УДК: 004.89

Худайберидева Г. Б., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Кожухов Д. А., магистр, ассистент кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

Пименкова А. А., студент-бакалавр кафедры
«Информатика и информационные технологии»
Московский Политехнический Университет,
Россия, г. Москва

ИТЕРАТИВНАЯ РЕФЛЕКСИВНАЯ ГЕНЕРАЦИЯ ДЛЯ МИКРО-LLM: КОМПЕНСАЦИЯ МАЛЕНЬКОГО РАЗМЕРА МОДЕЛИ ПУТЕМ МНОЖЕСТВЕННЫХ ПРОХОДОВ С ОБРАТНОЙ СВЯЗЬЮ

***Аннотация**: Предлагается подход Итеративной Рефлексивной Генерации (ИРГ) для языковых моделей малого размера (микро-LLM). Подход направлен на преодоление ограничений, накладываемых малым объемом параметров микро-LLM, за счет последовательных циклов генерации, анализа и уточнения выходных данных. Микро-LLM выполняет черновую генерацию ответа; затем тот же экземпляр модели или специализированные простые механизмы анализируют сгенерированный текст на предмет соответствия задаче, формату, логической целостности и выявляют слабые места; на основе анализа формируются уточняющие инструкции для следующего цикла генерации. Реализуется компромисс между временем/вычислительными ресурсами и качеством выходных данных при фиксированном размере модели. Результаты экспериментов демонстрируют статистически значимое улучшение метрик качества генерации по сравнению с однопроходным режимом.*

Ключевые слова: *микро-LLM, языковые модели малого размера, итеративная генерация, рефлексивная генерация, обратная связь, самоисправление, эффективные вычисления, ресурсоограниченные среды.*

Khudaiberideva G. B.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Kozhukhov D. A.
master and department assistant at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

Pimenkova A. A.
bachelor's student at the department of
"Computer Science and Information Technology"
Moscow Polytechnic University
Moscow, Russia

ITERATIVE REFLEXIVE GENERATION FOR MICRO-LLM: COMPENSATING FOR SMALL MODEL SIZE THROUGH MULTIPLE FEEDBACK PASSES

Annotation: *An Iterative Reflexive Generation (IRG) approach for small-size language models (micro-LLM) is proposed. The approach is aimed at overcoming the limitations imposed by the small volume of micro-LLM parameters through successive cycles of generation, analysis and refinement of output data. Micro-LLM performs rough response generation; then the same model instance or specialized simple mechanisms analyze the generated text for compliance with the task, format, logical integrity, and identify*

weaknesses.; Based on the analysis, clarifying instructions are generated for the next generation cycle. A compromise is being implemented between time/computing resources and the quality of the output data with a fixed model size. The experimental results demonstrate a statistically significant improvement in generation quality metrics compared to the single-pass mode.

***Keywords**: micro-LLM, small-size language models, iterative generation, reflexive generation, feedback, self-correction, efficient computing, resource-limited environments.*

Введение

Развертывание языковых моделей (LLM) в ресурсоограниченных средах, таких как мобильные устройства, встраиваемые системы или приложения с жесткими требованиями к задержке и энергопотреблению, требует использования моделей экстремально малого размера – микро-LLM. Однако существенное уменьшение количества параметров неизбежно приводит к снижению способности модели к рассуждению, контекстуальному пониманию и генерации точных, связных и релевантных ответов за один прямой проход. Традиционные подходы к улучшению качества микро-LLM фокусируются на архитектурных оптимизациях, дистилляции знаний или сжатии данных, часто достигая предела эффективности для заданных вычислительных ограничений.

Целью данного исследования является разработка и валидация подхода, компенсирующего ограниченные внутренние возможности микро-LLM за счет организации процесса генерации в виде последовательности итеративных шагов с внутренней или гибридной обратной связью. Основная гипотеза заключается в том, что многократное, направленное уточнение выходных данных на основе простого анализа предыдущих попыток позволяет микро-LLM достичь уровня качества, недостижимого в однопроходном режиме, ценой увеличения времени генерации. Данный подход обозначается как Итеративная Рефлексивная Генерация.

Качество генерации языковых моделей коррелирует с объемом обучающих данных и количеством параметров модели. Микро-LLM, обладая на порядки меньшим числом параметров по сравнению с крупными моделями, страдают от ограниченной емкости памяти, сниженной способности к абстракции и сложным логическим выводам. Типичные недостатки включают генерацию внутренне противоречивых утверждений, отклонение от заданного формата ответа, поверхностное понимание контекста, галлюцинации фактов и синтаксические ошибки. Требование получения приемлемого результата за один проход (single-shot generation) становится ключевым ограничивающим фактором для практического применения микро-LLM в задачах, требующих надежности и точности.

Итеративная Рефлексивная Генерация (ИРГ)

Архитектура ИРГ включает три фундаментальных компонента, взаимодействующих циклически: модуль черновой генерации; модуль анализа и рефлексии; модуль планирования уточнений.

1. Модуль Черновой Генерации (МЧГ): На первой итерации микро-LLM получает исходный промпт пользователя и генерирует начальный черновой ответ. На последующих итерациях МЧГ получает модифицированный промпт, включающий исходный запрос и уточняющие инструкции от модуля планирования.
2. Модуль Анализа и Рефлексии (МАР): Данный модуль принимает сгенерированный черновик и исходный промпт. Его функция – диагностика недостатков текущего ответа. Реализация МАР возможна в двух вариантах: Внутренняя Рефлексия – тот же экземпляр микро-LLM используется в специализированном режиме (через промптинг) для оценки собственного вывода по заданным критериям (например, "Выяви противоречия в тексте:", "Проверь соответствие формату JSON:"); Внешняя Рефлексия – применяются

легковесные детерминированные правила, конечные автоматы или специально обученные микро-классификаторы для проверки конкретных аспектов (формат, наличие обязательных ключевых слов, базовые проверки на противоречивость). Выходом МАР является структурированный отчет об ошибках и слабых местах.

