УДК 004.65:004.912

Морозов Д.А.

магистр, Самарский национальный исследовательский университет имени академика С. П. Королёва, г. Самара, Россия

МЕТОД ПРЕОБРАЗОВАНИЯ ТАБЛИЧНЫХ АНАЛИТИЧЕСКИХ ДАННЫХ В ДОКУМЕНТНОЕ ПРЕДСТАВЛЕНИЕ ДЛЯ ГИБРИДНОГО ИЗВЛЕЧЕНИЯ В РУССКОЯЗЫЧНОЙ КОРПОРАТИВНОЙ СРЕДЕ

Аннотация. Рассматривается задача извлечения сведений из табличных аналитических данных, предварительно преобразованных в документное представление для систем дополненной генерации. Предложен метод преобразования, включающий группировку записей по естественному ключу предметной области, разбиение на фрагменты, явное разграничение фактических и прогнозных значений, добавление поисковых маркеров и раздельное формирование текстов для плотного и разрежённого представлений. Выполнена экспериментальная оценка на корпусе из 600 документов с контролируемой долей трудных негативов и 60 запросах шести типов. Гибридное слияние обеспечило наилучшее ранжирование (MRR@10 = 0,48) и полноту (Hit@10 = 0,84), однако уступило чистому лексическому поиску по точности первой позиции (0,30 против 0,38); допущение о безусловном превосходстве гибридизации не подтвердилось. Установлено, что при принятом размере фрагмента 58,5 % документов превышают контекстное окно кодировщика и усекаются; на основании измерений выведено количественное требование к бюджету токенов на строку таблицы. Показано, что определяющим фактором качества извлечения является подготовка данных, а не выбор режима поиска.

Ключевые слова: дополненная генерация; RAG; векторный поиск; гибридный поиск; BM25; обратное ранговое слияние; чанкование; контекстное окно; лемматизация; русский язык.

Abstract. The paper addresses retrieval from tabular analytical data converted in advance into a document representation for retrieval-augmented generation

systems. A conversion method is proposed comprising grouping records by a natural domain key, splitting into fragments, explicit separation of factual and forecast values, addition of search markers, and separate construction of texts for dense and sparse representations. An experimental evaluation is carried out on a 600-document corpus with a controlled share of hard negatives and 60 queries of six types. Hybrid fusion achieved the best ranking quality (MRR@10 = 0.48) and recall (Hit@10 = 0.84) but lost to pure lexical retrieval in first-position accuracy (0.30 against 0.38), so the assumption of unconditional hybrid superiority was not confirmed. At the chosen fragment size, 58.5 % of documents exceed the encoder context window and are truncated; a quantitative requirement for the per-row token budget is derived from the measurements. Data preparation rather than retrieval mode is shown to be the decisive factor of retrieval quality.

Keywords: retrieval-augmented generation; RAG; vector search; hybrid search; BM25; reciprocal rank fusion; chunking; context window; lemmatisation; Russian language.

Введение

В системах дополненной генерации (retrieval-augmented generation, RAG) языковая модель формирует ответ на основе фрагментов, извлечённых из внешнего хранилища [1, 2]. Для сценариев, в которых аналитические показатели рассчитываются по регламенту и не требуют обновления в реальном времени, рациональной оказывается схема с предварительной материализацией результатов расчёта в виде документов и последующим поиском по ним, без обращения к СУБД в момент обслуживания запроса. Ключевым техническим элементом такой схемы является способ преобразования табличных данных в представление, пригодное для извлечения; ему и посвящена настоящая работа.

Задача нетривиальна, поскольку строка таблицы обладает свойствами, неблагоприятными для поиска. Она не содержит слов, которыми пользователь описывает свою потребность, вследствие чего лексическое пересечение с

запросом близко к нулю; её содержимое насыщено числовыми значениями, зашумляющими векторное представление; наконец, соседние строки различаются лишь отдельными подстроками, что затрудняет их разделение. Дополнительную сложность создаёт русскоязычная предметная область: выбор высокоточных кодировщиков для русского языка ограничен, а запросы логистической компании насыщены топонимами, внутренними сокращениями и артикулами — классами обозначений, на которых семантический поиск уступает лексическому.

