УДК 004.89

МЕТОД АВТОМАТИЧЕСКОЙ ОПТИМИЗАЦИИ ПРОМПТОВ НА ОСНОВЕ МЕТА-ПРОМТИНГА

Хайруллин А.Р.[1], Кулин Н.И.[1], Журавлёв В.Н.[1]
Научный руководитель – канд. техн. наук, доцент Муравьёв С.Б.[1]
[1]Университет ИТМО
anderbrine05@gmail.com

Работа выполнена в рамках темы НИР №625135 «Разработка программных модулей для эффективного обучения и применения модулей глубокого обучения».

Введение

Большие языковые модели (LLM) демонстрируют высокую эффективность в задачах NLP, однако их производительность зависит от качества входных инструкций — промптов [1]. Промпт-инжиниринг позволяет повысить точность генерации без изменения параметров модели. Традиционные техники (few-shot, chain-of-thought и др.) трудоемки и требуют ручной настройки, поскольку модели чувствительны к формулировкам [2].

Методы автопромптинга снижают зависимость от человека. Современные подходы включают эволюционные алгоритмы [3, 4], обучение с подкреплением [5], градиентные методы [6] и подходы на основе мета-промптинга [7]. Однако большинство методов имеют ограничения. Во-первых, многие требуют размеченного датасета, что ограничивает применение при дефиците данных. Во-вторых, алгоритмы вычислительно затратны и рискуют застрять в локальных оптимумах. В-третьих, переносимость промптов на новые задачи и модели часто низкая. В-четвёртых, чувствительность к инициализации вызывает высокую дисперсию результатов.

Проблема актуальна в том числе для мультиагентных систем, где разметка для отдельных агентов часто недоступна, а ручная настройка не масштабируется. Требуются гибкие методы, сочетающие скорость zero-shot генерации без размеченных данных и возможность итеративного улучшения при их наличии.

Основная часть

Предлагается подход к мета-оптимизации промптов, включающий базовый метод HyPE (Hypothetical Prompt Enhancer) и его итеративное развитие HyPER (HyPE with Refinement).

HyPE — одношаговая zero-shot техника мета-промптинга, не требующая размеченных данных. Модель побуждается сгенерировать гипотетический инструктивный промпт, который эффективнее исходного запроса решает ту же задачу.

Метод опирается на две ключевые гипотезы: по аналогии с методом HyDE [8], LLM способна преодолевать семантический разрыв путём синтеза гипотетического инструктивного промпта, воспроизводящего ключевые паттерны релевантности и приближающего к оптимальному решению задачи; модели, прошедшие SFT и RLHF, усваивают распределение высококачественных инструкций и способны самостоятельно генерировать такие инструкции при соответствующем побуждении без принудительного указания конкретных методик рассуждения.

Рабочий процесс: исходный запрос вместе с относящейся к задаче мета-информацией встраивается в мета-промпт с ролью эксперта и ограничениями формата; полученный промпт применяется как новая инструкция.

Эксперименты на бенчмарках GSM8K (Exact Match), TweetEval (F1), AG News

(F1), SQuAD v2 (BERTScore), XSum (BERTScore) и CommonGen (BERTScore) с моделями gpt-3.5-turbo и gpt-4o-mini показали превосходство над ручными промптами и техниками промптинга, а также конкурентные результаты с эволюционными методами. Метод также продемонстрировал хорошую переносимость оптимизированных промптов между разными моделями без дополнительной настройки.

HyPER — итеративное развитие метода HyPE для сценариев с валидационным датасетом. Метод позволяет последовательно улучшать качество генерируемых промптов путём адаптации мета-промпта на основе анализа ошибок, обеспечивая постепенное уточнение генеративной стратегии при сохранении интерпретируемости и совместимости с black-box моделями.

Рабочий процесс включает:

- генерацию популяции кандидатов с использованием мета-промпта HyPE;
- оценку каждого кандидата на валидационном наборе по выбранной метрике;
- отбор проблемных примеров и формирование рекомендаций по корректировке мета-промпта;
- интеграцию рекомендаций в мета-промпт;
- повторение цикла до сходимости (3–10 итераций).

Выводы

В работе предложено семейство методов мета-оптимизации промптов: одношаговый zero-shot метод HyPE и его итеративное развитие HyPER. HyPE демонстрирует заметное улучшение качества решения задач по сравнению с бейзлайном без использования размеченных данных. HyPER расширяет этот подход, позволяя последовательную адаптацию мета-промпта на основе анализа ошибок при наличии валидационного набора.

Метод HyPE успешно внедрён в проект по разработке мета-агентной системы для генерации мультиагентных систем в ИЦ «Сильный искусственный интеллект в промышленности» Университета ИТМО. HyPER проходит этап испытаний; планируется его интеграция в промышленные пайплайны и дальнейшее развитие в направлении многоцелевой оптимизации и синтетической генерации данных.

Литература

1. Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing //ACM computing surveys. – 2023. – Т. 55. – №. 9. – С. 1-35.
2. Schulhoff S. et al. The prompt report: A systematic survey of prompt engineering techniques //arXiv preprint arXiv:2406.06608. – 2024.
3. Guo Q. et al. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers //arXiv preprint arXiv:2309.08532. – 2023.
4. Fernando C. et al. Promptbreeder: Self-referential self-improvement via prompt evolution //arXiv preprint arXiv:2309.16797. – 2023.
5. Kwon M. et al. StablePrompt: Automatic prompt tuning using reinforcement learning for large language model //Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – 2024. – С. 9868-9884.
6. Shin T. et al. Autoprompt: Eliciting knowledge from language models with automatically generated prompts //Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). – 2020. – С. 4222-4235.
7. Yang C. et al. Large language models as optimizers //The Twelfth International Conference on Learning Representations. – 2023.
8. Gao L. et al. Precise zero-shot dense retrieval without relevance labels //Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2023. – С. 1762-1777.