Вопрос-ответные системы в облаке Yandex Cloud на основе продвинутого RAG и многоагентности

Яндекс

Дмитрий Сошников

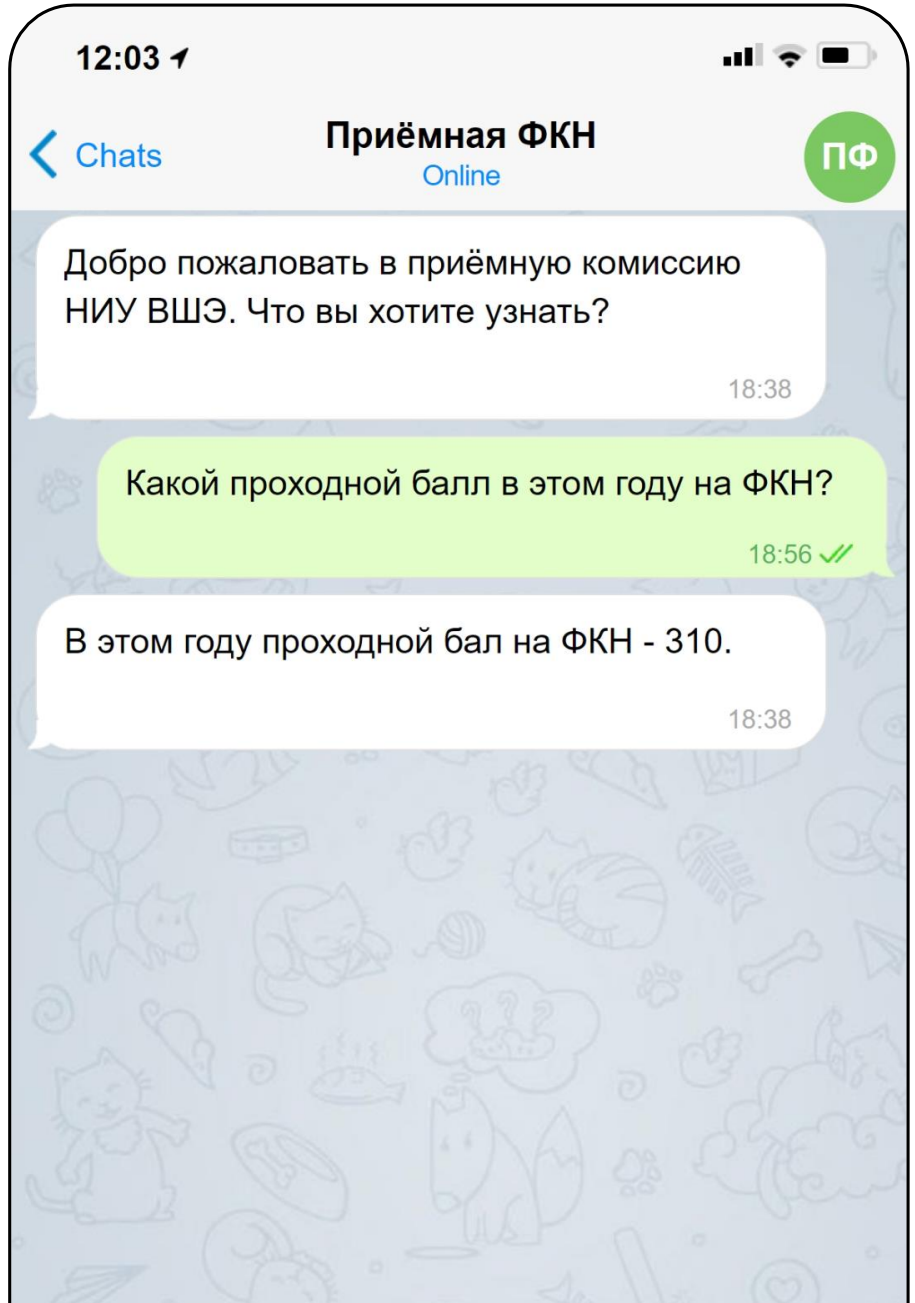
Доцент НИУ ВШЭ/МАИ,
консультант Yandex Cloud,
тех. руководитель AI Lab ШД НИУ ВШЭ

Tg: @shwarsico

ХАЛЯВА

Типовая задача

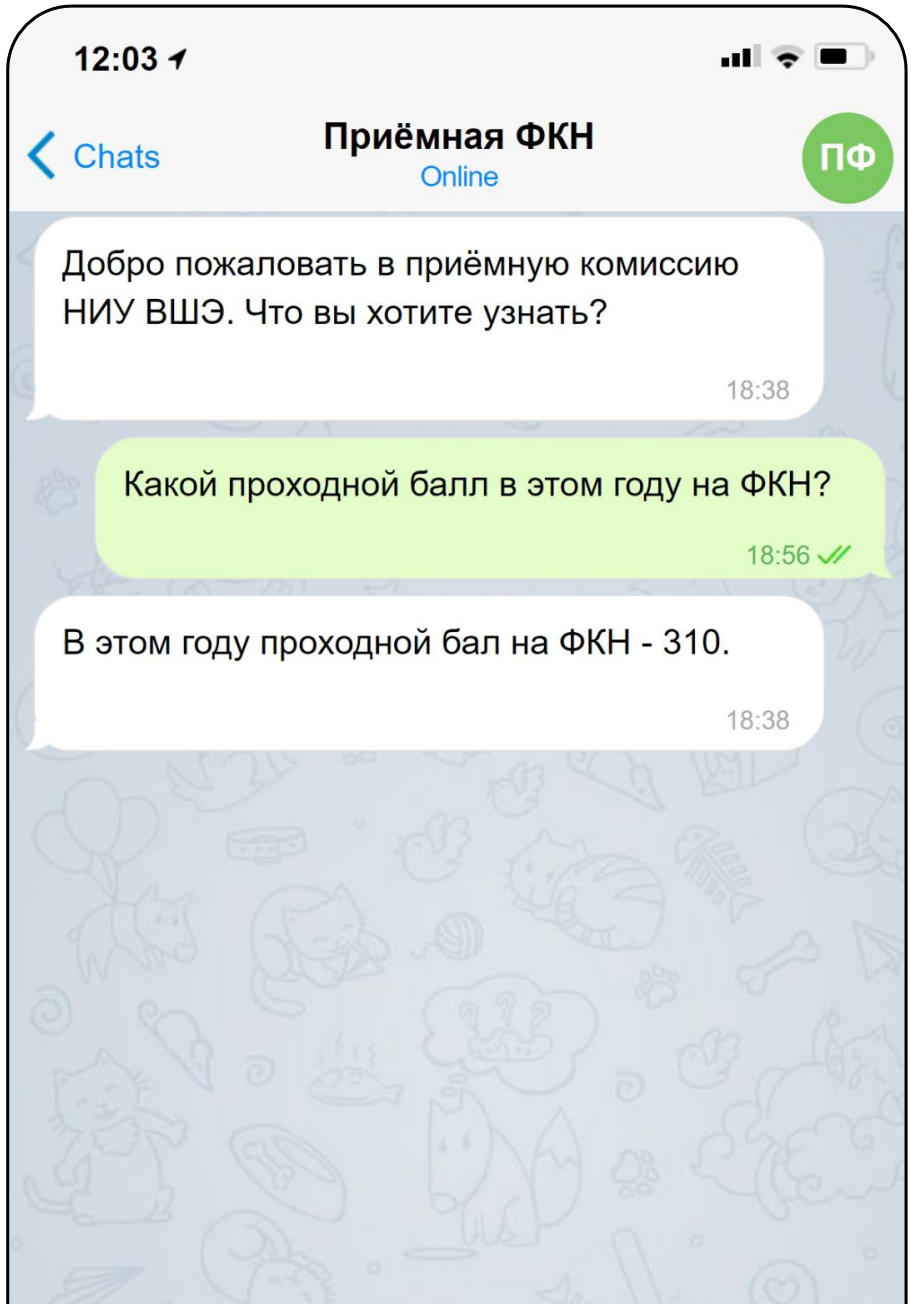
Специализированный чат-бот, который обладает предметными знаниями



