

Опыт и перспективы применения больших языковых моделей для оценивания заданий с открытым ответом

Иван Александрович Утешев

Старший методист Центра педагогического мастерства г. Москвы

11 июля 2025 г.

ИИ в сфере образования и развития талантов

Контекст

Активное внедрение ИИ для автоматизации рутинных задач

Масштабирование системы развития талантов в области естественных и точных наук

Проблема закрытых вопросов

«Генеративные модели ИИ способны решать геометрические задачи на уровне призёров Международной математической олимпиады» ([Trinh et al., Nature 2024](#))

«Нет сомнений, что подготовленный школьник или репетитор способен при помощи ИИ выполнить задания регионального этапа ВсОШ на уровне проходного балла по большинству предметов менее чем за полчаса» ([Ященко & Утешев, 2024](#))

Участники «масштабируются» лучше, чем эксперты

⇒ любыми силами (не во вред смыслу и содержанию) нужно помочь в проверке работ

Современные популярные инструменты типа ChatGPT-4o позволяют, в том числе, распознавать рукописные работы, рецензировать их по заданному шаблону и оценивать по критериям

Пример свободной ИИ-проверки

Работа участника Open World Astronomy Olympiad, 2024, задача Т3

Problem 3 ΩWAO Code BU-2

$T_s = 2900\text{K}$ $R_p = 6,10^4\text{km}$
 $R = 0,7R_\odot$ $P = 20d$
 $T = 3400\text{K}$

a) $L = \left(\frac{R}{R_\odot}\right)^2 \cdot \left(\frac{T}{T_\odot}\right)^4 \cdot L_\odot$

$$L_s = \frac{R^2 - R_s^2}{R_\odot^2} \cdot \left(\frac{T}{T_\odot}\right)^4 L_\odot + \frac{R_s^2}{R_\odot^2} \cdot \left(\frac{T_s}{T_\odot}\right)^4 L_\odot$$
$$\Delta L = L - L_s = \frac{R_s^2}{R_\odot^2} \cdot \frac{T^4 - T_s^4}{T_\odot^4} L_\odot$$
$$L_p = \frac{R^2 - R_p^2}{R_\odot^2} \left(\frac{T}{T_\odot}\right)^4 L_\odot$$
$$\Delta L = L - L_p = \frac{R_p^2}{R_\odot^2} \left(\frac{T}{T_\odot}\right)^4 L_\odot$$
$$\frac{R_p^2}{R_\odot^2} \cdot \frac{T^4}{T_\odot^4} L_\odot = \frac{R_s^2}{R_\odot^2} \cdot \frac{T^4 - T_s^4}{T_\odot^4} L_\odot$$

Page 6 of 21

Problem 3 ΩWAO Code BU-2

$$R_s^2 = \frac{T^4}{T^4 - T_s^4} \cdot R_p^2$$
$$R_s = \sqrt{\frac{T^4}{T^4 - T_s^4}} \cdot R_p$$
$$R_s = \sqrt{\frac{3400^4}{3400^4 - 2900^4}} \cdot 0,7R_\odot \cdot 6,10^4\text{km}$$
$$R_s \approx 1,02R_\odot \quad R_s \approx 8,7 \cdot 10^4\text{km}$$

~~$R_s \approx 7,10 \cdot 10^5\text{km}$~~

b) $t = ?$

$a \approx 0,6\text{au}$

$$\frac{a^3}{P^2} = \frac{GM}{4\pi^2}$$
$$t = \frac{\theta}{360^\circ} \cdot P$$

Page 7 of 21

Problem 3 ΩWAO Code BU-2

$$\theta = \frac{2 \cdot (0,7R_\odot + 6,10^4\text{km})}{0,6\text{au}}$$

Page 8 of 21

Пример свободной ИИ-проверки (рецензия)

✓ 1. Оценка радиуса звёздного пятна (6 баллов)

Анализ:

- Участник правильно начал с анализа изменения светимости (разности светимости без и с пятном).
- Использована формула:

$$\Delta L = L - L_s = \frac{R_s^2}{R_\odot^2} \cdot \frac{T^4 - T_s^4}{T_\odot^4} \cdot L_\odot$$

- Также получена аналогичная формула для планетного транзита, приравнены потери светимости, что корректно.
- Из формул выведено выражение для R_s , подставлены числа, получено:
 $R_s \approx 8.7 \cdot 10^4$ км — правильно.