Цель работы — предложить метод преобразования табличных аналитических данных в документное представление и экспериментально оценить достижимое качество извлечения, включая вклад отдельных элементов метода.

1. Каналы извлечения и их сочетание

Плотный (dense) поиск основан на кодировании запроса и документа в общее векторное пространство [3] и обеспечивает нахождение по смыслу при расхождении формулировок. Его систематическим недостатком является обобщение значений: редкие термины, коды и имена собственные не получают точного соответствия. Лексическое ранжирование по модели BM25 [4], учитывающее частоту термина, его информативность и нормировку по длине документа, обладает противоположными свойствами.

Объединение каналов выполняется методом обратного рангового слияния (reciprocal rank fusion, RRF) [5], оперирующего позициями документов, а не абсолютными оценками:

$$\text{score}(d) = \Sigma_i \, 1 / (\kappa + \text{rank}_i(d)), \qquad (1)$$

где $\text{rank}_i(d)$ — позиция документа d в i-м списке, к — константа сглаживания. Работа с рангами снимает проблему несопоставимости шкал косинусной близости и оценки BM25 и не требует калибровки.

Специфика интерпретации метрик извлечения применительно к RAG требует оговорки. В классических поисковых задачах приоритетна точность

первой позиции, поскольку пользователь просматривает выдачу сам. В RAG-контуре в контекст модели передаётся несколько верхних фрагментов, поэтому содержательной является полнота на глубине, равной бюджету контекста. Вместе с тем увеличение числа передаваемых фрагментов небезразлично: показано, что языковые модели хуже используют сведения, расположенные в середине длинного контекста, а присутствие правдоподобных, но нерелевантных фрагментов ухудшает ответ [6]. Метрики Hit@1 и Hit@10 следует поэтому рассматривать совместно: первая характеризует чистоту контекста, вторая — принципиальную достижимость ответа.

2. Метод преобразования табличных данных

Метод реализует отображение множества табличных записей в множество документов и состоит из шести этапов, каждый из которых обоснован ниже с точки зрения его влияния на качество последующего извлечения.

2.1. Группировка и разбиение на фрагменты

Записи группируются по естественному ключу предметной области — объекту учёта (складу, филиалу, направлению). Группировка отражает структуру типовых запросов и повышает связность контекста, предотвращая попадание показателей несвязанных объектов в один документ. Каждая группа разбивается на фрагменты ограниченного размера: крупные сегменты размывают релевантный сигнал в обоих каналах, снижая контрастность сходства, тогда как компактные уменьшают объём нерелевантного контекста, передаваемого модели. Обеспечивается самодостаточность фрагмента — наименование объекта, период и заголовки показателей воспроизводятся в каждом документе, поскольку при извлечении по k верхним позициям документы поступают модели вне исходного табличного порядка.

2.2. Разграничение фактических и прогнозных значений

Фактические и прогнозные величины разделяются явным полем-маркером. Их смешение ведёт к содержательным ошибкам вывода: модель может принять прогноз за свершившийся факт либо агрегировать разнородные по природе величины. Маркер выполняет двойную функцию — служит признаком для фильтрации выдачи на уровне метаданных и обеспечивает корректную вербализацию статуса показателя в ответе.

2.3. Текстуализация и поисковые маркеры

Табличные записи преобразуются в текст, где пары «показатель — значение» выражены в явной вербальной форме. Это переводит структурированные данные в модальность, для которой оптимизированы кодировщики, и делает аналитику однородной со справочной документацией, что позволяет обслуживать оба типа содержимого общим механизмом поиска.

Перед содержимым фрагмента помещаются поисковые маркеры — формулировки, отражающие вероятные постановки запроса («прогноз загрузки филиала», «ожидаемый объём груза», «план поступления на склад»). Приём направлен на устранение лексико-семантического разрыва (vocabulary mismatch) между терминологией пользователя и бедным контекстом табличной записи и повышает вероятность совпадения одновременно в лексическом канале, за счёт точных вхождений, и в плотном, за счёт сближения векторов.