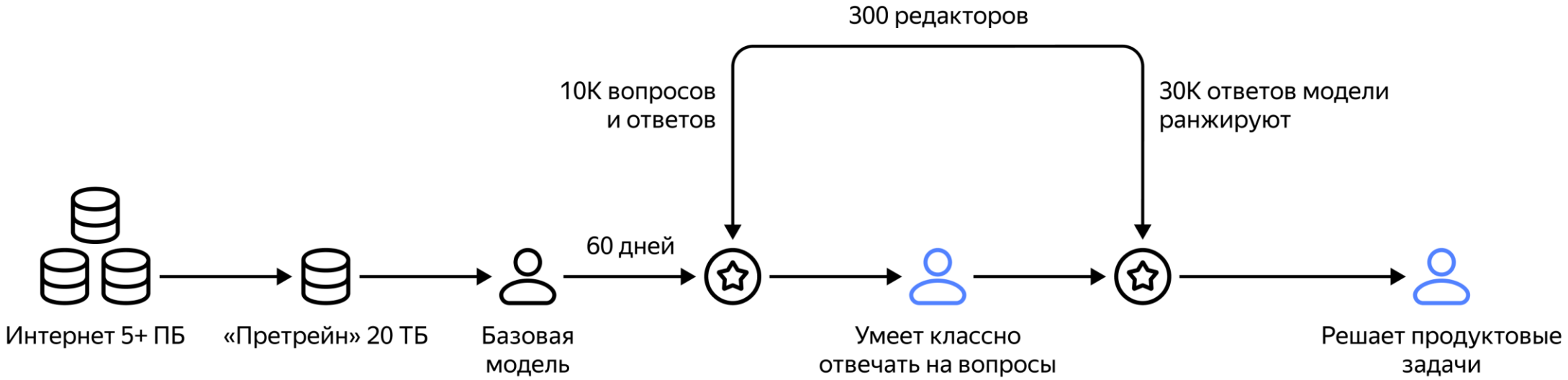
Типовая задача

Специализированный чат-бот, который обладает предметными знаниями

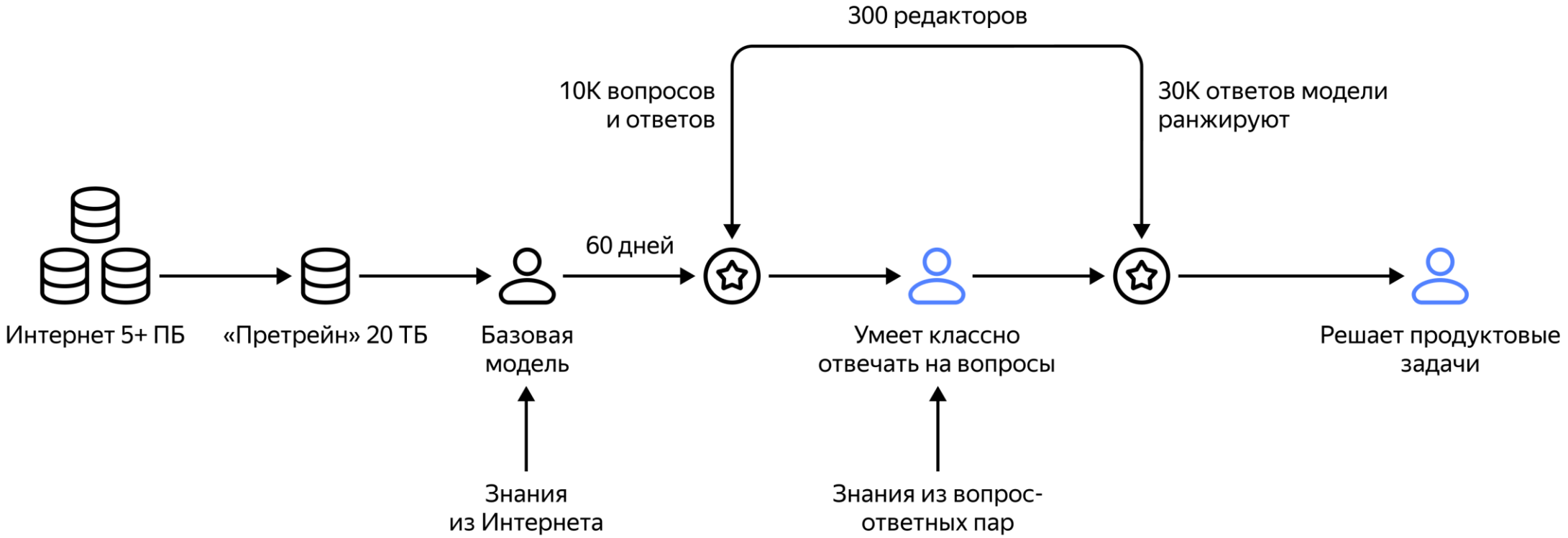
Употребление алкоголя вредит вашему здоровью...



Откуда берутся знания у LLM



Откуда берутся знания у LLM



Как добавить знания в LLM?

Дообучение модели

Full Fine-Tuning

- Много обучающих примеров
- Много ресурсов на обучение
- Изменяются веса исходной модели
- Получается новая языковая модель с большим числом параметров

LoRA Fine-Tuning

- Вместо обучения всех параметров обучается «добавка» к весам, которая раскладывается на матрицы меньшей размерности
- Из-за меньшего количества параметров нужно меньше данных и меньше вычислительных ресурсов

P-Tuning

- Параметры модели не меняются, но подбирается continuous prompt embedding, который помогает достичь нужных результатов
- Требует меньше всего вычислительных ресурсов и данных для обучения
- Хорошо работает на изменение формата ответа, но не знаний модели