Оценка: 6/6 баллов

⚠ 2. Орбитальная скорость планеты (5 баллов)

Проблема:

- Участник **не вычисляет скорость планеты явно**.
- Формулы закона Кеплера и выражение для углового смещения θ есть, но ни скорость V_p , ни период орбиты не указаны.
- Подсчёт длительности основан на угле поворота звезды и полном орбитальном периоде — метод допустимый, **но требует расчёта периода планеты** (по $a = 0.6$ а.е.) или хотя бы промежуточных величин, чтобы выйти на время в секундах или часах.

Оценка: 1/5 баллов — подход потенциально корректный, но не доведён до численного результата и без явного вывода скорости.

✗ 3. Скорость пятна (4 балла)

Проблема:

- Нет никакой информации о линейной или угловой скорости пятна.
- Это важная часть задачи, так как сравнение скоростей планеты и пятна критично для расчёта длительности.

Оценка: 0/4 баллов

4. Длительность транзита (5 баллов)

Плюсы:

- Участник рассматривает движение планеты по орбите и показывает геометрическую схему.
- Использует угол $\theta = \frac{2(R+R_p)}{a}$ и находит время как:

$$t = \frac{\theta}{360^\circ} \cdot P$$
- Метод в принципе применим.

Минусы:

- Нет вычисленного значения периода планеты, хотя указано $a = 0.6$ а.е.
- Результат для времени **не найден численно** (секунды или часы не представлены).
- Используемое обозначение $R + R_p$ — допустимо, но не соответствует обозначениям из условия.
- Нет явного вывода об оценке: $t \approx 3$ ч или аналогично.

Оценка: 2/5 баллов — метод обозначен, но не доведён до результата.

Пример свободной ИИ-проверки

 Итоговая оценка: 9 / 20 баллов

Критерий	Баллы
Радиус пятна	6 / 6
Орбитальная скорость	1 / 5
Скорость пятна	0 / 4
Длительность транзита	2 / 5
Итого	9 / 20

Сводная оценка жюри – $6 + 0 + 0 + 2 = 8 / 20$
При этом рецензия даёт важную обратную связь

Критериальное оценивание

Важно иметь калиброванный инструмент

Критерий — признак (утверждение), на основании которого производится оценка достижения некоторого результата.