2.4. Раздельные представления для плотного и лексического каналов

Для каждого документа формируются два независимых текста. Плотное представление кодируется моделью семейства multilingual-e5 [7] с предписанными префиксами для документов и запросов. Лексическое представление строится по отдельному текстовому полю. Разделение мотивировано различием требований: для точного поиска существенны

исходные написания терминов и кодов, для семантического — связность контекста. Ниже показано, что это разделение является не факультативной оптимизацией, а необходимым условием корректной работы.

2.5. Нормализация текста для лексического канала

Реализована нормализация, включающая токенизацию, приведение к нижнему регистру и лемматизацию русскоязычных словоформ, при которой кодоподобные токены (сокращения, артикулы, обозначения) сохраняются без изменений. Лемматизация принципиальна для русского языка: формы «отделение», «отделения», «отделению» должны приводить к одному результату, тогда как модель BM25 в исходном виде оперирует точным совпадением строк.

2.6. Реализация разрежённого представления

Лексический канал реализован как разрежённые векторы в векторном хранилище [8]: вес термина в документе задаётся насыщением частоты, вес термина в запросе — величиной обратной документной частоты, вследствие чего скалярное произведение воспроизводит оценку BM25. Такое представление позволяет обслуживать оба канала единым механизмом предварительной выборки с последующим слиянием по правилу (1).

3. Экспериментальная оценка

3.1. Методика

Построен синтетический корпус, воспроизводящий структуру данных логистической компании: 30 складов, включая намеренно схожие наименования, 36 месяцев, четыре типа груза — всего 4320 записей. Записи сгруппированы по складам и разбиты на фрагменты по 20 строк, что дало 240 целевых документов.

К целевым документам добавлен шум в доле 60 %, разделённый на два класса. Лёгкий шум (60 % объёма шума) — тематически посторонние

материалы: регламенты приёмки, инструкции, кадровые положения, классификаторы. Трудные негативы (40 %) воспроизводят то, что реально вытесняет правильный ответ: тот же объект за другой период, другой объект за тот же период, объект со схожим наименованием, факт вместо прогноза, иной показатель при той же структуре полей, почти-дубликат. Итоговый корпус — 600 документов.

Сформировано 60 запросов шести типов по 10 в каждом: с точным наименованием объекта и периода; в разговорной формулировке; через внутреннее сокращение; с опечаткой; в виде перифраза без лексического пересечения с документом; с несуществующим объектом и периодом. Первые пять типов (50 запросов) имеют эталонные документы, вычисленные непосредственно из данных; шестой служит для оценки ложных срабатываний. Характеристики стенда приведены в табл. 1. Выдача детерминирована: три повтора каждого прогона дали идентичные результаты.

Таблица 1 — Характеристики экспериментального стенда

Параметр	**Значение**
Исходных записей	4320
Целевых документов / шумовых / всего	240 / 360 / 600
Размер фрагмента	20 строк
Запросов (из них с эталоном)	60 (50)
Кодировщик	intfloat/multilingual-e5-small, 384 измерения
Словарь лексического канала / средняя длина документа	5790 термов / 409,8 токена
Векторное хранилище	Qdrant 1.18, локальный режим
Вычислительная платформа	8 ядер CPU, 15,7 ГБ ОЗУ, без GPU

Применение уменьшенной модели вместо полноразмерной обусловлено ограничениями стенда: замер показал 991 мс на документ против 89 мс, то есть более чем одиннадцатикратную разницу, делавшую полный прогон

неосуществимым. Абсолютные значения метрик плотного канала следует поэтому считать нижней оценкой.

3.2. Результаты по режимам извлечения

Сопоставлены четыре режима: плотный поиск, лексический и два варианта гибридного слияния. Необходимость двух вариантов выявилась в ходе работы: штатная реализация слияния в используемом хранилище не принимает константу сглаживания и применяет значение, эквивалентное $\kappa = 1$ в записи (1), тогда как рекомендованным в исходной работе [5] является $\kappa = 60$. Второй вариант реализован на стороне клиента поверх тех же ветвей предварительной выборки. Результаты приведены в табл. 2.

