Разработка адаптивного RAG подхода для создания эффективной вопросно-ответной системы

Антропова Е.Г.
antropova.eg@phystech.edu

Центр когнитивного моделирования МФТИ

2024 – Научно-исследовательский семинар

Введение

Мотивация

Вопросно-ответные системы (QA-системы) на основе больших языковых моделей (LLMs, Large Language Models) в последние годы привлекли значительное внимание благодаря их способности решать широкий спектр задач в области обработки естественного языка. Однако они сталкиваются с рядом проблем, включая высокие вычислительные затраты на поиск и генерацию ответов, а также недостаточную точность.

Цель работы

Разработка адаптивного подхода Retrieval-Augmented Generation (RAG), направленного на оптимизацию вычислительных затрат при обработке запросов. Это позволит создать эффективную вопросно-ответную систему.

Формальная постановка задачи

Входные данные:
q: текстовый запрос пользователя.
Цель:
Разработать метод *Method*, который выполняет выбор оптимальной стратегии обработки запроса q:

1. Определяет необходимость использования внешних данных с помощью ретривера (retriever);
2. Предлагает способ сужения области поиска во внешних данных;
3. Устанавливает, возможно ли обработать запрос на основе внутренних знаний LLM.

Результат работы метода:

$$decision = Method(q)$$

где *decision* — принятое решение для обработки запроса q.

Метрики оценки качества всей системы подхода

- $$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$
- $$EM = \begin{cases} 1, & \text{если } Predicted = GroundTruth \\ 0, & \text{иначе} \end{cases}.$$
- $$Accuracy = \begin{cases} 1, & \text{если } GroundTruth \in Predicted \\ 0, & \text{иначе} \end{cases}.$$
- Число шагов извлечения и генерации:

$$Steps_{total} = Steps_{retrieval} + Steps_{generation}.$$

- Среднее время ответа на запрос:

$$Time_{avg} = \frac{\sum_{i=1}^{N} Time_i}{N}.$$

Метрики оценки качества всей системы подхода

Метрики, которые вычисляются с помощью ответов LLM:

- Корректность (Correctness) – насколько сгенерированный ответ соответствует эталонному ответу (ground truth);
- Полнота (Completeness) – учитывает, содержит ли ответ все необходимые аспекты, которые затрагиваются в эталонном ответе;
- Релевантность (Relevance) – проверяется, насколько сгенерированный ответ соответствует сути заданного вопроса.
- Общий показатель (Overall)

$$\text{Overall} = \frac{\text{Correctness} + \text{Completeness} + \text{Relevance}}{3}$$

Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity

Классификатор присваивает запросу категорию:
"A"– только использование LLM;
"B" – один проход ретривера;
"C" – используется многоэтапный подход с ретривером.

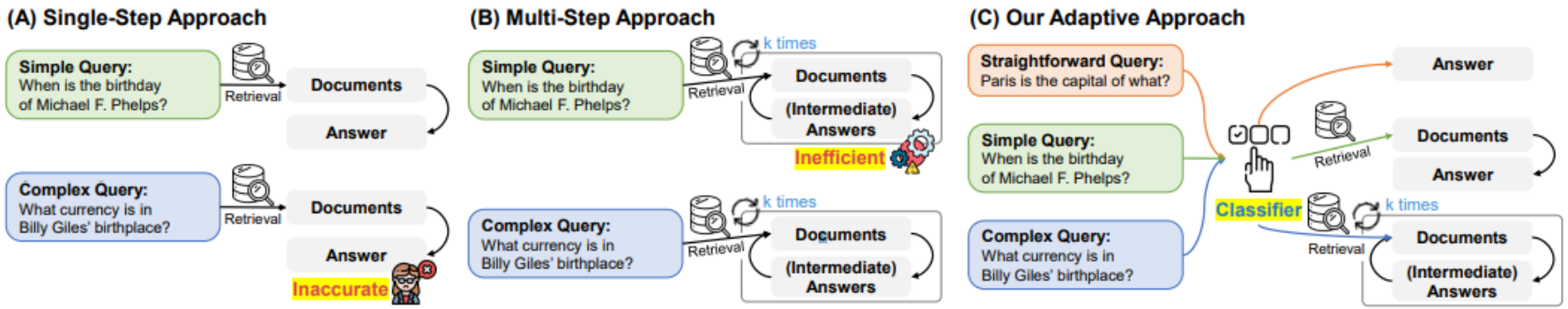

Рис.: концептуальная схема работы Adaptive-RAG

Soyeong Jeong et al. (2024). «Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity»

MBA-RAG: a Bandit Approach for Adaptive Retrieval-Augmented Generation through Question Complexity

Вместо классификатора используется алгоритм RL "Многорукий бандит".

