

## 1 - Пять частей стенда

|              |                                                                                 |
|--------------|---------------------------------------------------------------------------------|
| Golden set   | видео-тест-кейсы + известные хорошие исходы; берите трудные/состязательные.     |
| Тест функции | ваша реальная модель + промпт, прогнанные по каждому кейсу golden set.          |
| Судья        | LLM (или, для видео, VLM), ставящий балл каждому выходу по рубрике.             |
| Скоркарта    | разбивка прошёл/нет по кейсам - где качество держится и падает, не среднее.     |
| Ворота       | блокируй релиз, если балл упал, критичный кейс регрессировал или растут галлюц. |

## 2 - Почему метрики совпадения слов слепнут, и когда нужен VLM-судья

BLEU / ROUGE считают общие слова: наказывают синонимы (верный субтитр = 0,11) и не смотрят.  
LLM-as-judge оценивает смысл по рубрике; судья GPT-4 совпал с людьми >80% (= согласие людей).  
VLM-as-judge ещё читает ВЫБОРКУ КАДРОВ и сверяет с пикселями - ловит выдуманную деталь.

## 3 - Три решения по режиму оценки

|        |                                                                                     |
|--------|-------------------------------------------------------------------------------------|
| Форма  | pointwise (балл одному) - СЛЕДИТЬ за качеством; pairwise (A vs B) - ВЫБРАТЬ модель. |
| Эталон | с эталоном офлайн на golden set; без эталона онлайн в продакшене.                   |
| Шкала  | всегда грубая (прошёл/нет или 1-5). Тонкие (1-100) дают шум, не точность.           |

## 4 - Нейтрализуй три стоящих смещения судьи

|                      |                                                                         |
|----------------------|-------------------------------------------------------------------------|
| Смещение позиции     | судья хвалит первый -> оцени дважды, переставь порядок, засчитай в обе. |
| Смещение многословия | судья хвалит длинное -> лимит длины в рубрике; вели игнорировать длину. |
| Самопредпочтение     | судья хвалит своё сем.-> бери судью ДРУГОГО семейства, чем генератор.   |

## 5 - А судьи кто (сделай это ДО доверия любому автоматическому баллу)

Пусть люди разметят пару десятков кейсов golden set; прогони судью; измерь согласие.  
Каппа Коэна (с поправкой на случай) + Спирмен. Верь лишь при высоком согласии; точность != воспр.

## 6 - Перед выпуском функции спроси:

- ☐ Есть ли golden set реальных видео-кейсов - включая трудные и состязательные - по которым судят функцию?
- ☐ Читает ли судья кадры (VLM), когда выход - утверждение о видео, чтобы заземление реально проверялось?
- ☐ Откалиброван ли судья против человеческих меток (каппа Коэна), а не просто принят за верный?
- ☐ Нейтрализованы ли три смещения - перестановка-усреднение порядка, лимит длины, судья из другого семейства?
- ☐ Прогоняется ли стенд офлайн как ворота релиза И онлайн как сэмплер 1-5% продакшен-трафика?
- ☐ Текут ли провалы из продакшена обратно в golden set, чтобы стенд крепнул, а не ветшал?
- ☐ Выбран ли режим оценки осознанно - pairwise чтобы выбрать модель, pointwise чтобы следить за одной?
- ☐ Может ли команда показать доказательства оценки против фреймворка (NIST AI RMF / ISO/IEC 42001)?