12. Количественная оценка (llm-as-a-judge, доп тесты)

2.9

Качество прогноза: анализ корреляции human eval и автометрик

Цель теста. Определить возможность использования автометрики для автоматического скоринга ответов LLM (замены асессоров), скоринга альтернативных моделей и аугментаций.

Необходимые условия для проведения теста:

Необходимы данные: метки асессоров, истинно верные ответы (ground truth).

Алгоритм расчетов

- Рассчитываются метрики близости между ответами LLM и идеальными ответами (ground truth).
- Для каждой из автометрик рассчитывается корреляция Спирмена с метрками релевантности ответов, полученных от асессоров (human eval).

Наименование метрики	Значение корреляции
METEOR	0.9
BLEU	0.81
ROUGE-1	0.43
LLM-as-a-judge (gemma3)	0.8

https://pypi.org/project/jury/
https://github.com/Yale-LILY/SummEval

Традиционно (эволюционно) выделяются 3 метода автооценки

- Референсные (базирующиеся на сравнении статистик или эмбеддингов, посчитанных на ответах оцениваемой модели и GroundTruth)
- Безрефенсные (базирующиеся на сравнении статистик или эмбеддингов, посчитанных на ответах оцениваемой модели и исходным запросом)
- LLM-based (базирующиеся на внутренних знаниях LLM-оценщика для оценки корректности ответа испытуемой LLM, могут использовать или не использовать референсы).

https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

Постановка задачи

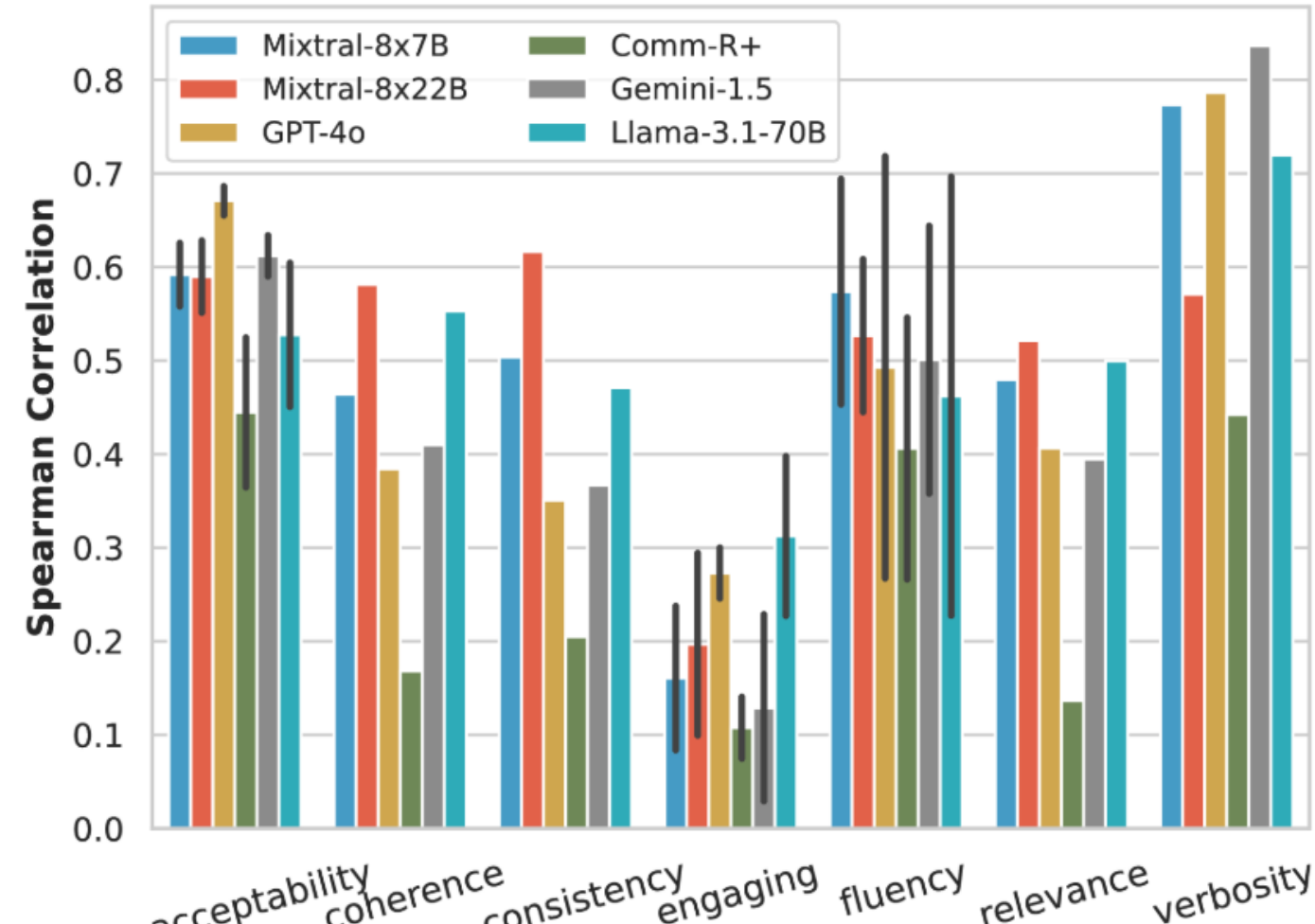
1. LLM1 стат. значимо хуже/лучше/также как LLM2 на тестовом сете? [частное]
2. Какая метрика качества LLM1 и LLM2 на тестовом сете? [общее]

Каковы наши шансы?

https://arxiv.org/html/2406.18403v2

https://arxiv.org/pdf/2410.10934

agent-as-judge –в it домене все хорошо



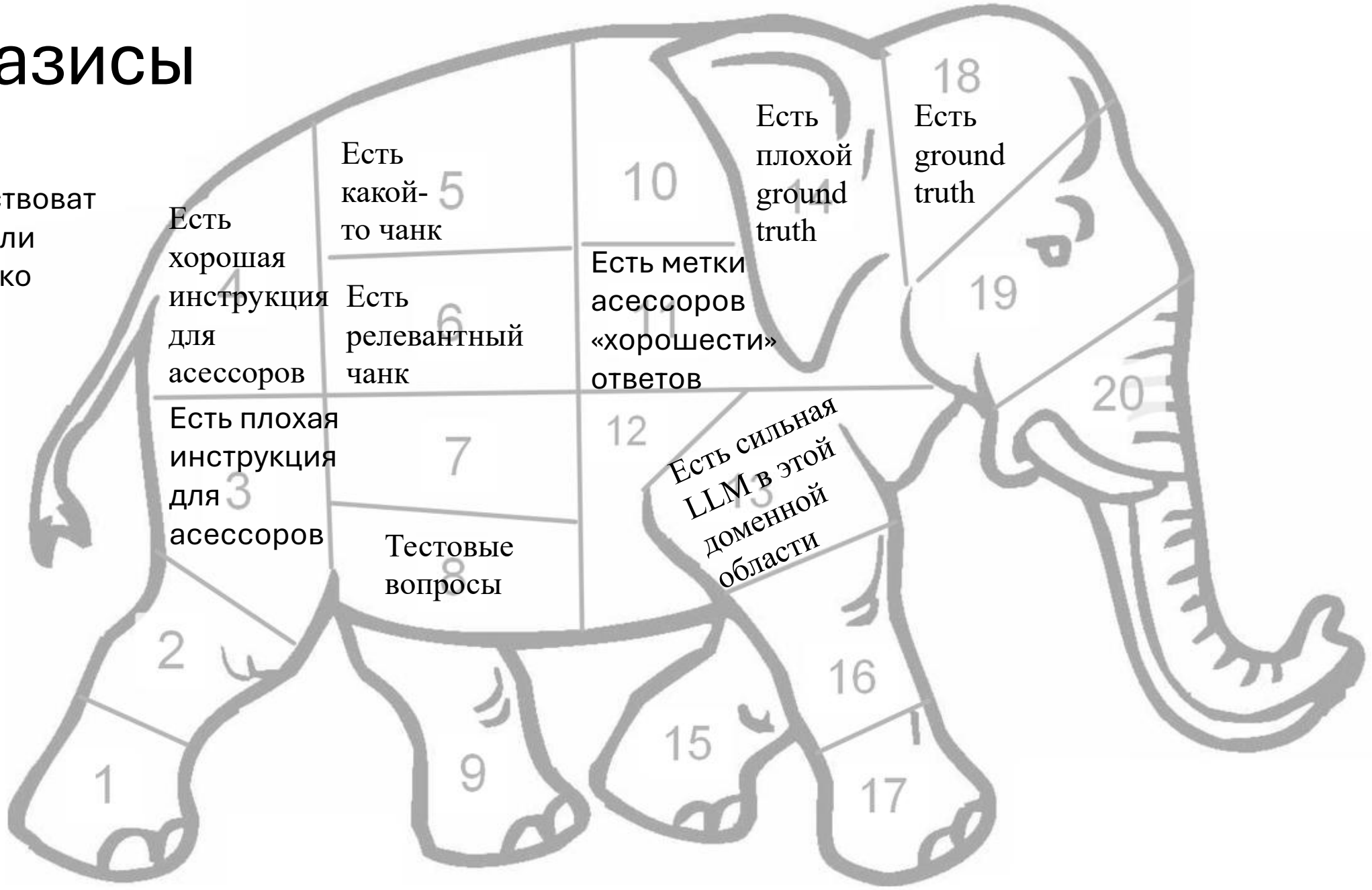
Характеристики тестового сета

- **Базис** (что-то, что берем за основу при оценке качества)
- Качество базиса
- Процесс порождения данных: пром / руками написали пром-подобные вопросы
- Объем тестового сета (чем больше сет, тем дороже генерить ground truth; чем меньше сет, тем хуже стат. значимость)
- Тип ml-задачи и особенности бизнес-задачи

https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics - про разделение метрик на референсные, безреыеренсные и модельные

2. Базисы

Может наличествовать один или несколько



Базисные точки для оценки as is – не зависит от архитектуры пайплайна (черный ящик)

