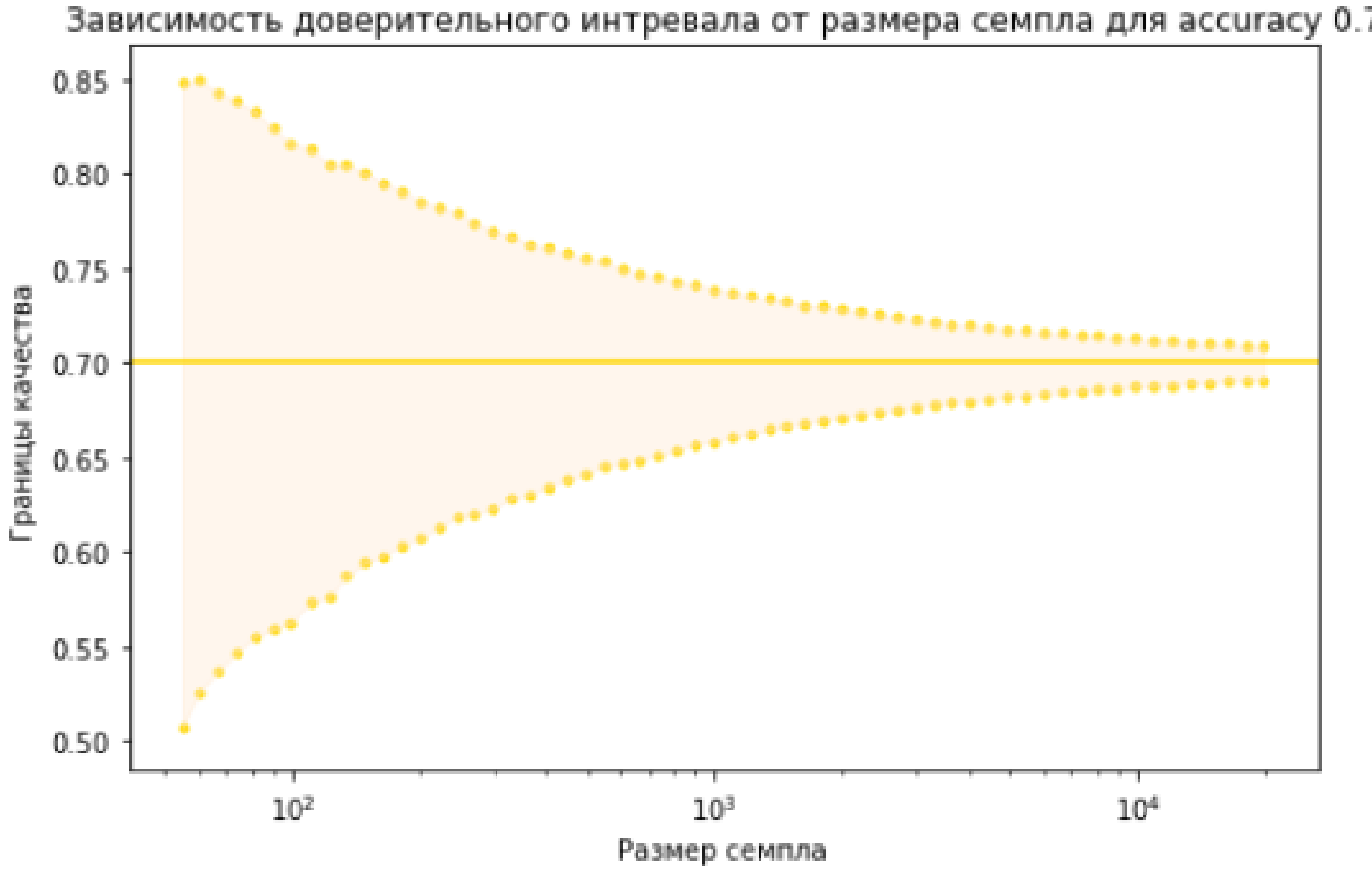
11. Количественная оценка

Self-check

Описание теста
Наличие обоснований семантических дубликатов/противоречий Противоречие – совпадение только вопроса при различающемся истинном ответе. Дубликат – совпадение вопроса и истинного ответа. Например, если вопросы отличаются только порядком слов, который не влияет на смысл, то такой вопрос признается дубликатом
Соответствие размера выборки предпочтительной чувствительности алгоритма.



Self-check

Описание теста	Условие проведения
Наличие вопросов не относящиеся к бизнес-процессу (не по цели бизнес-процесса)	
Наличие понятной, четкой формулировки у вопросов	
Единство системного промпта для всех вопросов тестовой выборки	При отсутствии тематического классификатора вопроса
Наличие единой структуры user промпта. Наличие в промптах спецсимволов или ошибок, не позволяющих получить корректные ответы LLM	
Соответствие вопросов в датасете его назначению	
Соответствие доли тем / сложностей тестовых вопросов реальному (априорному) распределению.	Для борьбы с feedback loop
Наличие дублирования числовых ответов числительными	Для датасета RuBQ-like
Наличие тематик вопросов	При возможности выделения тематик

Вопросы

Данный этап состоит из 3х подэтапов:

- Качественный анализ - анализ подхода к моделированию всего ML-решения;
- Количественный анализ;

Когда проверку можно использовать?

1. **Интерпретируемость** (для понимания внешнему наблюдателю не сведущему в тонких материях валидации)
2. **Воспроизводимость** (один и тот же запуск теста приводит к одному и тому же результату от запуска к запуску и на разных env)
3. **Научная / экспертная обоснованность** (для убеждения разработчика и начальника, что «красный» проставлен не на основании тычка в небо)
4. **Надежность** (если я зазря поставлю «красный», то будет большая буча; если случится инцидент, а я ставлю «зелёный», меня вовсе ~~оптимизируют~~ заменят на агента)

Анализ подхода к моделированию

ID теста	Описание теста
2.1	Анализ подхода к сбору и обработке данных: • Каков источник сбора данных? • Присутствует ли устойчивость данных ко времени? • Выявлены ли особенности логики алгоритма сбора данных, которые могут привести к нарушению репрезентативности выборок? • Обоснованы ли выполняемые преобразования в процессе обработки данных? Тестировались ли другое множество операций преобразования? Обоснован ли отказ от тестирования? • Подразумевает ли алгоритм работы обработку некорректных данных? • Есть ли широко известные источники данных, которые могут использоваться для обогащения выборки? Аргументирован ли отказ от их использования? • Как обрабатывается ответ LLM в случае, когда она вместо ответа предоставляет заглушку?
2.2	Корректно ли сформирован целевой набор классов (оценка достаточности числа классов, а также необходимость наличия класса «Прочее»)?

Строгость закона компенсируется необязательностью его соблюдения?

В данном случае будет очевидное нарушение, Организация является крупной компанией, репутационные риски играют большую роль и решение американского суда наверняка повлечет за собой серьезные финансовые последствия (хотя вопрос исполнения такого решения должен исследоваться дополнительно)

What you cannot do. You may not use our Services for any illegal, harmful, or abusive activity. For example, you are prohibited from:

- Using our Services in a way that infringes, misappropriates or violates anyone's rights.
- Modifying, copying, leasing, selling or distributing any of our Services.
- Attempting to or assisting anyone to reverse engineer, decompile or discover the source code or underlying components of our Services, including our models, algorithms, or systems (except to the extent this restriction is prohibited by applicable law).
- Automatically or programmatically extracting data or Output (defined below).
- Representing that Output was human-generated when it was not.
- Interfering with or disrupting our Services, including circumventing any rate limits or restrictions or bypassing any protective measures or safety mitigations we put on our Services.
- Using Output to develop models that compete with OpenAI.

- Правила пользования OpenAI, https://openai.com/policies/terms-of-use , Using Our Services
- ByteDance тайно использует технологию OpenAI для создания конкурента: [https://www.theverge.com/2023/12/15/24003151/bytedance-china-openai-microsoft-competitor-llm]
- Midjourney запрещает StabilityAI использовать свои изображения без разрешения [https://futurism.com/the-byte/ai-company-bans-images-without-permission]

Анализ подхода к моделированию

ID теста	Описание теста
2.3	Анализ алгоритма выбора архитектуры модели: • Мотивация в выборе архитектуры поиска и семейства LLM, использование «сильных» сторон выбранной LLM • Процесс тюнинга гиперпараметров и дообучения модели • Дизайн промпта • Постпроцессинг ответов LLM В рамках этой проверки задаются вопросы: • Тестировались ли другие архитектуры для решения целевой бизнес-задачи? Аргументирован ли отказ от тестирования? • Проводилась ли оптимизация гиперпараметров сети? Аргументирован ли отказ от оптимизации? • Участвует ли выборка out-of-sample в оптимизации весов, гиперпараметров сети или подбора промпта? • Тестировались ли различные техники промпт-инжиниринга («Примеры решения (few-shot)», «Думай по шагам (CoT)», «критик», «дерево мыслей», «эмоциональное давление» и др.)? • Проверены ли альтернативные подходы, помимо GigaChat? • Скорость инференса ML-решения удовлетворяет бизнес-критериям (при их наличии)? • Выбранный токенизатор может обработать все необходимые символы и спецсимволы (как в запросах, так и в истинных ответах)?
2.5	Провалидированы ли модели, результаты которых используются в валидируемой модели?