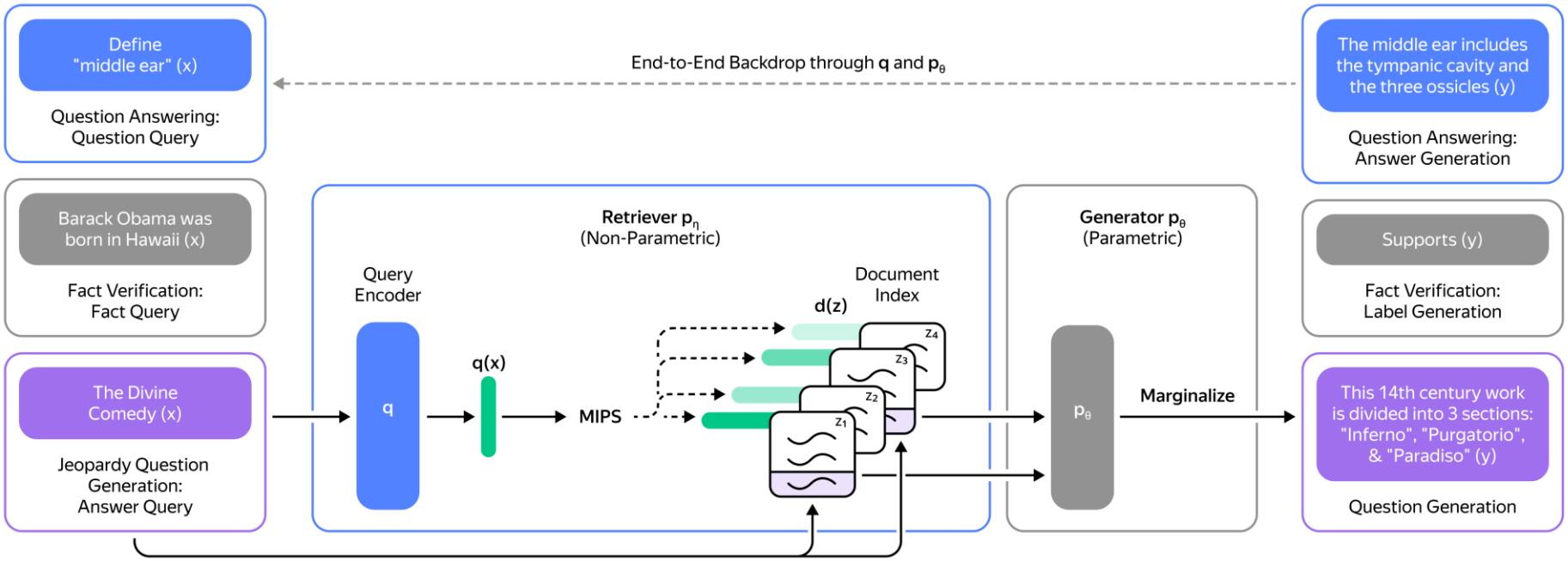
Как добавить знания в LLM: RAG

Retrieval-Augmented Generation

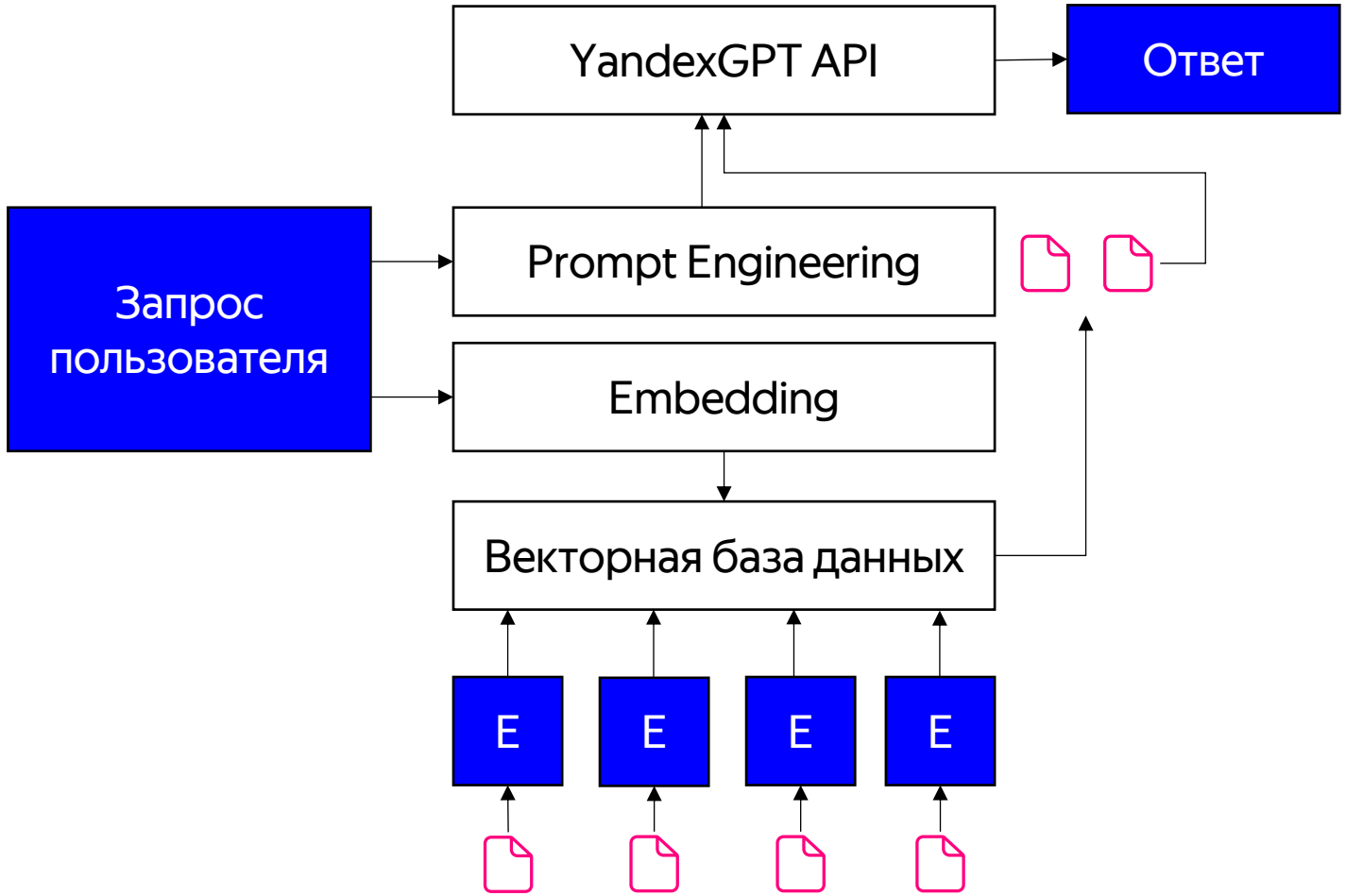
Lewis, Patrick, et al. «Retrieval-augmented generation for knowledge-intensive nlp tasks.» Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

RAG-Sequence: для генерации всей последовательности используются одни и те же документы

RAG-Token: для генерации каждого токена могут использоваться разные документы



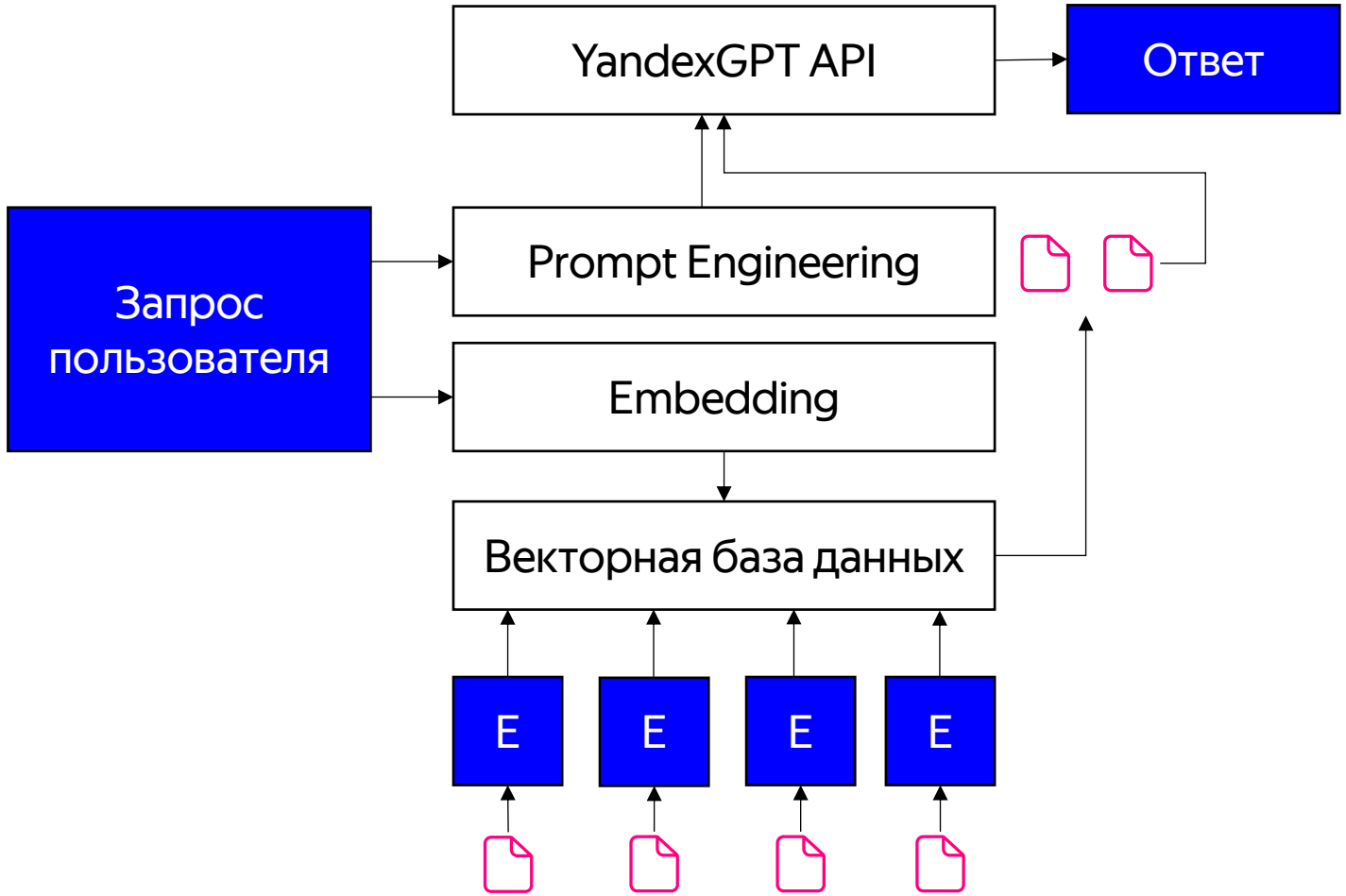
RAG на практике (Naïve RAG)