Базис	Опция	Наличие
Ground Truth	Хороший	Желательно, не всегда возможно
	Неоднозначный (один из возможных верных вариантов)	
	Плохой	
Инструкция для асессоров	Хорошая для LLM	Must have
	Хорошая для людей	
	Плохая (нечеткие критерии, нет few shot)	
Внутренние знания LLM	Для методов llm-as-judge	Желательно, не всегда возможно
Метки качества ответа	С высокой согласованностью (однозначную разметку можно использовать как ground truth)	Желательно, не всегда возможно
	С невысокой согласованностью	Must have
Тестовые вопросы	«классика» из ПРОМ-процесса(лучше отражают потребности)	Желательно, не всегда возможно
	«классика» - синтетика	Must have
	Затравки для диалога/конфиги для LLM-симуляции (Ping-Pong сет)	
Ответы на тестовые	Получены текущими, возможно неоптимальными, промтпами	Must have

Базисные точки для оценки as is – используем сильные стороны пайплайна

Методы дообогощения промптов (при хорошем качестве работы) могут также дать полезную инфу для оценки корректность финального ответа

Базис	Опция	Наличие
Чанк	С разметкой релевантности	Желательно
	Без разметки релевантности	Must have
Классификатор запроса*	С метками корректности тем	Желательно
	Без меток корректности тем	Must have

*Если классифиактор запроса отработал неверно, то последующий чанк будет искаться в неверной БД => сразу - итоговый ответ. Мб найдем противоречие с разметкой асессоров

Маппинг ML-задач и критериев

Критерии качества	L1	L2						
	Domain knowledge	Brainstorming	Generation	Question Answering	Chat	Rewrite и editing	Text translation	Summarization
Фактология	Нужно		Нужно?	Нужно	Нужно?			
Полнота			Нужно?					Авто
Релевантность			Нужно?					Авто
Краткость								Авто
Само согласованность					Авто			Авто
Семантическая близость		Авто			Авто	Авто	Авто	

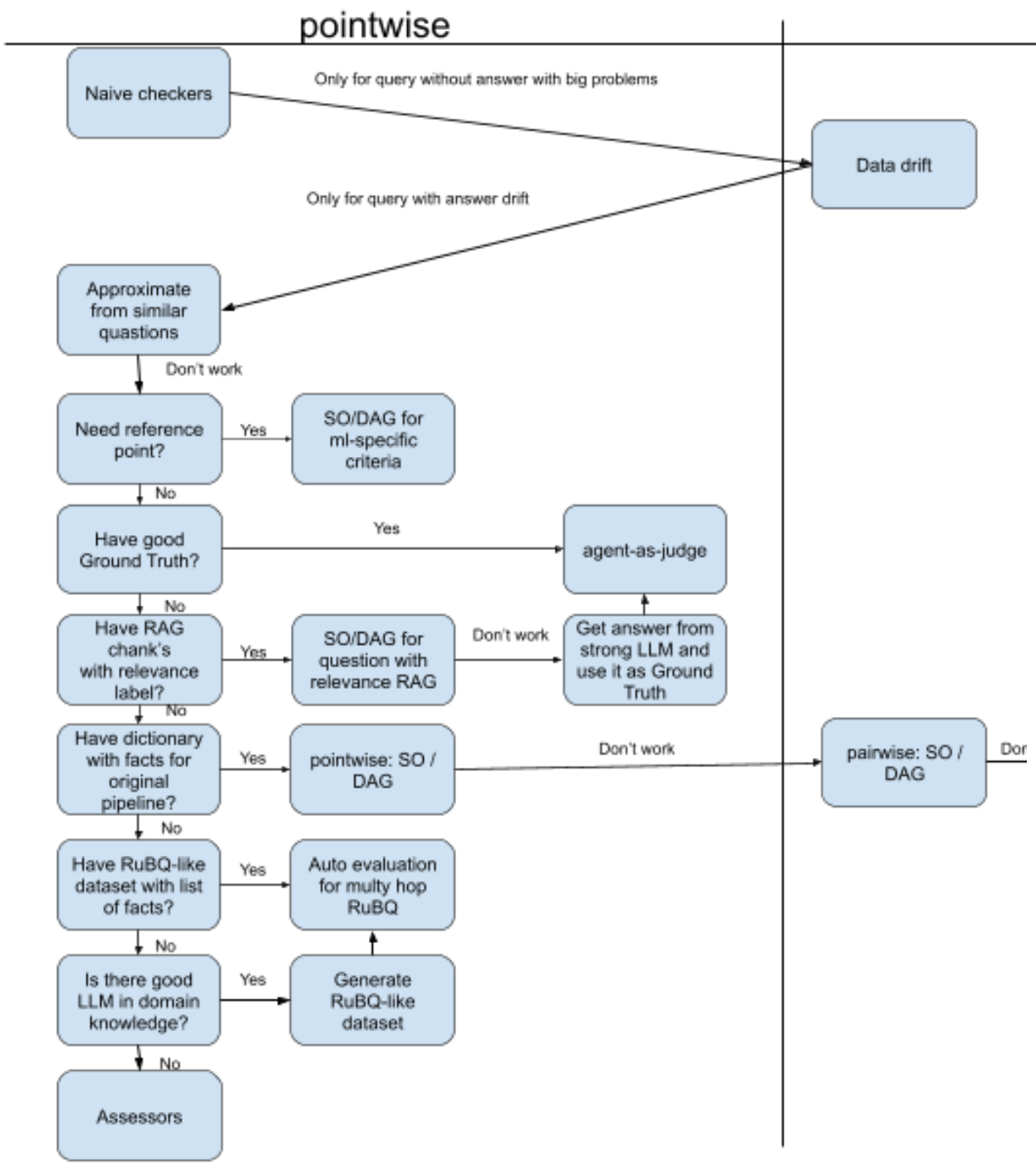
Какие критерии качества требуют базиса?

Aspect Relevance	Актуальность информации с исторической точки зрения
Missing	Полнота: все ли важные факты упомянуты в ответе
Hallucination	Отсутствие в ответе ложных (выдуманных) фактов
Оптимальный подход	Использован оптимальный подход к решению задачи.
Качество рассуждений	Логика и глубина рассуждений модели
Точность и аргументированность оценки	оценивает способность модели аргументировать и осуществлять оценку на должном экспертном уровне. — Оценка должна быть основана на проверке тезисов, фактов, гипотез, аргументов, их уместности, а также логической стройности и соответствия жанровой специфике. — Оценка должна быть подробно и качественно аргументированной. Между оценкой и аргументацией оценки не должно быть противоречий.

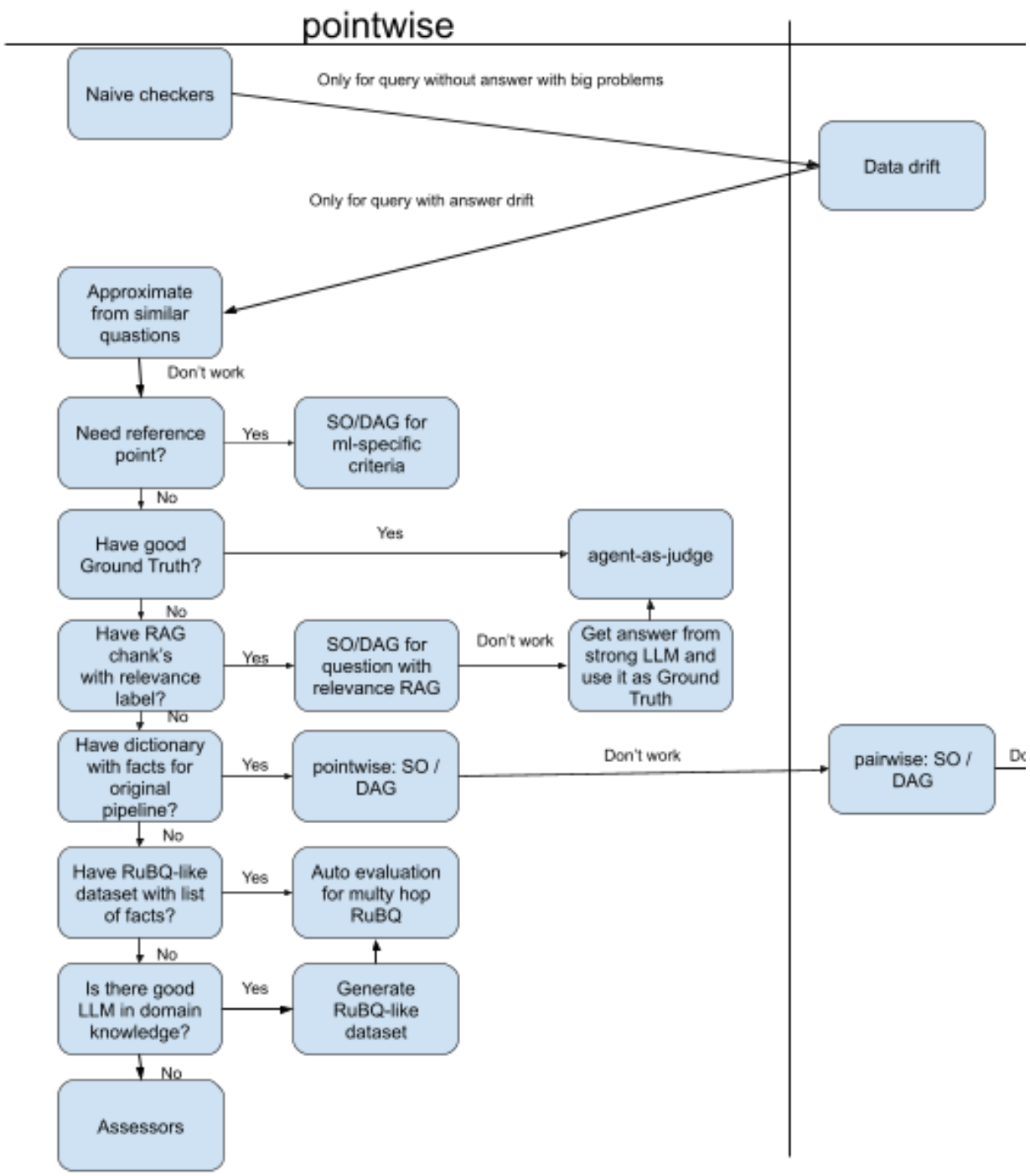
А какие критерии качества можно без базиса измерить?