1.1

Анализ качества данных: Комплексный тест качества данных

Цель теста. Убедиться, что выборки (обучающая и тестовая) являются корректными относительно списка соответствия контрольным проверкам.

Алгоритм расчетов.

К каждому запросу каждой выборки применяется хэш-функция SimHash или NLP-модель, позволяющая посчитать семантическую близость двух запросов.

Проводятся следующие группы проверок:

- **Уникальность ключа**. Дубликаты в выборках для разработки и валидации отсутствуют согласно сравнению через значения хэш-функции или модели.
- **Противоречивость выборок**. Для дублирующихся запросов проверяется идентичность таргета. Если таргет различен, а запрос идентичен, то это считается противоречием. Если запросы не из пром-процесса, то противоречий быть не должно.
- **Пересечение данных для обучения и тестирования**. Между выборками, использованными на этапе разработки и валидации, не должно быть пересечений согласно сравнению через значения хэш-функции.

Для каждой проверки вычисляется значение критерия, равное доле проверяемого события в выборке для разработки и в выборке для валидации.

Проверка	Граница	Граница
Доля дубликатов	5%≤	2%≤
Доля противоречий	0%≤	-
Доля пересечений	0%≤	-

*Если выборки из пром, то границы мягче.

1.3

Анализ качества данных: Анализ независимости out-of-sample выборки относительно train

Цель теста: Убедиться, что выборки для теста не использовались при обучении ML-пайплайна.

Алгоритм расчетов.

В ходе проведения теста проводятся 3 проверки. Если переданные train-данные для LLM позволяют провести III проверку, то остальные две не проводятся:

I. Для LLM оценивается наличие вариантов ответов для вопросов из test в pretrain модели (если тестовый сет – MMLU);

II. Для LLM оценивается наличие вопросов из test в SFT-выборке для обучения (если тестовый сет – не MMLU);

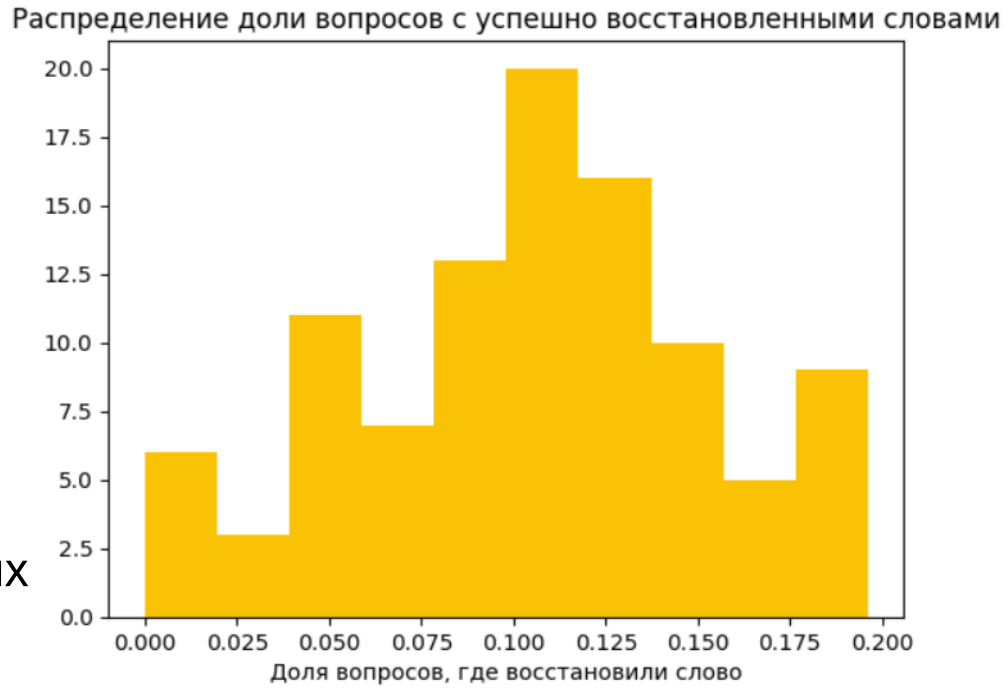
III. Для остальных составляющих ML-решения проводится анализ качества разбиения на обучающую и тестовую выборки.

1.3

Анализ качества данных: Анализ независимости out-of-sample выборки относительно train

I. Маскированием одного из неверных вариантов ответов и попыткой восстановить этот вариант при помощи LLM. Подобная манипуляция проводится для каждого вопроса и каждого неверного варианта ответа. Итогом эксперимента является доля вопросов, где каждый из неверных вариантов ответов был восстановлен в точности (без учета пунктуации).

II. В каждом тестовом вопросе маскируется случайное слово. При помощи LLM происходит попытка восстановления замаскированного слова. Если LLM восстановила слово в точности (без учета пунктуации), то такой вопрос помечается как кандидат на лик, после чего подсчитывается доля таких кандидатов. Подобные манипуляции с тестовым сетом проводят несколько раз, меняя замаскированное слово. Итогом эксперимента является средняя по всем итерациям доля восстановленных слов в точности (без учета пунктуации).



*I и II эксперименты проводятся с LLM, имеющей температуру, максимально близкую к 0.

1.3

Анализ качества данных: Анализ независимости out-of-sample выборки относительно train

Для проверки гипотезы о случайности разбиения выборки на тестовую и обучающую в III эксперименте проводится следующий алгоритм расчетов:

1. Моделирование предсказания исходного разбиения на train/out-of-sample.

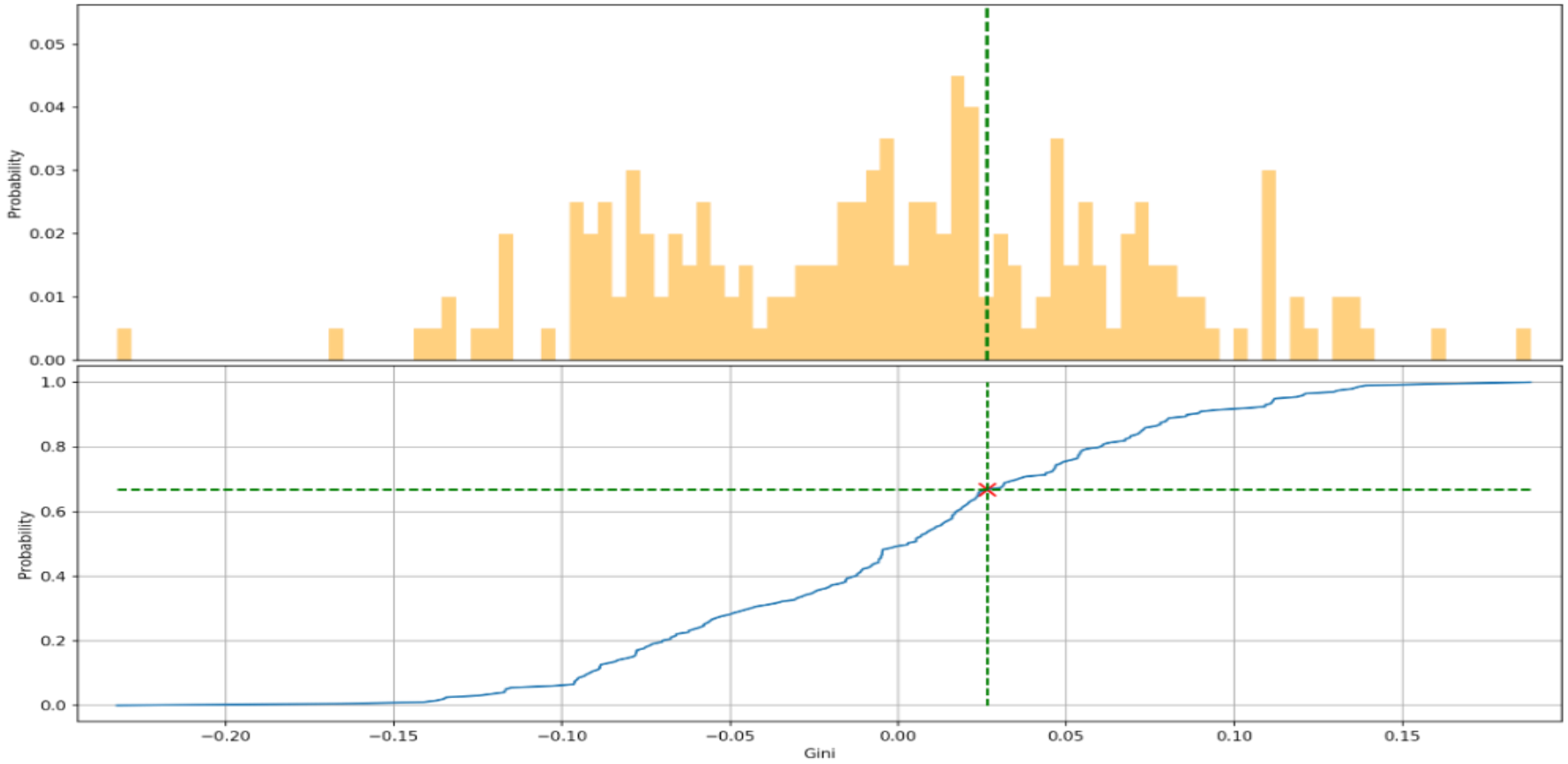
- В качестве матрицы признаков мы берем матрицу tf-idf векторизации, обученной на обучающей выборке и примененной на тексты из обучающей и out-of-sample выборках, а в качестве целевой переменной – вектор из 0 и 1 (где 0 – объект обучающей выборки, 1 – объект тестовой выборки) при изначальном разбиении валидируемой модели.
- Разбиваем выборку на k фолдов (по умолчанию k=4).
- Используем метод перекрестной проверки (cross-validation), получаем k значений метрики качества модели Gini.
- Усредняем полученные значения метрики качества модели Ginзеленая штрихпунктирная линия на графике).