Таблица 2 — Качество извлечения по режимам (50 запросов с эталоном)

Режим	Hit@1	Hit@3	Hit@10	MRR@10	Задержка, мс
Плотный поиск	0,20	0,46	0,84	0,369	1,4
Лексический (BM25)	0,38	0,56	0,62	0,467	9,4
Гибридный, штатное слияние	0,22	0,56	0,74	0,408	13,1
Гибридный, $\kappa = 60$	0,30	0,60	0,84	0,480	12,3

Ни один режим не доминирует по всем метрикам. Гибридное слияние при $\kappa = 60$ даёт наилучшее ранжирование (MRR@10 = 0,480) и наибольшую полноту наравне с плотным поиском (Hit@10 = 0,84), однако по точности первой позиции уступает лексическому режиму (0,30 против 0,38). Распространённое допущение о безусловном превосходстве гибридизации, таким образом, не подтверждается. Существенно и влияние константы сглаживания: переход от штатного значения к $\kappa = 60$ повышает Hit@10 с 0,74 до 0,84, а MRR — с 0,408 до 0,480, то есть параметр, обычно принимаемый по умолчанию, оказывает эффект, сопоставимый со сменой режима поиска.

3.3. Результаты по типам запросов

Разрез по типам запросов (табл. 3) объясняет усреднённые показатели и содержит два результата, противоречащих исходным ожиданиям.

Таблица 3 — Hit@1 / Hit@10 по типам запросов

Режим	Точное наимен.	Разговорный	Сокращение	Опечатка	Перифраз
Плотный поиск	0,30 / 1,00	0,30 / 1,00	0,10 / 0,60	0,20 / 0,90	0,10 / 0,70
Лексический	0,80 / 1,00	0,80 / 1,00	0,00 / 0,10	0,30 / 1,00	0,00 / 0,00
Гибридный, штатное	0,30 / 1,00	0,40 / 1,00	0,10 / 0,40	0,20 / 1,00	0,10 / 0,30
Гибридный, к = 60	0,50 / 1,00	0,60 / 1,00	0,20 / 0,90	0,20 / 1,00	0,00 / 0,30

Первый результат относится к запросам с внутренними сокращениями. Предполагалось, что они составляют слабое место плотного канала, однако провал показал лексический: Hit@10 = 0,10. Анализ ошибок выявил причину — верхние позиции занимает документ-классификатор, содержащий расшифровку сокращений и потому лексически совпадающий с запросом идеально, но не содержащий самих показателей. Плотный поиск, ориентируясь на смысл запроса целиком, этой ловушки избегает (0,60), а гибридное слияние при к = 60 даёт 0,90. Результат следует трактовать с оговоркой: в промышленном корпусе справочник сокращений является легитимным источником, и его извлечение не во всяком сценарии должно считаться ошибкой.

Второй результат касается перифразов, сформулированных без единого общего термина с эталонным документом (условие проверено автоматически). Здесь гибридный режим уступает плотному более чем вдвое по полноте: 0,30 против 0,70. Механизм прозрачен: лексический канал на таких запросах закономерно даёт нулевой результат, но слияние всё равно включает его кандидатов в объединённый список на равных правах, вытесняя корректные

плотные ответы. Это отрицательный результат для гибридизации, ограничивающий область её безусловного применения.

3.4. Статистическая значимость и анализ ошибок

Объём выборки требует осторожности при сопоставлении режимов. Доверительные интервалы Уилсона приведены в табл. 4.

Таблица 4 — 95 %-е доверительные интервалы Уилсона (n = 50)

Режим	Hit@1	ДИ для Hit@1	Hit@10	ДИ для Hit@10
Плотный поиск	0,20	[0,11; 0,33]	0,84	[0,72; 0,92]
Лексический	0,38	[0,26; 0,52]	0,62	[0,48; 0,74]
Гибридный, штатное	0,22	[0,13; 0,35]	0,74	[0,60; 0,84]
Гибридный, к = 60	0,30	[0,19; 0,44]	0,84	[0,72; 0,92]

Интервалы для Hit@1 у всех режимов взаимно перекрываются, поэтому наблюдаемое преимущество лексического поиска по точности первой позиции должно трактоваться как тенденция, а не как установленный факт. По полноте разделение выражено сильнее: интервал гибридного режима при к = 60 [0,72; 0,92] лишь незначительно пересекается с интервалом лексического [0,48; 0,74]. Наиболее надёжно установленным следует считать не превосходство какого-либо режима, а отсутствие доминирования.