Consistency	Согласованность (непротиворечивость) фактов внутри одного ответа
Factuality	Фактичность: ответ абстрактный или точный?
Thematic relevance	Совпадает ли тематика запроса и тематика ответа
Irrelevant	Отсутствие в ответе не выдуманных, но лишних (дополнительных) фактов, не относящихся к вопросу Отсутствие избыточной информации (воды)
Правильность извлечения информации	Данный критерий оценивает способность модели выписать из текста какую-либо информацию по запросу пользователя. Проверяем, что выписано только то, что требуется. Нет лишних выписанных элементов.
Непротиворечие фактам реального мира	Актуальность и точность фактической общеизвестной информации
Завершённость	План действий должен быть завершенным, приводить пользователя к запрошенной цели

В POLLUX – за большим количеством критериев



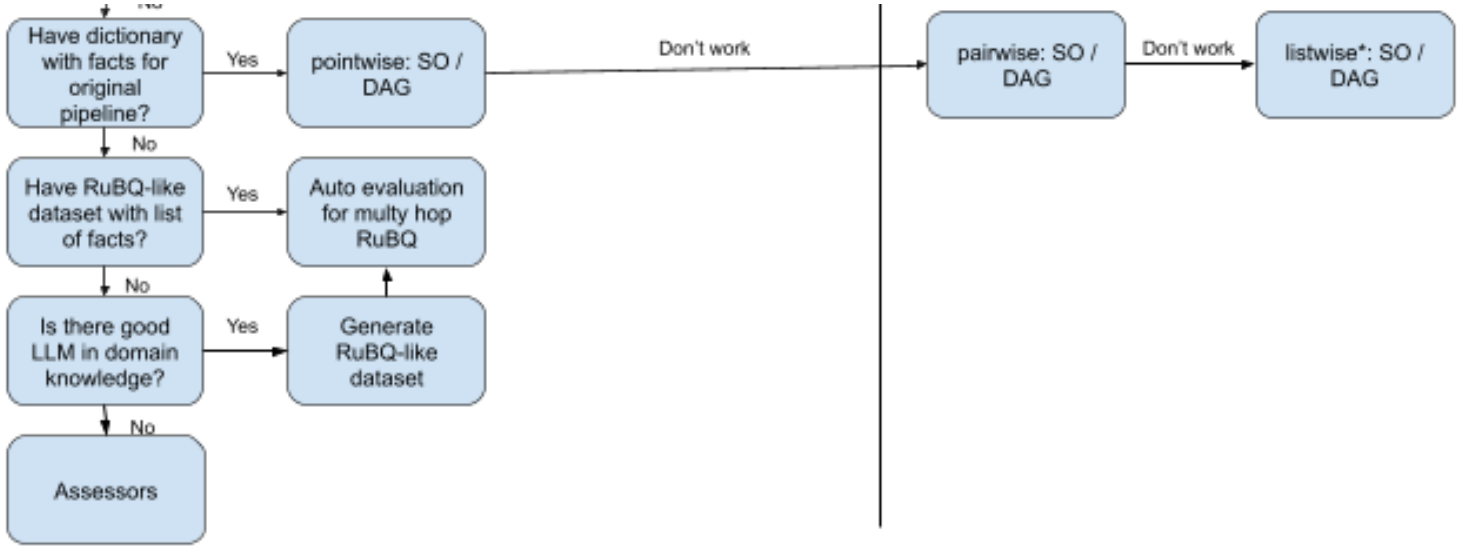
1. Применить **Naive checkers** (простые фильтры качества) [https://www.promptfoo.dev/docs/getting-started/#prompt-quality , https://arxiv.org/html/2309.07601v3] на ответы LLM. В дальнейший анализ попадают только те кейсы, которые удовлетворяют этим простым и быстрым фильтрам
2. Применить оценку **Data drift*** (Приложение 1) для ответов на одни и те же вопросы. Если ответы практически не отличаются, то нет необходимости анализировать их качество далее. Если мала** (Приложение 2) доля вопросов, для которых отличаются ответы, то для двух LLM признается паритет качества.
3. Если среди вопросов есть семантические дубликаты и ответы на эти вопросы похожи, то метки корректности этих ответов должны быть схожи. Эта гипотеза позволяет экстраполировать размеченные ответы на неразмеченные (**аппроксимировать ответов оценку между схожими вопросами**).
4. Для вопросов, ответы на которые отличны, нужно понять характер этих изменений. Дальнейшая логика работы зависит от критерия, который необходимо оценить. Критерии **без фактологии** оцениваются без опорных точек при помощи специальных моделей / LLM согласно методам в Приложении 4.



5. Для остальных вопросов, ответы на которые отличны, проводится сравнение ответов с **Ground Truth**, если они есть. На основании близости к ground truth определяется корректность ответа [https://arize.com/docs/phoenix/evaluation/how-to-evals/running-pre-tested-evals/ai-vs-human-groundtruth].
6. Если Ground Truth нет (когда составление полного и однозначного дорого, они меняются со временем или задача творческая), но есть р**азметка релевантности чанков от RAG**, то в этих релевантных чанках есть неявный Ground Truth с которым можно сравнить ответы моделей и получить корректность ответа LLM. Этот поиск проводится при помощи методов, указанных в приложени 3.
7. Если выделение неявного ground truth из релевантных чанков не приводит к качественной автооценке (например, в силу специфичной лексики), то предлагается **ответить на исходный вопрос при помощи сильной*** (LLM - SOTA в доменной области знаний) LLM и **использовать ее ответы как Ground Truth**, чем задача сводится к предыдущей.
8. При отсутствии разметки релевантных чанков прибегают к использованию **словарей фактов**. LLM при помощи четко описанных критериев качества в методе **G-Eval, DAG**: [https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation ,https://medium.com/@pedroazevedo6/how-to-use-deepeval-with-custom-llm-like-bedrock-c8c0c583abeb , https://arxiv.org/abs/2303.16634] оценивается **точечная** корректность ответа.

pointwise | pairwise

9. Если предыдущий метод недостаточно чувствителен, то прибегают к **попарному** сравнению ответов по конкретным критериям качества методами **G-Eval, DAG**: [https://arxiv.org/html/2306.05685v4 ,https://arxiv.org/html/2402.10524v1]
10. Если предыдущий метод недостаточно чувствителен, то проводят р**анжирование ответов модели** по качеству работы методами **G-Eval, DAG** [https://arxiv.org/html/2402.14860v2], в случае, когда среди вопросов есть семантические дубликаты (и именно с ответами на эти вопросы сравнивается оцениваемый ответ). Логика работы близка с 3 п., однако требует больше вычислительных ресурсов: каждая комбинация списка подается судье отдельно.
11. Однако не всегда есть подобный словарь, критерии качества могут быть прописаны недостаточно четко. В этом случае предлагается для каждого вопроса составить список фактов, что должны быть обязательно упомянуты для полного ответа.Если ответ LLM не содержит какого-то факта, то он признается неполным; если ответ LLM содержит какого-то факт вне списка фактов, то он признается ненадежным из-за потенциальных галлюцинаций. Данный процесс проверки аналогичен оценки **multi-hop query RuBQ** dataset.
12. В случае, когда нет подобного списка фактов, предлагается Генерируем закрытые вопросы к развернутому ответу (или рассуждениям модели). Так немного упрощаем задачку, контролируем, что факты корректны, которыми оперирует ллм (на основе знаний этих фактов другой ллм).
13. Если подобной LLM нет, то необходимо обращаться к **асессорам за side-by-side LAP** разметкой.



Limitations

- Garbage in - garbage out
- Если не получилось описать разницу в тонких материаях, то LLM ее также не почувствует
- LLM не обладает всеми экспертными знаниями, что мы от нее ждем, поэтому может потребоваться привлечение дополнительных инструментов (search, deep research, ...)
- Необходимы базисы для фактологических критериев

How to improve your judge

- Не используй судей там, где они излишни. Для сравнения качества двух ответов на один и тот же запрос можно воспользоваться методами поиска data drift. Если Data drift нет, то и изменения качества нет. Для тех вопросов, где data drift есть, можно использовать более умные методы оценки, чтобы понять характер изменений (положительный или отрицательный). Также не нужно оценивать через LLM те критерии, которые хорошо оцениваются специализированными моделями.
- Используй наивные чекеры качества ответа [https://www.promptfoo.dev/docs/getting-started/#prompt-quality , https://arxiv.org/html/2309.07601v3]. Если ответ стал менее надежен / менее уверен, то это может повлечь за собой и ухудшение качества.
- Используй независимо разных судей [https://arxiv.org/html/2404.18796v2]
- Формируй критерии качества максимально просто, атомарно. Лучше много простых критериев, чем мало составных.
- Избегай большой гранулярности критериев. Чаще всего, хватит шкалы 0 -1 (не критерий выполнен / выполнен) или 0-1-2 (критерий не выполнен / выполнен частично / выполнен полностью)
- Подробно определи (опиши) каждый критерий. Подробно определи, когда ставить 0, а когда 1.
- Few-shot и few-shot+chain-of-though помогают как llm, так и асессорам ставить более надежные метки. Предпочтительно иметь примеры для каждого критерия для каждой возможной метки.

Результаты: 1 кейс

- **0.73** – средняя корреляция по доверенным критериям
- 0.65 – согласованность асессоров
- DAG + переписанная инструкция

Критерий	Корреляция	Причины
Галлюцинации	0.4	Оценка занижена тем, что в системном промпте не указан текущий год, также у гигачата нет информации о том, является ли населенный пункт городом или ПГТ). Gemini также находила те фактические ошибки, которые не смогли определить асессоры, например, в инструкции сказано не учитывать МЦД и МЦК как метрополитен.
Полнота	1	
Наличие запрещенной информации	1	Мат, радикальные призывы
Самосогласованность	1	
Стилистическая согласованность	0.8	
Обоснованность эпитетов	1	
Новизна	0.6	Для оценки критерия использовался DAG на GigaChat, потенциально, использование регулярных выражений может увеличить коэффициент корреляции
Запрещенные формулировки о газе	0.56	Встречаются тонкие различия «в квартире проведено газовое оборудование» != «в квартире есть газ», которые не детектируются через ллм.
Орфография	0.2	Gemini плохо находит ошибки в использовании неправильного склонения и использования

Результаты: 2 кейс

- **0.63** – средняя корреляция по доверенным критериям
- 0.6 – согласованность асессоров
- DAG + переписанная инструкция

Критерий	Корреляция с метками асессоров	Комментарий
Галлюцинации	0.15	Проблема разметки от асессоров, по точечным наблюдениям оценка от llm более надежна
Полнота	nan	Асессоры как-будто бы закрыли глаза на её оценку, абсолютно все ответы отмечены полными
Оценка тональности	0.21	Асессоры оценивали, скорее не саму оценку тона, а то, что она правильно оформлена, нет двух разных оценок для одного текста
Оценка тональности (соответствие формату ответа)	1	Асессоры оценивали только соответствие тональности формату, в комментариях есть случаи, когда написано "тон скорее нейтральный, а не положительны1", и ошибка не отмечена
Безвредность	1	
Стиль	0.1	Плохая инструкция, не раскрыто понятие «делового стиля», за счет чего у асессоров и ллм разное понимание целевого стиля
Краткость	1	
Использование запрещенных терминов и сокращений	0.27	Плохая формулировка в инструкции. ЛЛМка не может понять, разрешен ли какой-то термин (уровень его общеупотребимости). Изменил промпт, добавил в список разрешенных те термины, которых там не было, но которые асессоры считали допустимыми (например, облигации).
Наличие сравнительных суждений в ответе	nan	Асессоры отметили, что в текстах нет сравнительных суждений. Инструкция сама по себе плохая, потому что сравнение может быть в самих новостях, из-за чего LLM их отмечает как ошибочные по этому критерию. Отбросить эти суждения нельзя, потому что теряется смысл исходной новости. Сейчас стараюсь исправить эту проблему изменением инструкции и промпта
Наличие в summary финансовых рекомендаций	1	
Общая оценка релевантности	0.15	Агрегация от остальных критериев, есть биас из-за разметсчиков

Вопросы

2.11

Качество прогноза: уверенность LLM в ошибочных ответах

Цель теста. Исследовать возможность отделения верных ответов LLM от неверных на основании уверенности LLM в ответе.

