

RAG Pipeline Project

Retrieval-Augmented Generation

Техническое задание для студентов

3 задания | 2 обязательных + 1 бонусное

Максимум: 100 баллов + 30 бонусных

Задание 1 (50 б.)	Задание 2 (50 б.)	Бонус (+30 б.)
Naive RAG + Advanced RAG	Golden Dataset + RAGAS + Улучшения	GraphRAG с Neo4j

Срок: 2 недели (1 марта)

Общая информация

Источники данных

В проекте используются два PDF-файла с годовыми отчётаами казахстанских компаний. Оба документа содержат сложную вёрстку: таблицы, диаграммы, многоколоночные блоки. Обычный текстовый парсинг (PyPDF2, pdfplumber) не подходит — необходим Visual Layout анализ.

Файл	Описание	Особенности
ktj.pdf	Интегрированный годовой отчёт АО «НК «КТЖ», ~368 стр.	Финансовые таблицы, ESG-данные, стратегические карты, диаграммы
matnp_2024_rus.pdf	Годовой отчёт АО «Матен Петролеум», ~20–30 стр.	Данные о добыче нефти, таблицы с объёмами реализации, текстовые блоки

Сами файлы:

[ktj.pdf](#)

[matnp_2024_rus.pdf](#)

Требования к парсингу

Для корректного извлечения данных используйте один из следующих инструментов Visual Layout анализа:

- DocLing — <https://github.com/docling-project/docling>
- Unstructured — <https://docs.unstructured.io/welcome>
- LlamaParse — https://docs.cloud.llamaindex.ai/llamaparse/getting_started

Таблицы должны быть преобразованы в Markdown-таблицы или JSON-объекты. Если таблица большая и разбивается на несколько чанков — дублируйте заголовок таблицы в каждом чанке.

Инструменты

Вы свободны в выборе конкретных библиотек, но рекомендуемый стек:

Категория	Рекомендация	Альтернативы

Оркестратор	LangChain	LlamaIndex
Векторная БД	ChromaDB	Qdrant, FAISS
Embedding	intfloat/multilingual-e5-large	BAAI/bge-m3, gte-multilingual
LLM	GPT-4o-mini (через API)	Qwen 2.5 (локально)
Оценка	RAGAS	—

Задание 1: Построение RAG-пайплайна (50 баллов)

Компонент	Баллы
Часть А: Naive RAG (базовый пайплайн)	20
Часть В: Advanced RAG (улучшенный пайплайн)	30

Часть А: Naive RAG (20 баллов)

Постройте базовый RAG-пайплайн, который может отвечать на вопросы по содержимому PDF-файлов.

Что нужно сделать:

1. Загрузить и распарсить оба PDF-файла с помощью Visual Layout инструмента.
2. Реализовать Naive Chunking: фиксированный размер чанков (1024 токена, overlap 200).
3. Создать эмбеддинги с помощью мультиязычной модели и загрузить их в векторную БД.
4. Реализовать Dense Retrieval (косинусное сходство).
5. Подключить LLM для генерации ответов по найденному контексту.
6. Протестировать на 5–10 вопросах вручную и зафиксировать типичные ошибки.

Ожидаемые проблемы Naive RAG: таблицы разрываются посередине; цифры из разных строк путаются; точные названия (например, «Достык – Мойынты») не находятся через семантический поиск.

Часть В: Advanced RAG (30 баллов)

Улучшите пайплайн, устранив проблемы Naive RAG. Каждое улучшение нужно обосновать и показать разницу.

Обязательные улучшения:

1. Продвинутый чанкинг (сравнить 3 стратегии)

Реализуйте и сравните минимум три стратегии разделения текста:

Стратегия	Описание	Когда использовать
Naive (Fixed Size)	Фиксированный размер 1024 токена, overlap 200. Базовый вариант.	Baseline для сравнения
Semantic / Recursive	Разделение по логическим границам: заголовки, концы параграфов. RecursiveCharacterTextSplitter из LangChain.	Текстовые блоки
Layout-Aware	Чанки по визуальным блокам документа. Таблица не разрывается; заголовки разделов — в метаданных.	Таблицы, диаграммы

1. Hybrid Search (обязательно)

Объедините Vector Search (семантика) и BM25 (ключевые слова) через Reciprocal Rank Fusion (RRF). Это критично для поиска точных названий и цифр.