Решение участника соотносится с критерием

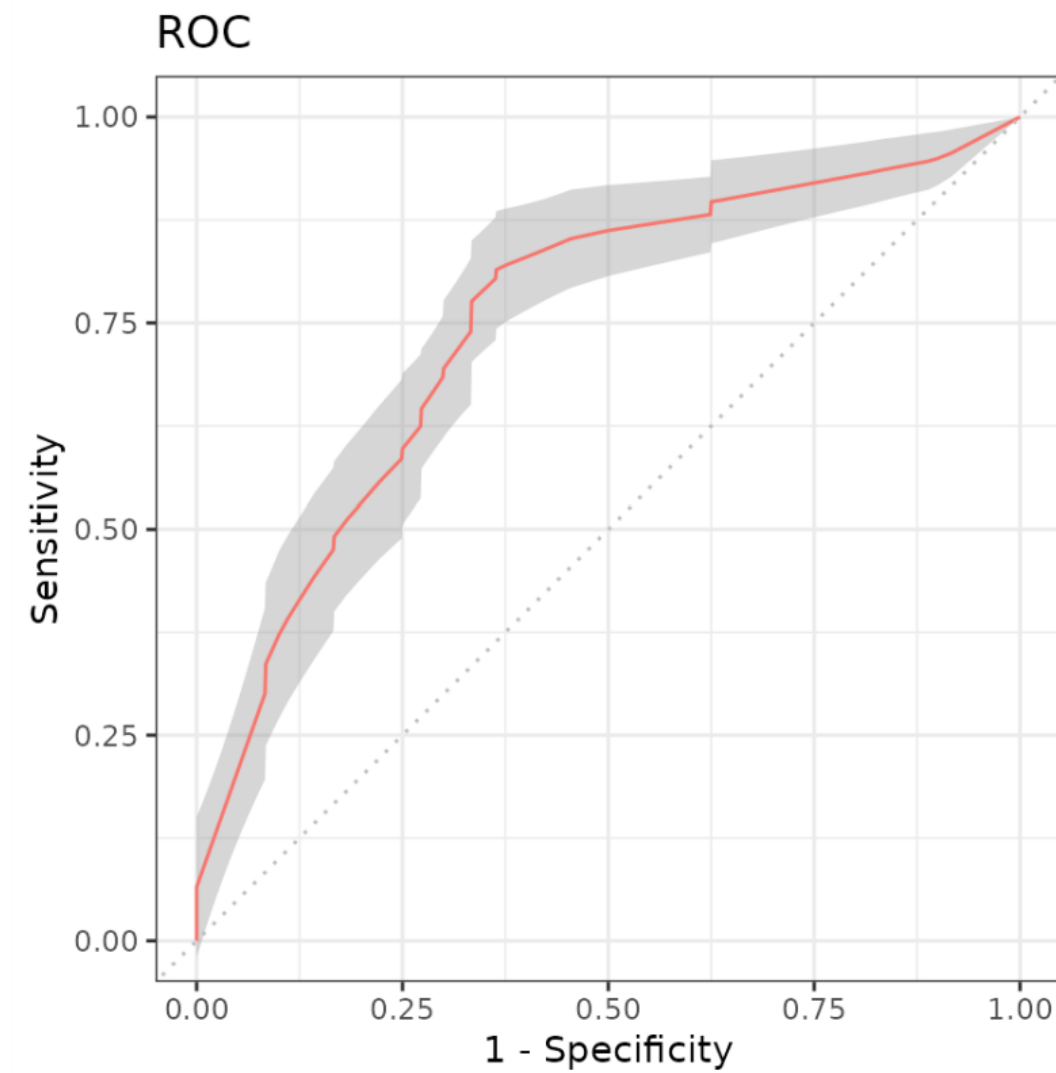
Оценка эксперта $J_i = 0$ или 1 , но оценка модели $S_i \in [0; 1]$

Конечный результат оценивания ИИ предполагается равным (округление)

$$\hat{J}_i = I(S_i > c) \equiv \begin{cases} 1, & S_i > c, \\ 0 & \text{иначе.} \end{cases}$$

Здесь $c \in [0; 1]$ — некоторое пороговое значение. [Как его определять?](#)

ROC-кривая



Semi-supervised и активное обучение

При калибровке полная разметка решений участников отсутствует.

Возникает задача **поиска наиболее удовлетворительных оценок для ROC-кривой и других характеристик классификатора** с использованием данных, для которых известны только «машинные» оценки

Случайная выборка не передаёт оптимальную информацию для принятия решения

Мотивация на примере средней оценки

С использованием только размеченных (вручную) данных:

$$\bar{J} = \frac{1}{n} \sum_{i=1}^n J_i$$

Можно попробовать извлечь дополнительную информацию о распределении оценок из «автоматических» оценок:

$$\mu = \mathbb{E}[J] = \mathbb{E}[\mathbb{E}[J | S]] = \int \mathbb{E}[J | S = s] \, d\mathbb{P}(S \leq s) \equiv \int m(t) \, dP(t)$$

$$\hat{\mu} = \int \tilde{m}(s) \, d\hat{P}(s)$$

Стратегия активной выборки

Критерий отбора для экспертизы

Максимизация информативности: $x^* = \operatorname{argmax}_x [I(J; \theta \mid S = x)]$,
где I — взаимная информация

