RAG
(Retrieval Augmented Generation)

Подготовили: Наумович Л., Цителадзе Г., Липинский Л.

План

1) Что это и зачем нужно?
2) Схема RAG
3) Векторизация
4) Поиск в RAG
5) Демонстрация
6) Возможные улучшения и итоги

Вход: произвольный вопрос
Выход: ответ на вопрос

какой курс доллара к рублю сегодня?

К сожалению, у меня нет доступа к актуальным данным в реальном времени, включая текущий курс доллара к рублю. Для получения свежей информации рекомендуется посетить финансовые сайты, такие как:
ЦБРФ
Investing.com
Yahoo Finance
или использовать приложение банка или финансового сервиса.

Проблемы

Требуется ответ строго по релевантным данным, но:

1) LLM не знает актуальных данных и отказывается отвечать

2) Модель придумывает факты

Возможное решение

SFT (Дообучение модели)

Проблема:
Очень дорого, времязатратно, возможно “катастрофическое забывание”, невозможность решения задачи в случае быстроменяющихся данных

Вход: произвольный вопрос, **релевантная внешняя информация**
Выход: ответ на вопрос **на основе внешней информации**

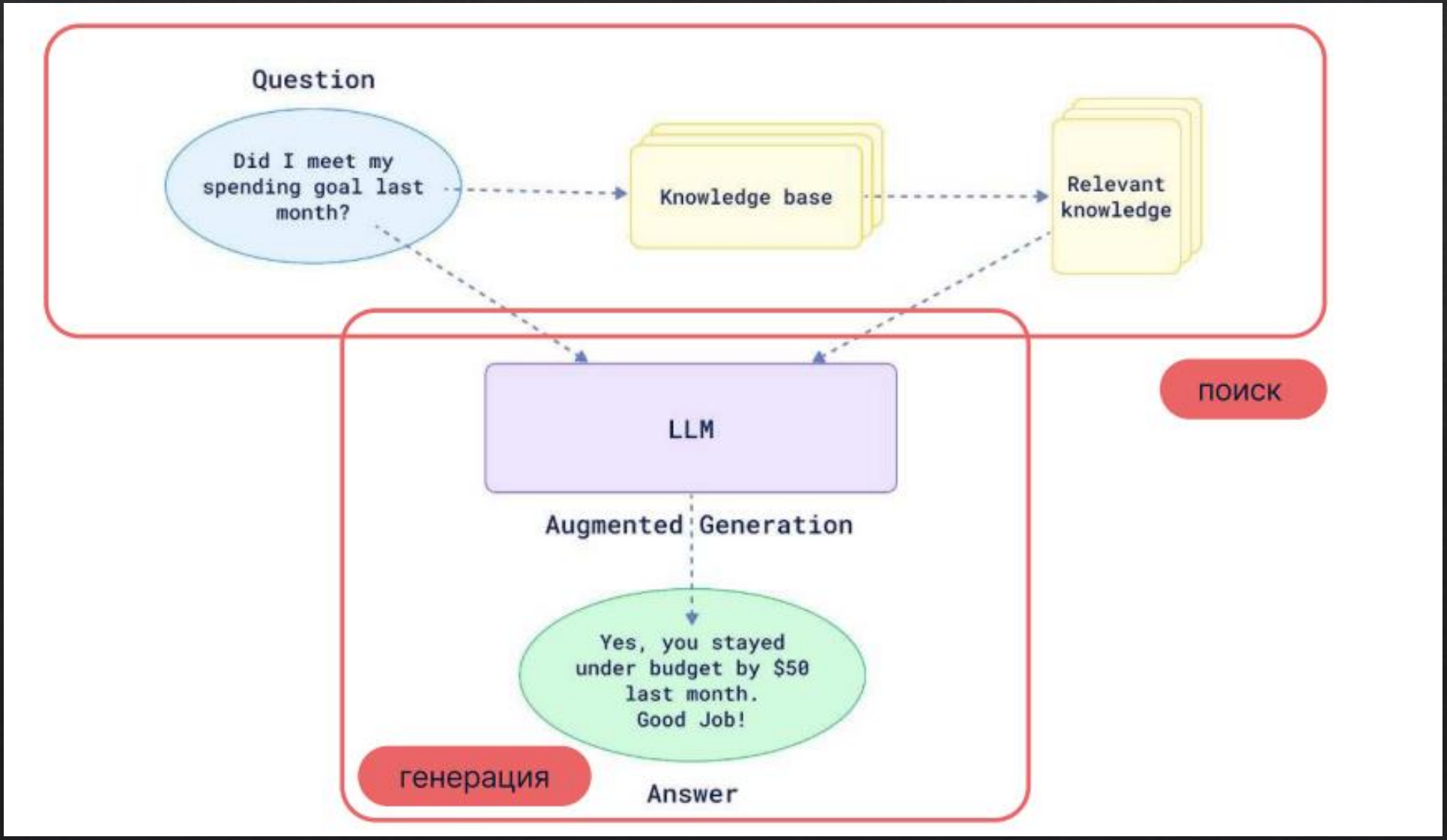
какой курс доллара к рублю сегодня?
Результат поиска: Центральный банк Российской федерации установил на сегодня следующие курсы иностранных валют к рублю Российской Федерации без обязательств Банка России покупать или продавать указанные валюты по данному курсу код Букв. Единиц код CB код Курс код Руб CNY, 1¥ 11,3561₽ USD, 1$ 81,2257₽ EUR, 1€ 93,8365₽