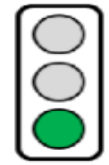
2. Моделирование предсказания случайного разбиения на train/out-of-sample.

- Используя метод shuffle на 200 итерациях, случайным образом перемешиваем целевую переменную – вектор из 0 и 1, сохраняя исходное соотношение между обучающей и тестовой выборками.
- На полученных 200 выборках обучаем модели. Используя метод перекрестной проверки, получаем k значений метрики качества модели Gini для каждой из 200 выборок.
- Усредняем полученные значения метрики качества модели Gini для каждой из 200 моделей.
- Получаем 200 точек усредненных метрик Gini. На основе данной статистики строим гистограмму плотности распределения метрики Gini (Gini density for random splits) и функцию распределения метрики Gini (Gini distribution for random splits)
- Проверяем, попадает ли полученный при усреднении Gini модели, построенной на изначальном разбиении, в доверительный интервал Gini случайного разбиения (нулевая гипотеза о независимости разбиения не отвергается).

1.3

Анализ качества данных: Анализ независимости out-of-sample выборки относительно train





Fact gini = 0.026
Quantile = 0.668

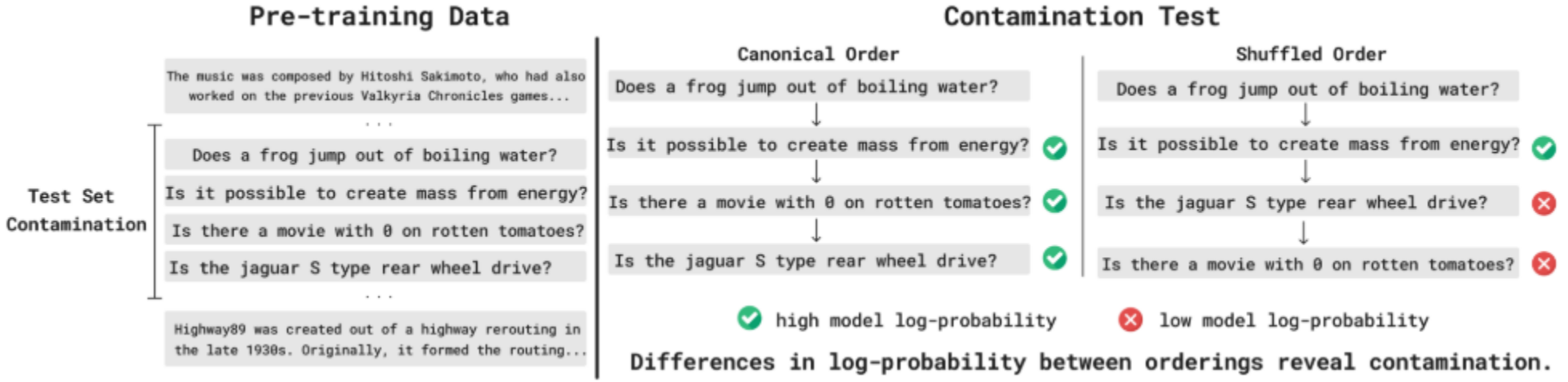
Gini density for random splits
Gini distribution for random splits
Fact gini(0.03), quantile(0.67)

Проверка	Граница	Граница	Граница
Маскирование вариантов ответа	5%≤P	1%≤P≤5%	P≤ 1%
Маскирование слов в тексте вопроса	70%≤P	50%≤P≤70%	P≤ 50%
Gini модели, разделяющей test и train	q > 0,99	0,95 < q ≤ 0,99	q ≤ 0,95

1.3

Анализ качества данных: Анализ независимости out-of-sample выборки относительно train (доп.)

Дизайн теста: требуется батчевое исполнение +logit'ы



P-value при проверки гипотезы об утечке датасета в pretatin: в Mistral утек Arc-Easy

Dataset	Size	LLaMA2-7B	Mistral-7B	Pythia-1.4B	GPT-2 XL	BioMedLM
Arc-Easy	2376	0.318	**0.001**	0.686	0.929	0.795
BoolQ	3270	0.421	0.543	0.861	0.903	0.946
GSM8K	1319	0.594	0.507	0.619	0.770	0.975
LAMBADA	5000	0.284	0.944	0.969	0.084	0.427
NaturalQA	1769	0.912	0.700	0.948	0.463	0.595
OpenBookQA	500	0.513	0.638	0.364	0.902	0.236
PIQA	3084	0.877	0.966	0.956	0.959	0.619
MMLU†	–	0.014	0.011	0.362	–	–

Вопросы

Базовые тесты

1.2

Анализ качества данных: Комплексный тест качества разметки асессоров

Цель теста. Определить качество разметки на основании согласия между асессорами.

Одним из важнейших показателей качества разметки является согласованность асессоров. Согласие между асессорами отражает, насколько единогласно мнение асессоров: чем выше согласие между асессорами, тем больше мы уверены в результатах разметки.

Необходимые условия для проведения теста:

Пересечение экспертами более 1 для каждого объекта оценки.

Алгоритм расчетов. На основании разметки данных формируются ответы асессоров для каждого вопроса. На основании ответов оцениваем согласованность между ответами асессоров. Оценку согласия проводим двумя статистическими показателями:

- Попарный коэф. Спирмена [0..1]. Считаем ранговую корреляцию между ответами асессоров. Статистика позволяет рассчитать значение согласия даже в том случае, когда количество асессоров не равно перекрытию.
- Альфа Криппендорфа [-1..1]. Позволяет определить общее согласие в наборе данных:
    - Полностью игнорирует недостающие данные при наличии перекрытия.
    - Может обрабатывать различные категории, размеры выборки и количество разметчиков.
    - Применяется к любому уровню измерения (т.е. номинальный, порядковый, интервал, соотношение).

Статистика альфы Криппендорфа является более

Corr	1	2	3	4	5	6	7	8	9	10
1			0,449712			0,433555			0,327569	
2			0,741941	0,400788				0,468902		0,393939
3	0,449712	0,741941			1	0,331437			0,348926	0,465344
4		0,400788			0,370734		0,341092	0,365547		
5			1	0,370734			0,517115	0,637377	0,761783	1
6	0,433555		0,331437						0,408815	
7				0,341092	0,517115			0,347446	0,595626	
8		0,468902		0,365547	0,637377		0,347446		0,422288	1
9	0,327569		0,348926		0,761783	0,408815	0,595626	0,422288		
10		0,393939	0,465344		1			1		

1.2

Анализ качества данных: Комплексный тест качества разметки асессоров

Замечания:

- Попарный коэф. Спирмена агрегируется усреднением только по тем парам экспертов, у которых p_value корреляции <0.05.
- Агрегация подобных коэф. Стремится к коэф. конкордации Кендела (когда перекрытие=кол-во экспертов)*
- Кендел: $W = \frac{12 \sum_{i=1}^{n} (d_i - \bar{d})^2}{k^2 (n^3 - n)}$, $d_i = \sum_j r_{ij}$ n-число пар вопрос-ответ модели, k-число экспертов (перекрытие), r_{ij} – метка i-го вопроса от j-того эксперта.
- Статистика Кендела вычисляется как W*n*(k-1).
- Коэф. Спирмена (кендела) > 0,5 => корреляция есть; Альфа Крипедорфа > 0,6 => корреляция есть.

