

Оценка AI-систем

как понять, что ваш ИИ прямо классный

2026.10.10 Дмитрий Колодезев, Промсофт



1. AI System Evaluation

- Данные, скрипты и процессы
- Позволяют сравнивать разные версии системы
- Делать выводы - стало ли лучше, не стало ли хуже
- Публичные бенчмарки, онлайн-оценка, приемка заказчиком - все это не сегодня
- Внутренняя офлайн-оценка
 - Тесты для детерминированной части
 - Eval для недетерминированной



2. Что такого сложного

- Он такой классный!
- Заказчик скажет, если что не так
- Примеры разметим
- Проверим тестами
- Проверим АБ-тестами
- Бенчмарки прогоним
- Свои бенчмарки сделаем



3. Такой непостоянный

- LLM дает разные ответы
- Внешнее API отвечает по-разному
- Данные в системе меняются
- Выбирает разные инструменты
- Ходит разными путями
- Время ответа меняется
- KV-cache шумит
- Отказ через границы подсистем
- Пользователи еще эти



4. Но зачем

чтобы

- быстро итерироваться
- двигаться при этом вперед
- отдать задачу кодовому агенту ★

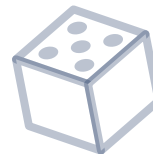
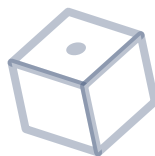
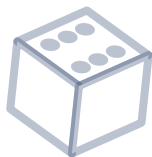
когда

- результат плохо определен
- дисперсия системы велика
- хаотичная система
- оценивать дорого
- обратная связь скупа



5. Кстати, Causal Inference

- Correlation does not imply causation, но:
 - поменяли/переобучили эмбеддер, стало лучше
 - поменяли модель, стало лучше
 - добавили агенту инструменты
 - нарисовали вокруг сервера пентаграмму и деплоим в полнолуние

