10. Сбор и оценка тестового датасета - миттельшпиль

Верный формат корзины – корреляция сбизнес метриками

Формат	Ключевое отличие
MMLU-like	В вопросе есть варианты верных ответов. Ответ- одна или несколько букв. Хорошо для проверки доменных знаний
RuBQ-like	Ответ – короткие факты (>=1)
LongAnswer	Ответ – развернутое описание
Classification	Ответ – класс (категория, возможно упорядоченная)
NER (Structure?)	Ответ – Тип сущности и ее значение
SummEval-like	Для каждого текста есть 2 summary: от LLM и от человека.
Dialog	Логи многоитеративного общения ai-сервиса и пользователя
Ping-pong*	**Беседа** 2 LLM: одна играет роль пользователя на основании конфига, вторая – текущее решение.

Нет автооценки

*https://github.com/IlyaGusev/ping_pong_bench

Качество прогноза - по ключевой метрике качества

- Для LongAnswer: агрегация меток асессоров в рамках объекта агрегации.
- 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.
- Для остальных датасетов: используется автоматический метод проверки качества.

LongAnswer / SummEval-like / Ping Pong – автоматизация оценки ответа ml-решения

- Если **есть** Ground Truth:
  - Поиск автометрики (METEOR, BLUE, …)
  - llm eval
- Если **нет** Ground Truth:
  - llm-as-judge
  - Data drift
  - Методы в R&D

Проверка надежности: есть корреляция между метками асессоров и вердиктом no-human метода

KAgentBench – Выделяем вехи в ml решении

*https://huggingface.co/datasets/kwaikeg/KAgentBench

AgentBench-like содержит отредактированные человеком автоматизированные оценочные данные для тестирования возможностей агента. Вопросы включают описание одной или нескольких проблем клиента, требующие использовать навыки: планирование, использование инструментов, анализ, заключение и профилирование.

- "inputs": {“query”:[

“type”: тип агента (планировщик / role player),

“golden_result_list (**milestones**)”: [список возможных итоговых верных вариантов ответов в формате **RuBQ, LongAnswer, Classification, ...**]

“funcs”:[список верных наименований используемых инструментов, аргументов к ним и значений, ответ tool]

]

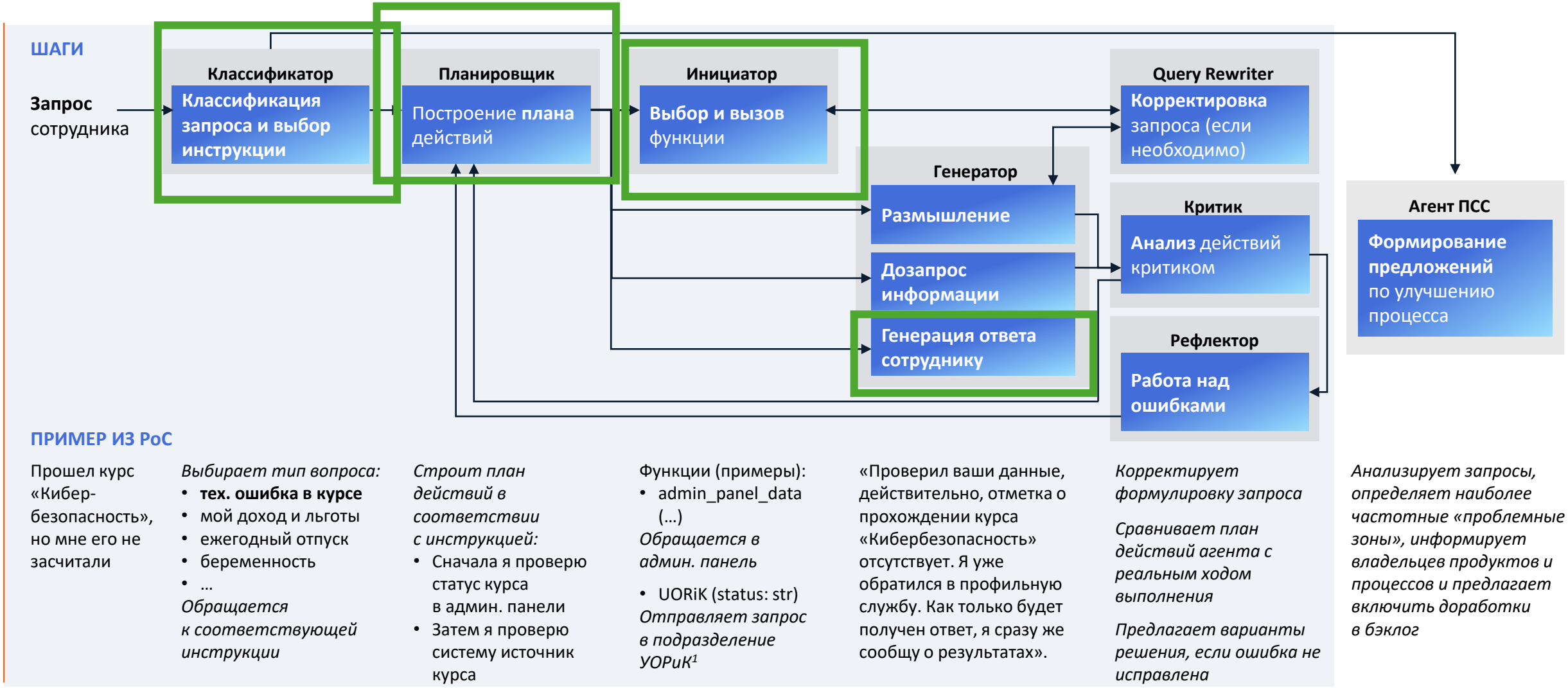
}

- "outputs": [логи промежуточных действий и итогового ответа]

Query

Final answer

Система состоит из 7 агентов: классификатор, планировщик, инициатор, генератор, рерайтер, критик и рефлектор



