багатогранними та складно-формалізованими цілями, такими як генерація діалогових відповідей чи оптимізація читабельності коду, первинна генерація часто не є оптимальною. Вирішенням цієї проблеми є впровадження механізмів ітеративного самовдосконалення, що є фундаментальною характеристикою людського когнітивного процесу [1], який дозволить моделі автономно підвищувати якість відповідей. Такий підхід базується на алгоритмічному циклі послідовних удосконалень, де, через серію ітерацій, модель генерує та корегує відповіді, досягаючи рівня якості, що перевищує можливості однократного виведення.

На рисунку 1 зображено механізм самовдосконалення у вигляді алгоритмічного циклу, що включає взаємодію трьох моделей:

–policy (M): генерує рішення-кандидата;

–feedback: аналізує результат, виявляє зони для покращення;

– refine: здійснює корекцію відповіді на основі відгуку з feedback.

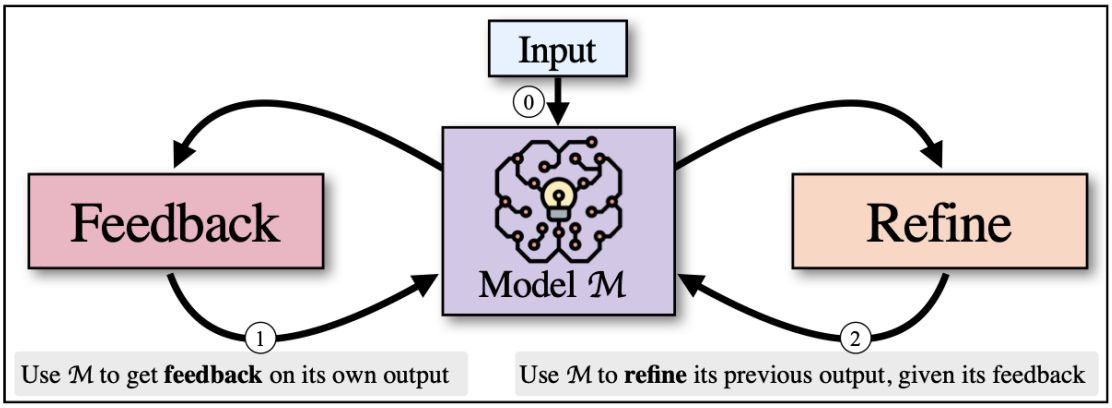

Рисунок 1 – Схема алгоритмічного циклу механізму самовдосконалення

Розглянемо кожен принцип роботи алгоритму окремо.

Згідно з (1), на вхід моделі policy (M) надходять вхідні дані та інструкція. Результатом роботи моделі є відповідь-кандидат.

$$y_0 = M\left(p_{gen} \,||\, x\right). \quad (1)$$

Відповідь передається до моделі feedback (2). На основі вхідних даних, відповіді-кандидата та інструкції для аналізу, модель формує відгук, що вказує на зони для покращення.

$$fb_t = M\left(p_{fb} \,||\, x \,||\, y_t\right). \quad (2)$$

Модель refine отримує комбінований вхід, що включає вхідні дані, попередню відповідь та отриманий відгук. Використовуючи інструкцію на вдосконалення, модель генерує наступну ітерацію відповіді:

$$y_{t+1} = M\left(p_{refine} \,||\, x \,||\, y_t \,||\, fb_t\right). \quad (3)$$

Процес повторюється ітеративно до моменту досягнення критерію зупинки.

Існує гіпотеза, що для створення агентів надлюдського рівня моделі потребують відповідного надлюдського відгуку для забезпечення якісних навчальних сигналів. Сучасні підходи, засновані на людських уподобаннях, стикаються з проблемою «вузького місця», зумовленого обмеженістю експертного рівня людини [2]. Крім того, постійна потреба у валідації даних людьми створює значні організаційні та фінансові бар'єри.

Еволюцією цього підходу стала парадигма «LLM як суддя» (рис. 2).

