

Почепцов Игорь Сергеевич

Использование Large language models при проверке
решений задач по математике

Стажер лаборатории "Математическое образование" НИУ ВШЭ

Москва 2024

Ключевые слова: Искусственный интеллект, прикладная статистика, проверка решений олимпиадных задач, преподавание математики, оценка знаний школьников.

Аннотация: В статье исследуется применение языковых моделей (Large Language Models), для помощи в проверке олимпиадных задач. Анализируется эффективность моделей в выявлении ошибок и оценке решений участников. Освещаются потенциальные преимущества использования ИИ для повышения объективности оценок и сокращения времени проверки работ.

We explore an applications of large language models for checking math Olympiad assignments. We analyze the effectiveness of the models in identifying errors and evaluating participants' solutions. The potential advantages of using AI to increase the objectivity of assessments and reduce the time spent checking works are highlighted.

1 Введение и постановка вопроса исследования

Сейчас происходит много прорывов в области развития и применения ИИ в различных сферах. В этой статье хотелось бы рассмотреть возможность применения ИИ в сфере образования.

1.1 Новые исследования в области решения задач по математике

Сейчас делается много прорывов в области использования искусственного интеллекта для решения математических задач. Если раньше верность трети решений на выборке количественных задач считалась достижением [1], то сейчас языковые модели уже умеют верно решать около 80% задач более высокого уровня [2].

Одна из самых престижных олимпиад для школьников - это Международная олимпиада по математике (ИМО). Нейросеть AlphaGeometry от Google удалось научить решать задачи по геометрии уровня Международной олимпиады по математике. При тестировании AlphaGeometry решила 25 из 30 геометрических задач уровня международной олимпиады. Для сравнения, золотой медалист Международной олимпиады в среднем решает 26 из 30 задач. Подробности исследования опубликовали в журнале nature[3].

Такие успехи в области решения задач при помощи искусственного интеллекта создают возможность нечестного участия в олимпиадах. Если нейросети научатся решать задачи уровня олимпиад, дающих поступление в лучшие вузы, то у участников будет соблазн использовать их для написания олимпиады. Получится ли проводить олимпиады честно? Можно ли как-то отследить недобросовестное использование искусственного интеллекта в обучении и написании олимпиад?

Эти вопросы важны, однако мы рассмотрим вопрос использования нейросетей с другой стороны. Подумаем о том, могут ли нейросети помочь преподавателям и жюри олимпиад.

1.2 Проблема проверки олимпиад

С каждым годом в олимпиадах по математике участвует всё больше школьников, что приводит к сложностям в проверке решений. Например, знаменитая олимпиада для 6-7 классов - Матпраздник, ранее проверявшаяся за один день, теперь требует

около месяца. В 2011 году количество участников составляло примерно 2.2 тысячи, тогда как к 2023 году это число выросло до 23.5 тысяч [4] [5].

Длительность проверки оказывает значительное влияние на мотивацию участников, которым иногда приходится ждать результатов до трех месяцев [6]. Это, в свою очередь, уменьшает интерес учащихся к олимпиадам, что противоречит одной из их основных целей - повышать заинтересованность в предмете.

Одним из возможных решений этой проблемы является внедрение компьютерной проверки. Например, с 2019 года на школьном этапе по математике была введена автоматизированная проверка ответов, что значительно упростило процесс проверки. Для данной олимпиады вопрос проверки стоит особенно остро, ведь в школьном этапе в 2023 году участвовало более 6 млн школьников [7]. Делаются исследования в области перевода количественных результатов проверки ответов в оценку [8]. Однако в контексте математических олимпиад, оценивание только финального ответа без учета рассуждений и методов, использованных учащимся для его нахождения, может не дать полной картины об их знаниях и умениях. Анализ логики, примененной для решения задачи, её креативности и системности, также важен, поскольку демонстрирует глубину понимания материала учащимся и их способность применять знания в нестандартных ситуациях.

Возможное решение этой задачи - применение искусственного интеллекта, который мог бы помочь в оценке как точности ответа, так и логики решения. Такой подход не только ускорит процесс проверки, но и обеспечит более всестороннюю оценку работ, помогая преподавателям и жюри в их задаче.

1.3 Основной вопрос исследования

В рамках данного исследования нашей основной целью является оценка потенциала языковых моделей в качестве инструмента помощи при проверке математических задач. В настоящее время мы сталкиваемся с растущим количеством участников математических олимпиад и соответственно с увеличивающейся нагрузкой на проверяющих. Это вызывает необходимость в эффективных и точных методах проверки, способных ускорить процесс, не теряя при этом в качестве оценки. Мы стремимся исследовать, насколько эффективно современные языковые модели, такие как GPT и их производные, могут анализировать и оценивать решения математических задач, предложенных участниками олимпиад. Особое внимание в исследовании уделяется способности ИИ не только определять правильность ответов, но и анализировать логику решения, что является ключевым фактором в оценке олимпиадных работ.

2 Подготовка к исследованию

2.1 Какие данные используем для исследования?

Исследование делаем по данным проверки задачи номер 7 олимпиады “Турнир Ломоносова” 2023 года.

Условие задачи: Существует ли число, которое может быть представлено в виде $\frac{1}{n} + \frac{1}{m}$, где m и n натуральные, не менее чем ста способами? Ответ объясните.

Всего у нас имеются данные по 3780 решениям задачи. По каждой задаче имеются данные: решение ученика, результат первой проверки, результат второй проверки и итоговый балл ученика. По каждой проверке также известны инициалы жюри, поставившего оценку.

За задачу ставилось от 0 до 2 баллов. В выборке 31 решение получило 1 балл, 661 решение получило 2 балла, 3088 решений получили 0 баллов.