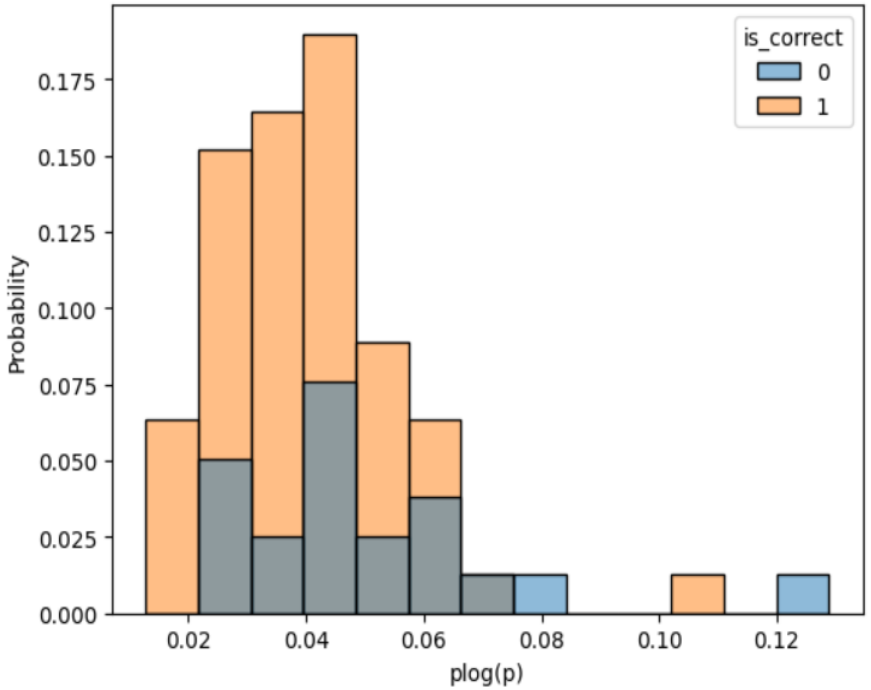
Необходимые условия для проведения теста: Тест применим тогда и только тогда, когда **оценка корректности** ответа LLM укладывается **в дискретную шкалу**.

Алгоритм расчетов. В рамках теста используется следующая гипотеза: уверенность LLM в своем ответе тем ниже, чем ниже вероятность того, что этот ответ верный (галюцинации друг друга не повторяют). Алгоритм:

- Производится инференс LLM на тестовом сете при гиперпараметрах разработчика.
- При увеличенной температуре производится инференс на тех же вопросах N раз.
- Подсчитывается уверенность p ответа LLM на каждый тестовый запрос.
- Для каждого тестового запроса подсчитывается энтропия как -p*log(p).
- Визуализируется распределение полученной энтропии в разрезе верности ответа LLM, полученного в п.1.

Для разных типов тестовых сетов уверенность LLM в ответе p считается по-разному:

- Для датасетов, предполагающих ответ, ограниченный дискретной областью принимаемых значений, подсчитывается количество раз, когда полученный при повышенной температуре ответ совпал с ответом LLM, полученным в п.1. Полученное значение нормируется на N.
- Для остальных датасетов подсчитывается косинусное расстояние между эмбеддингами GigaChat, построенными на каждой паре полученных ответов из п.1 и п.2. Данная статистика усредняется в рамках каждого запроса.



3.1

Спецификация RAG: ключевая метрика качества

- Что такое релевантный пассаж?
- Нужны метки релевантности.

Mean precision at top	MP@k = $\frac{1}{K*N}\sum_{i=1}^{k} target_i$, где i - номер позиции в топе-выдачи, k – гиперпараметр метрики, N – количество user-ов.
Mean average precision at top	MAP@k = $\frac{1}{K}\sum_{i=1}^{k} target_i * MP@i$, где i-номер позиции в топе-выдачи, k – гиперпараметр метрики.
Normalized discounted cumulative gain at K	Если target принимает значения только 0 или 1, то: DCG@k = $\sum_{i=1}^{K} \frac{target_i}{log_2(i+1)}$ Если область допустимых значений включает более широкий набор, то: DCG@k = $\sum_{i=1}^{K} \frac{2^{target_i}-1}{log_2(i+1)}$ IDCG@k – максимальное значение DCG@k. Если target принимает значение только {0, 1}, то IDCG@k=$\sum_{i=1}^{K} \frac{1}{log_2(i+1)}$. NDCG@k = $\frac{DCG@k}{IDCG@k}$
По-запросный AUC	auc@k = $\frac{1}{N} * \sum_{i=0}^{N} RocAuc_i$, где N – количество user-ов, $RocAuc_i$ – стандартный ROC-AUC, измеренный для i-того пользователя на k наиболее релевантных для него item-ов. Данную метрику можно рассчитывать только при достаточном количестве наблюдений для каждого запроса.
Expected reciprocal rank at k	ERR@k = $\sum_{i=1}^{K} \frac{p_i * \prod_{j=1}^{i-1}(1-p_j)}{i}$, где p_h- вероятность того, что пользователь будет удовлетворен объектом с рангом h. Для бинарного target {0, 1} $p_h = target_h$, для небинарного: $p_h = \frac{2^{target_h}-1}{2^{max(target)}-1}$
Mean reciprocal rank at k	MRR@k = $\frac{1}{k}\sum_{i=1}^{K}\{0 \; if \; \frac{1}{min_{p<i}(p:\; target_p=1)} is\ not\ None\}$

3.2

Спецификация RAG: качество ответа LLM в зависимости от качества найденного пассажа

- Нужны метки релевантности.
- Пусть локальное качество RAG - оценка качества поиска пассажей к каждому запросу в отдельности (до агрегации по всем вопросам).
- Подсчитывается корреляция Спирмена между метками релевантности ответов LLM и локальным качеством RAG

Релевантность пассажа	LLM	Accuracy 1 пассаж
0	GIGAR	0.266667
1	GIGAR	0.590164
0	GigaChat-13b-4k-base:1.0.19.2	0.307692
1	GigaChat-13b-4k-base:1.0.19.2	0.525424
0	GigaChat-70b-4k-base:1.0.19.2	0.307692
1	GigaChat-70b-4k-base:1.0.19.2	0.627119
0	GigaChat:v1.2.19.2	0.153846
1	GigaChat:v1.2.19.2	0.203390
0	Saiga	0.307692
1	Saiga	0.271186

3.3

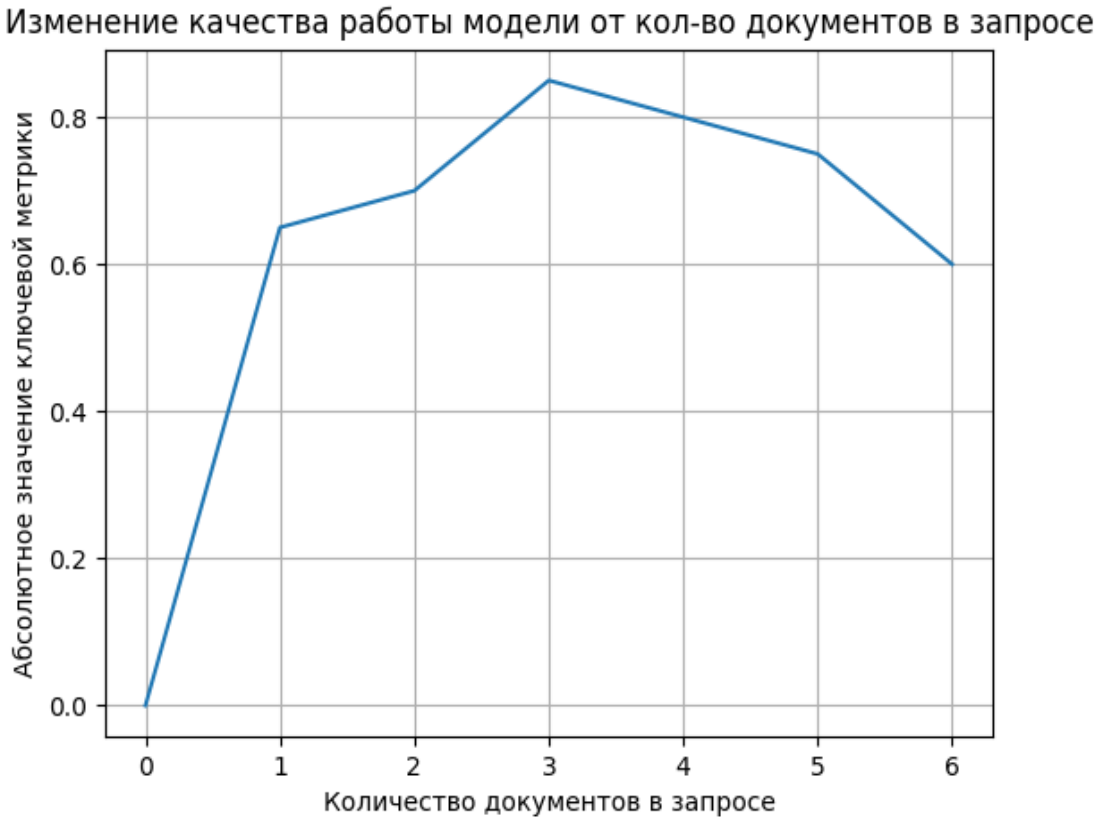
Спецификация RAG: динамики качества ответа LLM при увеличении размера ТОПа найденных документов

Цель теста. Определить динамику изменения ключевой метрики при увеличении топа найденных документов при возможности оценить качество работы.