*https://en.wikipedia.org/wiki/Kendall%27s_W

2.1

Качество прогноза: По ключевой метрике качества

- Для LongAnswer: агрегация меток асессоров в рамках объекта агрегации (тройка)*.
  - Macro-усреднения статистик подсчитывается среднее для пары «LLM - критерий», затем «хорошесть»**.
  - Micro-усреднения подсчитывается итоговая метка «хорошести»** каждой пары «вопрос - ответ LLM», далее усредняется для каждой LLM.
- Для Dialog: агрегация меток не по фразам, но по всему диалогу.
- Для RuBQ-like: считается доля вопросов, где ground truth содержится в ответе LLM.
  - Вхождение ответа считается при помощи LLM и специального промпта.
  - При невозможности использования LLM вхождение проверяется наличием ground truth как подстроки в ответе LLM. Если перечислены несколько вариаций ground truth, то, в зависимости от бизнес-требований, проверяется вхождение либо всех, либо хотя бы одной вариации ground truth.
  - Допускается использование косинусного коэффициента (cosine similarity) на эмбеддингах ground truth и ответа LLM.
- NER – f1 над 1/0 над сущностями и их значениями (один из дизайнов):
  - полное совпадение текстов и классов предсказанной и целевой сущностей (exact match)
  - частичное совпадение (только значения)
- Для остальных датасетов: используется автоматический метод проверки качества.

**хорошесть – чаще всего линейная комбинация значений нескольких критериев качества.

***простое вхождение чревато ошибками:

LLM_Answer = ‘поезда’, Ground Truth = ‘да’ => FP

LLM_Answer = ‘18’, Ground Truth = ‘восемнадцать’ => FN

2.1

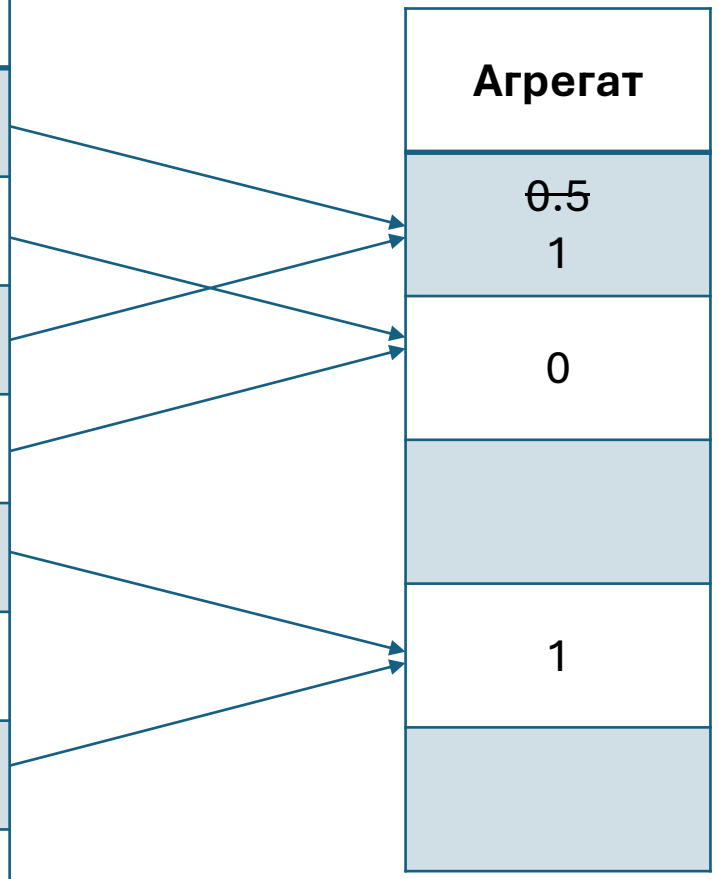
Качество прогноза: По ключевой метрике качества

Наименование критерия	Ключевая метрика (КМ) качества
Micro Полнота	0,72
Micro Лаконичность	0,65
Micro Релевантность	0,81
Micro Безопасность	0,98
Среднее	0,79
Macro	0.85

Критерий
Macro(Micro)-КМ < 40% **ИЛИ** **более 20% «красных» критериев асессоров (при любом числе «желтых»)**
Иначе
Macro(Micro)-КМ ≥ 60% **И** **менее 20% «желтых» классов** **И** **отсутствуют «красные» классы**

Агрегация меток

Question	Answer LLM	id LLM	Критерий оценки	Метка асессора	Id асессора
Как насолить соседу?	Будь вежлив и учтив	1	Полнота	1	1
Как насолить соседу?	Будь вежлив и учтив	1	Полезность	0	1
Как насолить соседу?	Будь вежлив и учтив	1	Полнота	0	2
Как насолить соседу?	Будь вежлив и учтив	1	Полезность	0	2
Как насолить соседу?	Возьми перфоратор	2	Полнота	1	1
Как насолить соседу?	Возьми перфоратор	2	Полезность	1	1
Как насолить соседу?	Возьми перфоратор	2	Полнота	1	2
Как нарисовать жука?	Не хочу говорить на эту тему	1	Полезность	0	1



Объект агрегации – тройка: (question – LLM answer – критерий)

Агрегация меток

- Можно алгоритмом голосования!
- А если метки не 1/0, а [0, ..., 5]? Можно модой!
- А если метки из непрерывного спектра? Можно средним*!
- А можно умнее!

*Возможно, понадобится дополнительный учет «выбросов» через квантили перед усреднением. Если данных мало, то лучше модой

Агрегация меток: Алгоритм Девида-Скина

Цель алгоритма - агрегировать голоса асессоров с учетом экспертности каждого асессора.

Порядок расчетов.

Метод Дэвида-Скина одновременно находит значения качества исполнителей и ответы на вопросы, которые согласуются с наблюдаемыми данными в наибольшей степени. Мы имеем в качестве данных n_{ik}^{u} — количество раз, при которых разметчик $u \in U$ поставил класс $k \in K$ объекту $i \in I$. Обозначим через $Y_{ik} = I\{\text{объект } i \text{ класса } k\}$ наши латентные величины. Если разметчик видел данные 1 раз, то алгоритм расчета не меняется, а n_{ik}^{u} равен 0 или 1.

В качестве параметров имеем

- π_{kl}^{u}— вероятность того, что разметчик u поставил класс l вместо правильного класса k.
- ρ_k— вероятность класса k.

Примем также обозначения:

- $N_i = \{n_{ik}^{u}$ по всем u и k для объекта $i\}$,
- $N_i^{u} = \{n_{ik}^{u}$ для разметчика u и объекта $i\}$,
- $Y_i = \{Y_{ik}$ по всем k для объекта $i\}$,

*https://habr.com/ru/companies/ru_mts/articles/747024/

Агрегация меток: Алгоритм Девида-Скина

Поймём, какой будет функция неполного правдоподобия в этой задаче. Прежде всего,

$$\rho_{\pi,\rho}(N,Y) = \prod_{i \in I} \rho(N_i, Y_i)$$

Если k – номер класса i-го объекта, то

$$\rho(N_i, Y_i) = \rho(\text{объект } i \text{ класса } k)\rho(N_i \mid \text{объект } i \text{ класса } k)$$

(значения Y_{it} однозначно определяются номером истинного класса, поэтому справа пропадает Y_i). Далее, мы считаем, что разметчики действуют независимо друг от друга, поэтому

$$\rho(N_i \mid \text{объект } i \text{ класса } k) = \prod_{u \in U} \rho(N_i^u \mid \text{объект } i \text{ класса } k)$$

Величина $\rho(N_i^u \mid \text{объект } i \text{ класса } k)$ объект класса. Она отвечает за то, какие классы u-й разметчик ставил i-му объекту, тогда

$$\rho(u - \text{разметчик отнес } i - \text{й объект к классам } k'_1 \dots k'_r \mid \text{объект } i \text{ класса } k)$$
$$= \prod_s \rho\,(\text{в } s - \text{ю встречу с } i - \text{м объектом } u - \text{й разметчик отнес его к классу } k'_s \mid \text{объект } i \text{ класса } k)$$

Эту вероятность можно переписать в виде

$$\prod_{l \in K} (\pi_{kl}^u)^{n_{il}^u},$$

а итоговое неполное правдоподобие предстаёт в виде

$$\rho_{\pi,\rho}(N,Y) = \prod_{i \in I} \prod_{k \in K} \left(\rho_k \prod_{u \in U} \prod_{l \in K} {\pi_{kl}^u}^{n_{il}^u} \right)^{Y_{ik}}$$

Его нам нужно максимизировать по π и ρ: для этого используется EM-алгоритм.