RAG на практике (Naïve RAG)

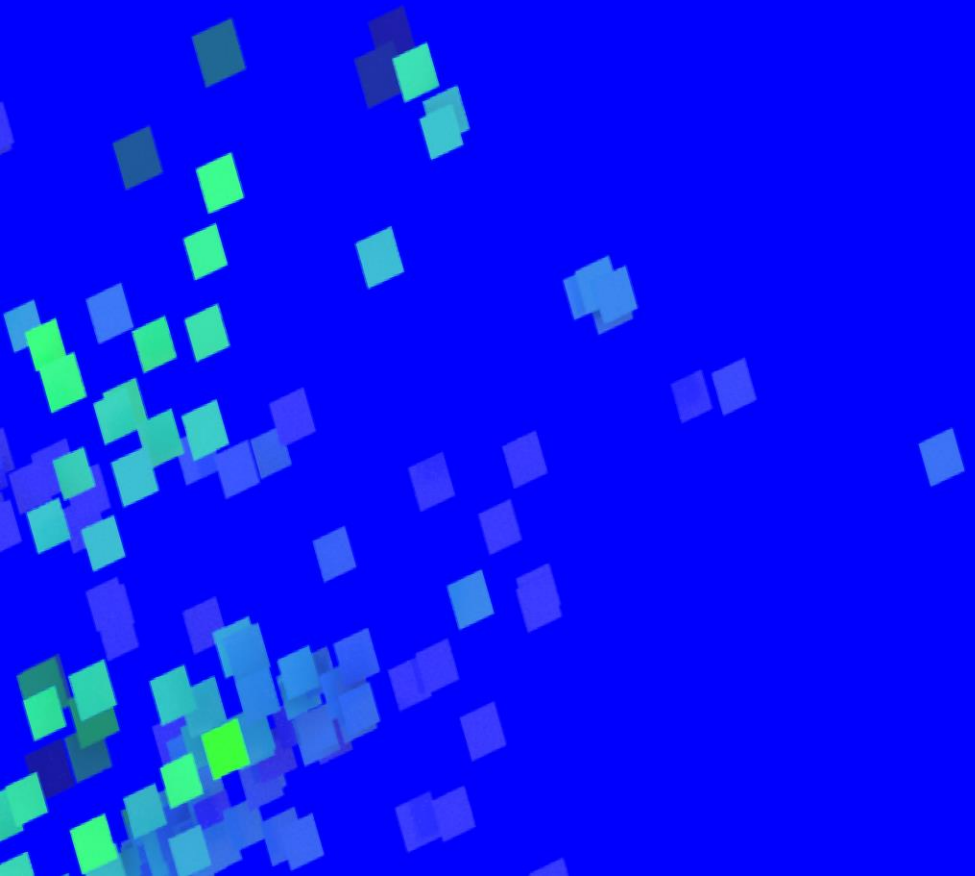
Гиперпараметры:

- Размер фрагмента текста
- Размер overlap
- Промпт
- Алгоритм поиска фрагментов
- Количество фрагментов в запросе
- Используемые эмбеддинги



Демонстрация

Яндекс

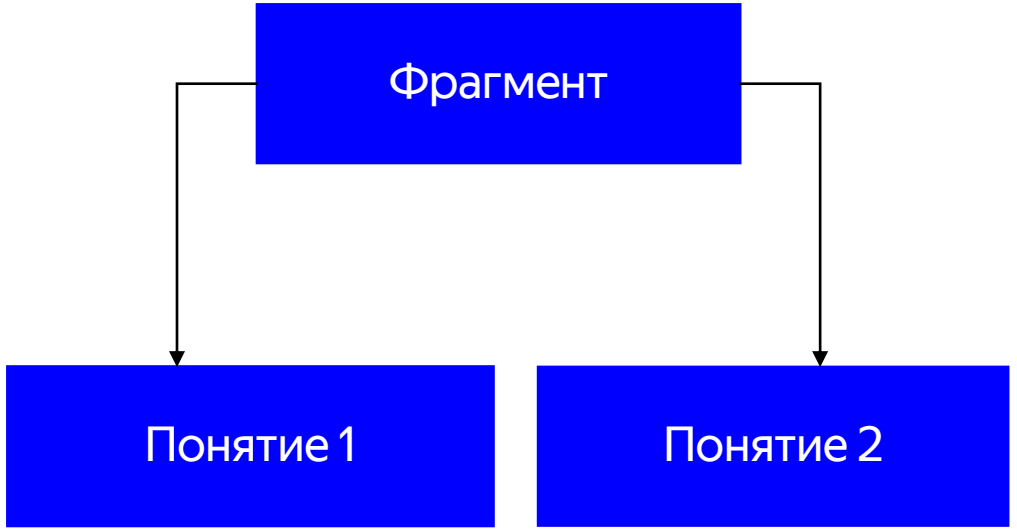


github.com/
yandex-datasphere/
advanced_rag

Продвинутые стратегии индексации

Parent Retriever

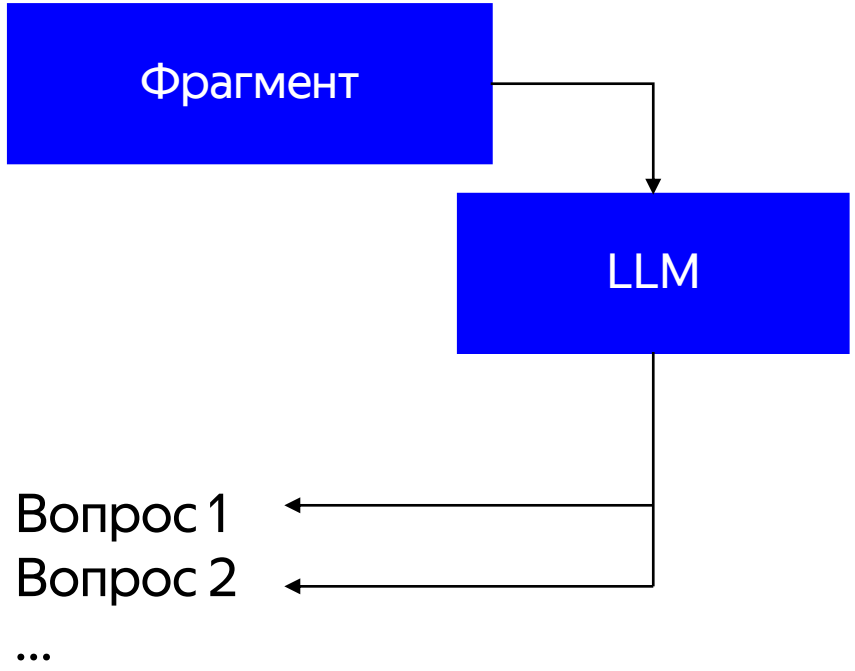
- Разбиваем фрагменты документов на ещё более мелкие фрагменты
- Parent-Child Relationship
- Мелкие документы уточняют конкретные понятия, родительские — задают более крупный контекст
- Находим ребенка, вбрасываем в контекст родителя