На сегодня курс доллара США (USD) к российскому рублю (₽) составляет **81,2257 ₽** за 1 доллар.

Что такое RAG?

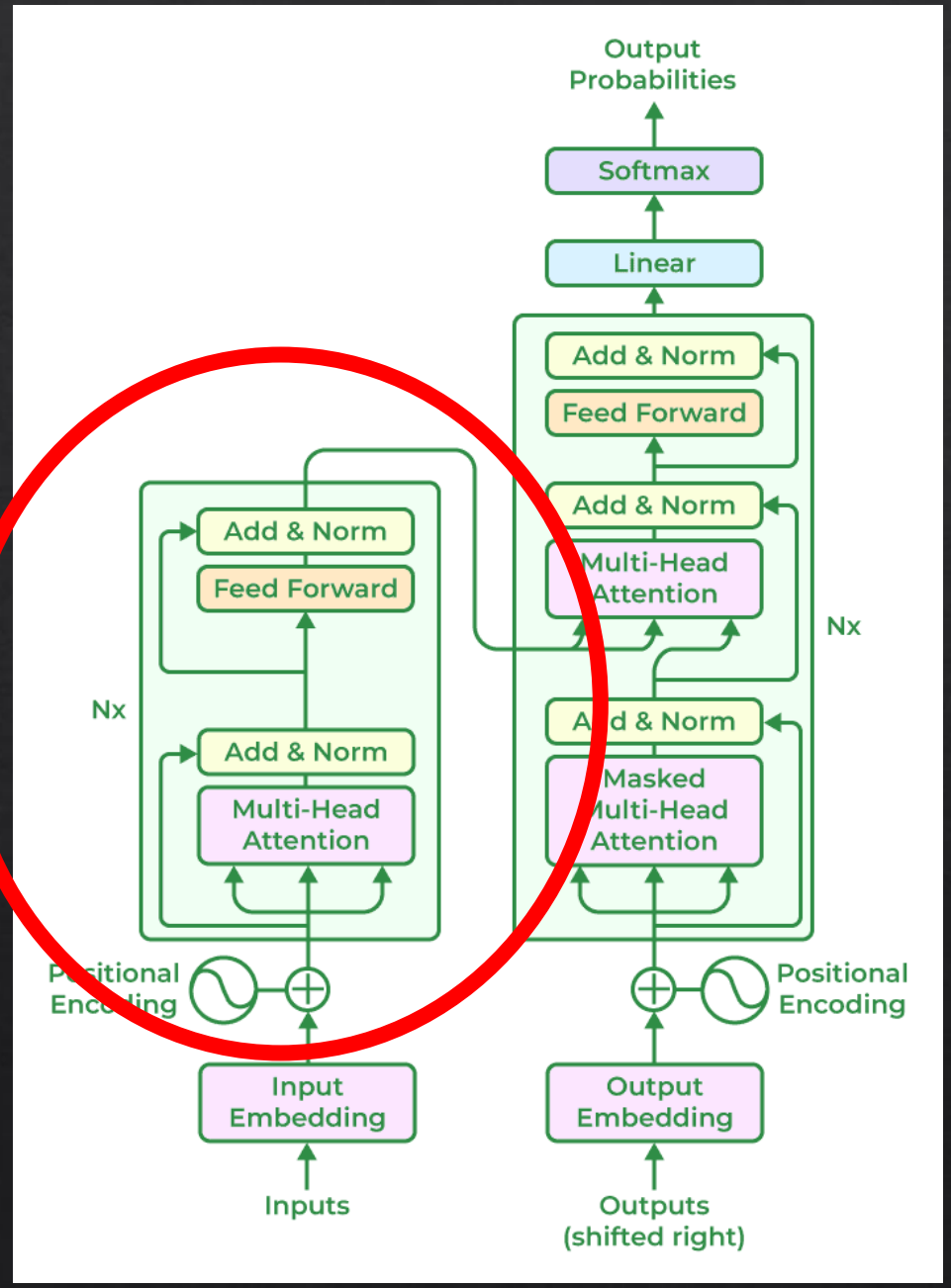
Это подход, который позволяет LLM использовать внешние данные в реальном времени. Модель не фантазирует, а подтягивает факты из базы знаний и генерирует ответ уже на основе найденной информации.