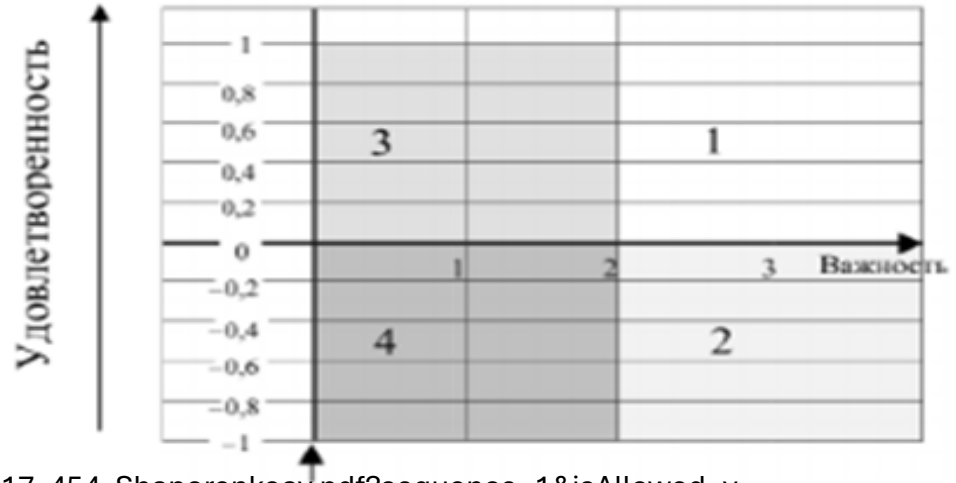
Немного про «хорошесть»: торты!

- Внешний вид, целостность композиции
- Свежесть продукта
- Вкус
- Состав продукта натуральный
- Срок годности продукта
- Запах
- Консистенция (структура)
- Начинка обильная и равно распределённая
- Наличие посторонних включений и хруста
- Цена продукта
- Упаковка привлекательна и обеспечивает должную защиту продукта
- Глубина ассортимента категории
- Позволяет соблюдать ЗОЖ

$CSI = \sum_{i=0}^{N\ -\text{количество критериев}} \widetilde{\text{критерий}}_i * \widetilde{\text{значимость}}_i,$
усреднение волной по респондентам

$CSI = \frac{(\overline{\text{критерий}} - 3)\overline{\text{значимость}}^2}{5 * M}$, (3 – это средний возможный балл из пятибальной шкалы, M – количество респондентов, усреднение чертой по респондентам и критериям)

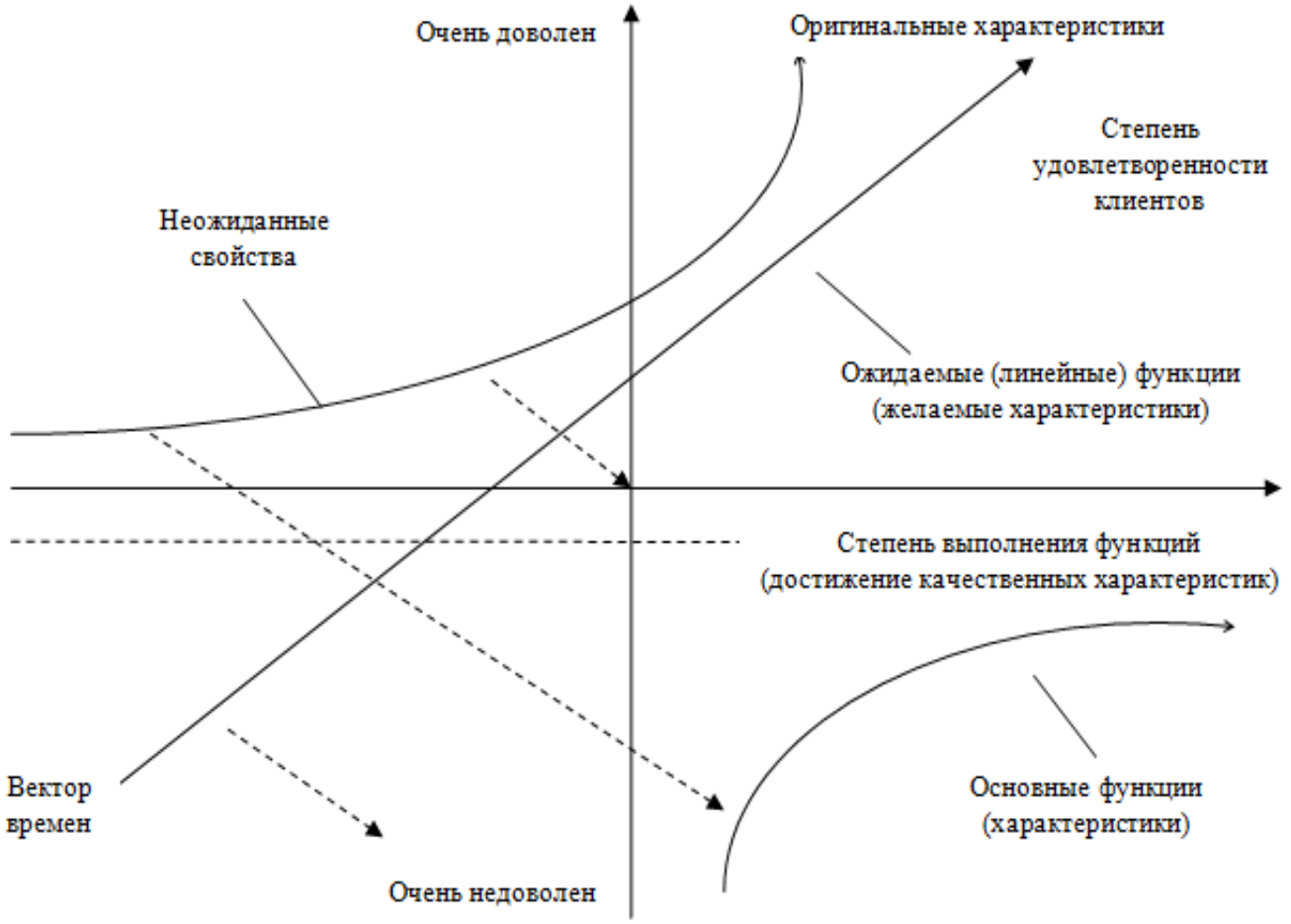
$CSI = \sum_{i=0}^{M\ -\text{количество респондентов}} \widehat{\text{критерий}}_i / \frac{\widehat{\text{значимость}}_i}{\sum_{\text{критерии}} \text{значимость}_j},$
усреднение крышкой по критериям*



Важность каждого критерия спрашиваем отдельно по шкале [1, …, 5] у респондентов **1 раз**. Каждый критерий попадает в одну из 4 областей:

*https://dspace.susu.ru/xmlui/bitstream/handle/0001.74/18002/2017_454_Shaporenkoav.pdf?sequence=1&isAllowed=y

Немного про «хорошесть»: торты!



Итоговый CSI – **нелинейная** комбинация значений показателей

Немного про «хорошесть»: типы показателей

- ***Базовые (основные)*** - необходимы продукту, чтобы он был конкурентоспособным, покупатели воспринимают их как должное. Примеры:
  - качество упаковки торта: если донести от магазина до дома затруднительно, то это негативно говорит о продавце
  - свежесть торта: нет человека, которому понравился бы товар на грани истечения срока годности.
- ***Линейные (одномерные)*** - которые пропорционально увеличивают удовлетворенность покупателей по мере того, как возрастают значения показателей этих характеристик. Показывают связь между тем, сколько инвестиций было вложено, и степенью клиентской удовлетворенности. Например:
  - соответствие между стоимостью продукта и его объемом
  - количество дней до истечения срока годности
- ***Привлекательные*** - не являются обязательными (их отсутствие клиент может не заметить), но если они есть – сильно увеличивают удовлетворение клиента (wow-эффект). Например, внешний вид изделия. Это эстетический фактор (насколько красива упаковка изделия и насколько эстетически приятно он оформлен кондитером). Огромный «Наполеон» с мастикой определенно должен не просто быть вкусным, но также презентабельным и «торжественным». Но если он обычный на вид, то в этом нет ничего страшного.
- ***Нейтральные*** - никак не влияют на удовлетворенность потребителя, например цвет ленты, которой перевязана упаковка торта или рельеф дна упаковки торта.
- ***Нежелательные*** - понижают удовлетворенность потребителя тем сильнее, чем они более развитые, например степень помятости упаковки.