Продвинутые стратегии индексации

Гипотетические вопросы

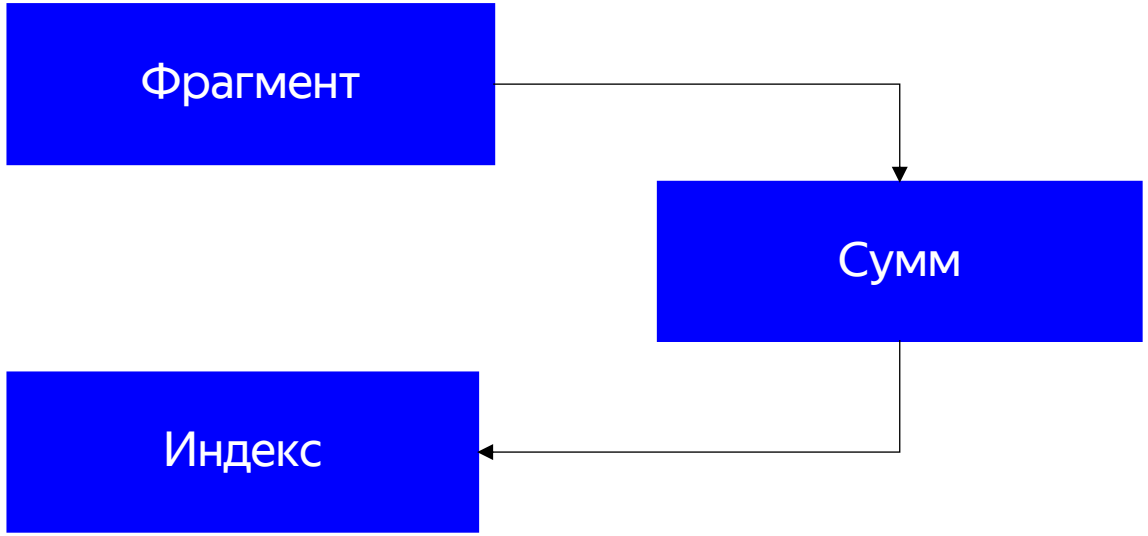
- Из фрагментов документов извлекаются вопросы, на которые эти фрагменты отвечают
- Индексируются гипотетические вопросы, вбрасываются в контекст исходные документы



Продвинутые стратегии индексации

Суммаризация

Индексируются суммаризованные фрагменты текста, вбрасываются в контекст исходные документы



YandexGPT поможет абитуриентам НИУ ВШЭ выбрать профессию

YandexGPT внедрили в чат-бот на сайте и в телеграм-канале НИУ ВШЭ. Абитуриенты могут поделиться с ботом информацией о себе и своих интересах, чтобы получить рекомендации по подходящим образовательным программам.

24 мая 2024 г. 5 минут чтения Поделиться 1

Одна из проблем наивного RAG

Мерло

Мерло хорошо сочетается с сыром Бри

Каберне

Каберне хорошо сочетается с сыром Бри

...

Сира

Сыр Бри отлично подходит к вину сорта Сира

Вопрос:

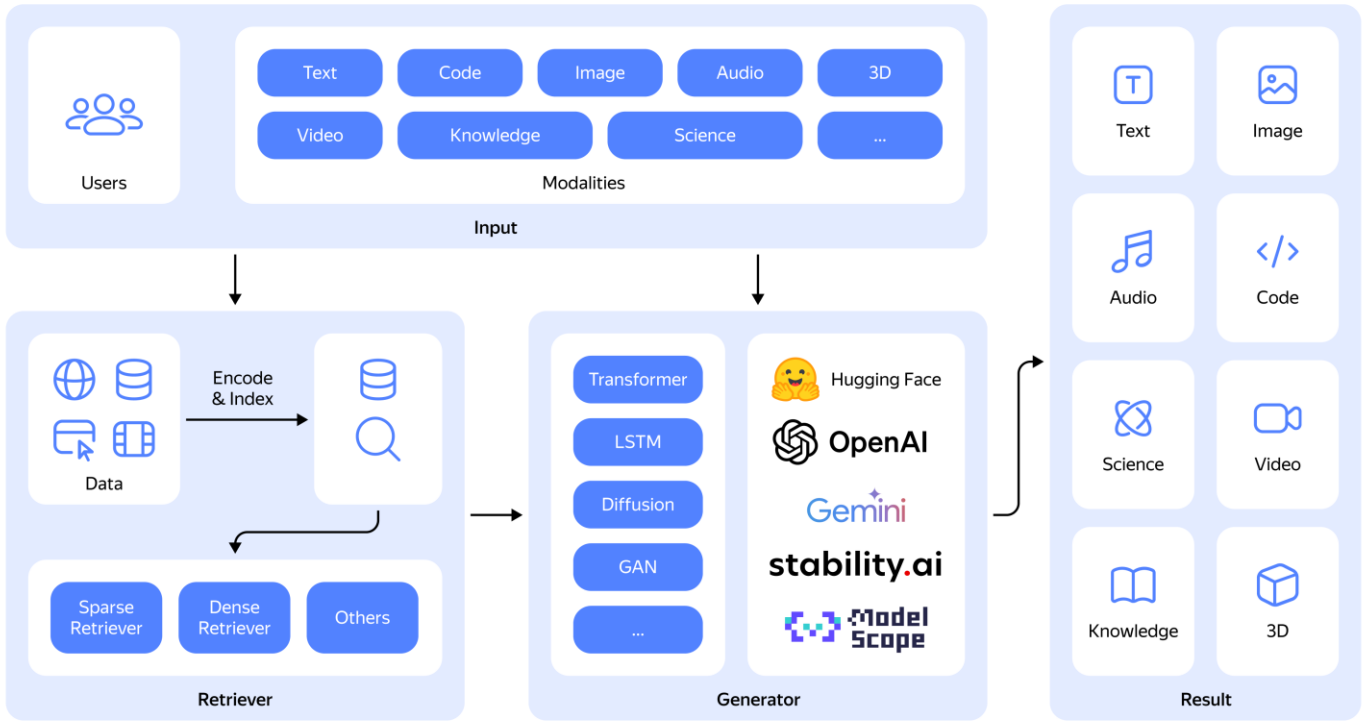
Какие вина хорошо подходят к сыру Бри?

Различные вариации RAG

1/ Передача Retrieval-Generator:
- Query-based
- Latent Representation
- Logit-based
- Speculative

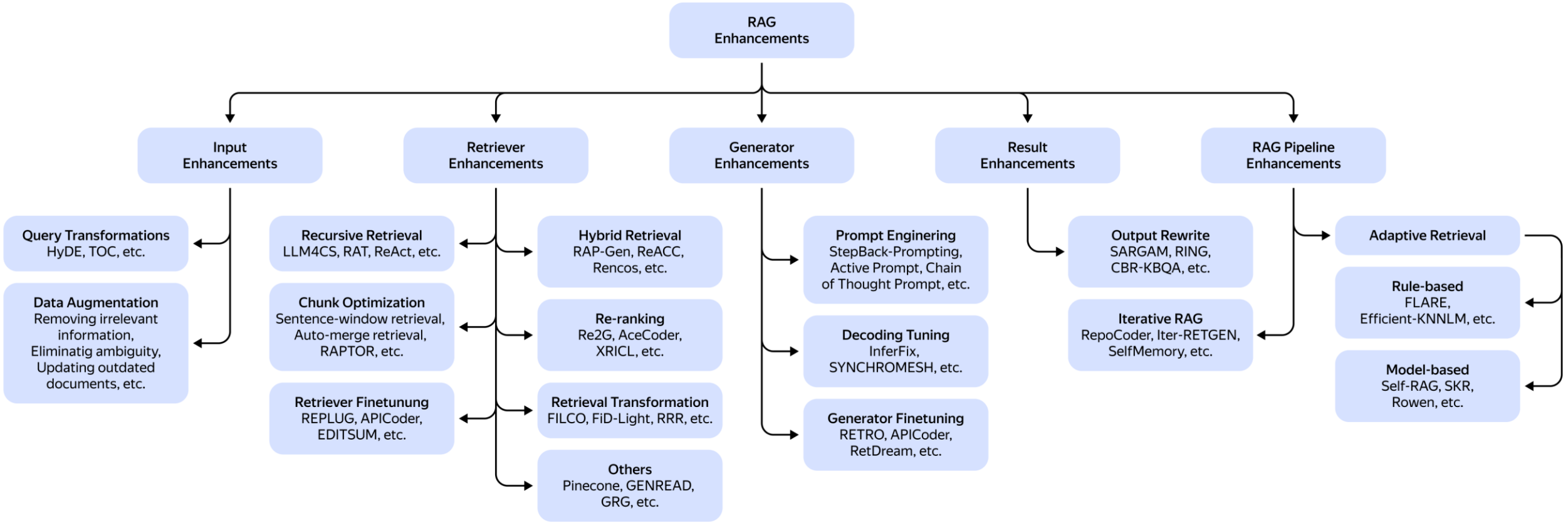
2/ Улучшения

3/ Нетекстовый RAG



Zhao, Penghao, et al. «Retrieval-augmented generation for ai-generated content: A survey.» arXiv preprint arXiv:2402.19473 (2024).

