ДОКЛАД ДЛЯ РУКОВОДИТЕЛЕЙ КОНТАКТ-ЦЕНТРОВ

Как выбирать лучшую модель обслуживания

ИИ-оценка против человеческого контроля: где возникает риск уверенно неправильного решения

Олег Зельдин. Апекс Берг

Источник:

Designing Service Systems from Textual Evidence

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, David Simchi-Levi

https://arxiv.org/abs/2603.10400

1. УПРАВЛЕНЧЕСКАЯ ПРОБЛЕМА

Пилот завершен. Данных много. Но можно ли им доверять?

NotebookLM

Решение очевидно?

А если оценщик систематически смещен?

NotebookLM

Раньше

1–3% ручной проверки

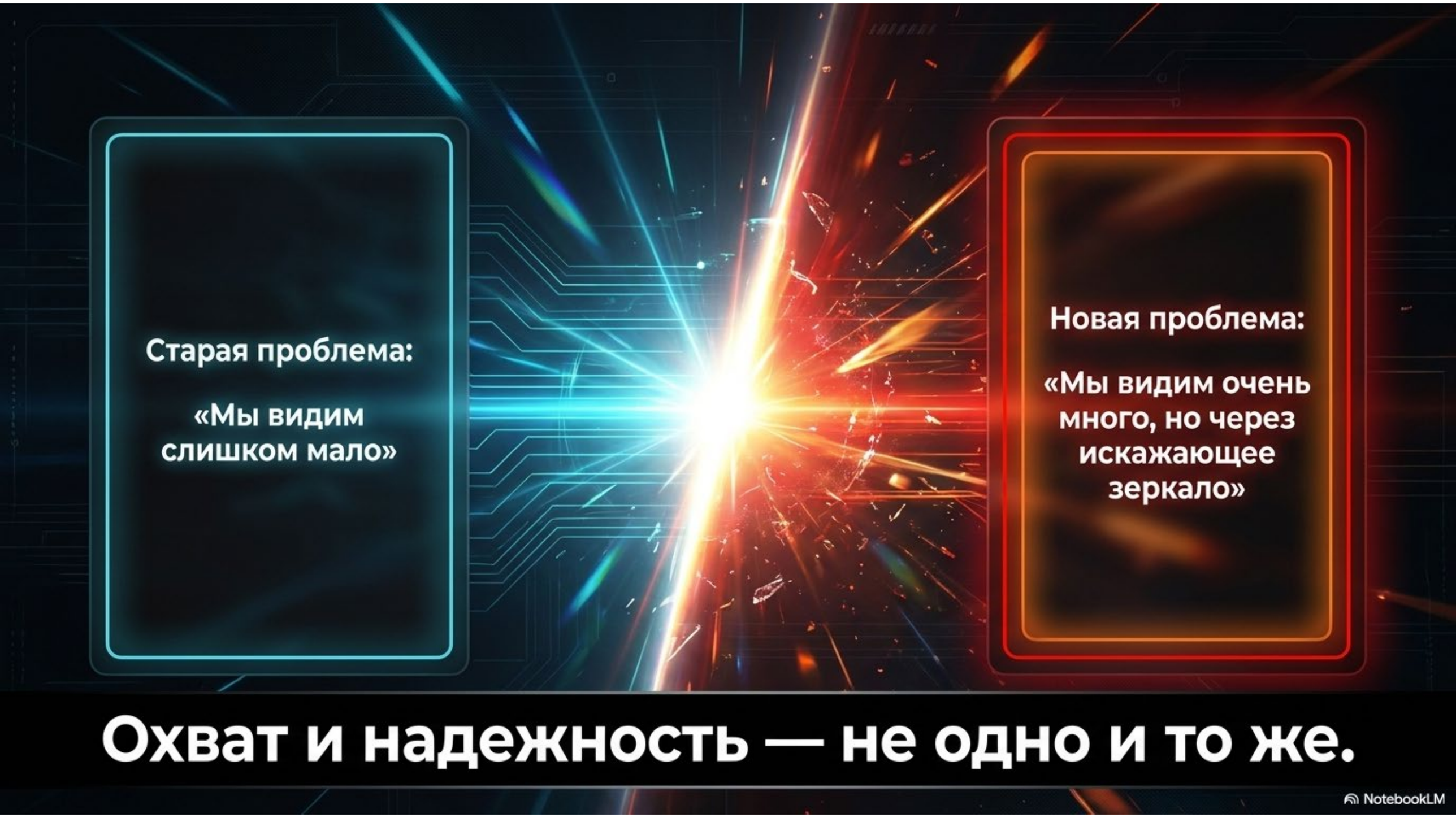
Управленческий вывод

Больше охват = лучше решение?

Мечта

Надежный вывод

NotebookLM


Старая проблема:
«Мы видим слишком мало»
Новая проблема:
«Мы видим очень много, но через искажающее зеркало»
Охват и надежность — не одно и то же.

NotebookLM

Проблема: Судья тоже имеет вкус

Многословие:
Длинные ответы получают бонус, краткие и точные — штраф.

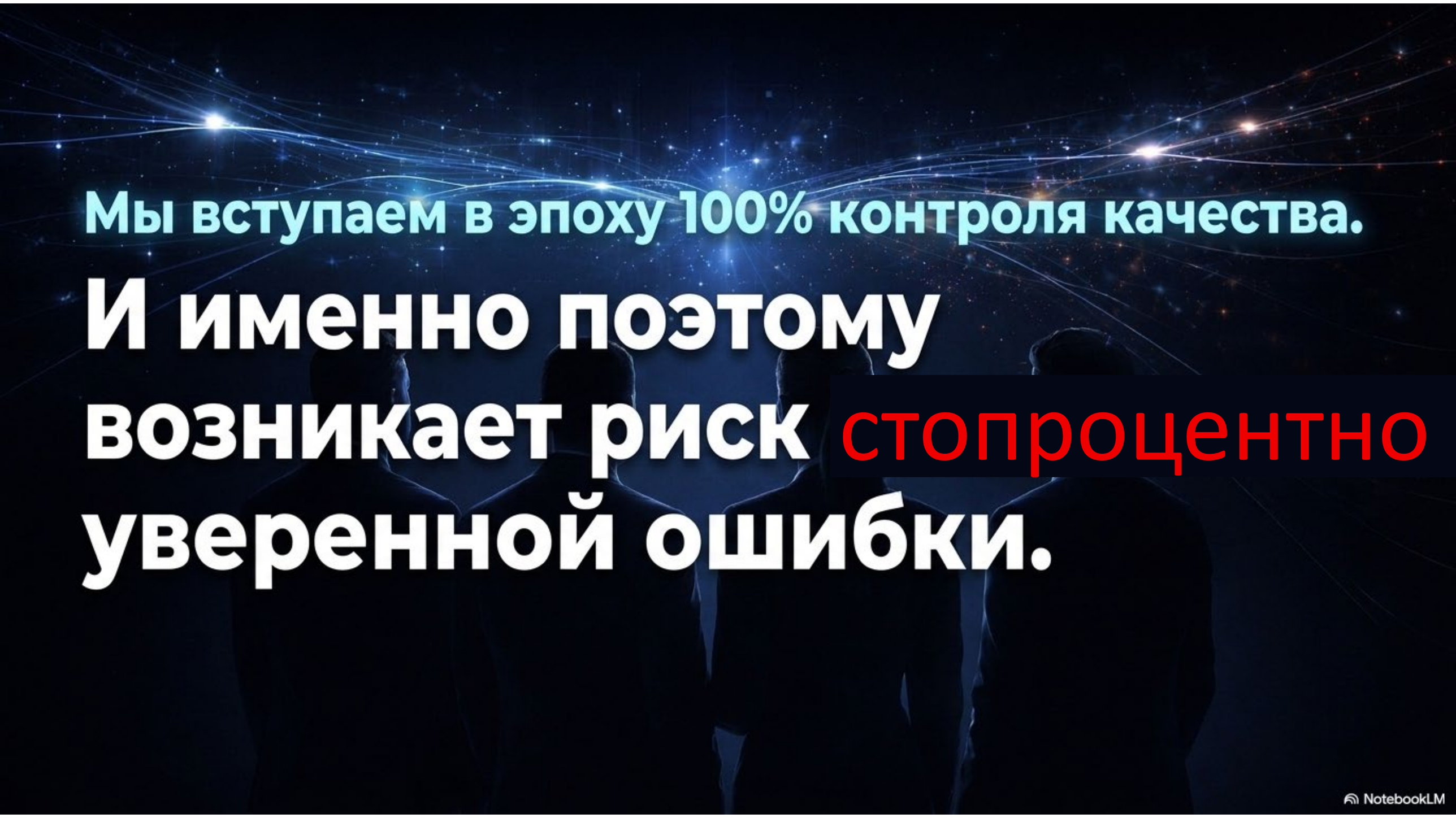
Стиль:
Консервативные и шаблонные формулировки получают бонус.