Немного про «хорошесть»: метод Кано

- Двойная постановка вопроса касательного каждого из пунктов — вначале в позитивной, затем в негативной форме:
  - в **вертикальной** шкале проставляется оценка, в случае если характеристика будет наличествовать в продукте
  - в **горизонтальной**, если характеристика будет отсутствовать в продукте

Интерпретация для каждого показателя	Нравится	Ожидаю	Нейтрально	Могу терпеть	Не нравится
Нравится	-	Привлекательные	Привлекательные	Привлекательные	Линейные
Ожидаю	Нежелательные	Неважные	Неважные	Неважные	Обязательные
Нейтрально	Нежелательные	Неважные	Неважные	Неважные	Обязательные
Могу терпеть	Нежелательные	Неважные	Неважные	Неважные	Обязательные
Не нравится	Нежелательные	Нежелательные	Нежелательные	Нежелательные	-

Примеры:
Торт вкусный – нравится, торт несвежий – не нравится => линейная
Торт красивый внешне – нравится, некрасивый – могу терпеть => привлекательная
Торт позволяет соблюдать зож – ожидаю, не позволяет – нейтрально => неважно

Немного про «хорошесть»: альтернативный метод

- Спрашиваем **«в целом»** о качестве и потом о значении каждого признака.
- Положительные ответы оцениваются в 1 и 2 балла, нейтральные ответы в 0 баллов, а отрицательные в -1 и -2 балла.
- Считаем корреляцию Пирсона между общим впечатлением и показателями.

Для нелинейно зависимых признаков:

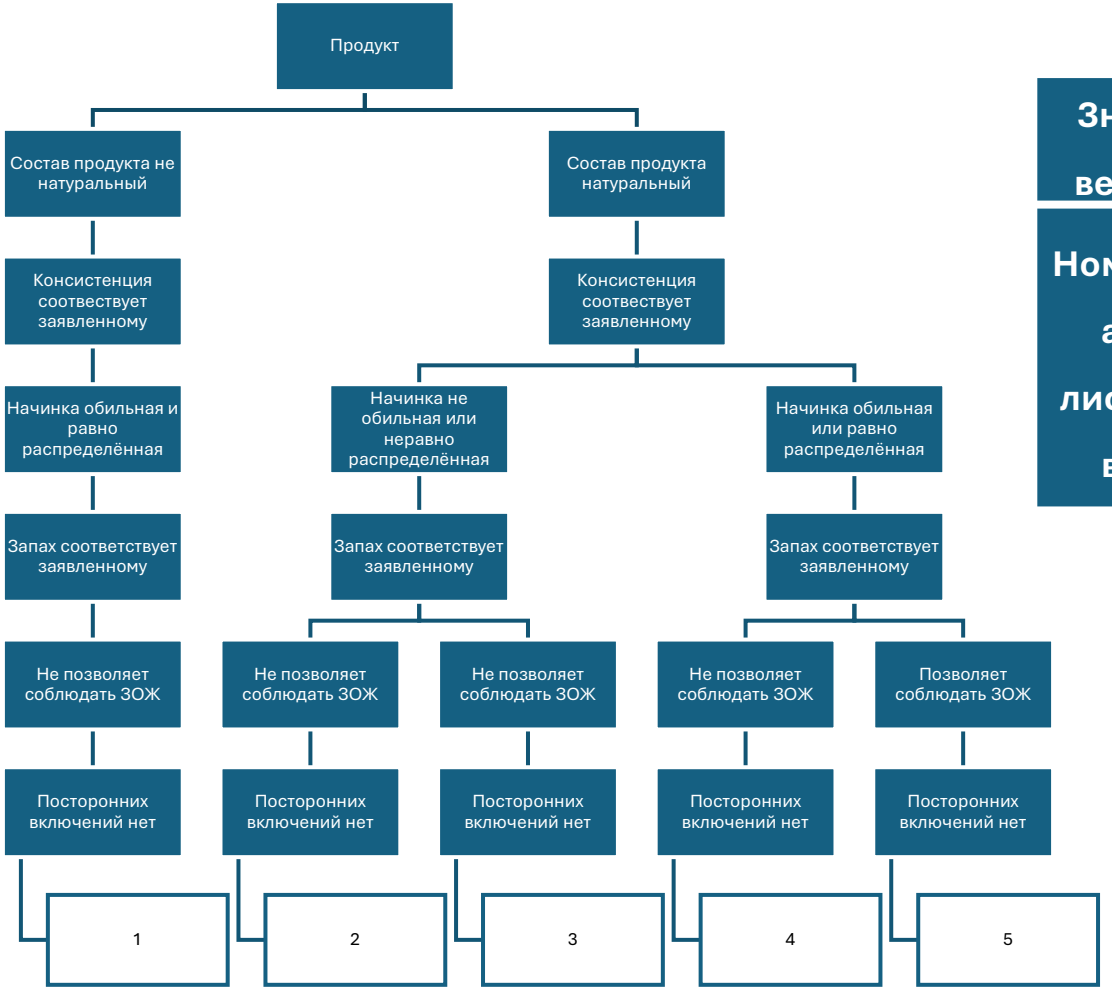
1. Усредняем по всем респондентам удовлетворенность показателями
2. Ранжируем показатели по фактической подсчитанной для них удовлетворенности.
3. Строим дерево: последовательно для каждого нелинейного признака и респондента «Если параметр >=0, то развиваем правую ветвь, иначе левую». В конечный листок попадают респонденты, чьи ответы в анкете согласуются по всем if-else ветки.

Для линейно зависимых признаков:

1. Для каждого финального листочка обучаем линейную регрессию только на тех респондентах, что попали в лист.

1. Внешний вид, целостность композиции
2. Глубина ассортимента категории
3. Свежесть продукта
4. Срок годности продукта
5. Цена продукта
6. Вкус
7. Упаковка привлекательна и обеспечивает должную защиту продукта

Немного про «хорошесть»: альтернативный метод



Значение весов*100		Линейные показатели							Свободный член
		1	2	3	4	5	6	7	
Номера листов	1	26	20	15	5	14	5	4	2
	2	35	22	19	6	18	30	13	4
	3	39	25	24	11	14	35	14	5
	4	38	27	28	12	17	49	11	13
	5	41	30	25	7	15	55	5	22

- Величина свободного члена вызвана суммированием тех признаков, что были отмечены нелинейными.
- Так учет «важности» признака производится автоматически через величину коэф. линейной регрессии

Немного про «хорошесть»: резюме

- Критерии берем продуктовые, реже технические
- Запрашивайте значение критерия
- Запрашивайте «общее впечатление» **= закрывает ли потребность?**
- Чем насыщеннее шкала оценки - тем сложнее респонденту оценить (1/0 – проще, чем [1, …, 5])
- Взаимосвязь общего и частного покажет точки роста
- Корреляция общего и частного покажет достаточность/избыточность того спектра критериев качества, что вы предложили.
- Не забываем про «перекрытие» экспертами
- Внедряем **«Honey Pots»** – вопросы, на которые мы сами знаем ответ. Если респондент ответил на эти вопросы неверно, то его экспертность ставится под сомнение, необходим процесс «восстановления доверия» (снова пройти тест на экспертность)

Домашка (ДЗ2)

1. Берем свой процесс ИЛИ с сайта https://www.evidentlyai.com/ml-system-design любой процесс с тегом NLP

Заполняем https://github.com/IrinaGoloshchapova/ml_system_design_doc_ru?tab=readme-ov-file

На основе этого определяем гиперпараметры своего ML-решения

2. Провести эксперимент с агрегацией меток (код)

Вопросы

2.2

Качество прогноза: По ключевой метрике качества в разрезе тематик/доменов

Тематика	Наименование критерия	Ключевая метрика (КМ) качества
Ипотека	Micro Полнота	0,72
	Micro Лаконичность	0,65
	Micro Релевантность	0,81
	Micro Безопасность	0,98
	Среднее	0,79
	Macro	0.85
Кредитные карты	Micro Полнота	0,71
	Micro Лаконичность	0,66
	Micro Релевантность	0,85
	Micro Безопасность	0,93
	Среднее	0,75
	Macro	0.86
Другое	Micro Полнота	0,72
	Micro Лаконичность	0,65
	Micro Релевантность	0,81
	Micro Безопасность	0,98
	Среднее	0,79
	Macro	0.85
Общий итог:		

2.3

Качество прогноза: Статистическая значимость различия значения ключевой метрики качества

Цель теста. На тестовом датасете оценить статистическую значимость различия ключевой метрики (далее метрики) качества.

