Применение самоуточнения для повышения точности и объективности генерации ответов в системах вопросно-ответного типа

Трофимова Маргарита Сергеевна

Московский физико-технический институт
Физтех-школа прикладной математики и информатики

20 мая 2025 г.

Оглавление

1. Мотивация

2. Цель исследования

3. Формальная постановка задачи

4. Выделенные подзадачи

5. Результаты

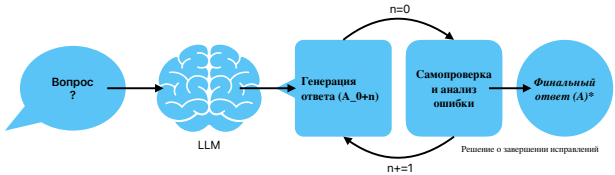
6. Заключение

7. Дальнейшие исследования

Мотивация

Почему важно решать проблему самообновления LLM?

- **LLM необратимо генерируют неточности и галлюцинации**, что снижает их надежность.
- **Отсутствие механизма самокоррекции** не позволяет моделям исправлять собственные ошибки.
- **Метод самоуточнения** дает моделям возможность анализировать, исправлять и улучшать свои ответы.
- **Метод особенно важен в задачах узкого домена** (медицина, юриспруденция, финансы), где высокая точность критична.



Цель исследования

Разработать и протестировать метод, основанный на принципе итеративного самоуточнения (Self-Refine), позволяющий языковым моделям автоматически исправлять ошибки и повышать качество генерации ответов.
Основные аспекты:

- Изучить, как модели могут **анализировать и исправлять свои ошибки**.
- Разработать **механизм самоуточнения**, снижающий зависимость от внешней обратной связи.
- Протестировать Self-Refine на задачах **QA, генерации текста и логических рассуждений**.

Формальная постановка задачи

- $\{M_i\}_i$ - множество языковых моделей
- $q = [x1, x2, ...]$ - запрос пользователя, x_i - токен
- $M_i(query) = A$, где A - ответ модели, $query$ - текстовый запрос
- $Agent_{k,role}(q) = M_k(role_prompt(q))$, агент - модель k с ролью $role$
- $role = (role_name, role_prompt)$ - роль
- $Agent_{k,role}(q) = A_0$, где A_0 - ответ модели
- В ответе A_0 могут быть ошибки
- $Agent_{k,role}(refine_prompt_{i,k,role}(A_i, q)) = A_{i+1}$ - итерация самоуточнения
- $\mu : A_i \to \{0, 1\}$ - критерий (метрика) остановки
- $Alg_{\{Agent\}_i}(q) = A_n$ - финальный ответ алгоритма, параметризованного множеством агентов

Формальная постановка задачи

Оптимизационная формулировка:

$$Alg^*_{\{Agent\}_i} = \arg\max_{Alg} F(A_n, q)$$

Где:

- A_n – исправленный ответ после n-й итерации.
- $F(A_n, q)$ – функция оценки качества ответа (**релевантность, полнота, логика**).

Выделенные подзадачи

1. **Исследование существующих методов самоуточнения на различных задачах**
   - Рассмотреть методы на различных датасетах
   - Исследовать параметры (количество итераций)
2. **Разработка своего метода, улучшающего классические методы**
   - Разработать и протестировать метод
   - Исследовантъ параметры метода
   - Исследовать модификации метода

Описание датасета TabularQA SemEval-2025

Для экспериментов был выбран DataBench Lite, представляющий широкий спектр задач по обработке табличных данных.

DataBench Lite – облегчённая версия DataBench, содержащая в себе таблицы размера 20 строк

Характеристики датасета:

- Содержит набор таблиц из различных доменов.
- Включает стандартизированные вопросы и аннотированные ответы
- Покрывает широкий диапазон сложности вопросов:
  - Простые булевые проверки
  - Вопросы, связанные с отдельными колонками
  - Запросы, требующие анализа нескольких строк и столбцов

Question	Answer	Type	Columns Used	Column Types
Is Lil Llama the oldest passenger?	false	boolean	Name, Age	category, number
What's the class of the oldest passenger?	first	category	Name, Age	category, number
What's the lowest fare paid?	10.2	number	Fare	number
Who are the passengers under 30?	[Lil Lama, Cody Lama]	list[category]	Name, Age	category, number
What are the fares paid by passengers under 30?	[30.25, 10.2]	list[number]	Age, Fare	number, category

Существующий Self-Refinement (DataBench Lite)

1. **Генерация исходного ответа (A_0)**
   - Модель формирует первоначальный ответ на основе данных таблицы.
2. **Самопроверка и анализ ошибки**
   - В модель подается A_0
   - Она возвращает анализ ответа на ошибки
3. **Исправление ответа (A_1)**
   - На вход подаются анализ ответа и ответ A_0
   - Если обнаружены расхождения, модель вносит исправления.