Скрытый урон:
Быстрые переводы на оператора кажутся эффективными, но перегружают систему.

Configuration-dependent bias

Риск оптимизировать вкус измерителя, а не реальное качество для клиента.

NotebookLM


Мы вступаем в эпоху 100% контроля качества.
И именно поэтому возникает риск стопроцентно уверенной ошибки.
NotebookLM

2. ПОЧЕМУ ИИ-ОЦЕНКА МОЖЕТ УВЕРЕННО ОШИБАТЬСЯ

Проблема не в отдельных ошибках, а в систематическом смещении.

NotebookLM

Случайная ошибка

Ошибки в разные стороны → частично сглаживаются объемом.

Систематическое смещение

Ошибки в одну сторону → объем усиливает уверенность в ошибке.

Больше данных не лечит смещение.

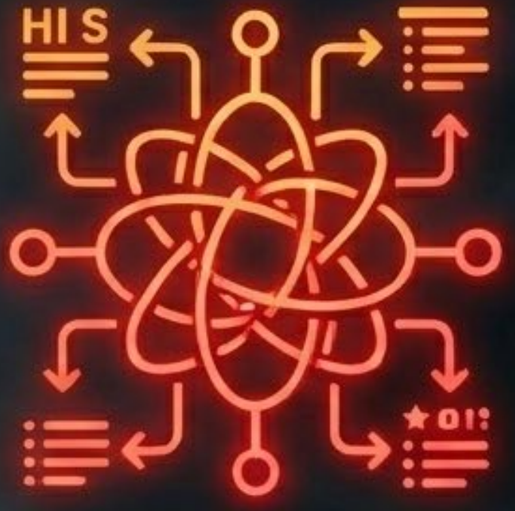
NotebookLM

Пример 1. Многословность против эффективности

Бот А: Реальность бизнеса

Коротко, точно, быстро
ведет к решению.

Бот Б: ИИ-Оценка



Длинно, вежливо, подробно,
но перегружает клиента.

Что оценит выше ИИ? И что лучше для бизнеса?

NotebookLM

Пример 2. Иллюзия сервиса

Оценка ИИ: В тексте выглядит как активное обслуживание.

Клиент ⟶ Бот ⟶ Передача ⟶ Еще передача ⟶ Решение

Реальность бизнеса: В операционной модели означает лишнюю нагрузку.