Алгоритм расчетов.

Итеративно увеличивается количество релевантных пассажей. Перебор размера топа начинается с 0, достигает размера, уставленного разработчиком, и продолжается до тех пор, пока рост ключевой метрики качества не перестанет наблюдаться три итерации подряд.

Интерпретация результатов. Тест позволяет определить оптимальный размер топа релевантных документов для решения данной задачи.



3.4

Спецификация RAG: Unsupervised оценка качества построения эмбеддингов от контента для векторной базы данных

Цель теста. Определение качества эмбеддинг-модели в разрезе поиска релевантных пассажей без целевой разметки для базы знаний.

Алгоритм расчетов. Выбирается случайное подмножество чанков из базы знаний. На основе каждого из выбранных чанков с помощью сторонней LLM генерируется вопрос, ответ на который содержится в этом чанке. Опционально с помощью второй LLM сгенерированные вопросы и соответствующие им чанки оцениваются от 1 до 5 на возможность полностью ответить на вопрос Q на основе чанка C.

Для каждого из вопросов (опционально - оценка которых более чем заданный порог) с помощью эмбеддингов модели происходит поиск top-n (5 по умолчанию) ближайших чанков из исходной базы данных.

Измеряются ключевые метрики качества. Для каждого из вопросов **целевым (т.е. с меткой 1)** является чанк, на основе которого был сгенерирован этот вопрос (родительский чанк):

- Hit rate @ 1;
- Hit rate @ 5;
- Average place – среднее место родительского чанка при условии, что родительский чанк есть в выдаче (Информативно).

Интерпретация результатов. Низкие значения Hit rate @ k и высокое значение Average place сигнализируют о построении эмбеддингов низкого качества для векторной базы данных.

Критерий для Hit rate @ 1	**Критерий для Hit rate @ 5**
Hit rate @ 1 < 0.3	Hit rate @ 5 < 0.5
0.3 ≤ Hit rate @ 1 ≤ 0.4	0.5 ≤ Hit rate @ 5 ≤ 0.7
Hit rate @ 1 > 0.4	Hit rate @ 5 > 0.7

3.5

Спецификация RAG: динамика качества ответа LLM при изменении размера пассажа и величины перекрытия

Цель теста. Оценить влияние размера пассажей, величины перекрытий и их комбинаций на качество финального ответа LLM.

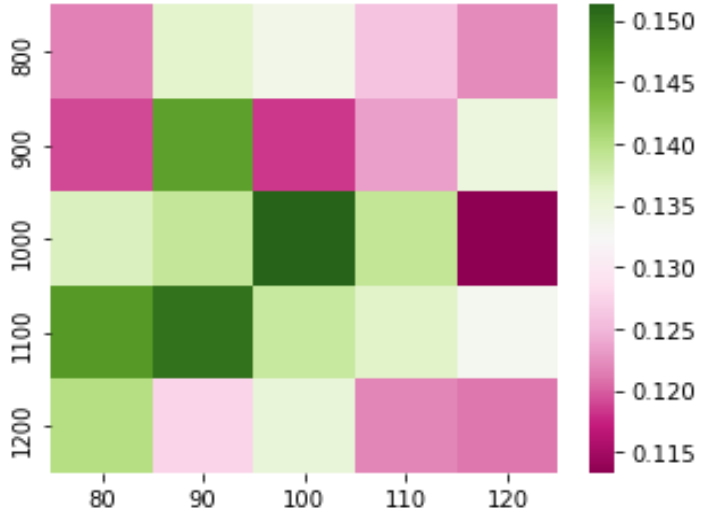
Необходимые условия для проведения теста:

1. Необходимо наличие исходных документов для RAG;

2. Должна быть возможность оценки качества результатов модели без участия асессоров;

3. Нарезчик должен позволять варьировать размер пассажей и/или величину перекрытий.

Алгоритм расчетов. При каждом размере пассажей и при каждой величине перекрытий из LLM получаются ответы на каждый тестовый вопрос, качество которых отдельно оценивается при помощи одной из существующих метрик качества, а затем усредняется на уровне размера пассажей и величины перекрытий. Это и будем считать качеством всего пайплайна.



ось x – размер перекрытия, ось y – средний размер пассажа, значение ячейки – среднее cos sim между ответом LLM и Ground Trurh

[2.5 , 97.5] квантили размера пассажей	Величина перекрытий				
	80	90	100	110	120
[791, 808]	0.122	0.136	0.134	0.126	0.122
[891, 904]	0.119	0.146	0.118	0.123	0.135
[990, 1014]	0.137	0.139	0.151	0.139	0.113
[1091, 1115]	0.147	0.150	0.139	0.136	0.133
[1189, 1209]	0.140	0.128	0.135	0.122	0.121

4.3

Спецификация NER: качество модели в разрезе конфликтующих предсказаний

Цель теста. Оценить устойчивости модели (с учетом постобработки) к генерации конфликтующих предсказаний для однозначной интерпретации результатов.

Необходимые условия для проведения теста:

Тест не проводится, если в бизнес-процессе не используются метки с позициями извлеченных сущностей в исходном тексте или можно для любой извлеченной сущности однозначно восстановить ее позицию в исходном тексте.

Алгоритм расчетов. В рамках данного теста каждый входной текст и предсказанные для него сущности исследуются на следующие конфликтующие предсказания:

- По классу и строке – для класса C количество предсказаний строки S меньше, чем число вхождений S в исходных текст.
- По тексту и строке – по всем классам количество предсказаний строки S больше, чем число вхождений S в исходный текст.

Тип конфликтующего предсказания	Пример
По классу и строке	Входной текст: «Paris Hilton visits Paris». Предсказание: LOC: [‘Paris’], PER: [‘Paris’]. Есть два конфликтующих предсказания: для каждого из классов строка «Paris» входит в исходный текст больше раз, чем была предсказана. При этом предсказание LOC: [‘Paris’, ‘Paris’], PER: [‘Paris’] влечет конфликтующее предсказание данного типа только для класса PER.
По тексту и строке	Входной текст: «Paris Hilton visits Paris». Предсказание: LOC: [‘Paris’, ‘Paris’], PER: [‘Paris’]. По всем классам строка «Paris» была предсказана больше раз (3), чем входит в исходный текст (2).

4.3

Спецификация NER: качество модели в разрезе конфликтующих предсказаний

Интерпретация результатов. Высокие средние показатели при большой доле текстов с конфликтующими предсказаниями указывают на возможность неверного выделения из исходного текста необходимого отрывка с предсказанной сущностью. Возникает риск выделения текста, который не принадлежит к какому-либо классу сущности или принадлежит к другому классу. Подмена одного класса сущности другим может быть критична в определенных областях применения NER-моделей, риск оценивается исходя из бизнес-процесса.

Тип агрегации конфликтующих предсказаний	Значение
Доля текстов с конфликтующими предсказаниями среди всех текстов	0.1
Средняя разность между предсказанным количеством строки и количеством ее вхождений в исходный текст для конфликтующего предсказания типа «По классу и строке» в текстах с конфликтующими предсказаниями	2
Средняя доля строк, признанных конфликтующим предсказанием типа «По тексту и строке», среди всех предсказанных строк в текстах с конфликтующими предсказаниями	0.14
Среднее количество вхождений строк, признанных конфликтующими, в соответствующий им входной текст	4

Спецификация NER: качество при распознавании всех типов сущностей/сущностей по отдельности за один запрос

4.4

Цель теста. Исследование влияния количества извлекаемых сущностей за один запрос (по одному или все классы за раз) на ключевые метрики качества.

Алгоритм расчетов. В рамках данного теста измеряются ключевые метрики качества NER по каждому из классов при изменении количества извлекаемых сущностей в одном запросе. При большом размере набора данных для проведения теста допускается использовать случайное подмножество текстов из исходного набора данных.

- Измеряются ключевые метрики качества при извлечении всех классов в одном запросе.
- Каждый из классов извлекается из текста отдельным запросом, предсказания объединяются, измеряются ключевые метрики качества.

В качестве основной рассматривается метрика F1. Можно и другие.

Кол-во классов	Пример запроса
Все классы	Задача состоит в том, чтобы извлечь все сущности и определить их классы. Выходные данные должны быть в виде списка кортежей следующего формата: [("сущность_1", "класс_сущности_1"), ...]
По одному классу	Задача состоит в том, чтобы извлечь все сущности, которые имеют класс “класс_сущности_N”.

Интерпретация результатов. Большой разброс качества при лучшем результате с типом запроса “по одному классу” потенциально может быть связан с распределением внимания и сложностью задачи. При извлечении всех классов за один запрос модель может распределить свое внимание по классам, что приводит к менее точному распознаванию каждого отдельного класса. В таком случае разбиение на несколько более простых подзадач может повысить значения ключевых метрик качества.

4.5

Спецификация NER: качество при при варьировании детализации описания распознаваемых сущностей.

Цель теста. Определение качества прогнозов NER-модели, измеряемое значениями ключевых метрик, в зависимости от степени детализации описания типов (классов) распознаваемых сущностей.

Алгоритм расчетов. В рамках данного теста измеряются ключевые метрики качества NER по каждому из классов при варьировании описаний классов сущностей.

В качестве основной рассматривается метрика F1: общая и по каждому из классов распознаваемых сущностей. Можно и другие.

Интерпретация результатов и пороговые значения. Повышение ключевых метрик качества при повышении уровня детализации сигнализирует о недостаточном количестве семантической информации в исходном описании о классах сущностей для качественного извлечения их из текстов. Снижение ключевых метрик качества при повышении уровня детализации может говорить об использовании нерелевантного описания, что может вызывать смещение в сторону неверного ответа. Если в наборе данных представлено мало примеров какого-либо из классов, то добавление даже релевантных примеров также может вызвать нежелательное смещение в их сторону.

Критерий
Абсолютный прирост более 3 п.п.
Абсолютный прирост от 1 п.п. до 3 п.п.
Абсолютный прирост менее 1 п.п.