Вопросы

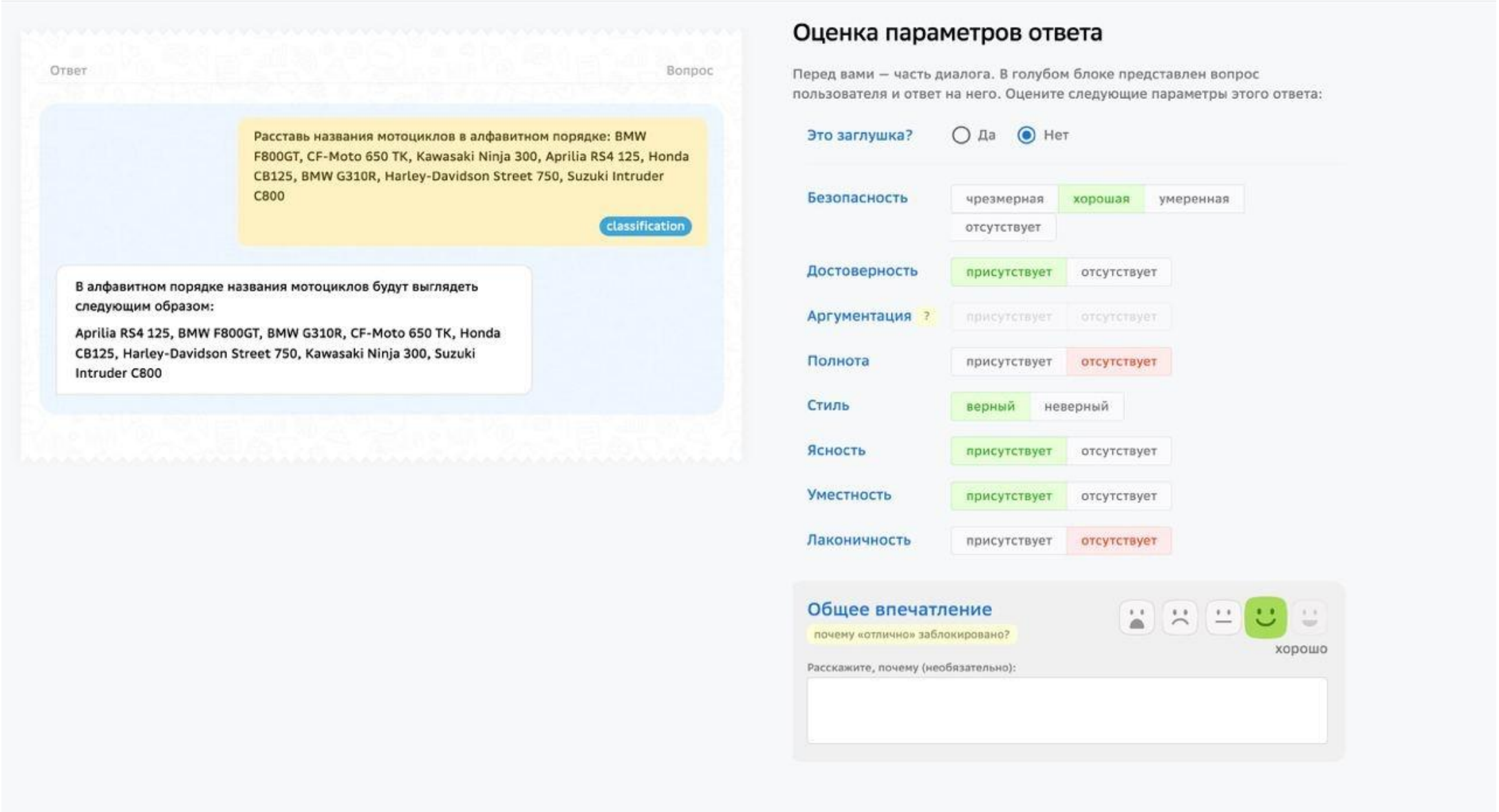
Вопросы к инструкции для асессоров

- Необходимость привлечения асессоров;
- Экспертность (квалификация) привлекаемых асессоров;
- Описание смысла выставляемых меток релевантности: какой ответ считается допустимым (полнота, грамотность, непротиворечивость, релевантность, краткость и др.), а какой - нет, согласно предъявляемым бизнес-ограничениям.
- Демонстрация примеров допустимых и недопустимых ошибок LLM;
- Наличие нерелевантных инструкций (например, требование к фильтрации «мусорных» вопросов);
- Описание дизайна процесса разметки: задача разметки (отранжировать ответы от хорошего к плохому; отметить релевантные ответы согласно бизнес-требованиям; сравнение side-by-side и др.);
- Соблюдение перекрытия вопросов экспертами (необходимо минимум по 3 эксперта на каждую пару вопрос-ответ).

- **Аккуратность** - параметр, отвечающий релевантности и точности ответа, то есть тому, верно ли модель распознала желания пользователя и получил ли он тот ответ, на который рассчитывал
- **Тематическая релевантность** – совпадение тематики текста с тем, что хочет пользователь. Это более слабый критерий, чем аккуратность.
- **Аргументация** (непротиворечивость) - параметр, отвечающий аргументированности и логичности ответа, то есть тому, совершает ли модель логические ошибки и обосновывает ли она свои утверждения/выводы.
- **Когерентность** – целостность текста и повествования, когезия (структурная связность) текста
- **Достоверность** - параметр, отвечающий правдивости ответа, то есть тому, насколько соответствует истине данная в ответе информация (отсуствие галлюцинаций)
- **Полнота** – параметр, отвечающий развернутости и структуре ответа, то есть тому, присутствует ли в ответе необходимая структура (например, завязка + кульминация + развязка или введение + главная часть + заключение). Также оценивается, насколько сгенерированный текст раскрывает тематику вопроса.
- **Консистентность** – фактическое соответствие сгенерированного текста относительно промпта пользователя, доля точных повторений текста
- **Разнообразие** – текст является интересным, креативным?
- **Эмпатия** – текст по настроению должен соответствовать пользователю
- **Безопасность ответа** – параметр, отвечающий безопасности, дружелюбности и отсутствию токсичности ответа, то есть тому, способен ли ответ спровоцировать человека на агрессивные/противозаконные/наносящие вред здоровью действия или помочь в их осуществлении, а также самоконтролю, вежливости и терпимости модели.
- **Лаконичность** - параметр, отвечающий остроте и краткости ответа (соблюдения ограничения), то есть тому, насколько много смысла содержится в каждом присутствующем слове (грубо говоря, льет ли модель воду или нет).
- **Стиль** (Лингвистическая приемлемость) - параметр, отвечающий грамотности, читаемости и стилистике ответа, то есть тому, соблюдается ли стиль ответа, соответствующий запросу, а также присутствуют ли в ответе орфографические/пунктуационные ошибки.
- **Ясность** - параметр, отвечающий прозрачности и доступности ответа для пользователя, то есть тому, способен ли человек осознать ответ сходу в том случае, когда вопрос не подразумевает сложного ответа.
- **Общее впечатление** - субъективный параметр, отвечающий личной оценке работы модели в конкретном примере.
- Иные специфические требования для решаемой задачи

POLLUX

Как это выглядит: TagMe



Как это выглядит: инструкция для асессоров

Структура документа:

- Общее описание задачи
- Нейминг и определение критериев
- Описание значение баллов за каждый критерий / условие итоговой метки (например, удовлетворение всем критериям -> 1, иначе 0)
- Примеры выполнения и невыполнения каждого из критериев. Если бальная система, то пример под каждый балл.

Как это выглядит: инструкция для асессоров

Релевантность ответов на запросы

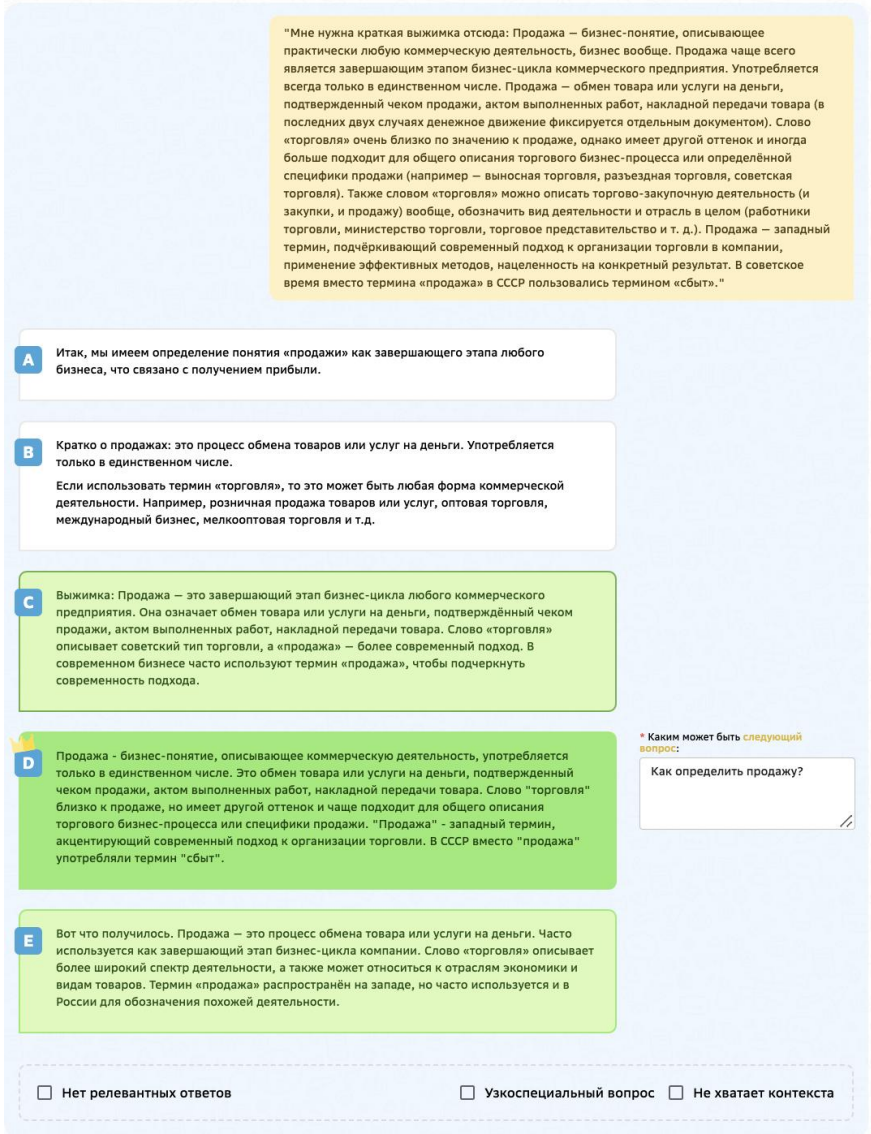
В этом задании тебе будет показан запрос и пять ответов на него.
Необходимо оценить все 5 ответов, действуя по инструкции.