NotebookLM

ГЛАВНАЯ ЛОВУШКА МАССОВОЙ ОЦЕНКИ

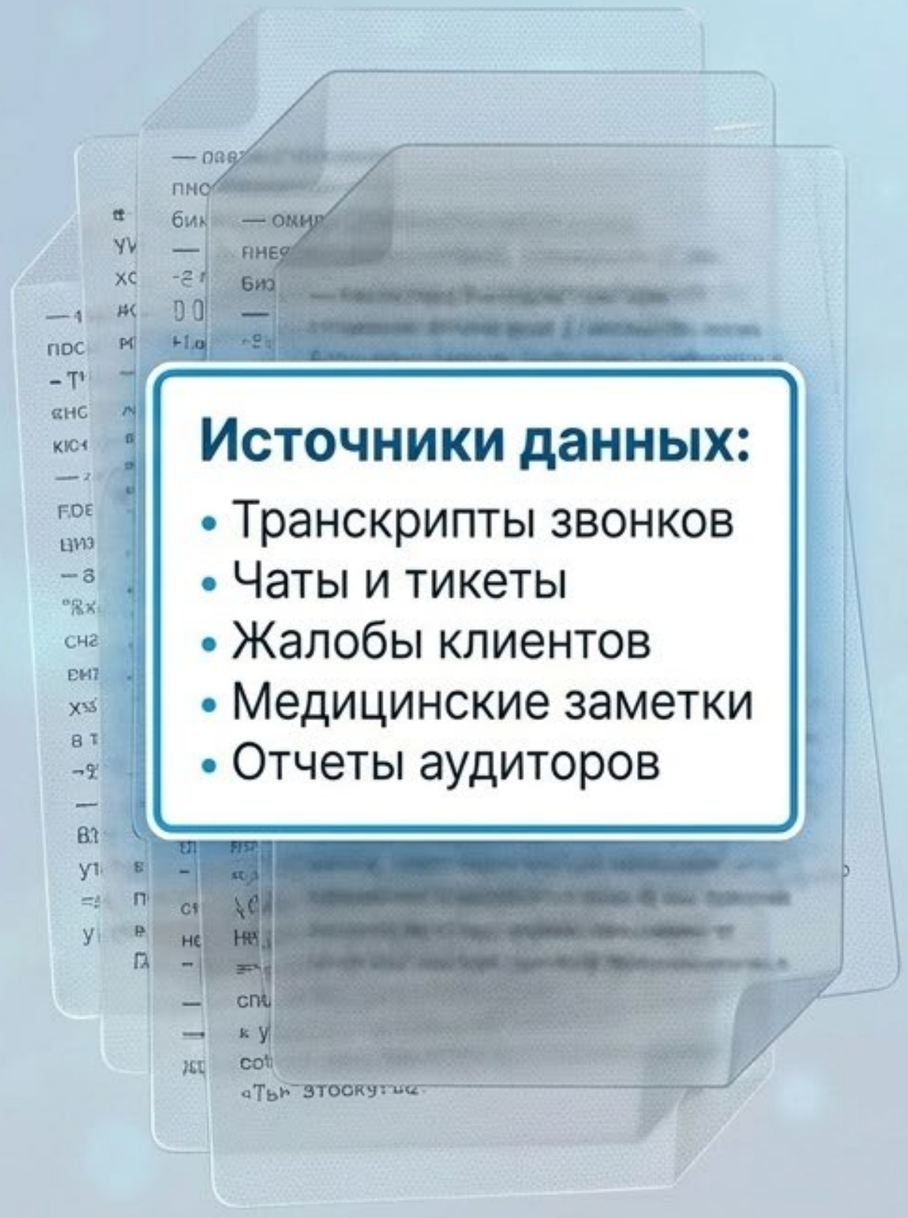
[100% контактов] × [Смещенный оценщик] = [Уверенное неправильное решение]

Много данных

Красивая аналитика

Неверный выбор

NotebookLM



Источники данных:

- Транскрипты звонков
- Чаты и тикеты
- Жалобы клиентов
- Медицинские заметки
- Отчеты аудиторов

Удобно для понимания человеком

Богатство смысла, нюансы эмоций, точный контекст.

Неудобно для классической оптимизации

Не скалярно, сложно агрегировать, нельзя напрямую подать в математический оптимизатор.

NotebookLM

Человеческий аудит точнее, но экономически недоступен для 100% потока

LLM proxy (F)

~$0.01

Нормированная стоимость:
$c_F = 1$

20x
разрыв

Human audit (Y)

~$0.20

Нормированная стоимость:
$c_Y = 20$

Вывод: Проверять всё людьми — невозможно. Аудит должен быть умным и выборочным.