Анализ вытеснения эталонных документов дал результат, важный для проектирования: основным источником помех оказались не синтетические трудные негативы, а другие документы самого целевого реестра — соседние объекты и периоды (21 позиция из 24 при плотном поиске, 14 из 24 при гибридном). Главным конкурентом правильного ответа выступает, таким образом, сам аналитический корпус. Отсюда следует, что период и объект учёта целесообразно выносить в фильтры по полям метаданных, а не полагаться на текстовое совпадение: документы, отличающиеся только периодом, почти идентичны по тексту и принципиально неразделимы средствами поиска.

Оценка ложных срабатываний показала, что порога отсечения пустой выдачи не существует. Запросы о заведомо отсутствующем объекте дают косинусную близость верхней позиции в диапазоне 0,863–0,915, тогда как содержательные запросы — 0,822–0,921; диапазоны перекрываются, а лучший достижимый порог обеспечивает сбалансированную точность лишь 0,77–0,78. Для гибридного режима абсолютный порог бессмыслен по построению, поскольку оценка (1) является функцией рангов. Отсечение нерелевантной выдачи требует, следовательно, отдельного механизма.

4. Согласование размера фрагмента с контекстным окном

Наиболее существенным для практики оказалось ограничение, обнаруженное при контроле длины документов. При размере фрагмента 20 строк медианная длина документа составила 1388 токенов при контекстном окне кодировщика в 512 токенов; усечению подверглись 58,5 % корпуса. Плотное представление фактически описывает поисковые маркеры и первые строки фрагмента, тогда как остальные периоды для него не существуют. Лексический канал усечения не имеет, чем и объясняется его преимущество на запросах с точным наименованием. Приведённые в табл. 2 значения плотного и гибридного режимов следует поэтому интерпретировать как полученные в условиях частично повреждённого индекса.

Измерения позволяют перейти от констатации к количественному требованию. При медианной длине 1388 токенов и блоке маркеров объёмом около 40 токенов на одну строку таблицы приходится приблизительно 67 токенов, откуда следует, что плотному представлению доступны маркеры и первые семь строк — оценка, совпадающая с наблюдаемой картиной выдачи. Чтобы фрагмент из 20 строк укладывался в окно целиком, бюджет одной строки не должен превышать примерно 23 токенов, то есть требуется сжатие текстового представления почти втрое.

Источник избыточности устанавливается разбором строки. Каждая строка воспроизводит полный набор имён полей, причём русскоязычные

наименования при субсловной токенизации дают по четыре-шесть токенов каждое. В результате около двух третей объёма фрагмента занимают не данные, а многократно повторённые имена столбцов. К этому добавляется дублирование значений, постоянных в пределах документа: наименование объекта повторяется в каждой строке, хотя группировка выполнена именно по нему.

Отсюда следуют три требования к конвейеру подготовки. Во-первых, размер фрагмента должен задаваться в токенах, а не в строках: число строк является косвенным параметром, значение которого зависит от ширины таблицы и потому не переносится между наборами данных. Во-вторых, для плотного представления целесообразен позиционный формат, при котором имена полей объявляются однократно в заголовке документа, а строки содержат только значения; постоянные в пределах документа атрибуты выносятся в заголовок. Оценка показывает, что этого достаточно для достижения требуемого бюджета. В-третьих, полная развёртка с явными именами полей сохраняется в тексте лексического канала, где ограничение контекстного окна отсутствует, а повторение терминов, напротив, работает на точность совпадения.

Существенно, что введённое в п. 2.4 разделение текстов для двух каналов, изначально мотивированное независимой оптимизацией, оказывается тем самым необходимым условием корректной работы. Каналы имеют принципиально разные ограничения — контекстное окно в одном случае и его отсутствие в другом, — и попытка обслуживать оба единым текстом неизбежно ведёт к потере содержимого. Ограничение контекстного окна в 512 токенов свойственно и полноразмерным моделям семейства, поэтому вывод не зависит от использованного на стенде уменьшенного кодировщика.