Уровни детализации	Пример
Создание описания при его отсутствии, дополнение существующего	Исходное описание: «Что описывает сущность ANATOMY в тексте?» Детализированное описание: «Что описывает сущность ANATOMY в тексте? ANATOMY – это сущности, описывающие части тела, органы, клетки и компоненты клеток.»
Добавление примеров	Исходное описание: «Что описывает сущность ANATOMY в тексте? ANATOMY – это сущности, описывающие части тела, органы, клетки и компоненты клеток.» Детализированное описание: «Что описывает сущность ANATOMY в тексте? ANATOMY – это сущности, описывающие части тела, органы, клетки и компоненты клеток. Например: артерия, мозг, кровь, клетка»

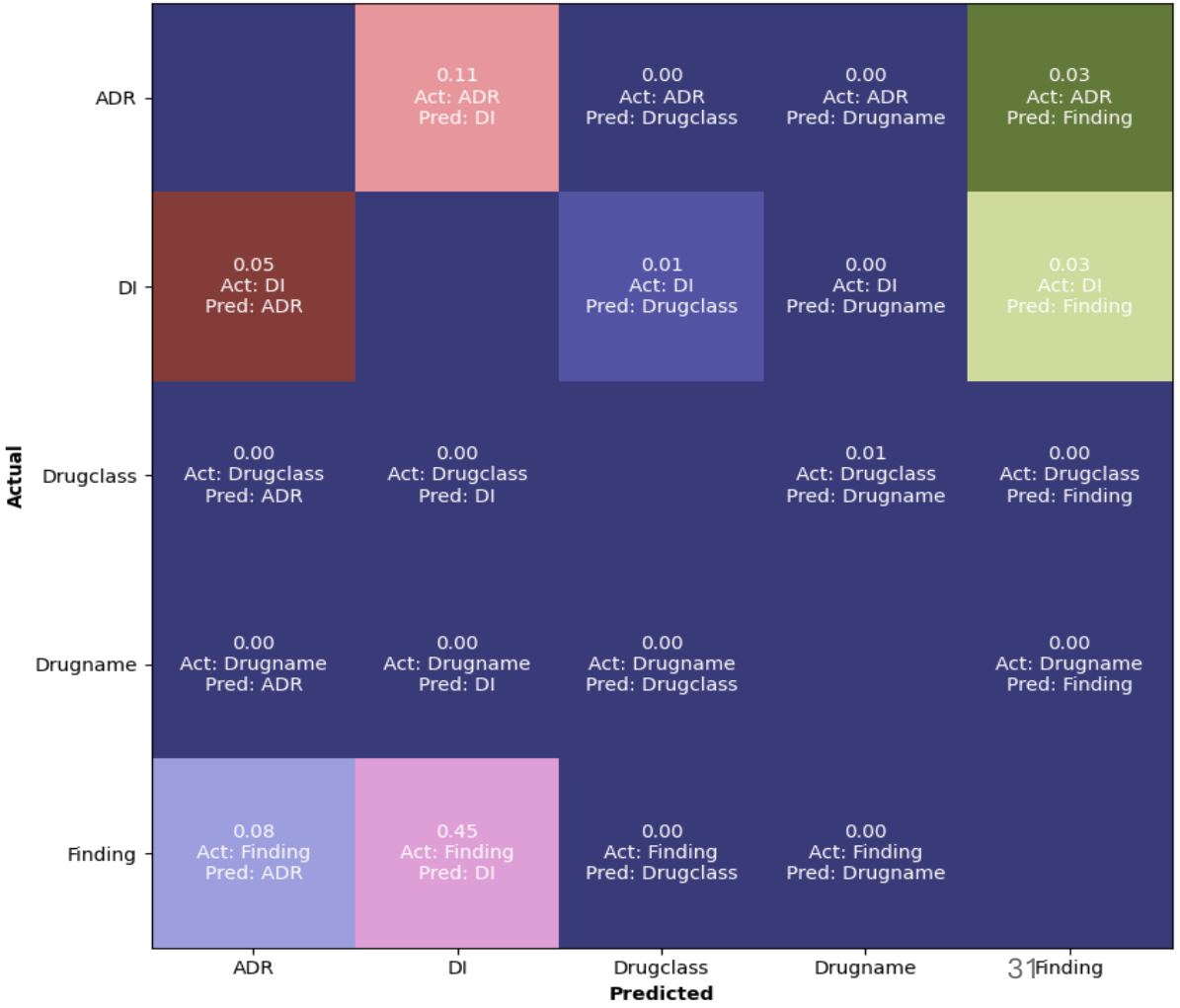
4.6

Спецификация NER: распределения ошибок предсказаний NER-модели в разрезе классов распознаваемых сущностей

Цель теста. Выявить пары классов распознаваемых сущностей, на которых предсказания NER-модели имеют существенное смещение в сторону одного из классов.

Алгоритм расчетов. Для всевозможных комбинаций пар классов распознаваемых сущностей ($class_1$, $class_2$) рассчитывается абсолютное и относительное количество предсказаний, когда при верном $class_1$ предсказывается $class_2$. Относительное количество (доля ошибок) считается как отношение количества ошибочных предсказаний к общему количеству сущностей верного класса.

Критерий на уровне классов	Критерий теста
Максимальная доля ошибок > 0.5	Количество классов с красным светофором >= 1
0.25 < Максимальная доля ошибок < 0.5	Количество классов с желтым светофором >= 2 и нет классов с красным светофором
Максимальная доля ошибок < 0.25	Количество классов с желтым светофором < 2 и нет классов с красным светофором



6.7

Спецификация Chat: средняя длина диалога

Цель теста. Убедиться, что LLM способна поддерживать беседу, когда в этом есть необходимость, а также способна оперативно решить задачу пользователя.

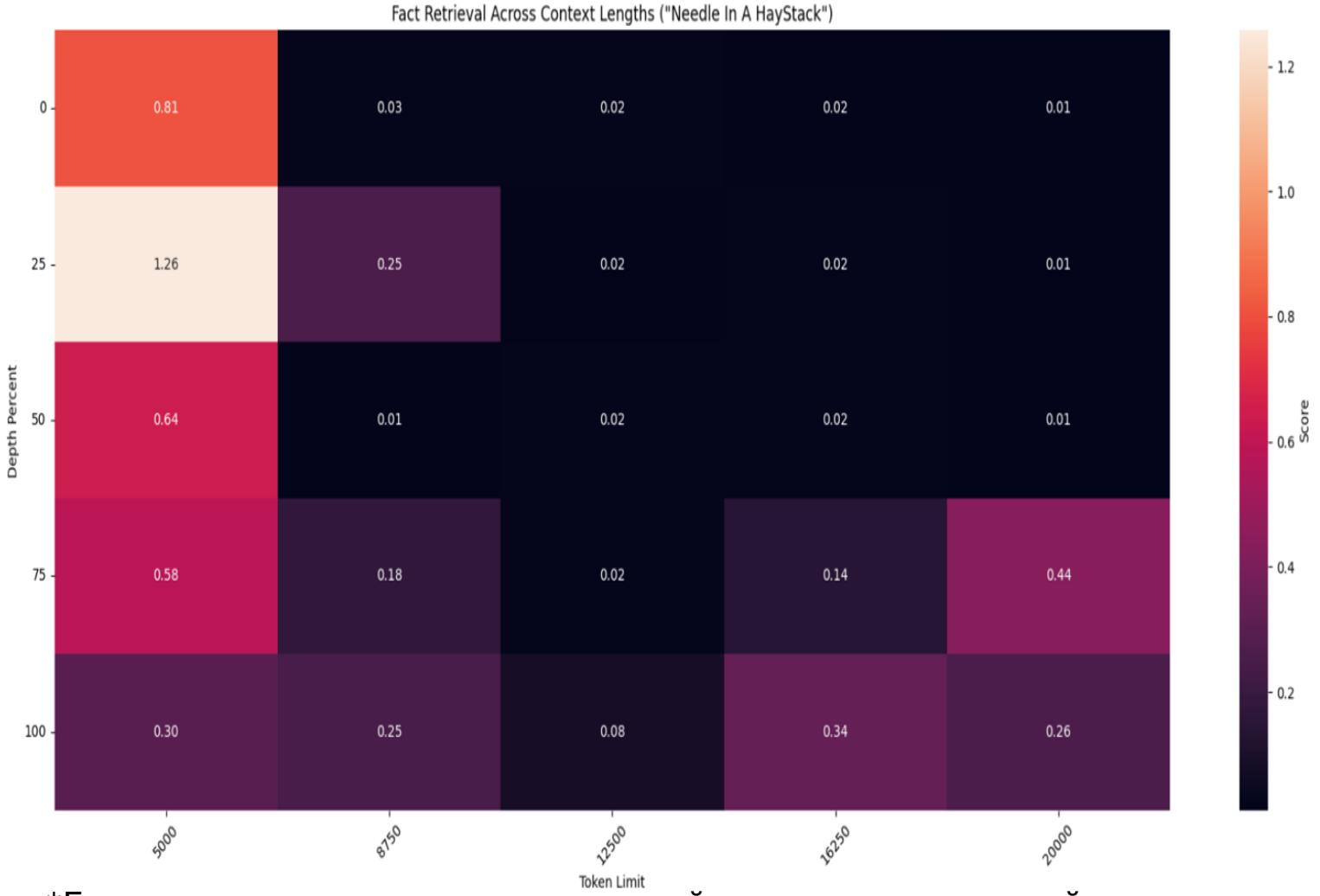
Алгоритм расчетов. Измеряется среднее количество ответов LLM в рамках одной сессии разговора с клиентом (на основе исторической информации).