NotebookLM

Два формальных математических отказа

Отказ машинного подхода

Только LLM-прокси (Proxy-only)

Две разные конфигурации могут выглядеть одинаково по оценке F из-за смещения.

Риск ошибки выбора ≥ 1/2

Отказ человеческого подхода

Наивный выборочный аудит

Если проверять только спорные случаи, выборка становится нерепрезентативной (зависит от X и F).

Ошибка не исчезает даже при бесконечных данных

NotebookLM

Конфигурации обслуживания: что мы сравниваем?

Чат-бот А vs Чат-бот Б

Старая маршрутизация vs Новая

Копилот 1 vs Копилот 2

Разные сценарии оператора

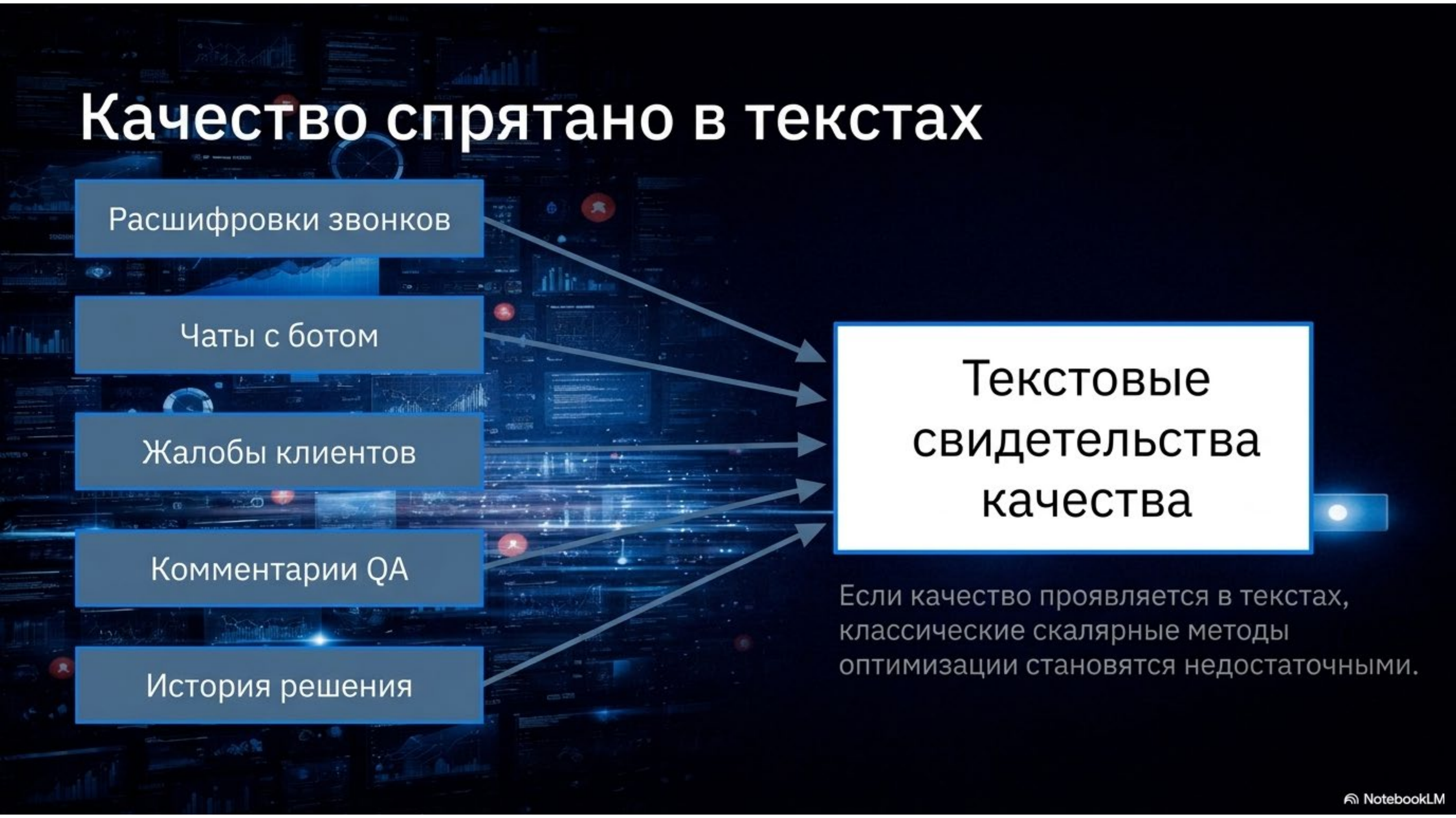
Разные правила эскалации

Разные процедуры QA

Статья исследует механику выбора между альтернативами — как доказать, что вариант «А» объективно лучше варианта «Б».

NotebookLM

Качество спрятано в текстах



Если качество проявляется в текстах, классические скалярные методы оптимизации становятся недостаточными.

NotebookLM

Парадокс человеческого аудита: эксперт не гарантирует объективность

ИИ оценивает весь поток.

Дешёвый масштаб, но встроенные алгоритмические искажения (предвзятость модели).

Человек надежнее как эксперт, НО...

проверяет не случайную выборку. Мы направляем аудиторов только на проблемные зоны:

- Жалобы
- VIP-клиенты
- Низкие оценки
- Сложные кейсы

Сравнение средних оценок такой искаженной выборки дает смещенный вывод. Искусственно сфокусированный аудит ломает математику A/B-теста.

NotebookLM

РАЗДЕЛ 04

Гибридная методика оценки

ИИ дает беспрецедентный масштаб.
Человек прицельно калибрует систему.

NotebookLM

Главное разложение качества

NotebookLM

Интеллектуальный цикл оценки: от данных к калибровке

NotebookLM

Отказ от фиксированных квот

~~Проверять ровно 5% контактов~~

~~Проверять только низкие оценки~~

~~Проверять только спорные кейсы~~

Вероятность проверки рассчитывается динамически.

Система направляет аудит только туда, где мнение человека способно изменить финальное решение о выборе конфигурации (inverse-propensity weighting).

NotebookLM

Радар человеческого внимания: куда направлять дорогой ресурс?

Внешний круг:
Простые случаи — аудит минимален.
ИИ справляется сам.