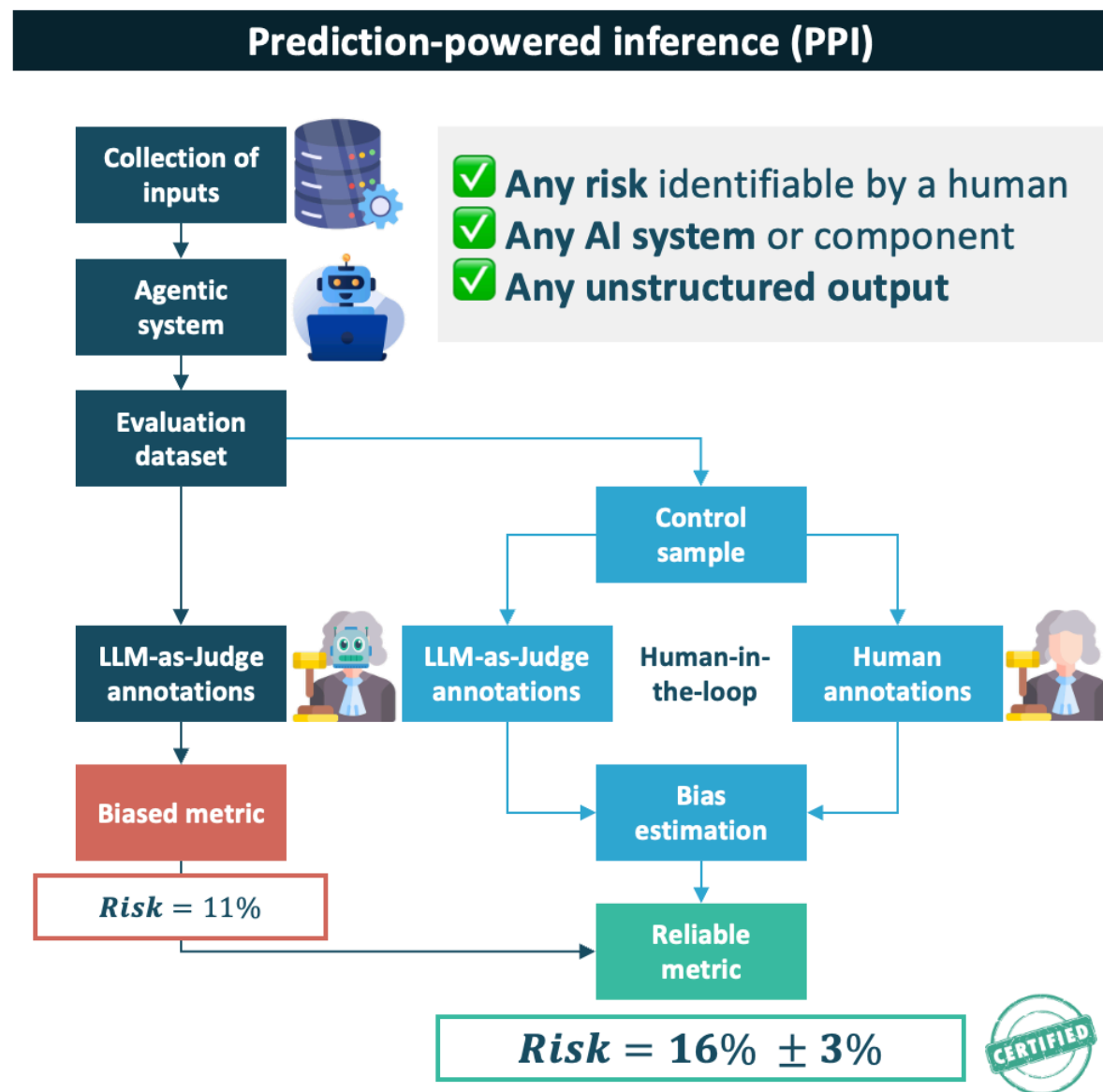


6. Оно работает? Работает, да?

	По правильным причинам	По неправильным причинам
Работает	 МОЖЕТ ЖЕ, КОГДА ХОЧЕТ Работает по правильным причинам Мы молодцы	 У НАС ВСЕ ХОРОШО, ЗУБ ДАЮ самый опасный Работает по неправильным причинам Мы думаем, что мы молодцы
Не работает	 БЛИН, НУ ЧТО ЕЙ ЕЩЕ НАДО Не работает по правильным причинам Мы исправимся	 СЛОМАНО ДАЛЬШЕ В ПАЙПЛАЙНЕ Не работает по неправильным причинам Мы думаем что ошибка тут

7. Разметки всегда мало

- GLIDE
- разметим, сколько сможем
- нагенерируем синтетики
- оценивать будем LLM-судьей
- судью откалибруем на разметке



8. DeepEval

- <https://deepeval.com>
- G-Eval
- Критерии → CoT → скор по токенам
- criteria или evaluation_steps
- evaluation steps генерируются
 - фиксируйте между запусками
 - criteria только для тестов
- logprob может не быть
- Рубрики
- PPI прикрутить **самим**

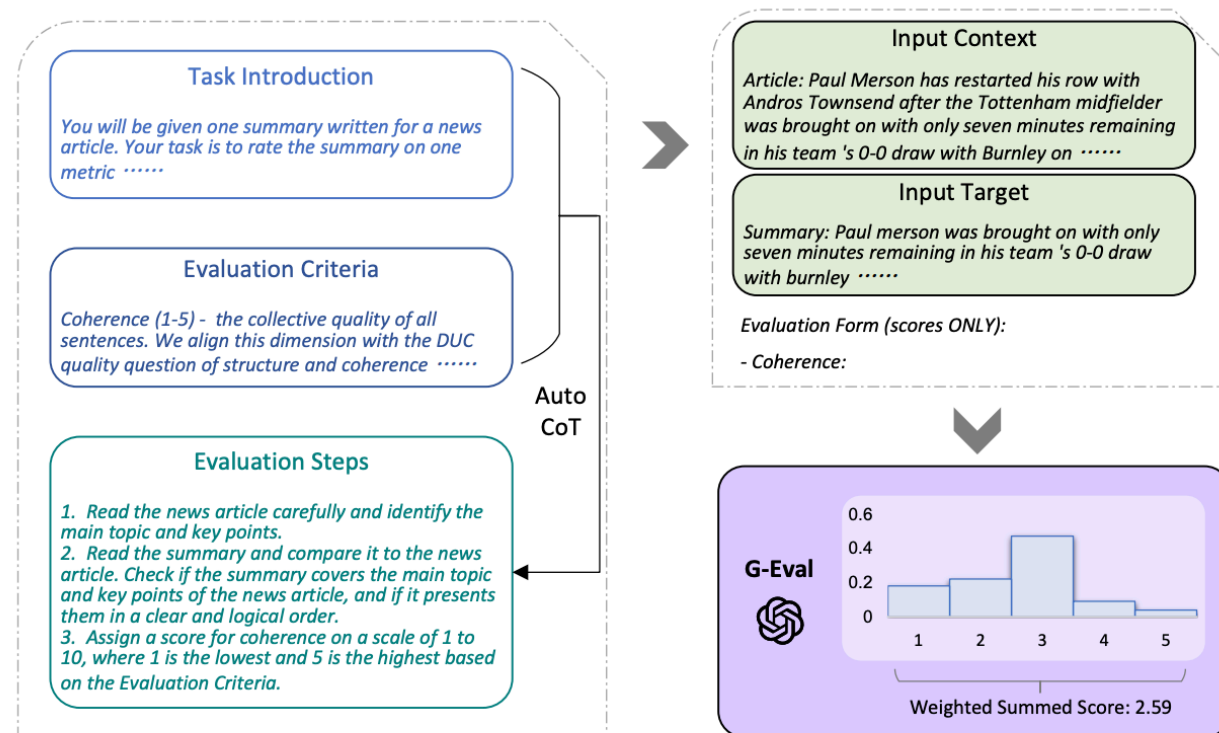
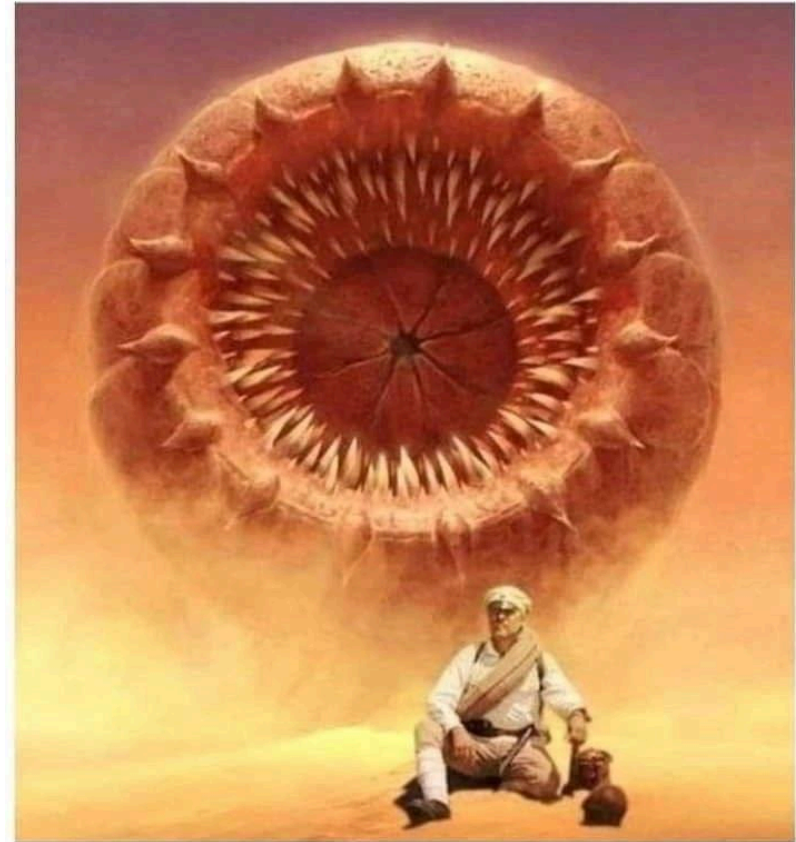