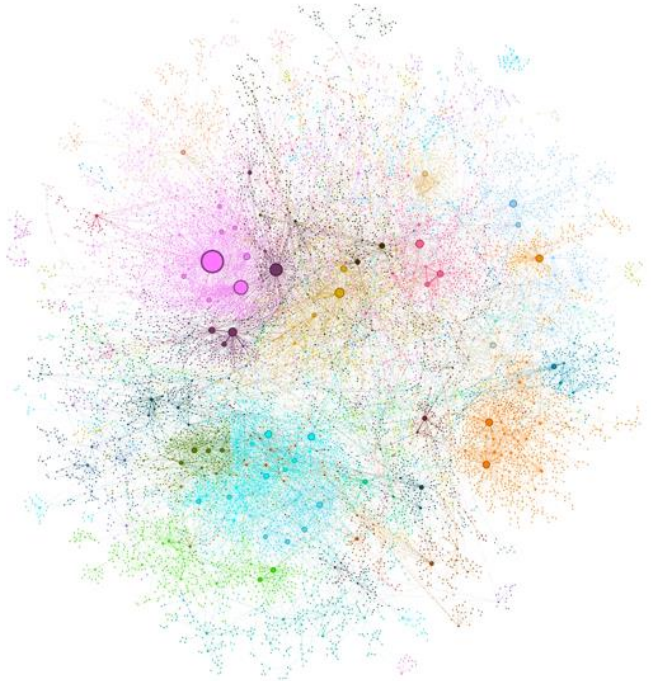
Различные подходы к улучшению RAG

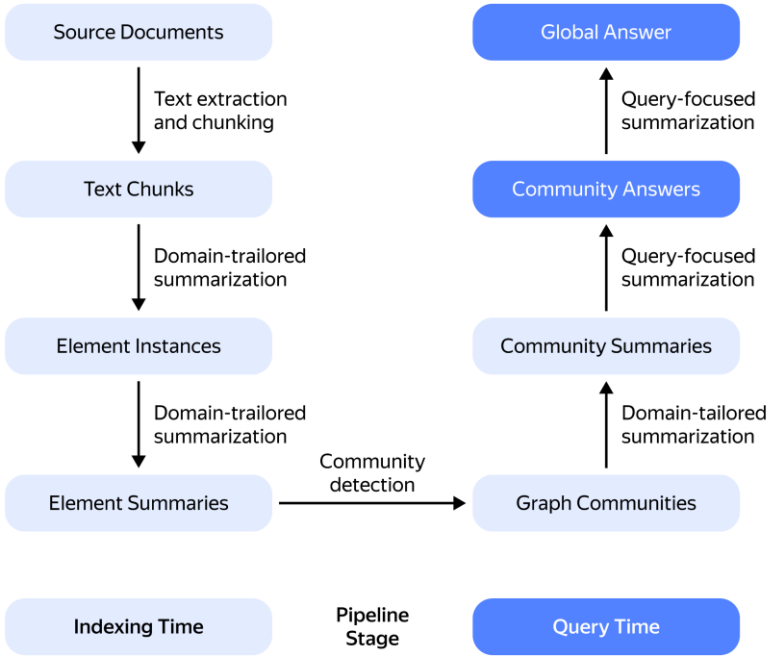


GraphRAG

github.com/microsoft/graphrag

Edge, Darren, et al. **«From local to global: A graph rag approach to query-focused summarization.»** arXiv preprint arXiv:2404.16130 (2024) — Microsoft Research





На стыке явного представления знаний и ML!

- Allemang, Dean, and Juan Sequeda. «Increasing the LLM Accuracy for Question Answering: Ontologies to the Rescue!.» arXiv preprint arXiv:2405.11706 (2024)
- Martino, Ariana, Michael Iannelli, and Coleen Truong. «Knowledge injection to counter large language model (LLM) hallucination.» European Semantic Web Conference. Cham: Springer Nature Switzerland, 2023

На стыке явного представления знаний и ML!

- Allemang, Dean, and Juan Sequeda. «Increasing the LLM Accuracy for Question Answering: Ontologies to the Rescue!.» arXiv preprint arXiv:2405.11706 (2024)
- Martino, Ariana, Michael Iannelli, and Coleen Truong. «Knowledge injection to counter large language model (LLM) hallucination.» European Semantic Web Conference. Cham: Springer Nature Switzerland, 2023

Food + Wine Ontology

```
<owl:Class rdf:ID="Merlot">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasColor" />
      <owl:hasValue rdf:resource="#Red" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasSugar" />
      <owl:hasValue rdf:resource="#Dry" />
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasFlavor" />
      <owl:allValuesFrom>
        <owl:Class>
          <owl:oneOf rdf:parseType="Collection">
            <owl:Thing rdf:about="#Moderate" />
            <owl:Thing rdf:about="#Delicate" />
          </owl:oneOf>
        </owl:Class>
      </owl:allValuesFrom>
    </owl:Restriction>
```

```
<owl:Class rdf:ID="PastaWithSpicyRedSauceCourse">
    <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#MealCourse" />
      <owl:Restriction>
        <owl:onProperty rdf:resource="#hasFood" />
        <owl:allValuesFrom rdf:resource="#PastaWithSpicyRedSauce" />
      </owl:Restriction>
    </owl:intersectionOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#hasDrink" />
        <owl:allValuesFrom>
          <owl:Restriction>
            <owl:onProperty rdf:resource="&vin;hasColor" />
            <owl:hasValue rdf:resource="#Red" />
          </owl:Restriction>
        </owl:allValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#hasDrink" />
        <owl:allValuesFrom>
          <owl:Restriction>
            <owl:onProperty rdf:resource="&vin;hasBody" />
            <owl:hasValue rdf:resource="#Full" />
          </owl:Restriction>
        </owl:allValuesFrom>
      </owl:Restriction>
    </rdfs:subClassOf>
```

Многоагентные системы

- Что подавать к Мерло?
- Сколько стоит литр Мерло?
- Из какого винограда делают Мерло?

Классификаторы на базе YandexGPT API

- Промт для YandexGPT API
- Zero-shot Endpoint
- Few-shot Endpoint
- Дообучение классификатора

Пример наполнения файла для обучения бинарной классификации:

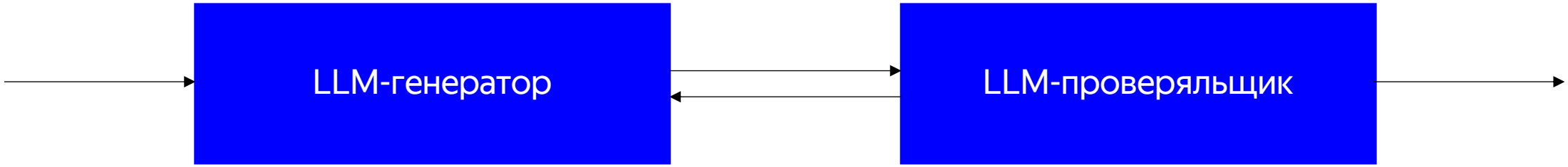
```
{"text":"у меня все в порядке","нейтральный":1,"эмоциональный":0}
{"text":"у меня все классно получилось ","нейтральный":0,"эмоциональный":1}
{"text":"вам не понять как тяжело бывает каждый день вставать на работу в шесть утра и два часа тащит
{"text":"все как всегда работа дом семья","нейтральный":1,"эмоциональный":0}
```

```
{
  "modelUri": "string",
  "taskDescription": "string",
  "labels": [
    "string",
    "string",
    ...
    "string"
  ],
  "text": "string",
  "samples": [
    {
      "text": "string",
      "label": "string"
    },
    {
      "text": "string",
      "label": "string"
    },
    ...
    {
      "text": "string",
      "label": "string"
    }
  ]
}
```