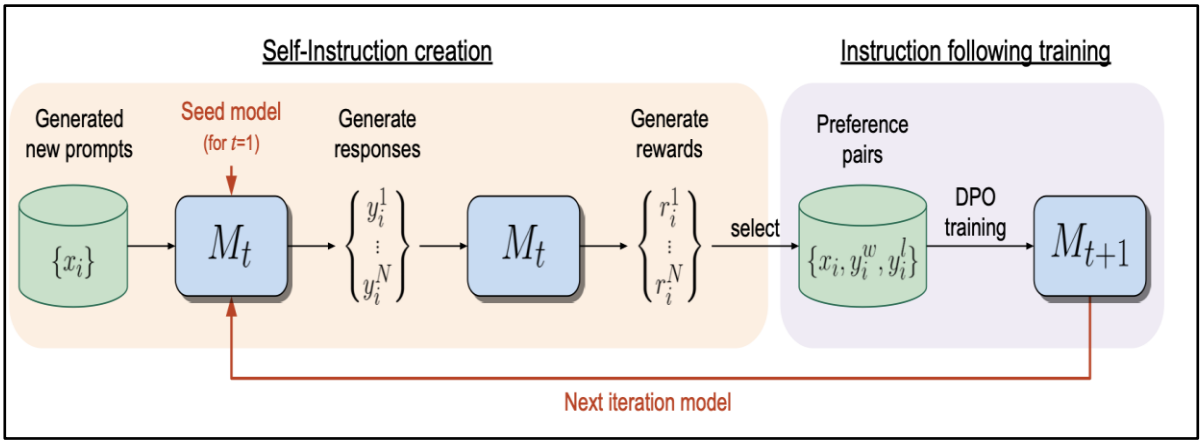

Рисунок 2 – Схема концепту LLM як суддя.

У межах цього концепту попередньо навчена за допомогою SFT модель, самостійно генерує множину варіантів відповідей та призначає їм оцінки. На основі цих оцінок формуються пари відповідь-оцінка, які використовуються для донавчання.

Використовуючи алгоритм DPO модель вчиться максимізувати ймовірність генерації відповідей з вищим рейтингом.

А хто оцінить суддю? Для подальшого підвищення автономності пропонується впровадження архітектури мета-судді (рис. 3).

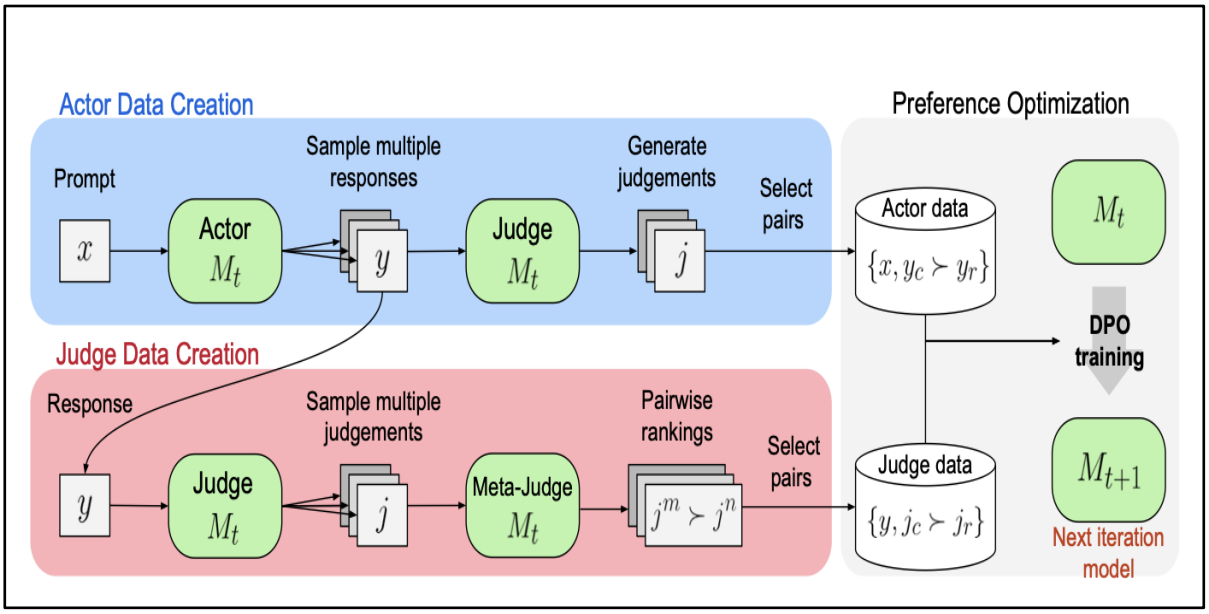

Рисунок 3 – Схема концепту мета-судді

У цій схемі LLM одночасно виконує три ролі в рамках однієї моделі:

– актор: генерує відповідь-кандидата;
– суддя: аналізує результат, виявляє зони для покращення;
– мета-суддя: аналізує якість та логіку оцінювання, наданого суддею.

Використання алгоритму DPO на рівні мета-судді дозволяє моделі максимізувати не лише якість контенту, а й точність власної аргументації та критеріїв оцінювання [3]. Це закладає основу для безперервного циклу самовдосконалення без участі зовнішнього вчителя.

Список використаних джерел:

1. Self-refine: Iterative refinement with self-feedback / A. Madaan, N. Tandon, P. Gupta [et al.] // Advances in neural information processing systems. – 2023. – Vol. 36. – P. 46534–46594.

2. Self-rewarding language models / W. Yuan, R. Y. Pang, K. Cho [et al.] // Forty-first International Conference on Machine Learning. – 2024.

3. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge / T. Wu, W. Yuan, O. Golovneva [et al.] // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. – 2025. – P. 11548–11565.