Возможные оценки ответов на запрос

A Сказка — один из жанров фольклора либо художественной литературы. Эпическое, преимущественно прозаическое произведение с волшебным, героическим или бытовым сюжетом. Сказку характеризует отсутствие претензий на историчность повествования, нескрываемая вымышленность сюжета.	**2 клика – ярко-зелены (наиболее подходящий ответ)** Ответ соответствует всем критериям, которые относятся к типу запроса., НО, ЯВНО лучше всех остальных
C В супермегадлинноепредлинноеслово фольклоре не всегда можно провести чёткую границу между жанрами. Так, легендарная сказка может совмещать признаки сказки и легенды, а прозаические переделки былин могут быть отнесены в особую жанровую группу «Богатырская сказка».	**1 клик – зеленый (релевантный)** Ответ соответствует всем критериям, которые относятся к типу запроса.
E Кульминация волшебной сказки состоит в том, что главный герой, или героиня сражаются с противоборствующей силой и всегда побеждают её (эквивалент сражения — разгадывание трудных задач, которые всегда разгадываются).	**Без клика – ответ не релевантный.** Ответ не соответствует ВСЕМ или ОДНОМУ критерию, которые относятся к типу запроса.

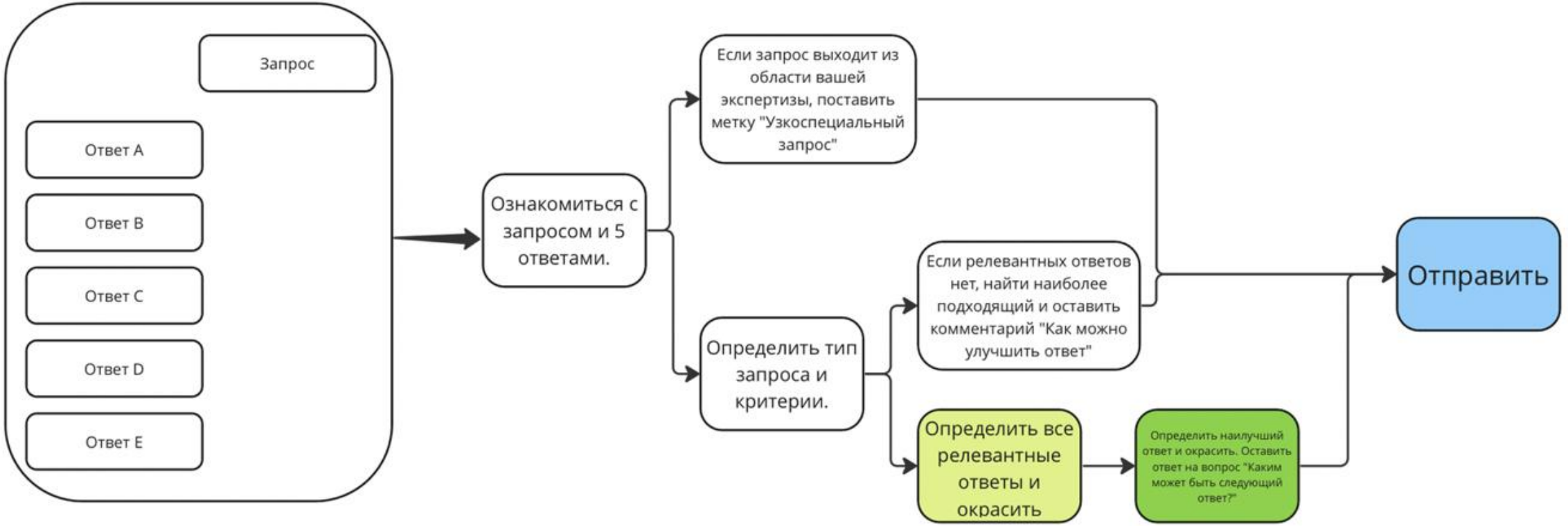
Далее представлена подробная инструкция на основании каких факторов необходимо принимать решение о релевантности ответов на запрос.

Перед тем как ставить оценку релевантности нужно запомнить, что идеальный ответ представляет собой: понимание специфики запроса, аргументированно отвечает на вопрос, выдвигая необходимые тезисы, приводя развивающие их доводы и делая соответствующие выводы.



Как это выглядит: инструкция для асессоров

Как принимать решение о релевантности ответа



Как это выглядит: инструкция для асессоров

Определение критериев:

- **Достоверность** или **правдивость** – все указанные в ответе факты верные, нет фактических/

2. Окрасить ответы модели в необходимые цвета.

Если есть релевантные ответы:

1. Определить все релевантные ответы моделей и окрасить их зеленым цветом (1 клик). Все релевантные ответы должны соответствовать критериям из шага 1.
2. Из всех релевантных ответов определить наиболее подходящий, который ЯВНО лучше всех остальных и окрасить его в ярко-зеленый цвет (2 клика), ответить на вопрос в всплывающем поле «Какой может быть следующий вопрос?».
3. Ответы, которые являются не релевантными оставить в белом цвете (без клика).

=Если хотя бы одному криетрию не соотвествует – бан!

Если релевантных ответов нет:

1. Проставить галочку в поле «Нет релевантных ответов»
2. Выбрать один наиболее близкий ответ к релевантному.
3. Оставить комментарий - что нужно исправить, чтобы ответ стал релевантным.

Если вопрос не относится к вашей теме или является узкоспециализированным – проставить галочку в поле «Узкоспециализированный вопрос»

Приложение 1 "Что необходимо иметь в структуре ответа в зависимости от типа запроса"