Типовая схема RAG:



Векторизация данных

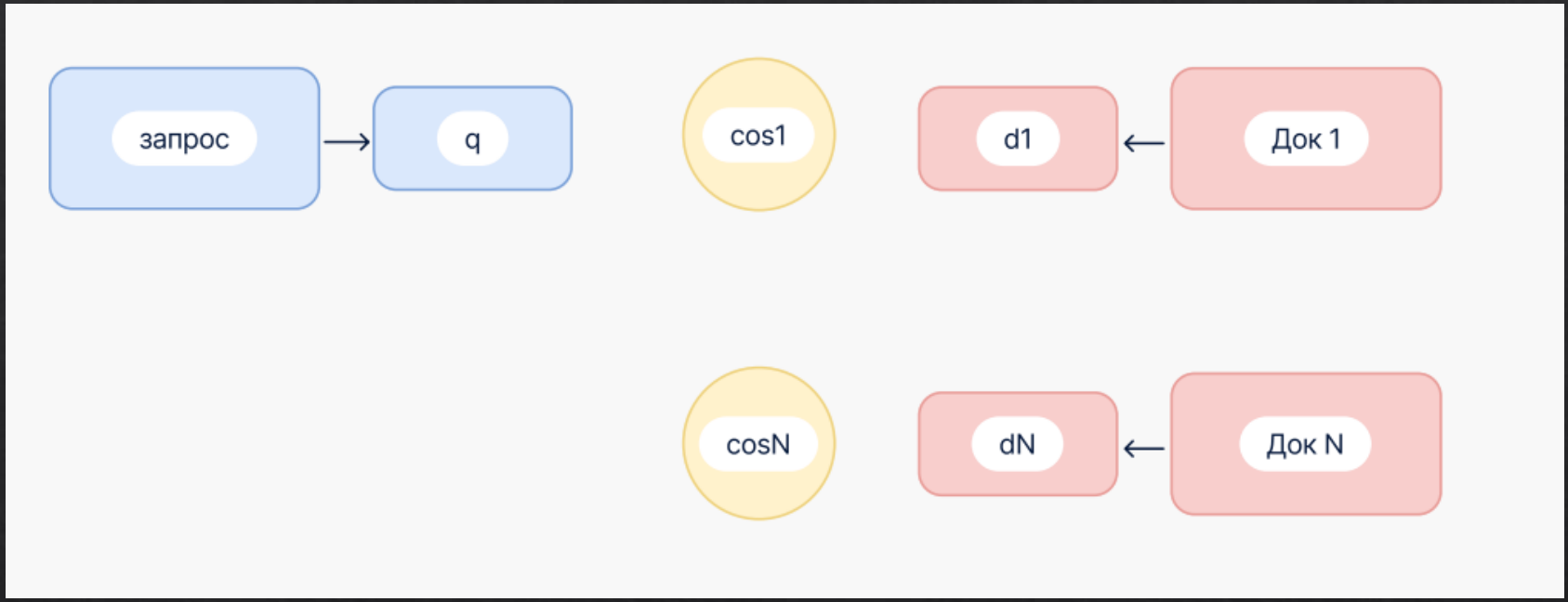
Embedding модель
(BERT/Sentence Encoder,
TF-IDF, BM25)