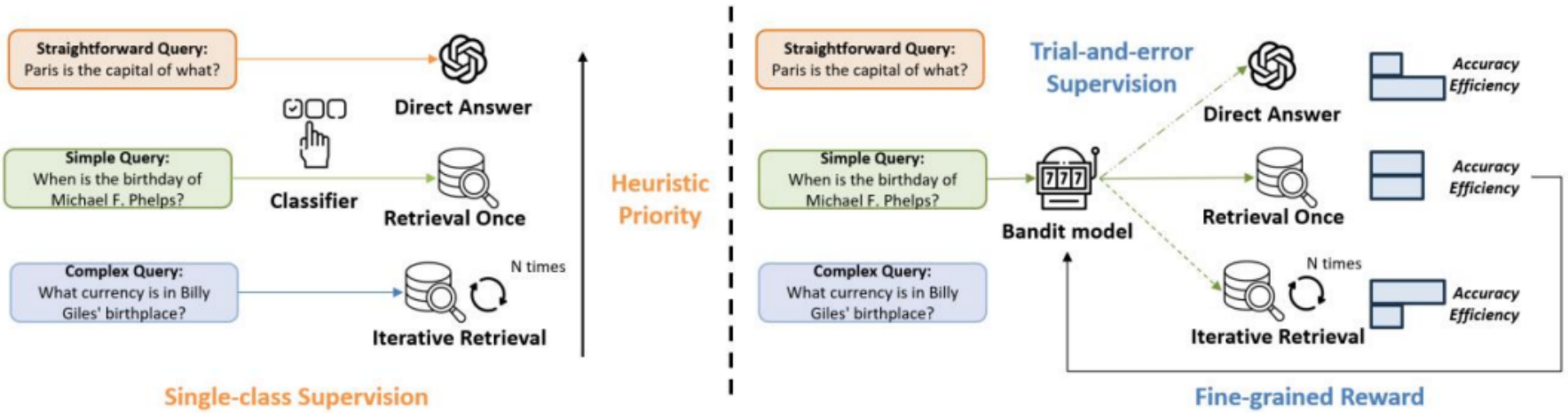

Рис.: Сравнение методов Adaptive-RAG и MBA-RAG

2

Xiaqiang Tang и др. (2024). *MBA-RAG: a Bandit Approach for Adaptive Retrieval-Augmented Generation through Question Complexity*. arXiv: 2412.01572 [cs.AI]. URL: https://arxiv.org/abs/2412.01572.

AT-RAG: An Adaptive RAG Model Enhancing Query Efficiency with Topic Filtering and Iterative Reasoning

Улучшает процесс многократного извлечения и генерации ответов (Multi-Step RAG) с помощью фильтрации тем для повышения эффективности извлечения релевантной информации.
Если начальный набор документов оказывается недостаточным (оценивает модуль Answer Grader), система может улучшить запрос и выполнить процесс извлечения снова, используя уточнённый запрос и накопленный контекст рассуждений.
Для реализации модели назначения темы используется BERTopic.[3]

Mohammad Reza Rezaei и др. (2024). «AT-RAG: An Adaptive RAG Model Enhancing Query Efficiency with Topic Filtering and Iterative Reasoning». В: *CoRR* abs/2410.12886. DOI: 10.48550/ARXIV.2410.12886. arXiv: 2410.12886. URL: https://doi.org/10.48550/arXiv.2410.12886.

RAG Approach Enhanced by Category Classification with BERT

Используется модель на основе BERT для классификации запросов по типу, тематике, и динамичности данных. Для сложных категорий система возвращает ответ "IDK" ("I Don't Know"), если не уверена в результате.

Три различных направления для классификации запросов:

- Domain (домен): определение тематической области текста;
- Question Type (тип вопроса): классификация типа вопроса (фактологический, причинно-следственный и т.д.);
- Dynamism (динамичность): оценка того, насколько часто контент изменяется (статические факты или динамические данные).

4

Yuki Taya и др. (2024). «RAG Approach Enhanced by Category Classification with BERT». В: *2024 KDD Cup Workshop for Retrieval Augmented Generation*. URL: https://openreview.net/forum?id=yzSkvdTYfm.

Self-DC: When to Retrieve and When to Generate? Self Divide-and-Conquer for Compositional Unknown Questions

Основная идея заключается в адаптивной обработке вопросов, где динамически выбирается стратегия работы в зависимости от уверенности модели. Диапазон уверенности делится на три области, выбор стратегии в зависимости от уверенности:

- $[0, \alpha - \beta]$: Низкая уверенность. Используется стратегия retrieve-then-read (извлечение внешних данных перед генерацией ответа).
- $(\alpha - \beta, \alpha + \beta)$: Средняя уверенность. Если вопрос вызывает неопределённость, он делится на несколько подзадач.
- $[\alpha + \beta, 1]$: Высокая уверенность. Применяется стратегия generate-then-read (генерация ответа без привлечения внешних данных).