1. Reranking

Добавьте Cross-Encoder reranker после retrieval. Рекомендуемая модель: BAAI/bge-reranker-v2-m3. Покажите результаты до и после reranking.

1. Pre-Retrieval техника (одна на выбор)

- Query Rewriting — переформулировка запроса для лучшего поиска
- HyDE — генерация гипотетического ответа и поиск по нему
- Query Routing — маршрутизация запроса к нужному источнику

Задание 2: Эксперименты и оценка (50 баллов)

Компонент	Баллы
Часть А: Серия экспериментов с гиперпараметрами	30
Часть В: Оценка RAGAS + итоговый вывод	20



Golden Dataset уже готов

В приложении к этому ТЗ дан набор из 30 эталонных пар «вопрос – ответ». Используйте его как есть для оценки. Создавать свой датасет не нужно.

Часть А: Серия экспериментов (30 баллов)

Цель — исследовать, как разные гиперпараметры влияют на качество RAG-пайплайна. Проведите **минимум 6 экспериментов**, варьируя параметры из таблицы ниже. Каждый эксперимент — это запуск пайплайна с изменённым параметром и оценка результатов на Golden Dataset.

Гиперпараметры для экспериментов

Параметр	Что это	Диапазон значений	Как влияет
Chunk Size	Размер чанка в токенах	256, 512, 1024, 2048	Маленькие — точнее, но теряют контекст. Большие — наоборот.
Chunk Overlap	Перекрытие между соседними чанками	0, 50, 100, 200	Больше overlap — меньше потерь на границах
Стратегия чанкинга	Способ разбиения документа	Fixed, Recursive, Layout-Aware	Layout-Aware сохраняет таблицы целиком
Top-K	Сколько чанков передаётся в LLM	3, 5, 10, 15, 20	Больше K — больше контекста, но и больше шума
Alpha (RRF)	Bec Vector vs BM25 в Hybrid Search	0.0 – 1.0 (шаг 0.1–0.2)	0.0 = только BM25, 1.0 = только Vector
Reranking	Переранжирование после retrieval	Вкл/Выкл	Улучшает точность, но добавляет latency
Embedding-модель	Модель для создания эмбеддингов	e5-large, bge-m3, mppnet и др.	Разное качество для multilingual

Как проводить эксперименты

Принцип: меняйте **один параметр за раз**, остальные фиксируйте. Так вы увидите эффект каждого параметра изолированно.

Пример серии экспериментов

#	Что меняем	Chunk	K	Alpha	Rerank	Faithf.	Вывод
0	Baseline	1024/200	5	0.5	Нет	?	Отправная точка
1	Chunk Size	512/100	5	0.5	Нет	?	?
2	Chunk Size	2048/200	5	0.5	Нет	?	?
3	Top-K	best	3	0.5	Нет	?	?

#	Что меняем	Chunk	K	Alpha	Rerank	Faithf.	Вывод
4	Top-K	best	10	0.5	Нет	?	?
5	Alpha	best	best	0.0	Нет	?	Только BM25
6	Alpha	best	best	0.3	Нет	?	?
7	Alpha	best	best	0.7	Нет	?	?
8	Alpha	best	best	1.0	Нет	?	Только Vecto
9	Reranking	best	best	best	Да	?	?

«best» = лучшее значение из предыдущих экспериментов. Вы каждый раз берёте победителя предыдущего шага и варьируете следующий параметр.



Это называется greedy search по гиперпараметрам. Вы не обязаны следовать именно этой таблице — можете добавлять свои эксперименты, менять порядок, пробовать разные embedding-модели и стратегии чанкинга. Главное — минимум 6 экспериментов и внятный вывод по каждому.