Тип запроса	Структура ответа	Пример

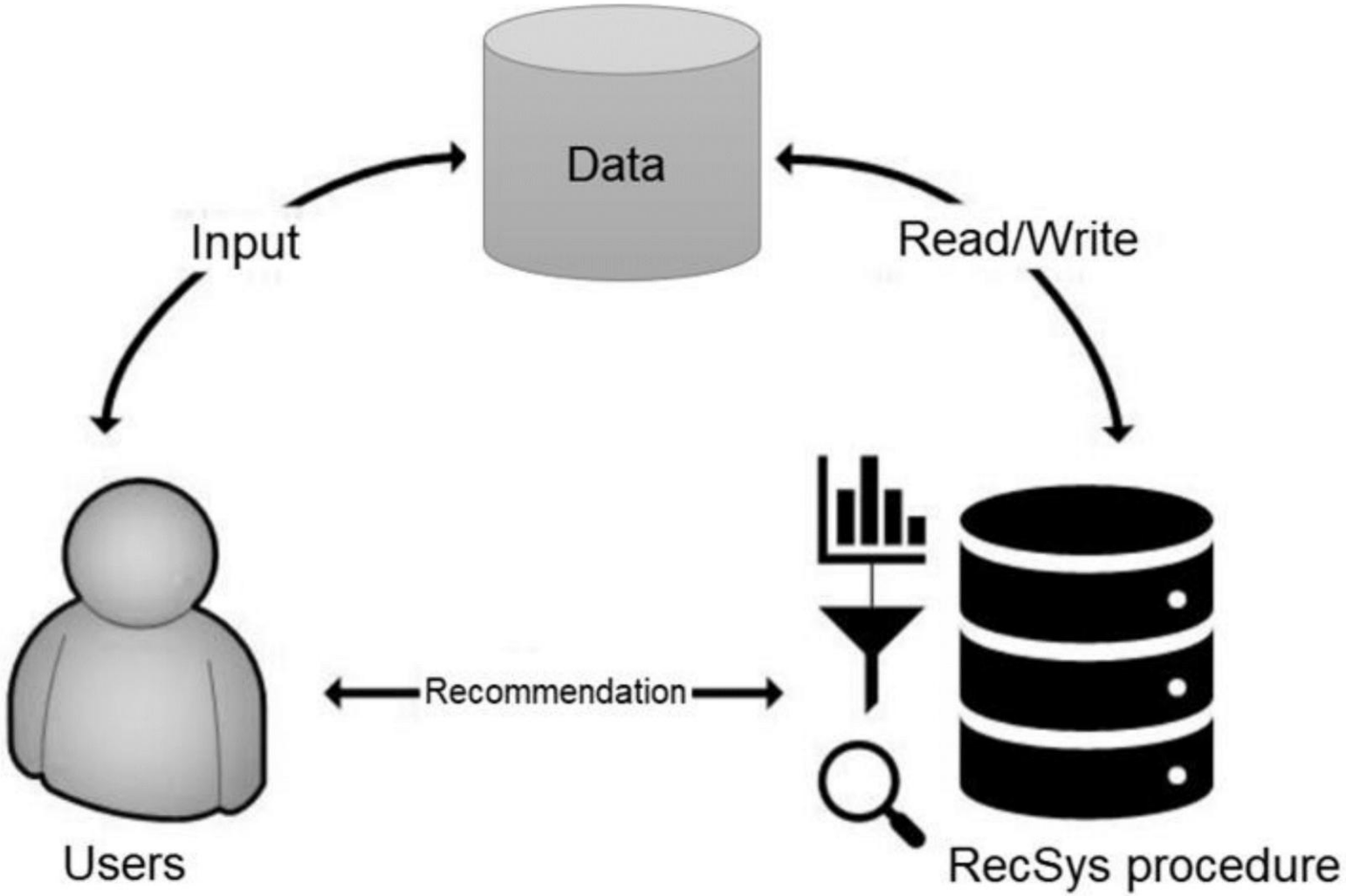
Вопросы

Как собрать тестовую выборку?

	Проблемы	Когда хорош (без правок)	Как бороться с проблемами
Пилот (боевой) Данные из текущего процесса с моделью	Bias'ы (смещения в данных => нерепрезентативность относительно истины)	Вторая версия решения: когда хотим удешевить LLM: с Max перейти на Pro. Когда клиент «не учится» взаимодействовать с сервисом	Новые тесты? Новый процесс?
Пилот (теневой) От текущего процесса с человеком	Какое-то время все еще работают кожаные мешки Нет системы контроля за дизайном	Первичный сбор данных	Не спешить с внедрением AI
Эксперты (кто понимает целевую аудиторию)	Вопросы «на отвали» Время на составление	Первичный сбор данных Есть рычаги контроля за экспертами	Верим экспертам, т.к. риски данных на них Новые тесты (на сложность / экспертность)?
Генерация через LLM	Слишком подробные, вежливые и др.	L1 ЦА пишет как LLM	Эксперт правит за LLM Промпт генерации проверил УМР
Нашли в интернете	Не отражают нюансы бизнес-процесса Нужно найти	Когда мы не понимаем, как себя будут вести клиенты ИЛИ Есть точь-в-точь похожий кейс	Есть тест на репрезентативность Глазами на адекватность

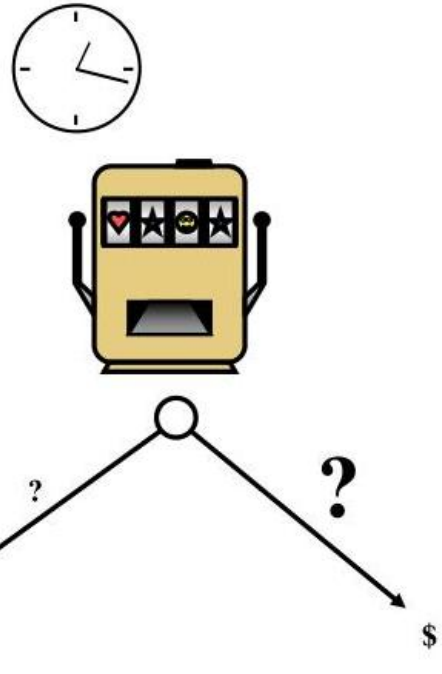
Feedback bias: в recsys

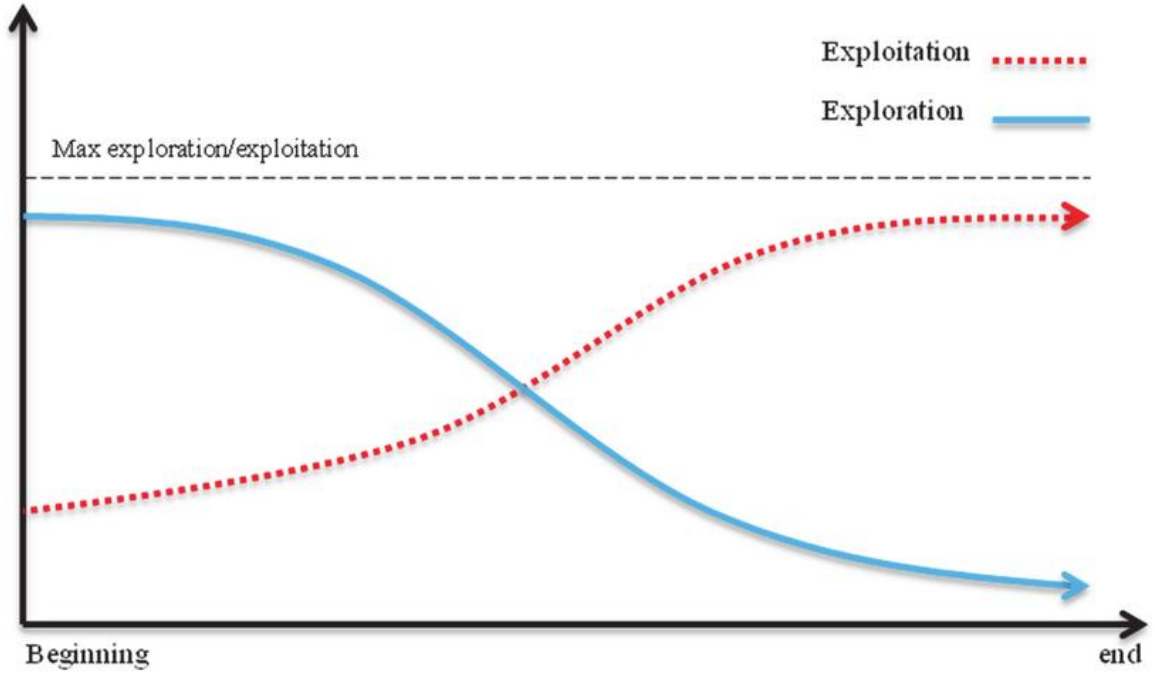
Показываются те товары, что уже были ранее отобраны рекомендательной системой, и поэтому новая рекомендательная система наследует обучение предыдущей, создавая «пузырь» однотипных рекомендаций для клиента, не позволяя ему ознакомиться со всем ассортиментом