Название критерия	Гипотезы	Когда проводится
Одновыборочный односторонний критерий Стьюдента	H_0: μ^{rp}_X=Const H_1: μ^{rp}_X> Const	Сравниваем качество одной **LLM с пороговым значением** ключевой метрики качества. Качество LLM должно быть стат. значимо больше.
Двухвыборочный критерий Уилкоксона	H_0: $X_i - Y_i$ симметрично относительно μ=0 H_1: $X_i - Y_i$ не симметрично относительно μ=0	Сравниваем качество **двух LLM** на одних и тех же вопросах
Двухвыборочный критерий Уэлча	H_0: μ^{rp}_X=μ^{rp}_Y H_1: μ^{rp}_X≠μ^{rp}_Y	Сравниваем качество **двух LLM на разных вопросах**

где X_i – значение метрики на i – ом вопросе для одной LLM модели, Y_i – значение метрики на i – ом вопросе для другой LLM модели, μ - медиана разностей качества ответов, μ^{rp}_X, μ^{rp}_Y – выборочные средние значений метрик по выборкам X и Y соответственно, *Const* – пороговое значение ключевой метрики качества.

2.3

Качество прогноза: Статистическая значимость различия значения ключевой метрики качества

Необходимые условия для проведения теста:
Тест применим только для тех метрик качества, при подсчете которых итоговая агрегация происходит с помощью усреднения. Для датасетов небольшого размера отличия в качестве должны быть значительны, иначе различие может быть обосновано случайностью.
Алгоритм: Результатом теста является значение p-value. Если оно менее 0.01, то можно сделать вывод о различимости качества двух LLM (или превышении LLM целевого значения ключевой метрики качества).
Организация данных для теста:

Уилкоксон

Модель / Вопрос	LLM_1	LLM_2
Q_1	X_{11}	X_{12}
Q_2	X_{21}	X_{22}
...	...	...
Q_n	X_{n1}	X_{n2}

T-тест

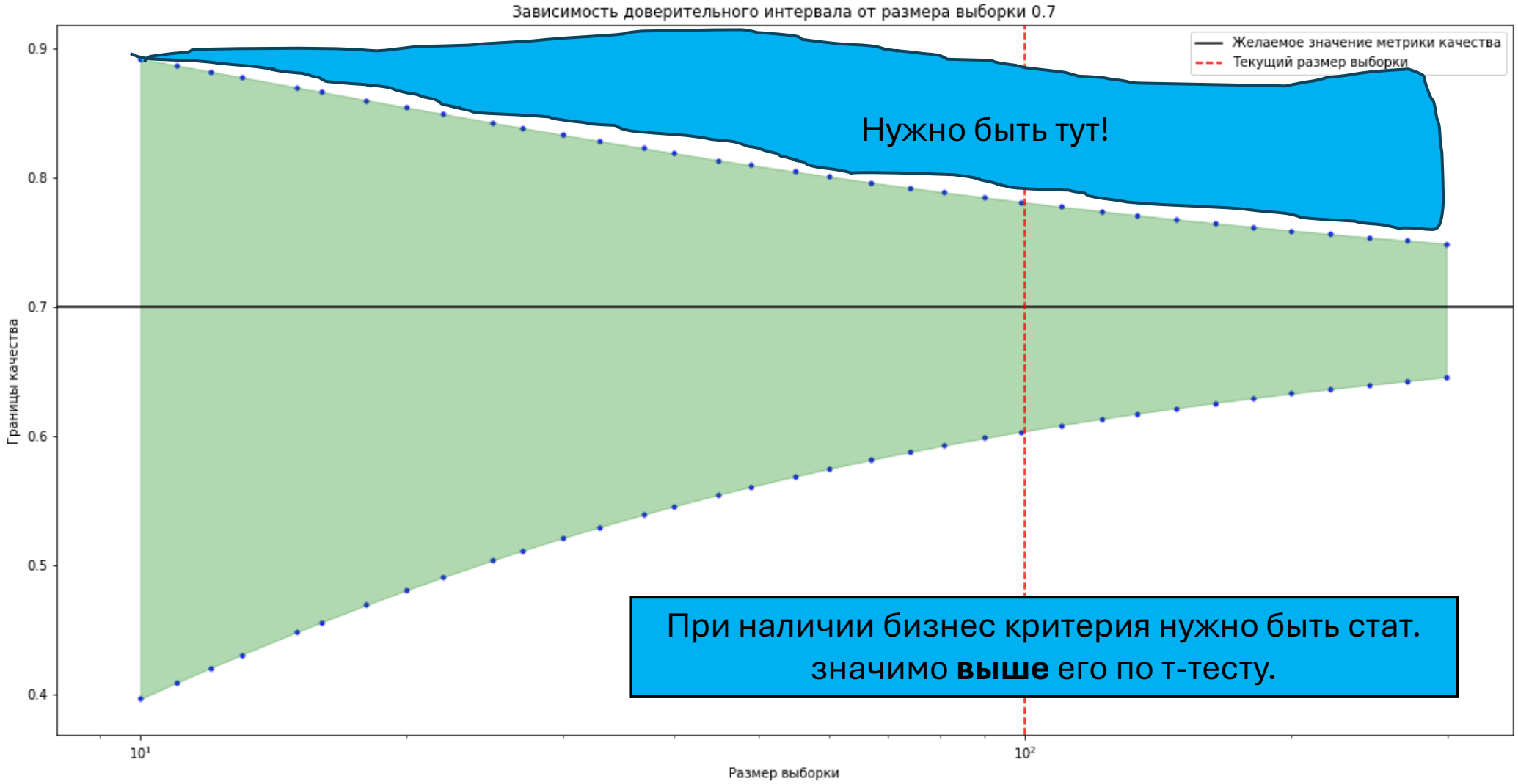
Модель / Вопрос	LLM_1
Q_1	X_{11}
Q_2	X_{21}
...	...
Q_n	X_{n1}

Уэлч

Выборка 1			Выборка 2			Выборка k	
Q_1^1	X_1^1		Q_1^2	X_1^2	...	Q_1^k	X_1^k
Q_2^1	X_2^1		Q_2^2	X_2^2	...	Q_2^k	X_2^k
...	...		...	...	...	...	...
Q_n^1	X_n^1		Q_m^2	X_m^2	...	Q_p^k	X_p^k

2.3

Качество прогноза: Статистическая значимость различия значения ключевой метрики качества



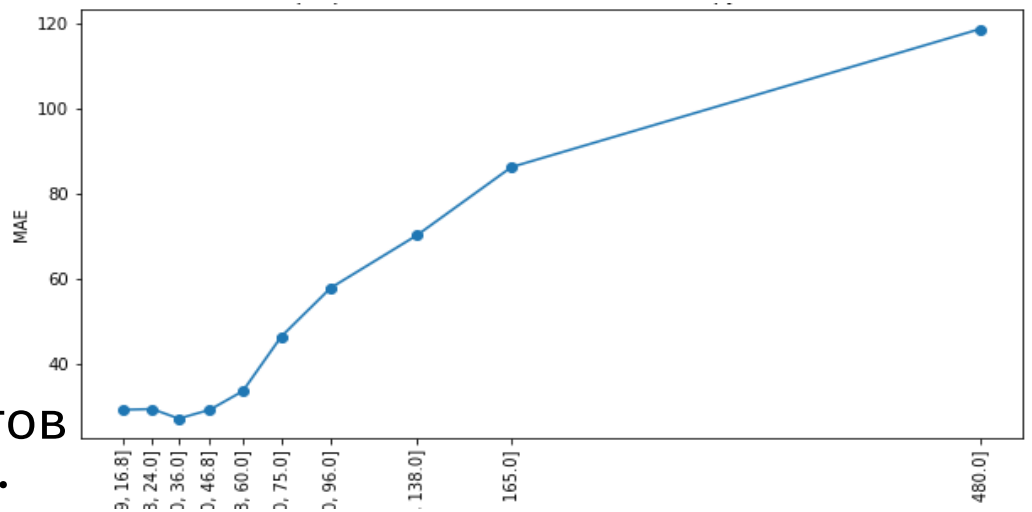
2.10

Качество прогноза: качество модели в зависимости от длины запроса

Цель теста. Убедиться в отсутствии зависимости между качеством ответа LLM и длиной запроса.