Figure 1: The overall framework of G-EVAL. We first input Task Introduction and Evaluation Criteria to the LLM, and ask it to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs in a form-filling paradigm. Finally, we use the probability-weighted summation of the output scores as the final score.

9. Знай пути их

- Сохранять трейсы
- Например Langfuse
 - **читерство**
 - хождение кругами
 - (не)выбор инструмента
 - странные ответы API
- Сравнивать длину траектории
- Сравнивать количество токенов
- Перепрогонять трейсы
- Claude, посмотри траекторию

Что Вам рассказать, дорогая Екатерина Матвеевна? Веду себя с ними, как знающий пути их, за это отсыплют мне щедро перца местного, но если честно, то уже в печёнках сидит это гадость, и видения будущего утомляют, что сил моих больше нет. Вижу постоянно червяка огромного и баркас. Ваш Лисухов аль-Гаиб.



10. Душно тестируем

- Парное сравнение (сравниваем оценки)
- Попарное сравнение (сравниваем ответы)
- Разметка и калибровка судей — на случайной выборке, не на сложных случаях
- Негативные примеры в eval-набор - ожидаемые отказы
- Сравнение на первом расхождении
- Владелец тестового набора
- Пополняем тестовый набор из прода, holdout заморожен
- Выбирайте фичи, которые проще оценить и измерить

11. Конкретно душно

- Единый LLM gateway с логированием
- Мета-информация трейса: версия промпта, модель, параметры, контекст
- Бюджеты токенов, времени, вызовов влияют на результат
- Микробенчмарк в CI/CD, на дашборде
- Пошаговые метрики в пайплайнах для идентификации проблем
- Песочница для агентов, имитирующая внешние системы (mock)
- Тесты для детерминированных частей
- Лестница проверок: сначала regex/if/assert, потом усложняем
- Повторные прогоны с долей успеха (3-5) для недетерминированных
- Min@K для небинарных метрик: худший результат из K прогонов

12. Забыли спросить

- Обратная связь пользователей пополняет eval-набор
- Точки встраивания обратной связи: 👍/👎, правки, отмены
- Шлюзы go/no-go: контролируем радиус поражения

13. Шпаргалка

Ситуация	Действие
У пользователя упало	Собрать примеры из реального трафика, пересобрать eval-набор
Доработали X, стало не хуже	Парное сравнение - что упало, что поднялось
Метрика хорошая, пользователи недовольны	Проверить метрику. Калибровать судью
Не понимаем, где ломается	Пошаговые метрики + полные трейсы сессий
Страшно деплоить	Порог качества + замороженный holdout + eval в CI
Разметка дорогая, данных мало	Синтетика, обратная связь с прода, LLM-судья с калибровкой

Неизмеренное качество вам только снится

Вопросы

kolodezev.ru

dk.promsoft@mail.ru

[@promsoft](#)

Нормально делай,
Нормально будет.