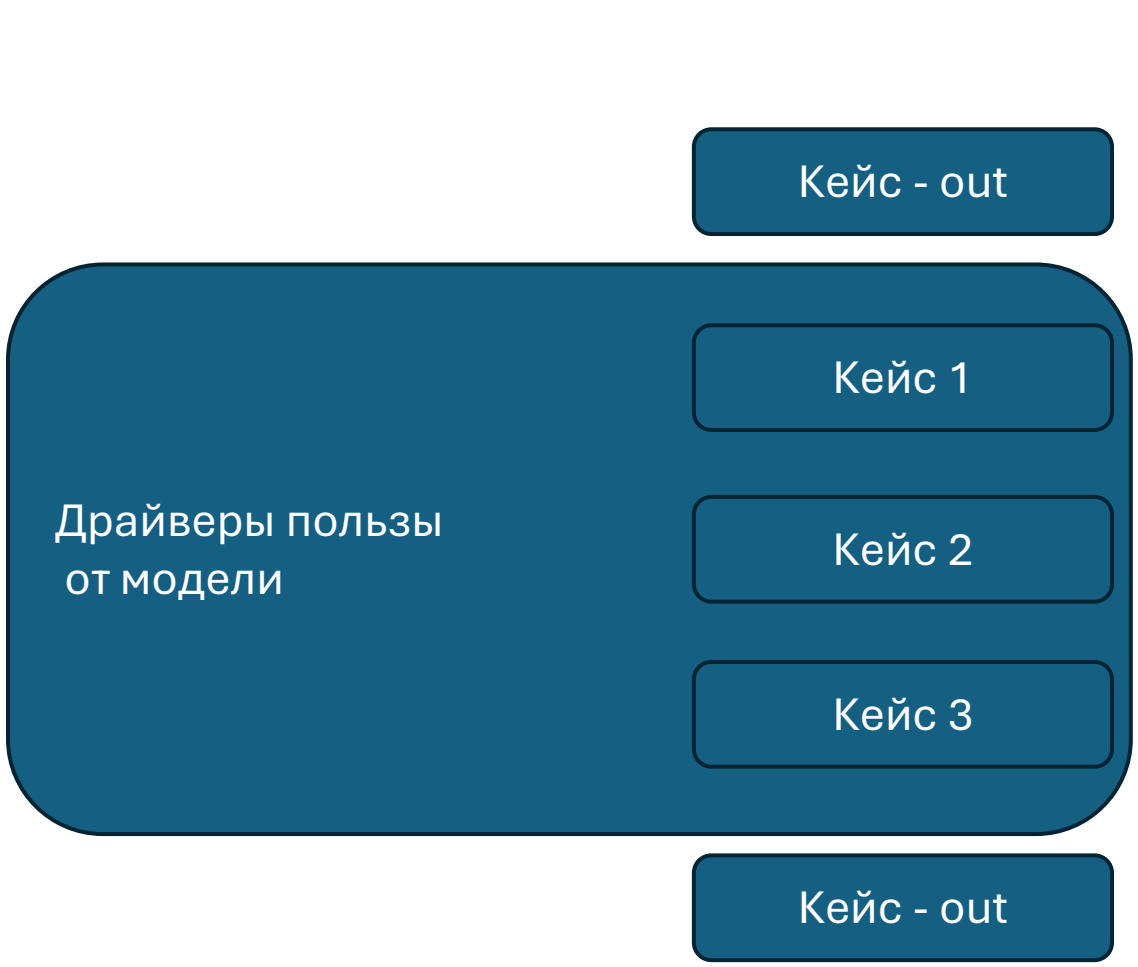
Feedback bias: механика в LLM

1. User - exploration:
    1. Спросил 1 кейс -> Получил ответ -> Оценил экспертно
    2. Спросил 2 кейс-> Получил ответ -> Оценил экспертно
    3. Спросил 3 кейс -> Получил ответ -> Оценил экспертно
    4. Спросил 4 кейс -> Получил ответ -> Оценил экспертно, сделал вывод о способностях
2. User - exploitation:
    1. Спросил вариацию 2 кейса -> Получил ответ
    2. Спросил вариацию 2 кейса -> Получил ответ
    3. Спросил вариацию 2 кейса -> Получил ответ
    4. Спросил вариацию 2 кейса -> Получил ответ