5. Угрозы валидности результатов

Корпус носит синтетический характер: он воспроизводит структуру и типовые наименования реальных данных, но не их естественное распределение. Частотность обращений к разным объектам, орфографическая вариативность формулировок и соотношение классов шума заданы конструктивно, а не измерены на промышленном хранилище. Абсолютные значения метрик поэтому характеризуют сопоставляемые режимы в равных условиях, но не переносятся на производственный корпус напрямую.

Оценка выполнена на уменьшенном кодировщике с размерностью 384 вместо применяемого в эксплуатации с размерностью 1024. Поскольку различие затрагивает именно плотный канал, результаты недооценивают плотный и гибридный режимы относительно лексического. Использована единственная реализация корпуса при фиксированном начальном значении генератора случайных чисел, что обеспечивает воспроизводимость, но не позволяет оценить дисперсию метрик по реализациям. Детерминированность подтверждена лишь для поисковой процедуры.

Индексация выполнена на корпусе из 600 документов, при котором приближённый поиск ближайших соседей практически вырождается в полный перебор; измеренные задержки и полнота не отражают поведение индекса на промышленном масштабе, где вступают в силу параметры графового индекса и связанный с ними компромисс между скоростью и полнотой. Наконец, конфигурация не включает этап переранжирования кросс-энкодером, применяемый в современных конвейерах поверх гибридной выдачи и способный изменить соотношение режимов, поскольку воздействует именно на верхние позиции — область, где в настоящем эксперименте различия наименее определённы.

Заключение

Предложен метод преобразования табличных аналитических данных в документное представление, включающий группировку по естественному

ключу, разбиение на фрагменты, явное разграничение фактических и прогнозных значений, добавление поисковых маркеров, раздельное формирование текстов для плотного и лексического каналов и нормализацию с лемматизацией русскоязычных словоформ.

Экспериментальная оценка на корпусе из 600 документов с контролируемой долей трудных негативов показала, что ни один режим извлечения не доминирует: гибридное слияние обеспечивает наилучшее ранжирование (MRR@10 = 0,480) и полноту (Hit@10 = 0,84), уступая лексическому поиску по точности первой позиции. Получены два результата, ограничивающие распространённые допущения: гибридизация ухудшает выдачу на запросах без лексического пересечения с документом, а оценка близости не позволяет построить порог отсечения нерелевантной выдачи. Установлено также, что константа сглаживания при слиянии влияет на качество не менее существенно, чем выбор самого режима.

Основным практическим результатом является выявленное рассогласование размера фрагмента с контекстным окном кодировщика, приводящее к усечению 58,5 % корпуса, и выведенное из измерений количественное требование к бюджету токенов на строку таблицы. Определяющим фактором качества извлечения оказывается, таким образом, подготовка данных, а не выбор режима поиска. Направлениями дальнейшей работы являются подбор размера фрагмента с непосредственным контролем длины в токенах, введение фильтрации по объекту учёта и периоду на уровне метаданных, а также оценка на полноразмерном кодировщике с этапом переранжирования.

Список литературы

1. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.

2. Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2024. arXiv:2312.10997.

3. Karpukhin V., Oguz B., Min S. et al. Dense Passage Retrieval for Open-Domain Question Answering // Proceedings of EMNLP. 2020. P. 6769–6781.

4. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. 2009. Vol. 3, No. 4. P. 333–389.

5. Cormack G.V., Clarke C.L.A., Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods // Proceedings of the 32nd International ACM SIGIR Conference. 2009. P. 758–759.

6. Liu N.F., Lin K., Hewitt J. et al. Lost in the Middle: How Language Models Use Long Contexts // Transactions of the Association for Computational Linguistics. 2024. Vol. 12. P. 157–173.

7. Wang L., Yang N., Huang X. et al. Multilingual E5 Text Embeddings: A Technical Report. 2024. arXiv:2402.05672.

8. Qdrant: Vector Database Documentation [Электронный ресурс]. URL: https://qdrant.tech/documentation/ (дата обращения: 27.07.2026).

9. Malkov Yu.A., Yashunin D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020. Vol. 42, No. 4. P. 824–836.

10. Barnett S., Kurniawan S., Thudumu S. et al. Seven Failure Points When Engineering a Retrieval Augmented Generation System. 2024. arXiv:2401.05856.