Мой метод самоуточнения (DataBench Lite)

1. **Самоуточнение списка колонок таблицы**
   - Модель возвращает VALID или исправленный список в формате JSON
   - При возвращении списка шаг (1) повторяется
2. **Генерация ответа A_0**
3. **Самопроверка и анализ ошибки**
   - Модели подается запрос и правила генерации ответа
   - Правила генерации влкючают в себя: необходимость в размышлении (CoT) и подсказках, правила форматирования
   - Модель размышляет (CoT) и генерирует VALID при правильном ответе или скорректированный ответ при необходимости исправления
   - Скорректированный ответ модель возвращает в формате JSON
   - Если необходимо исправление, то повторяем шаг (3)

Эвалюация: обучение в контексте (ICL)

Метрика оценки:

- **Accuracy (точность)** – доля полностью корректных ответов, совпадающих с эталоном.

Результаты сравнения Self-Refinement с базовой моделью (CAG):

Модель	базовый(CAG)	+Self-Refine	+Самоуточнение (мое)
Llama-3.1-8B	37.93%	41.95%	47.13%
Llama-3.3-70B	64.37%	64.75%	69.36%

Таблица: Сравнение точности

Самоуточнение для кодового решения (DataBench Lite)

1. **Генерация кода A_0 с помощью LLM**
   - Модель генерирует SQL или Pandas код
2. **Запуск кода на таблице**
   - Возвращает данные или ошибку исполнения кода
3. **Самоуточнение**
   - Код, данные или ошибка подается модели на анализ
   - Модель анализирует и генерирует новый код

Эвалюация: кодовое решение

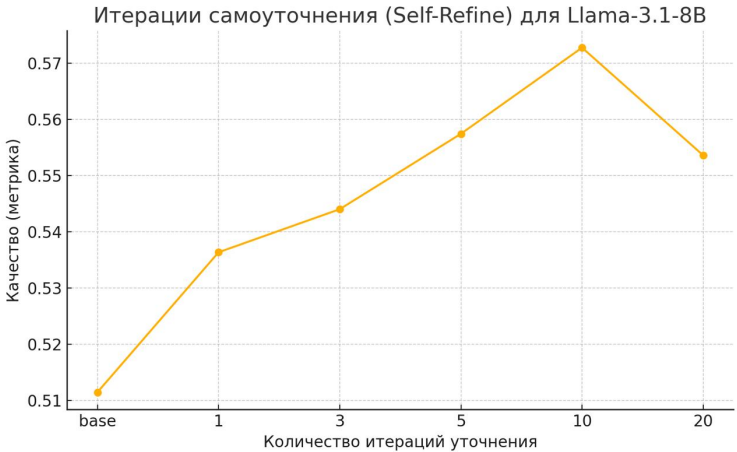


Рис.: Итерации самоуточнения для кодового решения

Расширения и Модификации

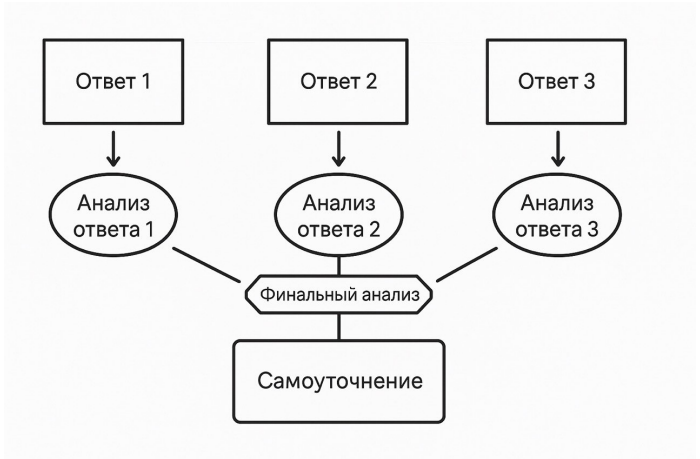


Рис.: Применение параллельного уточнения

Эвалюация: сравнение с self-consistency

Метрика оценки:

- **Accuracy (точность)** – доля полностью корректных ответов, совпадающих с эталоном.

Модель	Метод	Точность
Llama-3.1-8B	CoT	54.59
	CoT + self-cons.(5)	63.40
	CoT + паралл. уточн. (5)	61.15

Таблица: Сравнение с self-consistency

Основные результаты

- **Применение Self-Refinement продемонстрировало улучшение точности**
  Улучшение с 37.93% до 41.95% на Llama-3.1-8B подтверждает эффективность итеративного исправления ошибок, но для крупной модели значительного прироста не наблюдалось
- **Моя модфикация метода самоуточнения привела к улучшению качества по сравнению с классическим метод Self-Refinement**
  Был получен прироста как на малой, так и на большой модели
- **Были исследованы гиперпараметры Self-Refinement**
  Была рассмотрена зависимость качества от количества итераций
- **Были опробованы расширения и модификации**

Дальнейшие исследования

Основные направления будущей работы:

- **Разработка подходов самоуточнения с использованием многоагентных систем**
- **Разработка методов автоматического определения качества исправлений**, что позволит избежать избыточных итераций
- **Применение самоуточнения в других доменах данных**

Спасибо за внимание