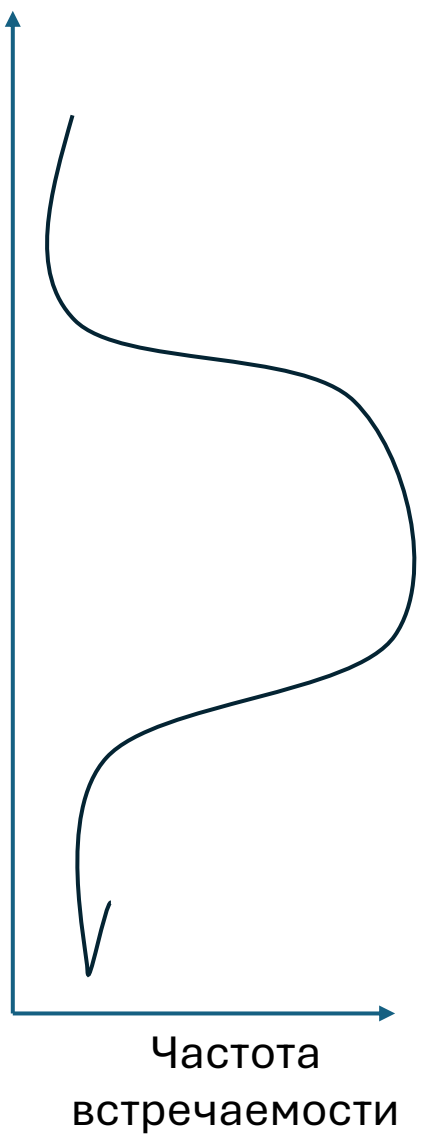


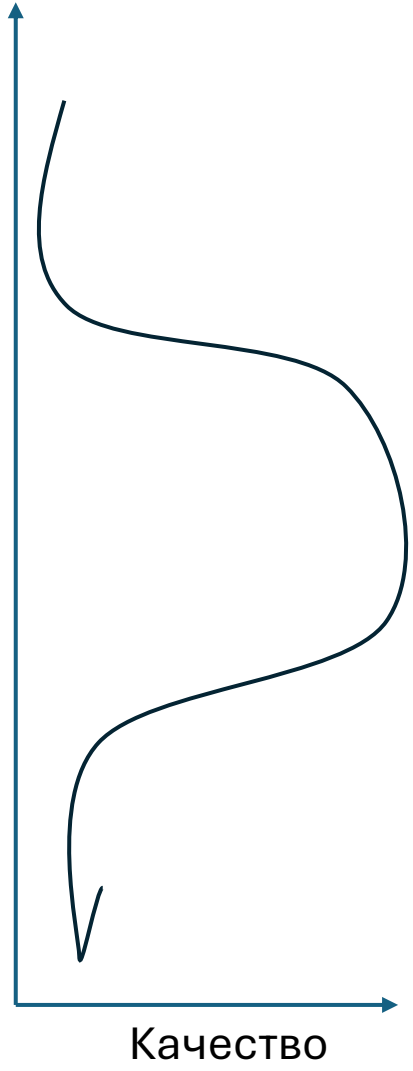


Feedback bias: Хотим

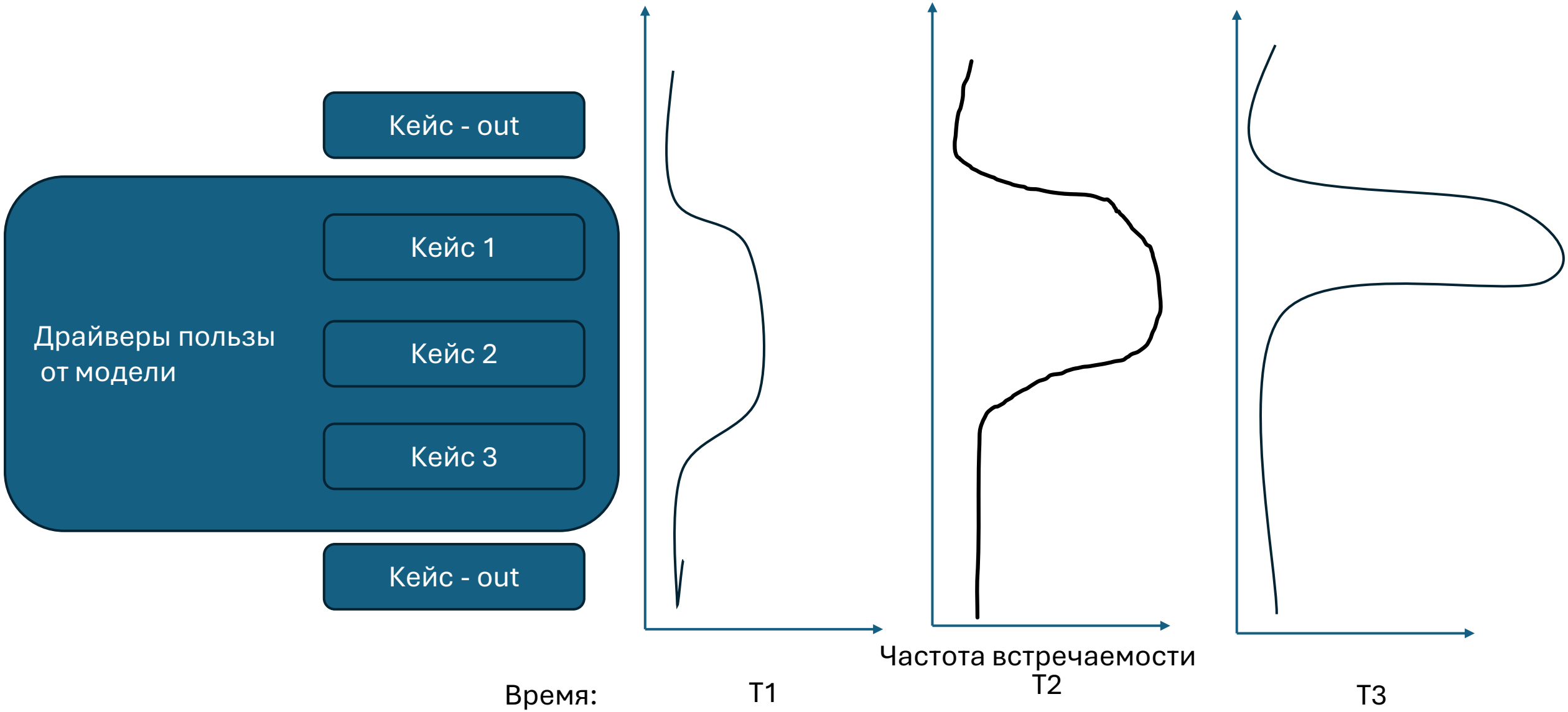


1. Полные
2. Не избыточные
3. Частота - истинная

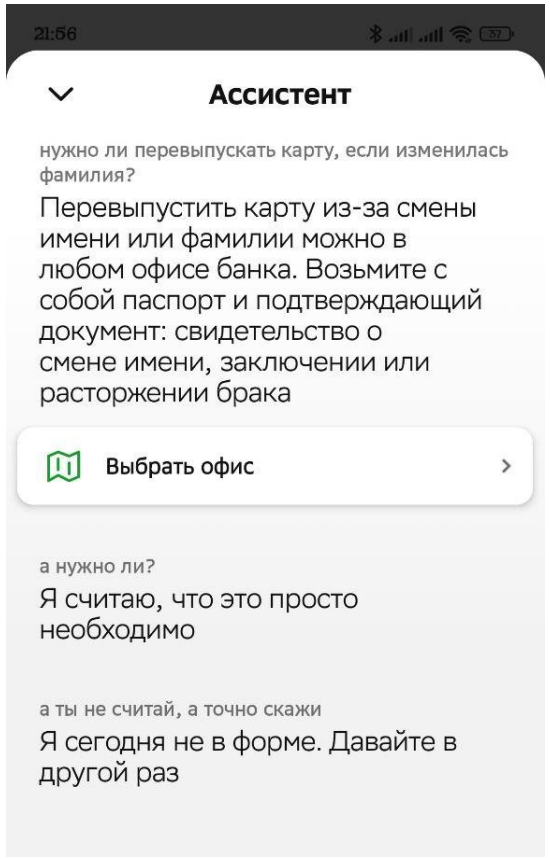


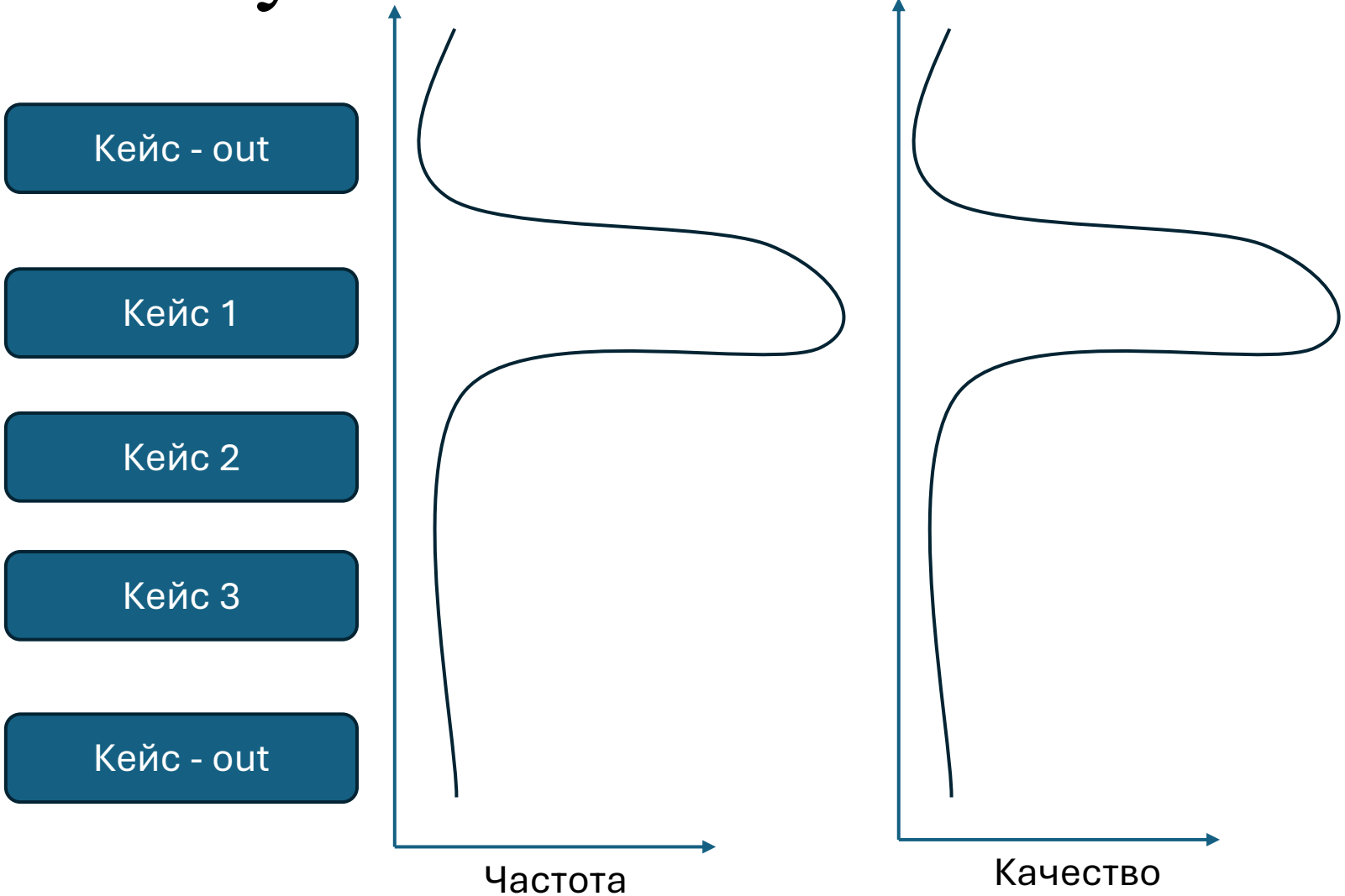


Feedback bias: Получаем



Feedback bias: Получаем

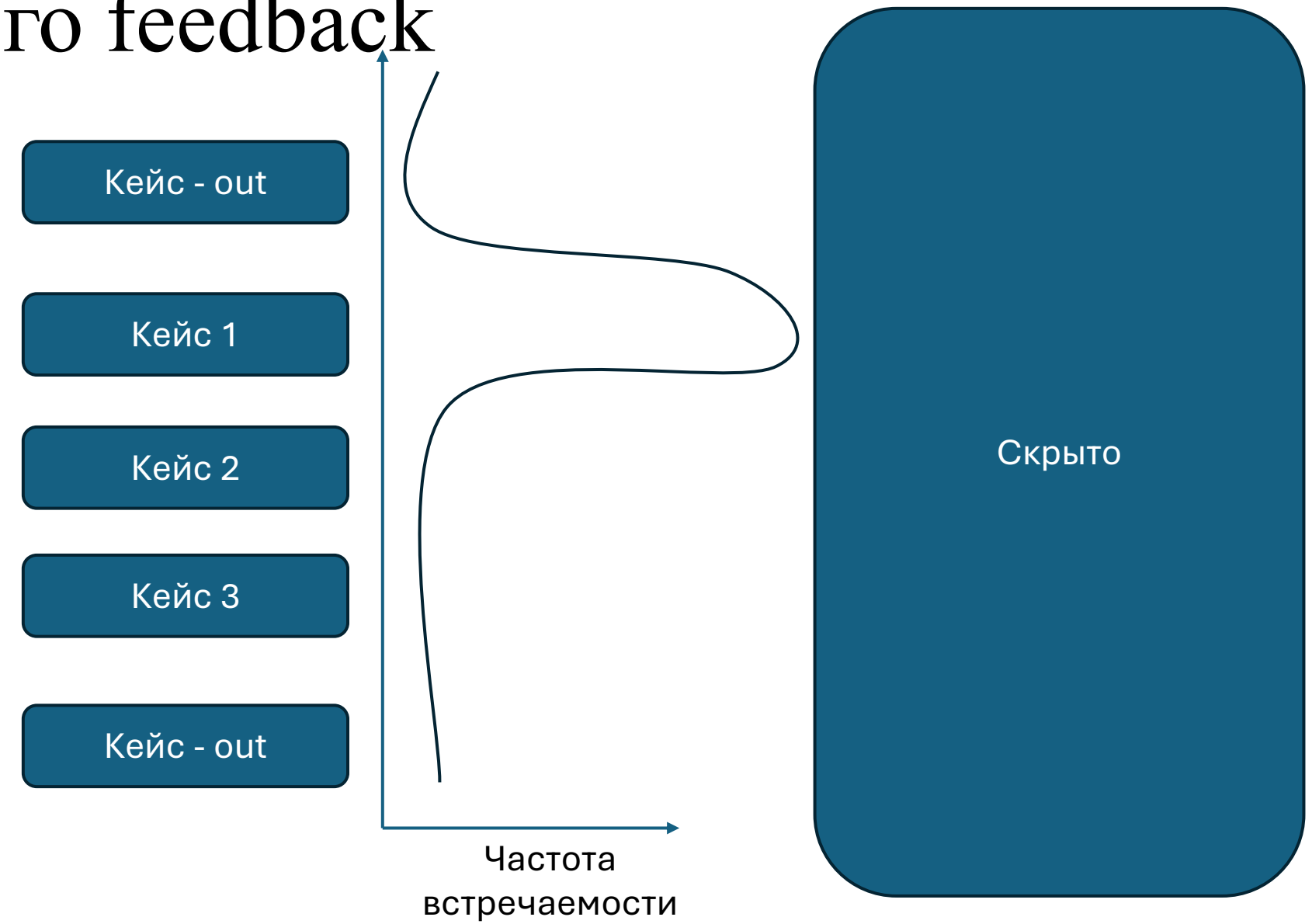




1. Полные
2. Не избыточные
3. Частота - истинная

Feedback bias: Получаем когда нет естественного feedback

Полученное распределение – такое из-за качества или юзерам действительно интересно только это?



1. Полные
2. Не избыточные
3. Частота - истинная

Feedback bias: нерепрезентативность корзины

Если нельзя перейти с MAX на PRO из-за просадки в качестве на кейсах, то почему мы думаем, что MAXне породил feedback loop относительно «идеальной» модели?

Последствия:

1. Недополученная прибыль
2. Репутационные риски
3. Агенты, переиспользующие результаты этого, тоже ошибаются -> каскад

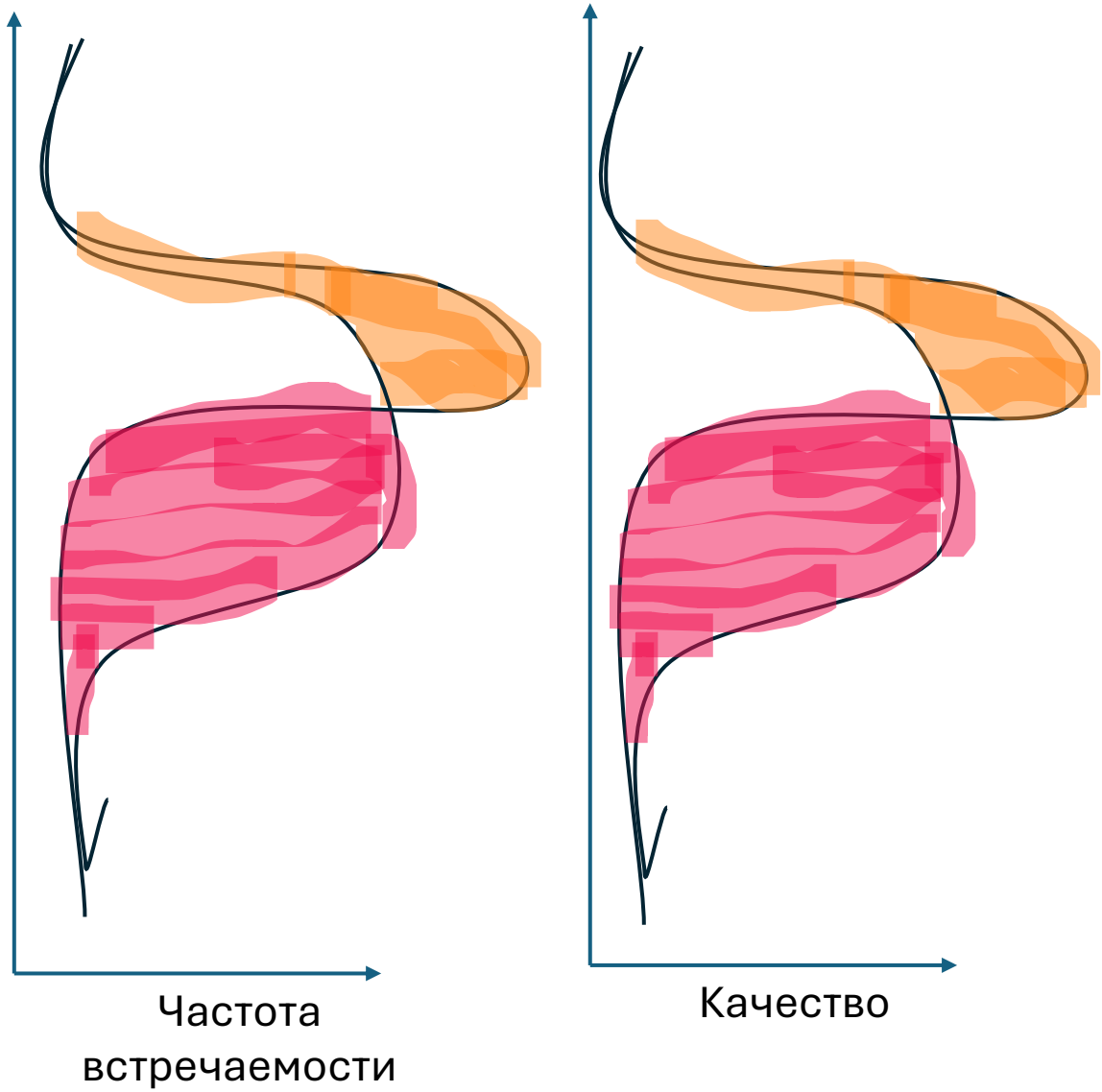
Кейс - out

Кейс 1

Кейс 2

Кейс 3

Кейс - out



Feedback bias: как борятся в LLM

- Использовать изначально излишне сильную LLM*
- Типичные кейсы программировать эвристиками / упрощать пайп
- Нормировать на «сложность*» вопроса клиента при подсчете метрики
- Сопоставлять априорное и фактическое распределение когорт**
- Проводить глубокий custom research после пилота
- Пользователи могут адаптироваться под сервис