Что нужно по каждому эксперименту

1. Чётко указать, какой параметр изменён и какое значение выставлено.
2. Запустить пайплайн на всём Golden Dataset (10 вопросов).
3. Записать значения RAGAS-метрик (Faithfulness, Answer Relevancy, Context Recall, Context Precision).
4. Написать краткий вывод: стало лучше или хуже? Почему?

Часть В: Оценка RAGAS + итоговый вывод (20 баллов)

После всех экспериментов подведите итоги.

1. Итоговая таблица метрик

Соберите результаты всех экспериментов в единую таблицу. По каждой метрике отметьте лучший и худший результат.

2. Анализ метрик RAGAS

Метрика	Что измеряет
Faithfulness	Подтверждены ли утверждения в ответе найденным контекстом? (галлюцинации)
Answer Relevancy	Отвечает ли ответ на заданный вопрос?
Context Recall	Найдены ли все нужные факты из эталонного ответа?
Context Precision	Релевантны ли найденные чанки? (нет ли мусора в контексте)

По каждой метрике: числовое значение (среднее), интерпретация, 2–3 конкретных примера (вопросы с высокой/низкой метрикой и объяснение).

3. Итоговый вывод (в Markdown внутри .ipynb)

В конце ноутбука создайте Markdown-ячейку с заголовком «Итоговый вывод». В ней ответьте на вопросы:

- Какая комбинация гиперпараметров оказалась лучшей? Приведите конкретные значения.
- Какой параметр оказал наибольшее влияние на качество? Почему вы так считаете?
- Какой параметр оказал наименьшее влияние?
- При каком значении Alpha (баланс Vector/BM25) результат лучше всего? Почему?
- Помог ли reranking? На каких типах вопросов особенно?
- Какие вопросы из Golden Dataset остались «сложными» даже при лучшей конфигурации? Что бы вы попробовали ещё?



Оценивается именно вывод. Баллы ставятся не за «правильные» метрики, а за качество анализа. Низкий Faithfulness — это нормально, если вы объяснили почему и что попробовали. Главное — продемонстрировать осознанный подход к настройке пайплайна.

Бонусное задание: GraphRAG (+30 баллов)

Необязательное задание для тех, кто хочет получить дополнительные баллы и углубить понимание RAG.

Суть задания

Постройте Knowledge Graph из тех же PDF-документов и используйте его для ответов на вопросы, которые плохо решаются обычным vector search.

Что нужно сделать:

1. Развернуть Neo4j (локально через Docker или Neo4j Aura Free).
2. Извлечь сущности и связи из документов (компании, показатели, проекты, даты).
3. Загрузить граф в Neo4j.
4. Реализовать Cypher-запросы для ответов на вопросы.
5. Сравнить результаты GraphRAG с Vector RAG по тем же вопросам.

Пример вопросов для GraphRAG:

- «Какие инфраструктурные проекты связаны с направлением Достык – Мойынты?»
- «Какие компании упоминаются в обоих отчётах?»
- «Покажите связь между ESG-рейтингом КТЖ и конкретными экологическими проектами.»

Что представить:

- Код извлечения сущностей и загрузки в Neo4j
- Примеры Cypher-запросов с результатами
- Таблица сравнения: Vector RAG vs GraphRAG на одних и тех же вопросах
- Краткий вывод: когда GraphRAG лучше, а когда нет

Справочные материалы

Embedding-модели

Для русско- и казахскоязычных документов нужна мультиязычная модель. Ниже — рекомендуемые варианты:

Модель	Размер	Dim	Tokens	Примечание
intfloat/multilingual-e5-large	560M	1024	512	SOTA на MIRACL. Рекомендуем начать с неё.
intfloat/multilingual-e5-base	278M	768	512	Баланс скорости и качества
BAAI/bge-m3	568M	1024	8192	Длинный контекст, Dense+Sparse
Alibaba-NLP/gte-multilingual	305M	768	8192	Быстрый, 70+ языков
paraphrase-multilingual-mppnet	278M	768	512	Классика, проверено временем

Рекомендация: начните с intfloat/multilingual-e5-large. Если работаете с большими таблицами — попробуйте BAAI/bge-m3 (8192 токенов).