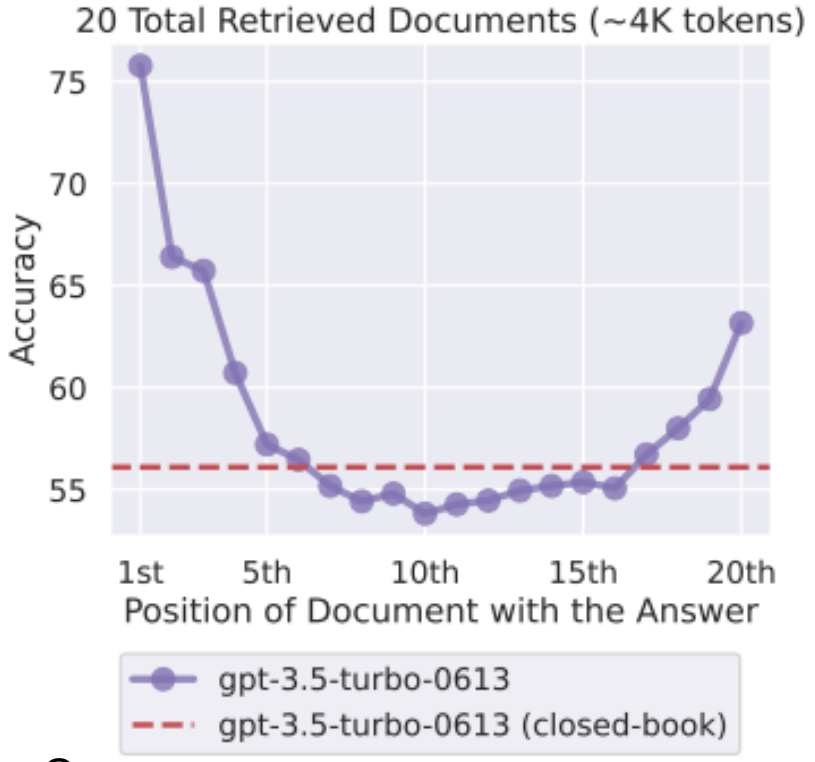
Интерпретация результатов. Эксплуатация LLM в режиме chat может преследовать две цели: решение конкретной проблемы клиента или беседа «по душам». LLM должна оперативно (т.е. за малое количество итераций диалога) за счет уточняющих вопросов решить проблему клиента. Если же клиент заинтересован в беседе, то чем продолжительнее беседа, тем выше качество LLM.

Критерий goal-orientated LLM	**Критерий discussion-orientated LLM**
Средняя длина диалога > 5	Средняя длина диалога < 3
5 ≥ Средняя длина диалога ≥ 3	
Средняя длина диалога < 3	Средняя длина диалога > 5

7.1

Спецификация Extract: «Иголка в стоге сена»





Смотрим качество ответа при вариации позиции 1 релевантного пассажа: https://arxiv.org/pdf/2307.03172

*Если стог тематически нерелевантный вопросу – «ванильный тест» – проверяет attention. Если стог тематически релевантный – проверка релевантна для RAG и более продукто-ориентирована.

10.1

Стабильность: качество в зависимости от аугментаций текстов на уровне символов

Необходимые условия для проведения теста:

Тест проводится только при возможности изменения ключевой метрики качества на новых ответах LLM.

Алгоритм расчета. Предсказания модели собираются на исходных текстах и аугментированных **human level augmentations**, **word swap**, **butter finger**.

- Human-level augmentation – случайно выбранная аугментация с низкой вероятностью опечататься на клавиатуре, поменять соседние места словами, добавить случайно пробел и пр.
- Word swap – перемещение двух слов (Пример: «перемещение двух слов» -> «перемещение слов двух»)
- Butter finger – случайная замена символов в слове на соседние по клавиатуре (к примеру: слово -> слрво)

Критерий
По всем аугментациям абсолютное снижение менее 5 п.п.
Хотя бы по одной аугментации абсолютное снижение от 5 п.п. до 10 п.п., по остальным менее 5 п.п
Хотя бы по одной аугментации абсолютное снижение более 10 п.п.

10.2

Стабильность: качество в зависимости от аугментаций текстов на уровне слов

Необходимые условия для проведения теста:

Тест проводится только при возможности изменения ключевой метрики качества на новых ответах LLM.

Алгоритм расчета. Предсказания модели собираются на исходных текстах и аугментированных перефразированием, сравнивается ключевая метрика.

- **Paraphrase** – перефразирование запроса с сохранением контекста
- **Special token** – разделение слов с помощью символом (к примеру: слово -> с-л-о-в-о)
- **Transliterate** – транслитерация предложения (слово -> slovo)
- **Translation** – перевод вопроса. Переводим с русского на английский, с русского на украинский и с русского на украинский и обратно на русский.

Критерий
По всем аугментациям абсолютное снижение менее 5 п.п.
Хотя бы по одной аугментации абсолютное снижение от 5 п.п. до 10 п.п., по остальным менее 5 п.п
Хотя бы по одной аугментации абсолютное снижение более 10 п.п.

11.3

Стабильность без таргета: Стабильность топа RAG при аугментациях

Цель теста. Оценить изменение подобранного топа пассажей при аугментациях вопроса.

Алгоритм расчета. Запросы аугментируются в соответствии с методами, указанными в тестах 10.1 и 10.2. Для каждого запроса сравниваются два топа по двум стратегиям:

- Строгая: если хотя бы один пассаж из топа изменился, то засчитываем такой топ измененным (ставим 1), иначе ставим 0, далее считаем среднее по всем вопросам.
- Мягкая: в рамках каждого вопроса считаем долю замененных пассажей, после считаем сумму этих долей по всем вопросам.

Размер топа	Строгая стратегия	Мягкая стратегия
1	0,8	0,8
2	0,4	0,7
3	0,2	0,65
4	0,1	0,4
5	0,05	0,1

Критерий
Мера по мягкой стратегии ≤ 0.1 ИЛИ Мера по строгой стратегии ≤ 0.05
Иначе
Мера по мягкой стратегии ≥ 0.8 И Мера по строгой стратегии ≥ 0.7

11.4

Стабильность без таргета: воспроизведение ответа

Цель теста: оценить смысловое и языковое разнообразие/стабильность генераций.

Алгоритм расчетов. Производится оценка разнообразности ответов исходных вопросов на каждом из предоставленных тестовых сетов:

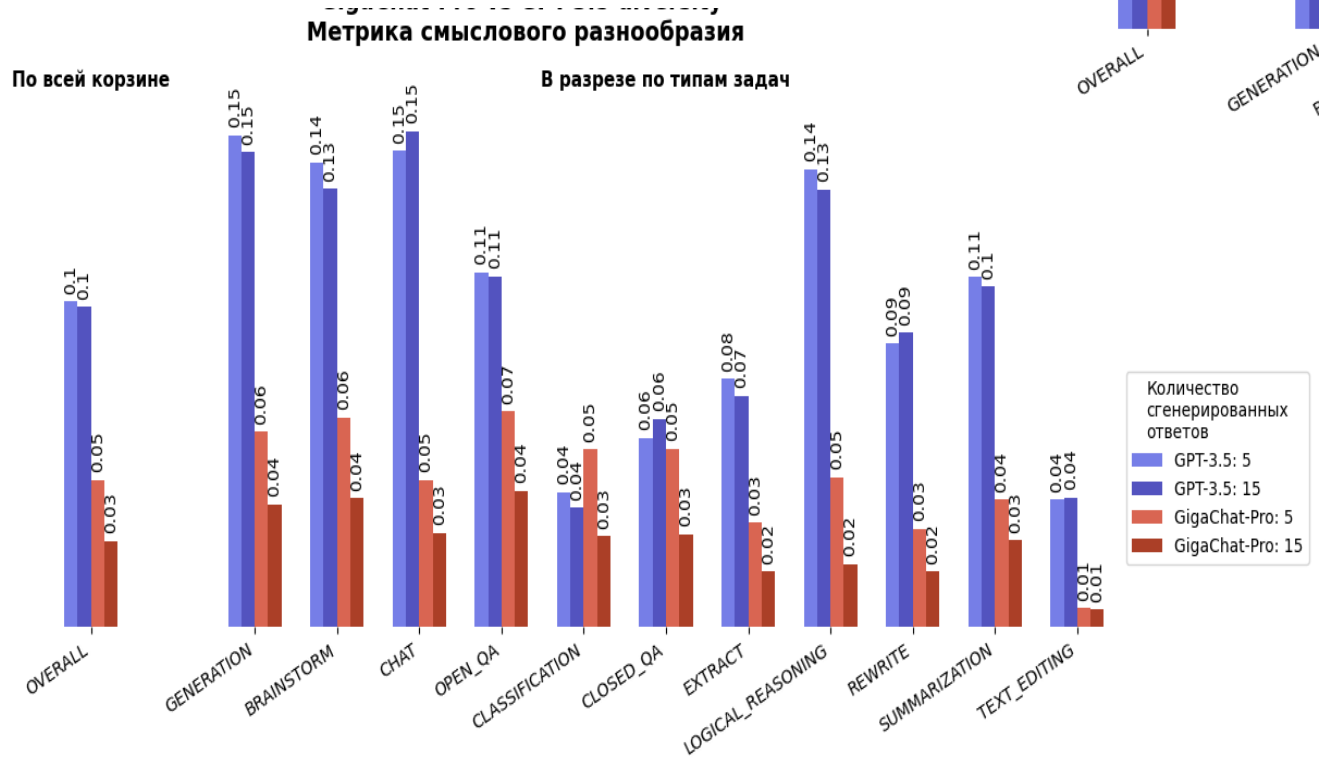
- Прогоняем код генерации ответов N раз
- Для каждой генерации считаем метрики схожести (AverageNgrams, SentBert*) каждой генерации с каждой (сочетание из N по 2).
- Усредняем метрики между парами генераций. AverageNgrams отвечает за языковую похожесть генераций (рассчитываем для данных типа MMLU, RuBQ), SentBert - за смысловую (рассчитываем для задач творческого типа).