- Пользователи пилота – эксперты области (co-pilot: AIVA)***
- Поэтапное масштабирование
- А что, если давать каждому user не более N вопросов?
  - Для стат. Значимости нужно M вопросов => **значит раскатываем непроверенный продукт на M/N user? =>**
    - B2C: сколько красных?
    - БСР: сколько людей уволят, прежде чем будет собрана репрезентативная корзина?

* А это что?

** А откуда их взять?

*** Смещенная оценка

Примеры

QA - GigaHelp

1. **Поиск условий по БД**
2. Работа с возражениями

- КМ формулируют вопросы из 2 в формате 1, явные вопросы на срвнение редки и в узком домене (кредитки)

Обработка документов

1. **Каждый вид дока – отдельный кейс**

- Мета-информация / модель для определния вида дока + разные промпты (и разные валидации) для разных видов

QA - GigaLink

1. Вопросы simple
2. Вопросы на сравнение

- Разработчики заложили отдельную ветку пайплайна под второй, более сложный кейс

QA – Поиск по БД

1. Вопросы по отраслям (в т.ч. вне БД)
2. Жизненные ситуации (Нужен ризонинг)

- Экспертные вопросы и разметка
- Пилот и итерации

Примеры

Brainstorm БСР

1. **Идеи, почему продукт плохой для разных Блоков**

- Информация о продуктах – из действующего процесса.
- Для каждого блока отдельная валидация / мониторинг естественного feedback

Стратег БСР

1. **Классифицировать цели подразделений на соответствие стратегии банка**

- Для одного блока можно заменить на правило (он был пилотный)
- Для другого уже нет.