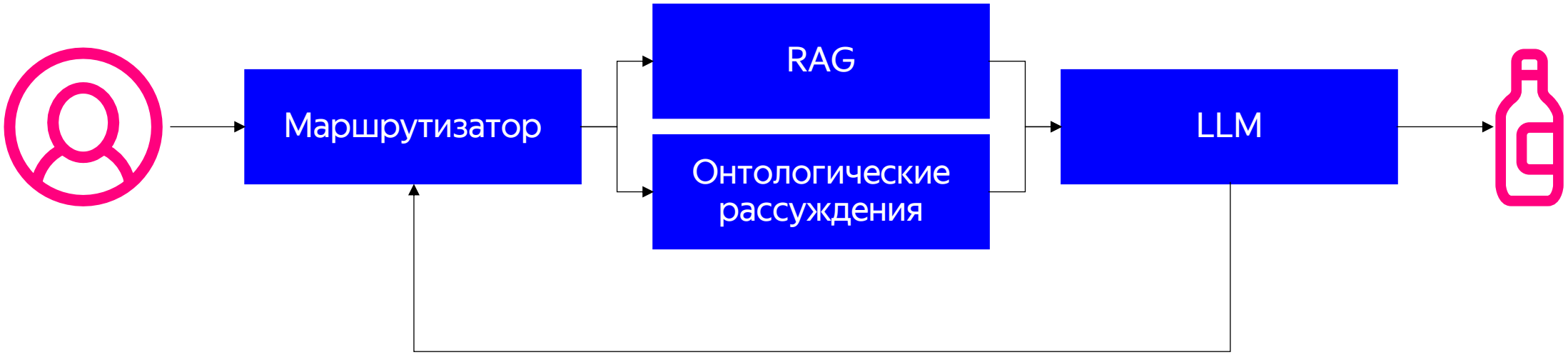
Проблема с рассуждениями у LLM

Sébastien Bubeck, Varun Chandrasekaran, et.al. Sparks of Artificial General Intelligence: Early experiments with GPT-4 arXiv:2303.12712 [cs.CL]

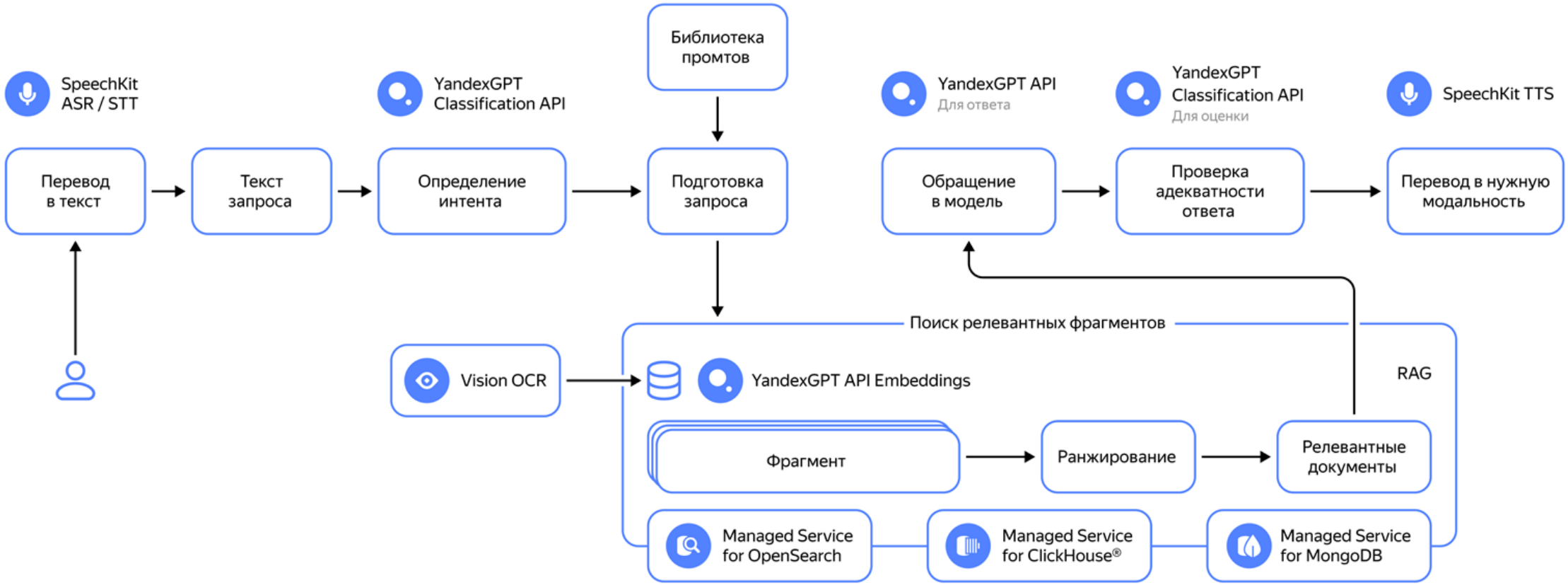
Основная проблема — линейный характер модели предсказания слов, отсутствие возможности размышлять с возвратами, в стиле «генерация и проверка»