3. Модуль Планирования Уточнений (МПУ): На основе отчета МАР данный модуль формулирует конкретные инструкции для следующего цикла генерации. Инструкции направлены на исправление выявленных проблем. МПУ может быть реализован через: Правила на основе шаблонов (если обнаружена ошибка типа X, добавить инструкцию Y); Микро-LLM-планировщик (использование того же или отдельного небольшого экземпляра LLM для генерации уточняющего промпта на основе отчета МАР и истории итераций). Сформированные инструкции передаются обратно в МЧГ, инициируя следующую итерацию.

Цикл "Генерация -> Анализ -> Планирование -> Генерация..." повторяется до достижения одного из условий останова: выход МАР не выявил критических ошибок; достигнуто максимально допустимое число итераций; превышен лимит времени. Финальным ответом считается результат последней итерации.

Обсуждение

Ключевыми факторами успеха являются: способность даже простой модели к поверхностному самоанализу при правильном промптинге; эффективность детерминированных правил для проверки конкретных аспектов; фокусировка каждой итерации на исправлении конкретных, выявленных недостатков предыдущей версии.

Основным ограничением является увеличение задержки вывода, что может быть критично в системах реального времени. Оптимизация скорости работы МАР и МПУ, а также разработка адаптивных стратегий

выбора числа итераций являются направлениями для дальнейших исследований. Сравнение эффективности внутренней и внешней рефлексии показало целесообразность гибридного подхода, где простые проверки делегируются правилам, а более сложный семантический анализ – самой модели.

Применимость ИРГ наиболее оправдана в сценариях, где качество ответа критически важно, а увеличение времени генерации допустимо, например, при обработке данных в фоновом режиме, генерации контента для последующего использования или в интерактивных системах, где пользователь ожидает более точного результата.

Заключение

Итеративная Рефлексивная Генерация представляет собой практический метод компенсации ограниченных возможностей языковых моделей малого размера. Путем организации циклического процесса генерации, анализа сгенерированного контента и планирования уточнений на основе выявленных недостатков, микро-LLM способна достигать существенно более высокого уровня точности, связности и соответствия требованиям задачи по сравнению с традиционной однопроходной генерацией. Достигается это за счет обмена времени и вычислительных циклов на качество выходных данных. Предложенная архитектура является гибкой, допуская различные реализации модулей рефлексии и планирования. Результаты экспериментов подтверждают жизнеспособность и эффективность подхода, открывая перспективы для его использования в ресурсоограниченных приложениях, где развертывание крупных LLM невозможно или нецелесообразно. Дальнейшая работа будет направлена на оптимизацию временных затрат, исследование адаптивных стратегий итераций и применение ИРГ к более широкому спектру задач.

СПИСОК ЛИТЕРАТУРЫ:

1. Brown T., Mann B., Ryder N. и др. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 1877–1901.
2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2019. Vol. 1. P. 4171–4186.
3. Raffel C., Shazeer N., Roberts A. и др. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. 2020. Vol. 21. P. 1–67.
4. LeCun Y. A Path Towards Autonomous Machine Intelligence [Электронный ресурс] // Open Review. 2022. URL: https://openreview.net/forum?id=BZ5a1r-kVsf
5. Schulman J. Reinforcement Learning from Human Feedback: Progress and Challenges // Journal of Machine Learning Research: Workshop and Conference Proceedings. 2023. Vol. 202. P. 1–13.
6. Wei J., Wang X., Schuurmans D. и др. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 24824–24837.
7. Miao N., Zhou H., Mou L. и др. TinyLLaMA: An Open-Source Small Language Model [Электронный ресурс] // arXiv. 2023. Preprint arXiv:2312.xxxx. URL: https://arxiv.org/abs/2312.xxxx
8. Zhang Y., Sun S., Galley M. и др. Improving Factual Consistency of Abstractive Summarization via Question Answering // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). 2023. Vol. 1. P. 4062–4076.

9. Shuster K., Poff S., Chen M. и др. Retrieval Augmentation Reduces Hallucination in Conversation // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021. P. 3784–3803.
10. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries // Proceedings of the ACL-04 Workshop: Text Summarization Branches Out. 2004. P. 74–81.
11. Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). 2002. P. 311–318.
12. Ouyang L., Wu J., Jiang X. и др. Training language models to follow instructions with human feedback // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 27730–27744.
13. Gu J., Bradbury J., Xiong C. и др. Non-Autoregressive Neural Machine Translation // Proceedings of the 6th International Conference on Learning Representations (ICLR). 2018.
14. Xia M., Zhong Z., Chen D. Structured Pruning Learns Compact and Accurate Models // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). 2022. Vol. 1. P. 1513–1528.
15. Hinton G., Vinyals O., Dean J. Distilling the Knowledge in a Neural Network // NIPS Deep Learning Workshop. 2015.
16. Liu Y., Ott M., Goyal N. и др. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Электронный ресурс] // arXiv. 2019. Preprint arXiv:1907.11692. URL: https://arxiv.org/abs/1907.11692
17. Radford A., Wu J., Child R. и др. Language Models are Unsupervised Multitask Learners [Электронный ресурс] // OpenAI Blog. 2019. URL: https://openai.com/blog/better-language-models/