Как это выглядит на практике

$$\sum_{i=1}^k \|S_i - \hat{c}\| + \lambda \operatorname{Var}(S_i) \rightarrow \operatorname{opt}$$

Возможная интерпретация:

Первый член — близость к порогу

Второй член — разнообразие выборки

Совместный проект МФТИ и Яндекса (2024)

Вызов

Турнир имени М. В. Ломоносова

35 тыс. работ → **3 месяца** ручной проверки

Ложные срабатывания (FPR) vs пропуск решений (FNR)

Цель

Минимизация экспертной нагрузки

Сохранение точности оценивания и удовлетворённости участников



Вклад автора

Разработка и корректировка методик и критериев оценивания
Конвейер оценивания и способ выборки решений

Ключевые результаты

Впервые — массовая олимпиада без «ручной» проверки всех работ
Ускорение проверки в 3 раза
Качество сохранено

«Критерии оценивания должны включать в себя утверждения, наличие которых необходимо для выставления баллов.

Утверждения необходимо формулировать однозначно и ёмко, делая акцент на смысловой составляющей. Хорошее эталонное утверждение должно показывать понимание происходящего»

Примеры вопросов и критериев

Удачные примеры

Q: Есть ли такое созвездие?

A: На Земле в официально принятой номенклатуре нет созвездия с таким названием

Q: Как можно определить, повышался или понижался уровень мирового океана в далёком прошлом?

A1: Уровень океана определяется по изменению состава слоев (фаций) осадочных пород

A2: Уровень воды можно определить по наличию ископаемых

Q: Почему в грунте Луны гелия больше, чем в грунте Земли?

A1: На Луне много гелия благодаря Солнцу и солнечному ветру

A2: Гелий — легкий газ. Более тяжелые газы вытесняют гелий в верхние слои атмосферы. Гравитация Земли не может удержать быстро движущийся газ. Гелий солнечным ветром уносится в космос

Примеры вопросов и критериев

Неудачные примеры

Q: Как вы думаете, настанет ли когда-нибудь такое время, что юлианский календарь будет более точным и нужно будет переходить обратно?

A1: Вращение Земли замедляется

A2: Количество дней в году уменьшается

Q: Можно ли уничтожить созвездие? Что мог иметь в виду персонаж фильма?

A1: Созвездие — это участок на небе

A2: Созвездие можно переименовать или изменить его границы

A3: Персонаж мог иметь в виду звёздную систему

Q: Как можно определить, повышался или понижался уровень мирового океана в далёком прошлом?

A: Повышение уровня океана называется трансгрессией

Фильтрация — отбрасываются заведомо бессмысленные ответы

Автоматическое оценивание при помощи большой языковой модели

Выборка решений для калибровки модели

Выборочная ручная перепроверка работ

*Дообучение модели и повторное автоматическое оценивание

Интерпретация результатов

Выборочная ручная валидация оценок

Результаты: эффективность стратегии

Экономия времени $\sim 10^3$ человеко-часов

Точность оценки $c^* \sim 0.03$

Ручная валидация — $FNR < 5\%$, $FPR < 10\%$

Доля апелляций $\sim 3 \div 5\%$ (не увеличилась г/г)

Этические аспекты

Динамическая калибровка:

$$(\{S_{t-1}\}, \{J_{t-1}\}) \mapsto c_t$$

Мультимодальные решения

Инструменты для экспертов

*Нейросеть ничего не придумывает:
она пользуется предыдущим «опытом».
А мы?*

Спасибо за внимание!

Обсудить: Иван Uteshev@phystech.edu