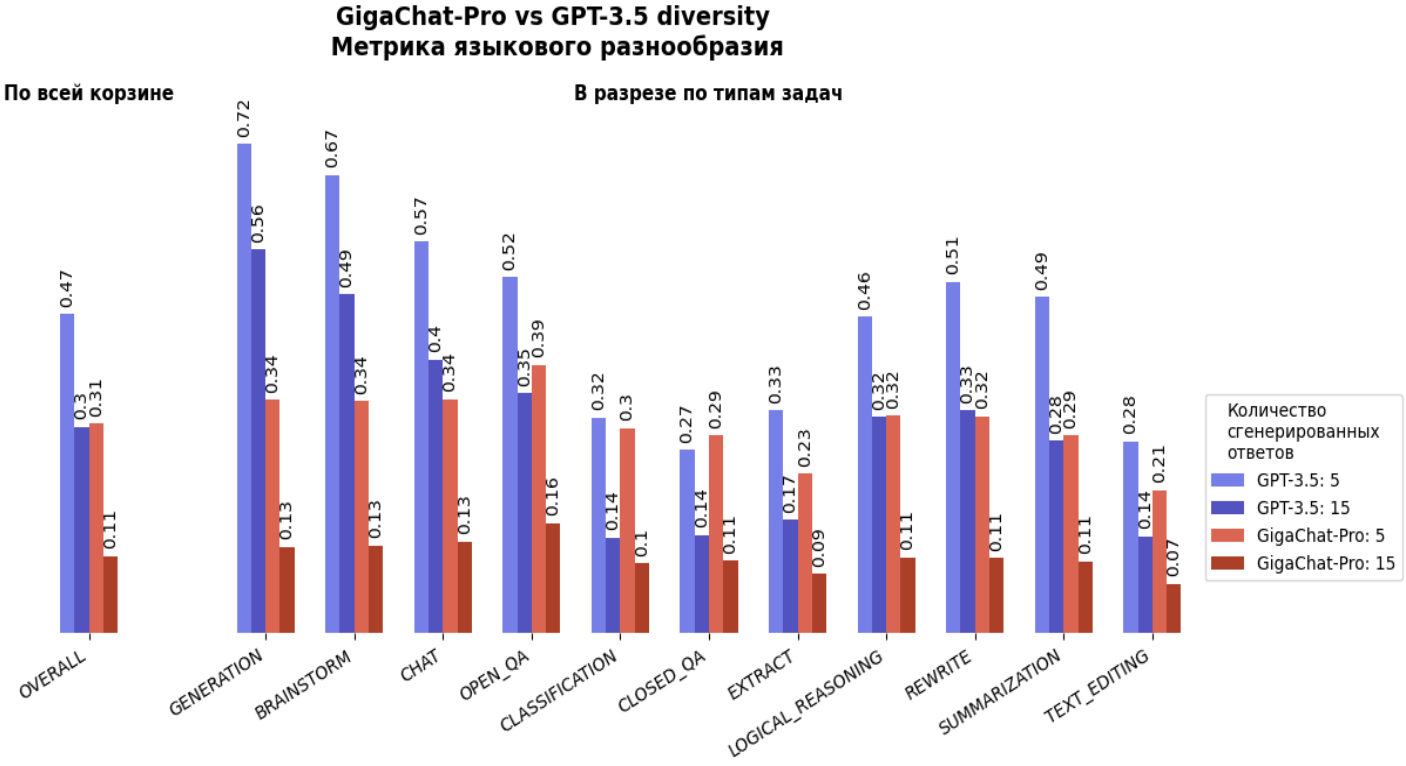
https://arxiv.org/pdf/2004.02990.

11.4

Стабильность без таргета воспроизведение ответа

К творческим задачам относятся те, для которых собраны датасеты LongAnswer либо SummEval-like.





Тип задачи	
Творческая	Точная
Смысловое разнообразие < 3 п.п.	Языковое разнообразие > 5 п.п.
3 п.п. < Смысловое разнообразие < 5 п.п.	3 п.п. < Языковое разнообразие < 5 п.п.
Смысловое разнообразие > 5 п.п.	Языковое разнообразие < 3 п.п.

Вопросы

Мониторинг

Нужно наладить:

- Сбор естественного feedback пользователей
- Базу данных для хранения логов: id клиента, запросы к сервису от клиентов, ответы LLM и этапов сервиса, feedback пользователя
- Регулярную поставку данных на вычислительный кластер
- Запуски тестов по расписанию и по триггеру на кластере
- При проведении первичной валидации определить коррелирующие автометрики (чтобы избавиться от selection bias 1 буллета)

10.6

Стабильность: OOS vs OOT

Среднее относительное изменение ключевой метрики при перестановках в абсолютных значениях
Абсолютное снижение ключевой метрики качества более 25 п.п. ИЛИ Результат теста 2.1, применённого к OOT - «красный»
Абсолютное снижение ключевой метрики от 15 п.п. до 25 п.п. И Результат теста 2.1, применённого к OOT - «Желтый» или «Зеленый»
Абсолютное снижение ключевой метрики менее 15 п.п. И Результат теста 2.1, применённого к OOT - «Зеленый»

11.1

Стабильность без таргета: динамика стандартных статистик

Цель теста. Оценить изменение выборки OOT по сравнению с OOSв разрезе стандартных метрик, а также определить, какие из них коррелируют с ключевой метрикой качества.

Интерпретация результатов.

- Сильное изменение распределения частот стоп-слов в предложениях свидетельствует о различии выборки out-of-time относительно валидационной выборки out-of-sample, что может в числе прочих факторов свидетельствовать о потенциальной нестабильности модели.
- Значительное изменение распределений свидетельствует о том, что в модель начали приходить тексты, значительно отличающиеся от тех, которые были в процессе обучения. В этом случае есть риск падения качества.
- В случае, если длины текстов значительно изменились, нормы векторов эмбеддингов на основе частотных методов (TF-IDF, PPMI и др.) также претерпят значительные изменения, к чему модель может быть не готова.

Критерий
3 ≤ «Красных» светофоров для стандартных статистик ответов LLM
1 ≤ «Красных» светофоров для стандартных статистик ответов LLM < 3 ИЛИ 3 ≤ «Желтых» светофоров для стандартных статистик ответов LLM ИЛИ 3 ≤ «Красных» светофоров для стандартных статистик запросов
Иначе

Статистика	Значение PSI запросы	Значение PSI ответы	Корреляция Спирмена PSI ответов с ключевой метрикой качества
Доля стоп-слов	0,05	0,57	0.2
Частота токенов	0,04	0,08	0.22
Количество токенов	0,03	0,09	0.08
Доля новых токенов	-	0.1	0.05
PSI предсказаний ответов LLM	-	0.09	0.4

Статистика	Границы для	Границы для	Границы PSI для
Доля стоп-слов	PSI<0.5	0.5<PSI	-
Частота каждого токена, деленная на общее количество токенов в корпусе	PSI<0.5	0.5<PSI<1	1<PSI
Количество токенов	PSI<0.5	0.5<PSI<1	1<PSI
Доля новых токенов	Относительное увеличение <0.3 п. п.	0.3 п. п. < Относительное увеличение <0.5 п. п.	0.5 п. п. < Относительное увеличение
PSI предсказаний ответов LLM	PSI<0.5	0.5<PSI<1	1<PSI

11.2

Стабильность без таргета: Разделение OOS и OOT

Цель теста. Оценить изменение выборки out-of-time по сравнению с валидационной выборкой out-of-sample посредством разделения выборок стандартной моделью.

Алгоритм расчета. Аналогично 1.3, метки присваиваются по следующему правилу: 0 – OOS, 1 – OOT (без случайного перемешивания).

Критерий
Gini > 0,8
0,8 > Gini > 0,4
0,4 > Gini

11.5

Стабильность без таргета: качества ответа LLM в зависимости от локального drift запросов

Цель теста. Оценить ключевую метрику качества LLM исходя из степени изменения запросов к ней.

Алгоритм расчета. Для каждого экземпляра запроса из OOS и OOT строятся эмбеддинги через актуальную версию GigaChat. Для каждого из запросов OOT при помощи ANN происходит подбор наиболее близких N запросов, заданных модели на OOS, после чего определяется ожидаемое качество ответа на выбранный запрос в OOT по формуле: $score_j = 1 + \frac{1}{2N}\sum_{i=0}^{N} scoreANN_i * target_i$,

- $score_j$ – score для запроса из OOT, принимающий значения [0, 1]
- $scoreANN_i$ – score для запроса из OOS, принимающий значения [0, 1]
- $target_i$ – метка релевантности ответа LLM запроса из OOS, принимающая значения {-1, 1}.

Финальное качество на OOT подсчитывается как средний $score_j$ для каждого запроса из OOT. Надежность теста оценивается как среднее значение $scoreANN_i$, попавших в топ-N.

Критерий
Абсолютное снижение ключевой метрики качества более 25 п.п. ИЛИ Согласно критериям качества из теста 2.1 для OOT «красный»
Иначе
Абсолютное снижение ключевой метрики менее 15 п.п. И Согласно критериям качества из теста 2.1 для OOT «зеленый»
Средний scoreANNi < 0.2

11.6

Стабильность без таргета: качества ответа LLM в зависимости от глобального drift запросов

Цель теста. Оценить ключевую метрику качества LLM исходя из степени изменения запросов к ней.

Необходимые условия: Тест применяется только при наличии достаточного количества данных в OOS.

Алгоритм расчета. Для каждого из запросов OOS и OOT строятся эмбеддинги актуальной версией GigaChat. После происходит обучение линейной регрессии для предсказания качества на OOT:

- OOS делится на две равные части OOS_{base} и OOS_{add}.
- Из OOS_{add} выделяются подвыборки OOS_{add_i}.
- Между эмбеддингами каждого запроса каждой подвыборки OOS_{add_i} и эмбеддингами запросов OOS_{base} рассчитывается расстояние. Таким образом, каждой OOS_{add_i} соответствует набор расстояний $\{D_{ij}\}$. Для каждого запроса из OOS_{add_i} происходит усреднение посчитанных расстояний до запросов OOS_{base} и образуется набор $\{AD_i\}$.
- Для каждой подвыборки OOS_{add_i} рассчитывается ключевая метрика качества (на основе известных меток релевантности ответов LLM).
- Для каждой подвыборки OOS_{add_i} рассчитывается набор фичей: mean(D_{ij}), std(D_{ij}), max(D_{ij}), mean(AD_i), std(AD_i), max(AD_i), sum({5 наибольших значений в AD_i }), sum({10 наибольших значений в AD_i }), sum({15 наибольших значений в AD_i }) и др.
- Для каждой из сгенерированных фичей смотрится корреляция Пирсона с ключевой метрикой качества. На тех фичах, которые имеют p_value < 0.05 и корреляцию по модулю > 0.5, строится линейная регрессия. Если таких фичей нет, то тесту присваивается «серый» светофор.

Критерий
Абсолютное снижение ключевой метрики качества более 25 п.п. ИЛИ Согласно критериям качества из теста 2.1 для OOT «красный»
Иначе
Абсолютное снижение ключевой метрики менее 15 п.п. И Согласно критериям качества из теста 2.1 для OOT «Зеленый»
Нет статистически значимо коррелирующих фичей

Заметки о LLM ops

- Validation inside
- Ci/cd
- Тесты на века, генерализуемые
- Data drift
- Микросервисы / монолит
- Документация
- Грамотный а/б: разбиение групп,а/а, снижение дисперсии, …
- Нагрузочное тестирование (новый sonnet в 3 раза больше токенов генерит)

Вопросы

КОНЕЦ ЛЕКЦИИ 18.02