Reranker: BAAI/bge-reranker-v2-m3

Готовый Golden Dataset (30 пар)

Ниже — стартовый набор.

```
[  
  {  
    "question": "Каков был доход от основной деятельности АО «НК «КТЖ» в 2024 году?",  
    "ground_truth": "Доход от основной деятельности составил 2 163,9 млрд тенге.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "На сколько вырос доход от основной деятельности КТЖ в 2024 году по сравнению с 2023 годом (в абсолютном значении)?",  
    "ground_truth": "Доход вырос на 229,8 млрд тенге.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какой объем доходов от грузовых перевозок получила КТЖ в 2024 году?",  
    "ground_truth": "Доходы от грузовых перевозок составили 1 875,6 млрд тенге.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "На сколько процентов увеличились доходы от грузовых перевозок КТЖ в 2024 году?",  
    "ground_truth": "Доходы увеличились на 11,5%.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какого числа был утвержден годовой отчет КТЖ Советом директоров?",  
    "ground_truth": "Отчет был утвержден 30 мая 2025 года (протокол №6).",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какие международные стандарты ISO внедрены в системе управления КТЖ?",  
    "ground_truth": "Внедрены стандарты ISO 9001:2015, ISO 14001:2015, ISO 45001:2018 и ISO 50001:2018.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какую цель по грузообороту ставит перед собой КТЖ на 2025 год?",  
    "ground_truth": "Планируется увеличить грузооборот до 273,8 млрд т-км.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какой процент роста грузооборота запланирован на 2025 год по сравнению с 2024?",  
    "ground_truth": "Запланирован рост на 4,6%.",  
    "file_source": "ktj.pdf"  
  },  
  {  
    "question": "Какой целевой показатель по объему контейнерного транзита установлен на 2025 год?",  
    "ground_truth": "Целевой показатель составляет 1 553 тыс. ДФЭ.",  
    "file_source": "ktj.pdf"  
  },  
]
```

```
        "question": "Какие крупные инфраструктурные объекты планирует ввести в эксплуатацию КТЖ в 2025 году?",  
        "ground_truth": "Планируется ввод вторых путей на участке «Достык – Мойынты» и железнодорожной линии в обход станции Алматы.",  
        "file_source": "ktj.pdf"  
    },  
    {  
        "question": "Сколько километров железнодорожных путей планируется модернизировать в рамках новых проектов?",  
        "ground_truth": "Планируется модернизация порядка 3 000 километров путей.",  
        "file_source": "ktj.pdf"  
    },  
    {  
        "question": "До какого года рассчитана Стратегия развития АО «НК «КТЖ»?",  
        "ground_truth": "Стратегия развития рассчитана до 2032 года.",  
        "file_source": "ktj.pdf"  
    },  
    {  
        "question": "Каков был доход от грузовых перевозок в 2023 году (для сравнения)?",  
        "ground_truth": "В 2023 году доход составил 1 682,6 млрд тенге.",  
        "file_source": "ktj.pdf"  
    },  
    {  
        "question": "Каков был совокупный объем добычи нефти АО «Матен Петролеум» и его дочерней компании в 2024 году?",  
        "ground_truth": "Совокупная добыча составила 685 тысяч тонн.",  
        "file_source": "matnp_2024_rus.pdf"  
    },  
    {  
        "question": "Как называется 100% дочерняя компания АО «Матен Петролеум»?",  
        "ground_truth": "Дочерняя компания называется АО «Кожан».",  
        "file_source": "matnp_2024_rus.pdf"  
    },  
    {  
        "question": "Какой объем реализации нефти был зафиксирован у «Матен Петролеум» в 2024 году?",  
        "ground_truth": "Объем реализации нефти составил 677 тысяч тонн.",  
        "file_source": "matnp_2024_rus.pdf"  
    },  
    {  
        "question": "Какие внешние вызовы для развития компании упоминает руководство «Матен Петролеум»?",  
        "ground_truth": "Упоминаются волатильность мировых цен на нефть и сложная геополитическая обстановка.",  
        "file_source": "matnp_2024_rus.pdf"  
    },  
    {  
        "question": "На что были направлены значительные инвестиции «Матен Петролеум» в 2024 году?",  
        "ground_truth": "На модернизацию производственных мощностей, внедрение инновационных технологий и повышение экологической безопасности.",  
        "file_source": "matnp_2024_rus.pdf"  
    },  
    {
```

```
"question": "Сколько средств перечислило АО «Матен Петролеум» на благотворительность в 2024 году?",  
    "ground_truth": "Было перечислено 626,2 миллиона тенге.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Каким категориям граждан оказывает поддержку «Матен Петролеум»?",  
    "ground_truth": "Поддержка оказывается ветеранам ВОВ, пенсионерам, детским домам и благотворительным учреждениям.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Были ли выявлены факты коррупции в АО «Матен Петролеум» в 2024 году?",  
    "ground_truth": "Нет, фактов коррупции со стороны работников в 2024 году выявлено не было.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Когда в «Матен Петролеум» была разработана политика по противодействию коррупции?",  
    "ground_truth": "Политика была разработана в июне 2019 года.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Применялись ли к «Матен Петролеум» штрафы или санкции за нарушение законодательства в 2024 году?",  
    "ground_truth": "Нет, существенные штрафы и нефинансовые санкции не применялись.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Были ли случаи нарушения антимонопольного законодательства со стороны «Матен Петролеум»?",  
    "ground_truth": "Факты, связанные с препятствием конкуренции и нарушением антимонопольного законодательства, отсутствовали.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Какова основная продукция «Матен Петролеум» и как она реализуется?",  
    "ground_truth": "Единственной продукцией является нефть, которая реализуется по системе нефтепроводов.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "В каком городе базируется АО «Матен Петролеум» согласно отчету?",  
    "ground_truth": "В городе Атырау.",  
    "file_source": "matnp_2024_rus.pdf"  
},  
{  
    "question": "Чему был равен доход от основной деятельности КТЖ в 2023 году?",  
    "ground_truth": "1 934,1 млрд тенге.",  
    "file_source": "ktj.pdf"  
},  
{  
    "question": "Какова разница в абсолютном значении между доходами от грузовых перевозок 20
```

```

        "24 и 2023 года в КТЖ",
        "ground_truth": "Разница составляет 193 млрд тенге.",
        "file_source": "ktj.pdf"
    },
    {
        "question": "Какие проекты планирует начать реализовывать КТЖ помимо ввода новых линий?",
        "ground_truth": "Начать реализацию новых проектов по строительству ж/д линий и модернизации существующих путей.",
        "file_source": "ktj.pdf"
    },
    {
        "question": "Каково процентное отношение планируемого роста контейнерного транзита в 2025 году к 2024 году?",
        "ground_truth": "Планируемый рост составляет 11,3%.",
        "file_source": "ktj.pdf"
    }
]