Параметры α и β задаются заранее и определяют пороги уверенности. [5]

[5] Hongru Wang et al. (2024). «Self-DC: When to retrieve and When to generate? Self Divide-and-Conquer for Compositional Unknown Questions».

Схема взаимосвязи рассмотренных подходов

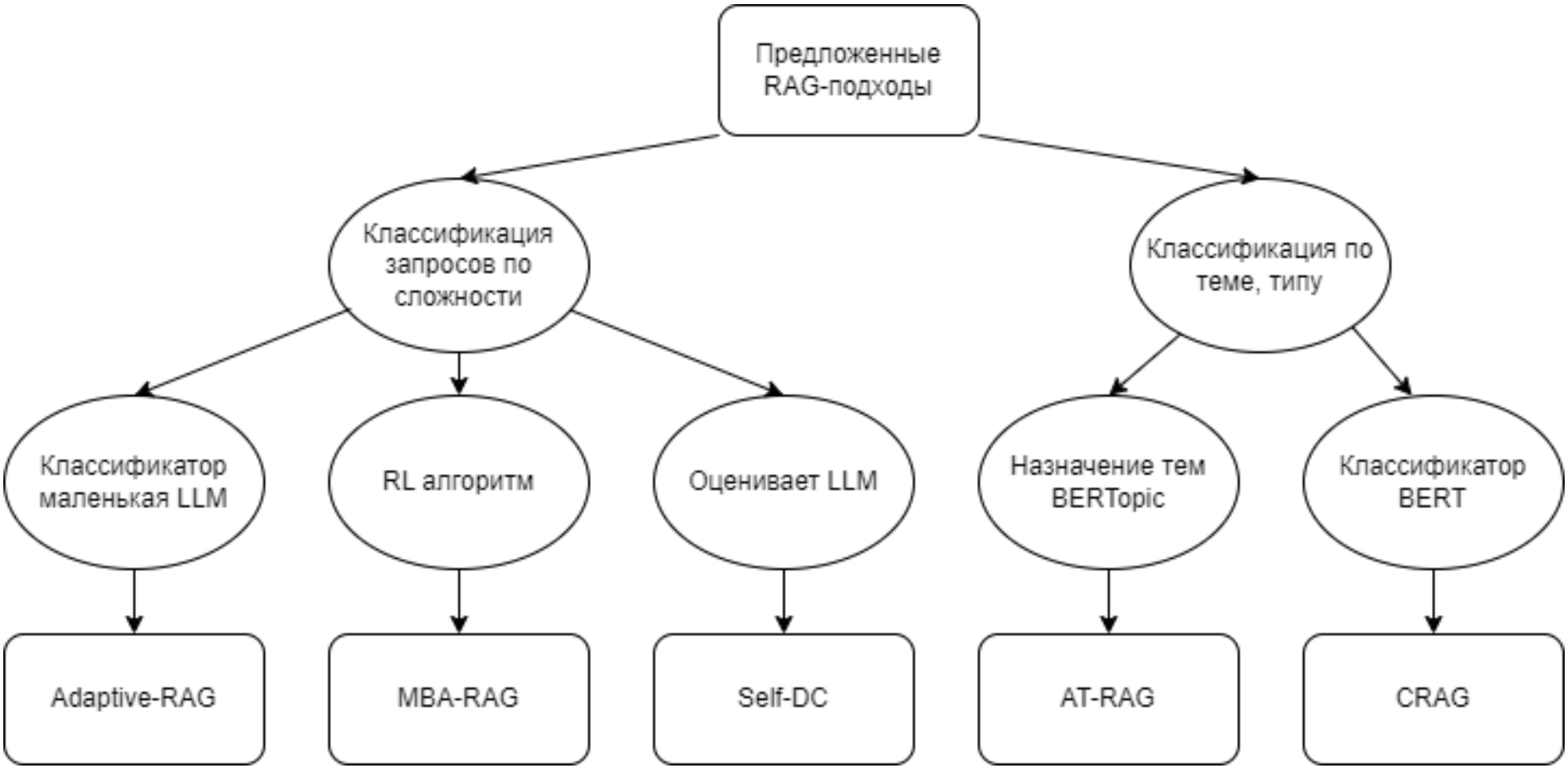

Рис.: Схема всех подходов

Спасибо за внимание!

antropova.eg@phystech.edu