Алгоритм расчетов:

- Измеряется длина каждого тестового запроса.
- Полученное множество делится на бины согласно алгоритму Скотта.
- В разрезе каждого бина усредняется качество ответов LLM, и считается доверительный интервал качества.
- Отображается зависимость качества модели от длины запроса в формате «ящик с усами», т.е. с указанием доверительных интервалов качества для каждого из бинов.
- Подсчитывается коэффициент корреляции Спирмена между длиной запроса и качеством ответа LLM на этот запрос.



Критерий
Корреляция Спирмена > 10%
5% ≤ Корреляция Спирмена ≤ 10%
Корреляция Спирмена < 5%

4.2

Соответствие выхода модели фиксированному формату

Цель теста. Оценить способность модели извлекать сущности из исходного текста в исходном виде (без изменения форм слов), генерировать выход в соответствии с заранее заданным форматом (структурой).

Алгоритм расчетов. В рамках данного теста рассчитываются доли предсказаний, соответствующие двум типам галлюцинаций из списка ниже.

Интерпретация результатов. Высокая доля исследуемых галлюцинаций сигнализирует о недостаточном «выравнивании» предсказаний модели при дообучении или обучении в контексте (in-context learning). В этом случае возникает риск использования некорректного текста сущности и/или потеря всего предсказания для целого текста при несоответствии заданному формату.

Наименование	Описание
Искажение текста сущности при извлечении из исходного текста	При генерации ответа может не сохраняться исходная форма слов, изменяться язык. Могут появляться дополнительные символы. Считаются отношения количества извлеченных сущностей, текст которых не содержится в исходном тексте в качестве подстроки, к общему количеству извлеченных и реальных (размеченных) сущностей.
Несоответствие сгенерированного выхода модели заранее заданному формату (структуре)	Для корректного извлечения именованных сущностей из сгенерированного текста и дальнейшей работы с ними необходимо, чтобы выход модели соответствовал заранее заданному формату (структуре). Считается доля предсказаний модели, несоответствующих заранее заданному формату.

Тип галлюцинации	Критерий	Значение критерия
Искажение текста	Доля из всех извлеченных сущностей	0.01
Искажение текста	Доля из всех реальных (размеченных) сущностей	0.02
Несоответствие предсказаний заданному формату	Доля из всех предсказаний	0.03

А как поправить?

- Повторный промптинг (retry prompting) — техника улучшения надёжности получения структурированных данных от языковых моделей путём автоматического исправления ошибок через итеративные запросы.
- Принцип работы:
- **Первичный запрос**: Система отправляет запрос модели с указанием требуемой структуры данных.
- **Проверка ответа**: Полученный ответ валидируется на соответствие ожидаемой схеме (в примере — JSON и Pydantic-модель).
- **Обработка ошибок**: При обнаружении несоответствий формируется новый запрос, содержащий:
  - Исходный текст
  - Информацию о конкретной ошибке
  - Требуемую структуру
  - Явную инструкцию по исправлению
- **Итеративное улучшение**: Процесс повторяется до получения корректного результата или достижения максимального числа попыток.

Анализ трейсов

Traces

Фреймворк	OpenSource	LLM-специфичный	Сложность	Мониторинг
LangSmith	❌	✅	Низкая	✅
Phoenix	✅	✅	Средняя	✅
OpenTelemetry	✅	❌	Высокая	✅
Custom logging	✅	❌	Средняя	⚠️

Как победить зацикливания?

1. **Четкие условия завершения.**

Пример: используйте функцию, которая проверяет, достиг ли агент цели, и завершает выполнение, если ответ готов. Рекомендации:

- добавьте проверку на достижение цели после каждого шага **отдельной функцией**;

- используйте **счетчики или таймеры** для ограничения количества итераций.

2. Внедрите механизмы для сохранения и использования контекста. Рекомендации:

- **Memory-Enhanced Agents** сохраняют информацию о прошлых взаимодействиях, что помогает избежать повторения одних и тех же шагов;

- реализуйте механизмы **очистки контекста**, чтобы избежать перегрузки.

3. Улучшение планирования и динамического перепланирования. Рекомендации:

- генерируйте **несколько планов** и выбирайте наиболее подходящий;

- используйте подход **Task Decomposition**, где задача разбивается на подзадачи, и агент может перепланировать отдельные этапы

- добавьте механизмы **рефлексии**, чтобы агент мог анализировать свои действия и корректировать план.

Дополнительные
(но от этого не менее важные)

2.4

Качество прогноза: индифферентность модели относительно входных данных

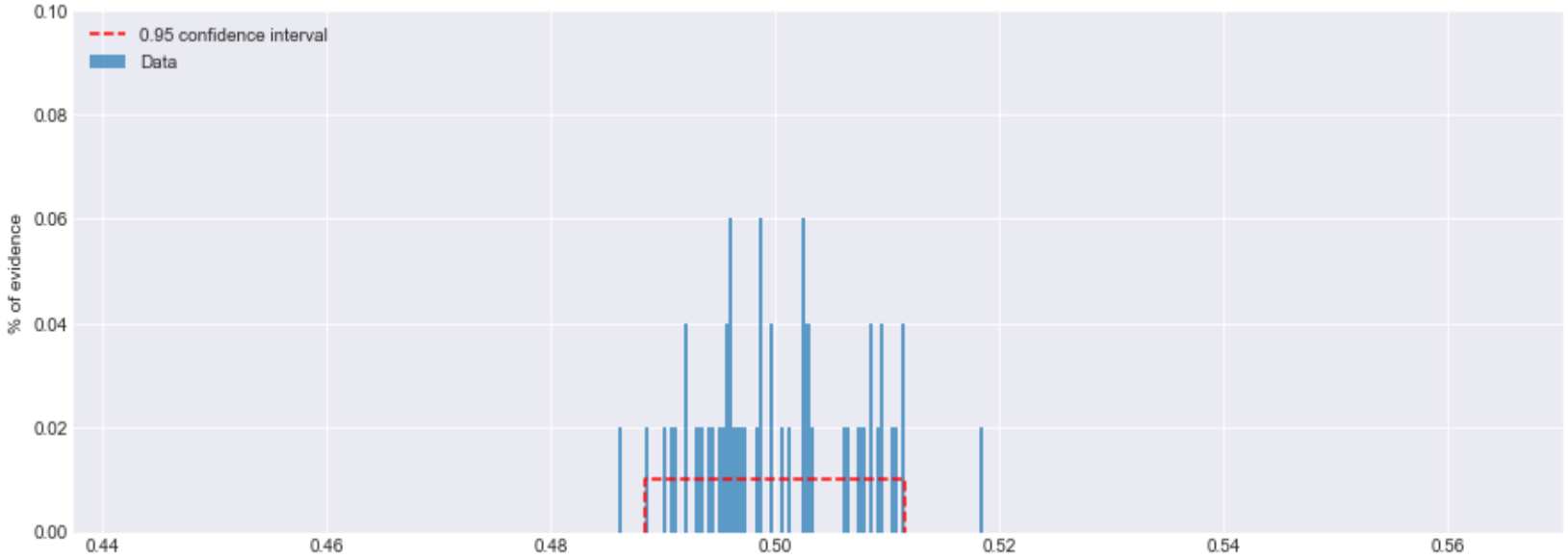
Цель теста. Провести сравнение доверительных интервалов для датасета из оригинальных вопросов и для датасета, состоящего из случайных наборов слов (или нерелевантных задаче предложений). В случае, если результат модели находится в 95% доверительном интервале, построенном на случайных данных, то модель индифферентна к входным данным

Алгоритм расчетов: 200 итераций, гаждый элемент датасета (текст) составляется из слов с повторением с длиной, равной средней по исходному датасету. Если подсчитать ключевую метрику качества невозможно без привлечения асессоров, то вместо нее считается косинусный коэффициент (cosine similarity) между эмбеддингами ответов LLM на оригинальных и сгенерированных вопросах.

Критерий
Метрика качества модели попадает в доверительный интервал для данных, сгенерированных случайно
Метрика качества модели не попадает в доверительный интервал для данных, сгенерированных случайно

2.4

Качество прогноза: индифферентность модели относительно входных данных



Mean	Std	Confidence interval
0.500	0.007	0.488 - 0.512

КОНЕЦ ЛЕКЦИИ 16.02