```

Что сдавать

Артефакт	Описание
Jupyter Notebook	С пояснениями, визуализациями и результатами
README.md	Архитектура, схема пайплайна, инструкция запуска

Quick Install

pip install langchain langchain-community langchain-openai chromadb qdrant-client ragas sentence-transformers docling unstructured FlagEmbedding

Для бонусного задания (GraphRAG):

```
pip install neo4j
```

```
docker run -d --name neo4j -p 7474:7474 -p 7687:7687 -e NEO4J_AUTH=neo4j/password neo4j:latest
```

Полезные ссылки

Ресурс	URL
LangChain	https://python.langchain.com/docs/introduction/
LlamaIndex	https://docs.llamaindex.ai/en/stable/
ChromaDB	https://docs.trychroma.com/
Qdrant	https://qdrant.tech/documentation/
RAGAS	https://docs.ragas.io/en/stable/
DocLing	https://ds4sd.github.io/docling/
Unstructured	https://docs.unstructured.io/welcome
LlamaParse	https://docs.cloud.llamaindex.ai/llamaparse/getting_started
MTEB Leaderboard	https://huggingface.co/spaces/mteb/leaderboard

Sentence Transformers	https://www.sbert.net/
FlagEmbedding (BGE)	https://github.com/FlagOpen/FlagEmbedding
Neo4j	https://neo4j.com/docs/