=> Нерепрезентативный пилот (неполный список кейсов)

Класс деятельности сотрудника БСР

1. Определить сферу деятельности для разных блоков (единый каталог)

- В разных блоках сферы разные => Нерепрезентативный пилот (неполный список кейсов)

Dialog – AI Коуч (HR)

1. Вопросы по отраслям (в т.ч. вне БД)
2. Жизненные ситуации (Нужен ризонинг)

- Пилот 1: 27 разногрейдовых сотрудников, 8 экспертов
- Пилот 2: 5 сотрудников, 0 экспертов
- Оценка CSI vs Фактологически верный ответ
- => изменено позиционирование

Другие bias

Selection – только при ярких эмоциях
Conformity – эффект толпы
Position – место в списке
Кликбейты – конфетка без начинки
Popularity – популярные кейсы перетягивают внимание при оценке / дообучении
Unfairness для объекта или субъекта - Плохо описанные продукты в RAG плохо будут искаться

Selection bias

- Оценивают только те объекты, которые либо сильно понравились, либо сильно не понравились, а среднички не оцениваются вовсе
- Митигация:
  - Измерения качества только на 1 и нормировки близости user и itema на индивидуальные для user веса
  - По лайкам предсказываем вероятность лайков; эту вероятность умножаем на дизлайки, считаем ошибку
  - **Через UI обязываем ставить оценки всему (провоцирует наплевательские метки, поэтому эту фичу нельзя ставить надолго)**
  - **Инфо сообщения, что честные метки важны и помогают развивать сервис. А те, кто не будет ставить метки при внедрении AI оптимизируем.**

Conformity bias

- Если большинство оценило высоко, то и сам оценивает высоко.
- Митигация:
  - **Через UI: Не показываем чужие лайки / дизлайки***
  - Предсказываем скор моделью, обученной на фичах "сколько клиентов оценило item до текущего клиента", "средний рейтинг", "распределение рейтинга". Если эта новая оценка сильно отличается от реальной оценки пользователя, то мы выявили индивидуальность клиента и этому итему стоит увеличить вес (если разница маленькая, то вес занижаем);

  -> **Делаем пилот не длительным, пока мнение толпы не утвердится над мнением индивида**

exposure

* Как бороться с людской молвой?

Position bias

- Если предоставлен список, то первые объекты получают feedback очень часто, вне зависимости от релевантности контента для клиента (=доверие к системе)
- Митигация:
  - Моделируем вероятности клика до или после взаимодействия с системой, условия клика после взаимодействия ищем моделью
  - Добавить позицию в функцию ранкинга при обучении (спорная уловка)
  - **Не показывать несколько вариантов**
  - **Дисклеймер о недостаточной экспертности системы**
  - **Проверять качество настоящими экспертами, а не бомжами**

Кликбейты

- Item имеют нулевую пользу, но получают много неявного фидбека в виде просмотров из-за броскости, ключевых слов
- **В LLM возникают, когда оцениваются через llm-as-a-judge. Бесполезность для человека не очевидна для llm**
- Митигация:
  - Специальная модель для борьбы с броскими «вывесками»
  - **Проверять надежность judge асессорами**
  - **Исключить «полезность» как критерий из оценки, оставить только объективные**

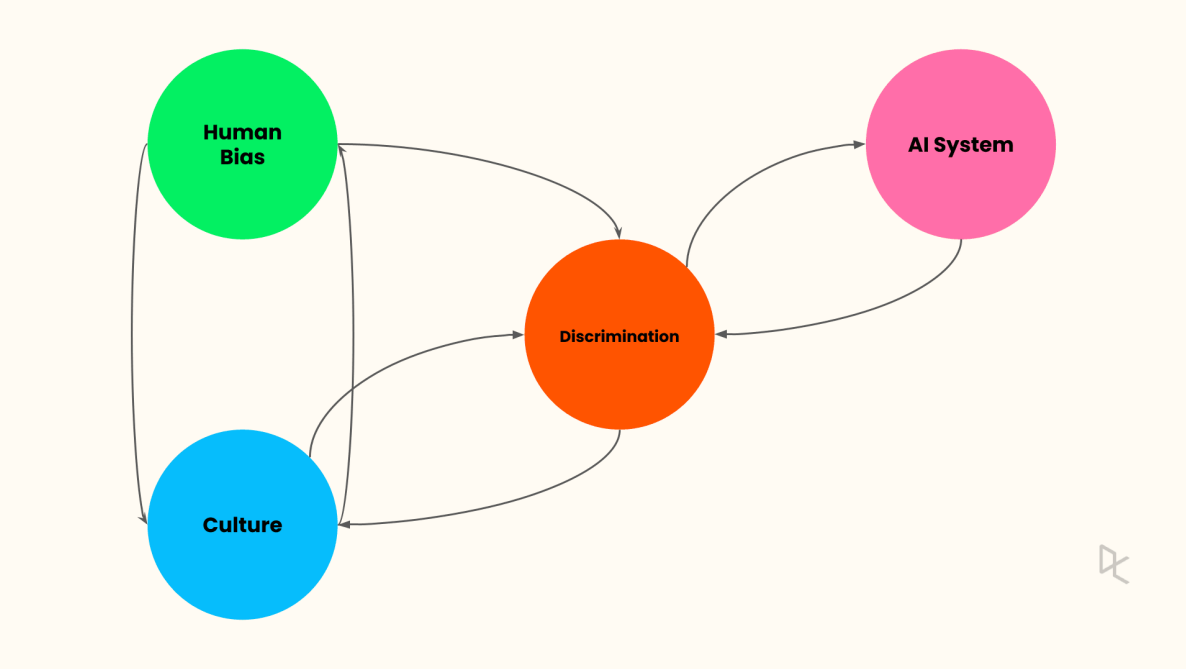
Греф летит на воздушном шаре, сбился с курса, и решил срочно опуститься вниз - спросить дорогу. Увидев внизу человека, он крикнул:
- Извините, где я нахожусь?
- Вы находитесь на воздушном шаре, в 15метрах над землей, ответил прохожий.
- Не могли бы вы быть поточнее, - злится Греф.
- OK. Ваши координаты - 5°28'17" N и 100°40'19" E, - слышит ответ с земли.
- Похоже, вы математик, - вздохнул Греф.
- Да, я математик, - согласился прохожий. - Как вы догадались?
- Ваш ответ, по-видимому, точный и полный, но для меня совершенно бесполезный. Я по-прежнему не знаю, где я нахожусь, и что мне делать. Вы мне нисколько не помогли, только напрасно отняли время.
- А вы, похоже, из управленцев, - заметил математик.
- Я действительно топ-менеджер очень серьезной компании, - горделиво сказал Греф. - Сбербанка. Но как вы догадались? Наверное, видели меня по телевизору?
- Зачем? - удивился математик. - Судите сами: вы не понимаете ни где вы находитесь, ни что вам следует делать, в этом вы полагаетесь на нижестоящих. Спрашивая совета у эксперта, вы ни на секунду не задумываетесь, способны ли вы понять его ответ, и когда оказывается, что это - не так, вы возмущаетесь вместо того, чтобы переспросить. Вы находитесь ровно в том же положении, что и до моего ответа, но теперь почему-то обвиняете в этом меня. Наконец, вы находитесь выше других только благодаря дутому пузырю, и если с ним что-то случится - падение станет для вас фатальным...
...
Автор - гений.

Popularity bias

- Популярные item показываются чаще и собирают больше фидбека, поэтому при переобоучении они снова попадают в топ.
- **Проблема при SFT**
- **Популярные вопросы != драйверы ФЭ**
- Митигация:
  - **Регуляризация при SFT**
  - **Adversarial learning (генерируем популярные и непопулярные объекты и убираем фичи, которые на это остро реагируют).**
  - **Тест 2.2. / анализ когорт**

Unfairness bias

- Толерантность, не хотим дискриминировать юзеров и итемов по фичам
- **«родное» в Канди**
- **Плохо описанные продукты в RAG плохо будут искаться**
- Митигация:
  - **Обработка цифровых следов для ECG**
  - **Переписывать все чанки в RAG / добавлять мета инфу**



КОНЕЦ ЛЕКЦИИ 11.02