Поиск в RAG

В RAG используется векторное представление фрагментов документов из базы знаний

$q \epsilon \mathbb{R}^n$



$d_i \epsilon \mathbb{R}^n$

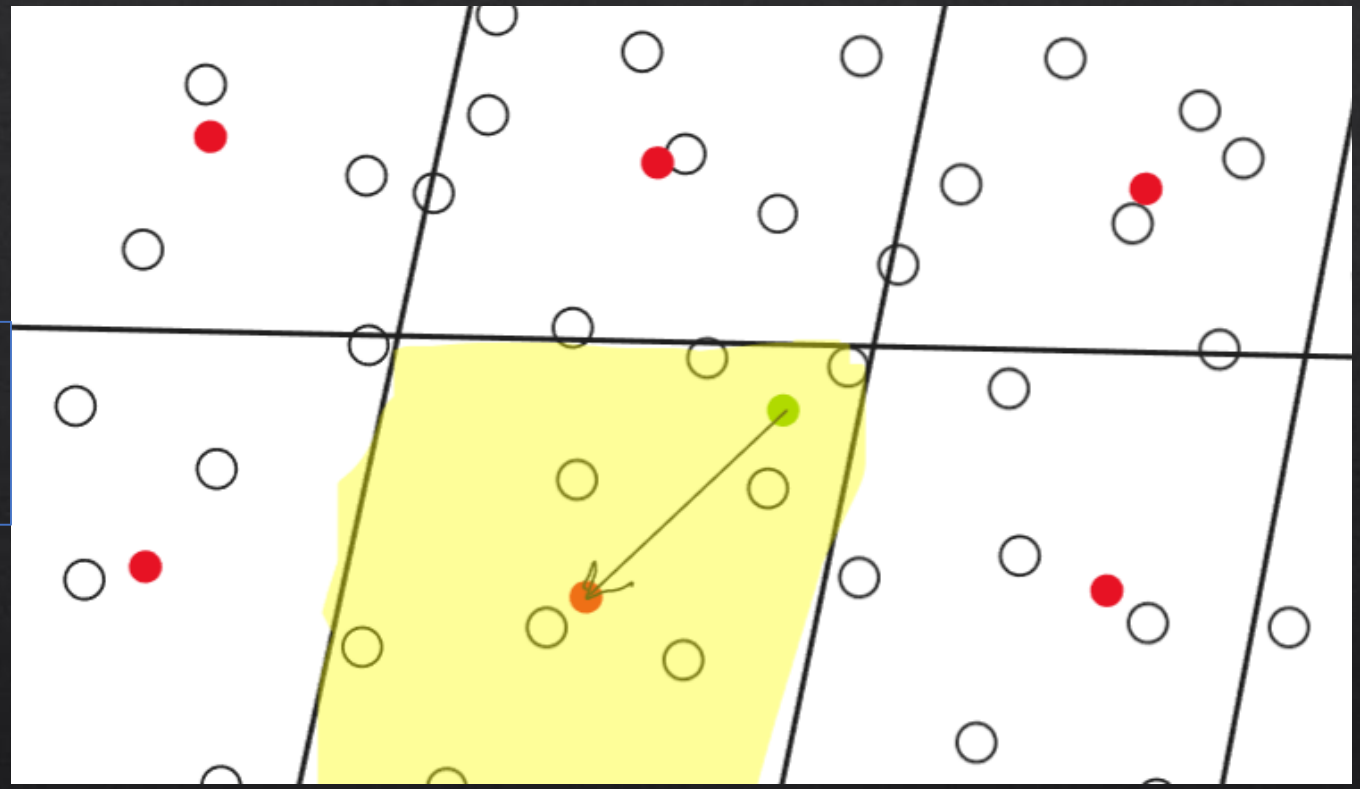
$$sim(q, d_i) = \frac{q \cdot d_i}{\| q \| \| d_i \|}$$

$$(q, d_i) = \sqrt{\| q - d_i \|}$$

Поиск в векторных базах данных

$$\underset{d_i \in \mathcal{C}(q)}{\operatorname{argmin}}[1 - sim(q, d_i)]$$

$$\underset{d_i \in \mathcal{C}(q)}{\operatorname{argmin}}[\sqrt{\| q - d_i \|}]$$



Кластеризация

Демонстрация

Методы улучшения

1) Reranking
2) Регенерация запросов
3) Гибридные подходы

Итог

RAG даёт:

Точные ответы основанные на релевантной информации

Экономия. Не требует дообучение модели

Проверяемость ответов

Простая актуализация базы знаний