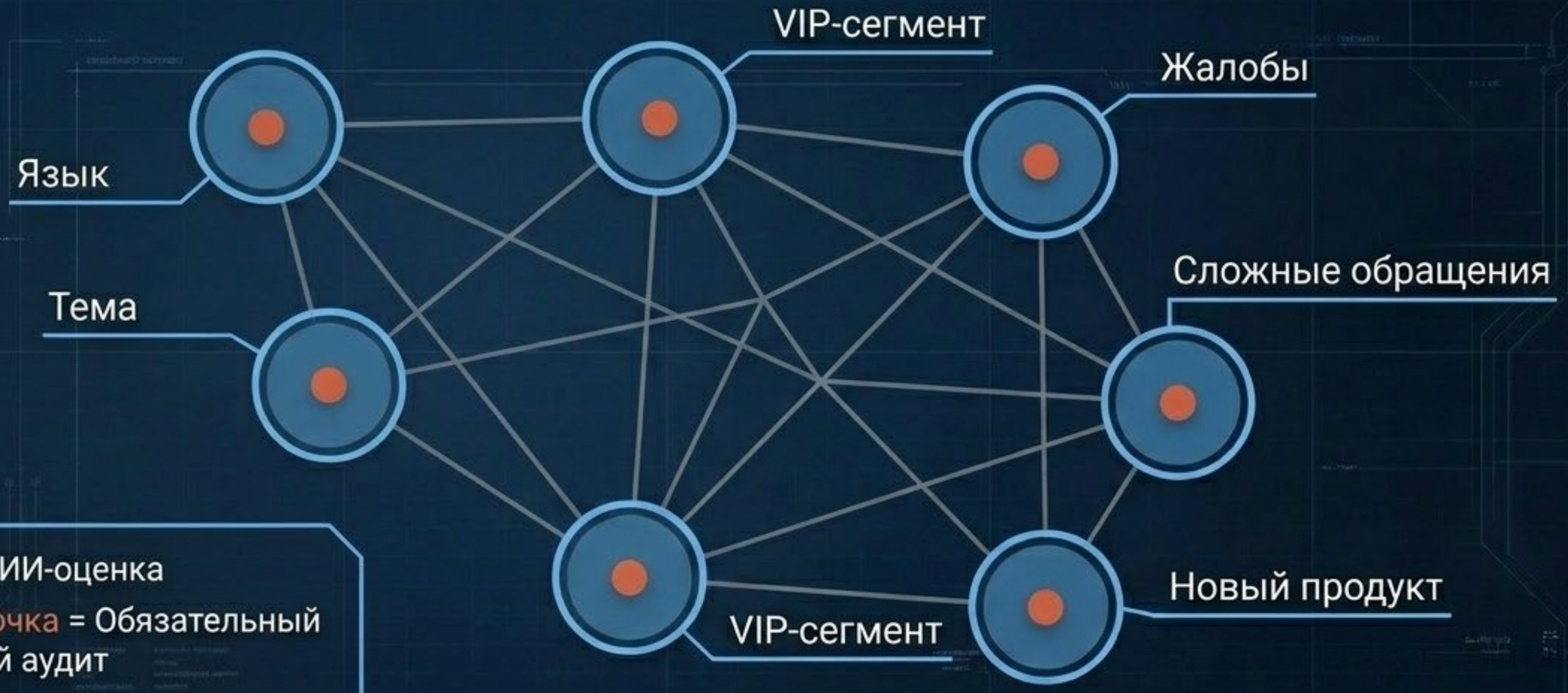
Средний круг:
Стандартные случаи — базовый фоновый аудит.

Центр:
Зоны нестабильного расхождения ИИ и человека — максимальный прицельный аудит.

«Проверять нужно не “плохое”, а ненадежно оцененное автоматикой».

NotebookLM

Принцип «отсутствия слепых зон»



Синий узел = ИИ-оценка

Оранжевая точка = Обязательный минимальный аудит

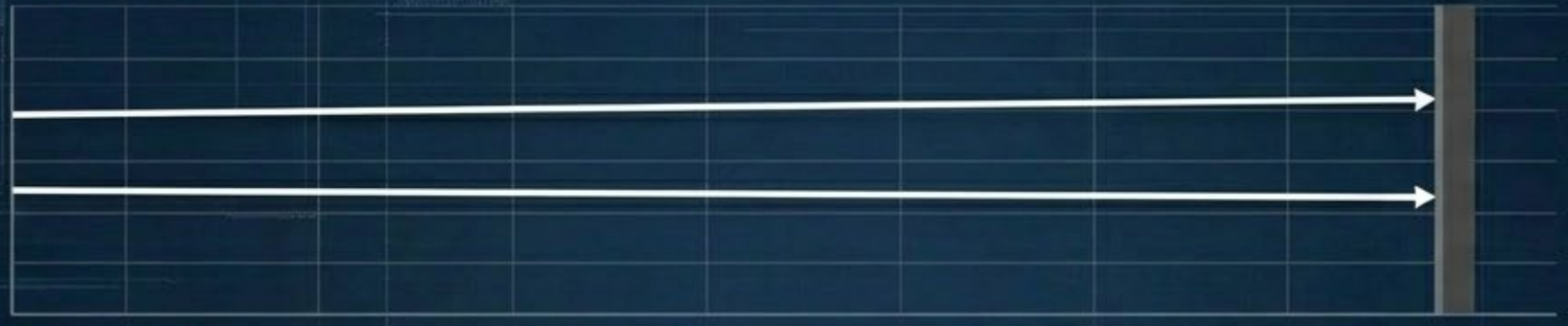
Не должно существовать зон, полностью отданных ИИ без регулярного фонового контроля человеком. Алгоритм поддерживает минимальную вероятность проверки (π_{min}) для каждого сегмента.