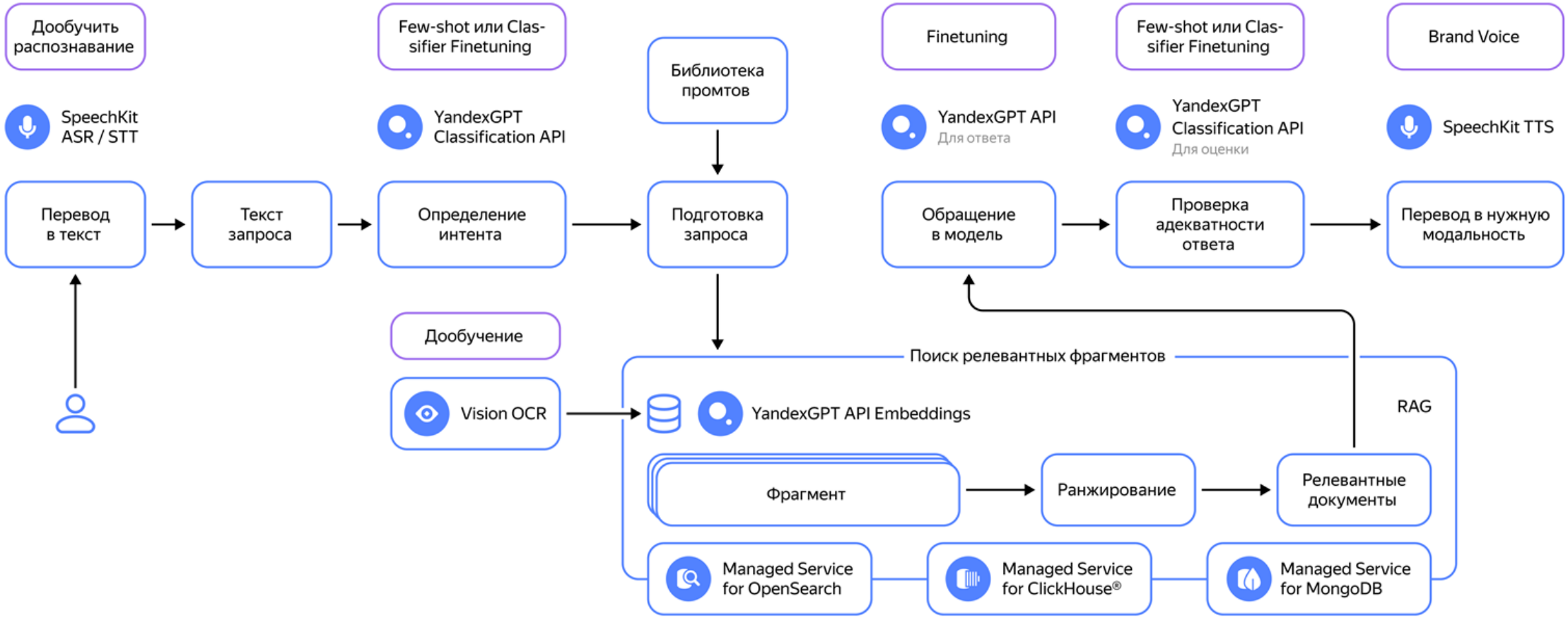
Многоагентные системы с верификацией



Архитектура

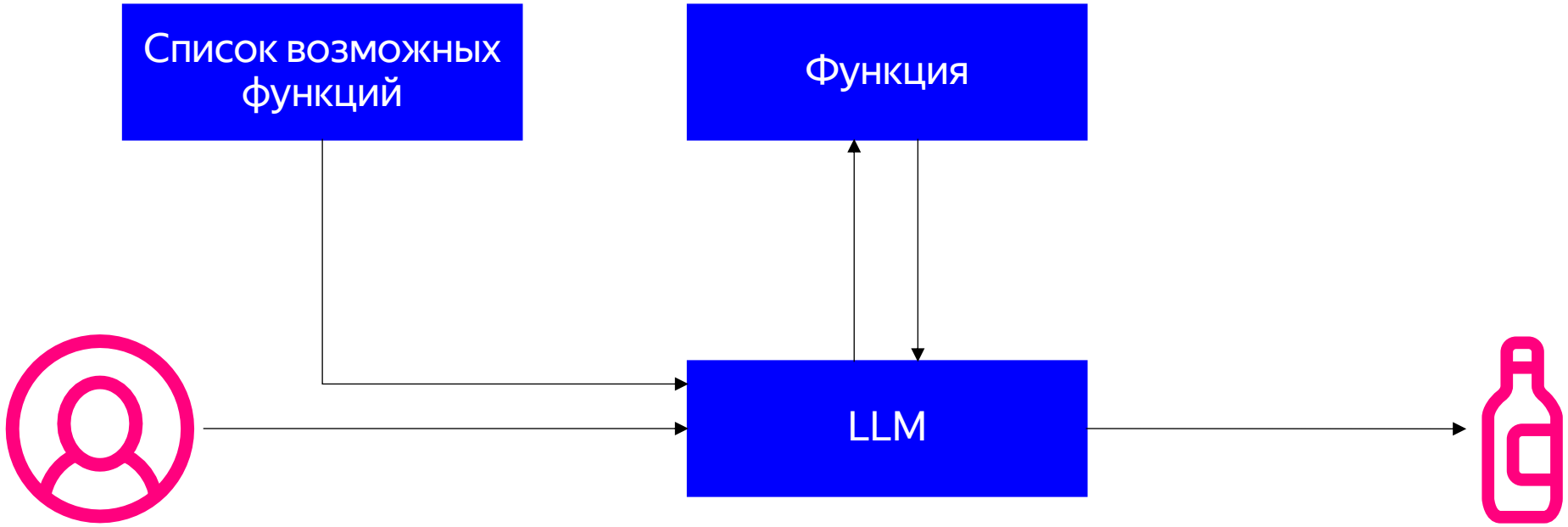


Архитектура



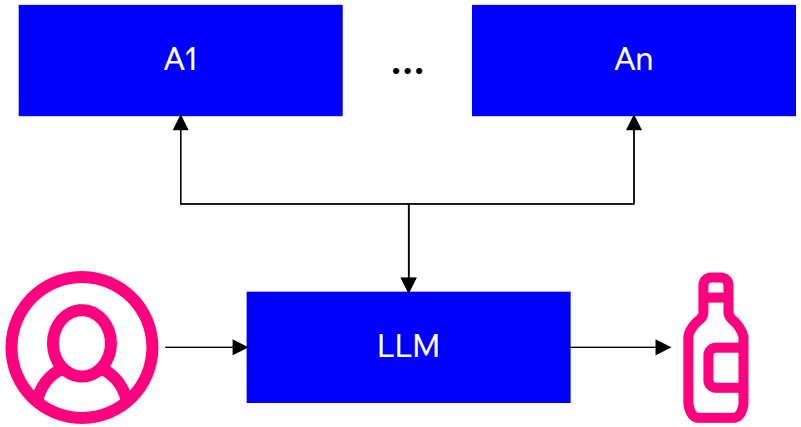
Function Calling

Что лучше есть с самым известным вином из Австралии?

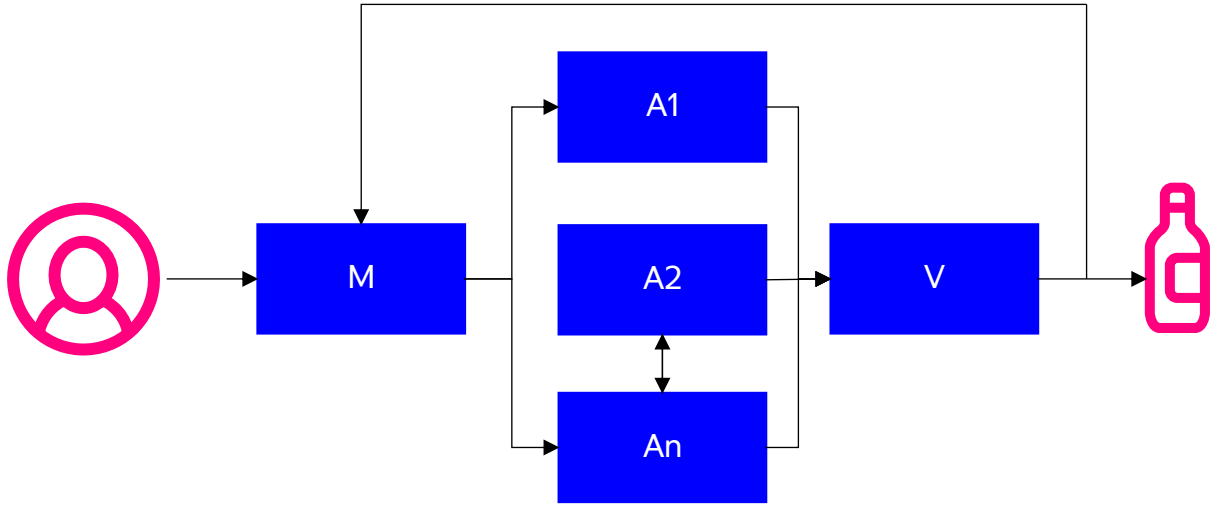


Два подхода к оркестрации MAC

LLM-оркестрация (Function Calling)



Ручная оркестрация



Обзор разных подходов к аугментированию LLM

01/

Reasoning [=Problem Decomposition]

- CoT
- Recursive Prompting

02/

Using Tools

- Information Retrieval
- Navigating the Web
- Symbolic Models and Code Interpreters
- Acting
- Calling Other Models

03/

Learning to Reason and Use Tools

- Suprevision
- Reinforcement Learning

Mialon G. et al. Augmented language models: a survey //arXiv preprint arXiv:2302.07842. — 2023

Эволюционная кибернетика

Принцип метасистемного перехода и эмерджентность

Турчин, Валентин Федорович. «Феномен науки: Кибернетический подход к эволюции». (1993): 2000 г.

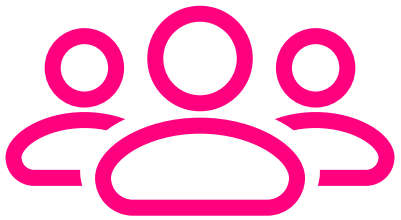
Эволюционная кибернетика

Принцип метасистемного перехода и эмерджентность

Турчин, Валентин Федорович. «Феномен науки: Кибернетический подход к эволюции». (1993): 2000 г.

Эволюционная кибернетика

Принцип метасистемного перехода и эмерджентность



Турчин, Валентин Федорович. «Феномен науки: Кибернетический подход к эволюции». (1993): 2000 г.

Мораль

01/

Различные подходы
к обогащению LLM
с помощью внешних знаний
позволяют подобрать подход,
оптимальный для решения
конкретной задачи

02/

Graph RAG — это движение
в сторону объединения
неявного (LLM) и явного
(графы, онтологии,
семантические сети)
представления знаний,
в большей
интерпретируемости

03/

LLM вдыхают новую жизнь
в теорию мультиагентных
систем. За счет синергетичес-
кого эффекта MAC позволяют
улучшать результаты LLM,
хотя иногда за счет
существенного увеличения
стоимости

Спасибо за внимание!

Яндекс

Дмитрий Сошников

Доцент НИУ ВШЭ/МАИ, консультант Yandex Cloud, тех. руководитель лаборатории ГенИИ Школы дизайна НИУ ВШЭ

t.me/shwarsico