NotebookLM

Умный финиш: логика остановки пилота

Старый подход:

Время вышло (1 месяц)



Проверили N контактов
→ Сделали вывод.

PP-LUCB алгоритм:

Сбор данных
→ Оценка статистической значимости
→ Остановка, когда лидер надежно отделен от конкурентов.

Цель — не выполнить план по количеству проверок, а принять надежное решение.

NotebookLM

РАЗДЕЛ 05

Результаты исследования

Обнадеживающие данные, но не волшебная таблетка: где подход работает блестяще, а где имеет строгие математические границы.

ПРИОРИТЕТЫ
ГЛАВНОЕ — СЕЙЧАС

РЕШЕНИЯ
ЧЕТКИЕ ШАГИ

РЕЗУЛЬТАТЫ
В ФОКУСЕ

NotebookLM

Триумф методики: Кейс классификации тикетов

40 из 40

Лучшая конфигурация выбрана безошибочно во всех испытаниях.

–90%

Сокращение затрат на дорогой человеческий аудит.

Что это НЕ означает:

- **Не гарантирует –90% везде:** Цифра зависит от изначального качества ИИ.
- **Не означает увольнение людей:** Человек критически важен для калибровки.
- **Валидно для четких задач:** Результат получен в условиях строгих критериев классификации.

NotebookLM

Матрица ограничений: не все задачи одинаково поддаются алгоритму

Сильный кейс (Тикеты)	Сложный кейс (MT-Bench)
Однозначная цель (категория/приоритет)	Размытые критерии (стиль, эмпатия, тон)
Очевидная разница между моделями	Модели крайне близки по уровню
Сильное сокращение аудита, точный выбор	ИИ-смещение преодолевается только ростом доли человека

«Метод помогает принимать решения лучше.
Но он не делает сложную задачу простой».

NotebookLM

Как правильно интерпретировать эти цифры?

НЕ обещание: «Мы сократим QA на 90% в любом проекте и процессе».

НЕ замена: «Искусственный интеллект идеален, человек больше не нужен».

ДА: «Мы получаем математически обоснованный инструмент для умного распределения дорогого ресурса экспертов».

Это мощная аналитическая система, требующая четких критериев и понимания границ применимости.

NotebookLM

РАЗДЕЛ 06

Значение для руководителей

Служба контроля качества эволюционирует: от проверки звонков к системе аудита самой алгоритмической оценки.

NotebookLM

Роль человека меняется фундаментально

Было

Фабрика по оценке контактов

Человек слушает звонок →
Человек ставит балл
по чек-листу.
Ресурс тратится на рутину.

Стало

Аудитор и калибратор AI-системы

Человек находит слепые зоны →
Человек обновляет логику
модели. Ресурс тратится
на обучение системы.

Контролер качества проверяет не только работу оператора или бота.
Он проверяет надежность автоматической оценки.

NotebookLM

5 вопросов, которые стоит задать завтра утром (Executive Action Plan)

- [ ] Сравниваем ли мы варианты обслуживания, опираясь лишь на одну среднюю ИИ-оценку?
- [ ] Знаем ли мы точные зоны, где наш ИИ систематически расходится с мнением эксперта?
- [ ] Остались ли у нас «слепые зоны», которые человек вообще перестал проверять?
- [ ] Понимаем ли мы системно, почему конкретный контакт попал в аудит (или это случайность)?
- [ ] Есть ли у нашего QA-пилота строгий критерий остановки, кроме наступления календарной даты?

NotebookLM

Грань между наукой и бизнес-практикой

NotebookLM

Раньше мы боялись, что видим слишком мало...

Теперь мы должны бояться, что видим очень много, но неправильно

Масштаб ИИ + Калибровка человеком = Надежное решение.

NotebookLM