2.2 Какую языковую модель мы используем в исследовании?

На момент исследования активно развиваются как Российские, так и зарубежные языковые модели.

Среди Российских можно выделить GigaChat от Сбербанка и YaLM от Яндекса.

Среди зарубежных – Gemini от Google и ChatGPT от OpenAI.

На момент написания статьи ChatGPT 4.0 работает значительно лучше аналогов, так что для пилотного исследования будет использоваться именно эта нейросеть.

При этом модели Яндекса и Сбербанка очень быстро развиваются, так что можно сделать вывод о том, что в ближайшем будущем результаты исследования можно будет переложить и на российские нейросети.

2.3 Проблема защиты персональных данных

Помимо статей об использовании и развитии языковых моделей, также появляются статьи об обходе системы безопасности и извлечении обучающих данных [9]. В связи с этим все данные в нейросеть мы будем передавать в обезличенном виде.

2.4 Распознавание текста и способы запросов

В нашем исследовании мы анализируем только печатные решения задач. Распознавание рукописного текста — это отдельная область исследований, которая пока не достигла высокого уровня точности. Мы ожидаем улучшения в этой сфере в будущем, что позволит нам использовать результаты таких исследований.

В ходе нашей работы мы также вручную задавали вопросы, отправляя запросы на проверку пакетами по десять задач. Отметим, что в других областях уже существует интеграция ChatGPT через API, которая позволяет использовать нейросеть на больших объемах данных.

Наша цель — оценить потенциал использования языковых моделей в образовательной сфере.

3 Как проверяются олимпиады?

Проверка задачи олимпиады проходит в три этапа:

Первый и второй этап – проверка работы. На каждое решение нужно по 2 проверки от различных членов жюри. Каждый из проверяющих ставит оценку за решение задачи.

Третий этап – третья проверка жюри. Если две оценки совпадают, то эта оценка ставится ученику. Если оценки не совпадают, то кто-то из жюри смотрит на решение и определяет оценку за задачу.

Результаты выдаются участникам, после чего у них есть возможность подать апелляцию. Ознакомившись с результатами, ученики пишут текст апелляции, если они не согласны с полученными баллами. Жюри рассматривает апелляцию и меняет оценку за задачу в большую или меньшую сторону или оставляет оценку без изменений.

В этом процессе очень много ресурсов уходит на проверку и сверку результатов. Ведь на каждую задачу нужно от 2 до 3 проверок квалифицированного эксперта.

4 Методика исследования

Результат работы языковой модели может отличаться в зависимости от того, какой запрос мы ей зададим. В связи с этим мы протестируем модель на различных запросах.

Выборка неверных (S_0) – репрезентативная выборка из 50 случайных решений, получивших 0 баллов

Выборка верных (S_2) – репрезентативная выборка из 50 случайных решений, получивших 2 балла

Выборка случайных (S^*) – репрезентативная выборка 50 случайных решений

Выборка единичек (S_1) – 31 решение, получившее оценку 1

Выборка расхождения (SD) – выборка из 50 случайных решений, получивших разные оценки от жюри

Для каждого метода X и каждой пары $i, j \in \{0, 1, 2\}$ посчитаем количество $X_{i,j}$ случаев, когда нейросеть ставит за задачу i баллов, а итоговой оценкой жюри является j

5 Сбор данных, различные запросы в модель

5.1 Метод A

Для начала отправим в модель запрос на проверку решений задачи, содержащий условие задачи, пример верного решения задачи, описание требований проверки и критерии жюри с сайта олимпиады.

Текст запроса в языковую модель:

Тебе нужно проверить решения задач

Условие: Существует ли число, которое может быть представлено в виде $1/n + 1/m$, где m и n натуральные, не менее чем ста способами? Ответ объясните.

Пример верного решения: Пусть a, b такие что $a = 100 - v, b = 100 + v$. тогда для любого $1 \leq v \leq 100$: $x = (a+b)/200!$. x будет равен $1/199!$. $(a+b)/200! = a/200! + b/200!$. $n = 200!/a, m = 200!/b$. так-как $a, b < 200 \Rightarrow$ числа целые, всего 100 пар.

Критерии проверки: Максимум за задачу — 2 балла. Верными считаются решения, которые приводят пример явно или иным способом доказывают его существование.

Критерии:

- в решении пары (n, m) и (m, n) считаются различными — не влияет на оценку;
- только ответ «да, существует» или «нет, не существует» без доказательства — 0 баллов;
- приведено число, но не доказано, что его можно представить в нужном виде 100 способами — 0 баллов;
- в решении написано, что подходящее число найдено с помощью программы, но не приведён её код или не приведены 100 подходящих пар m и n — 0 баллов.

Готов ли ты проверять решения?

Выборка	A00	A01	A02	A10	A11	A12	A20	A21	A22	другое
Выборка неверных	47	-	-	-	-	-	1	-	-	2
Выборка верных	-	-	6	-	-	11	-	-	33	-
Выборка случайных	43	-	-	-	-	2	1	-	4	-
Выборка расхождений	14	12	10	3	1	10	-	-	-	-
Выборка ответов 1	-	9	-		19	-	-	3	-	-

Таблица 1: Результаты проверки метода A

При проверке выборки неверных решений модель 2 раза не смогла поставит оценку решению. Вместо оценки нейросеть сообщила, что не может понять, соответствует ли решение критериям и попросила сделать дополнительную проверку решения.

Текст решения участника олимпиады (орфография и пунктуация сохранены):

Такое число существует, пусть это будет число $q = 1/(3^1 + 3^2 + 3^3 + \dots + 3^{100})$. Тогда представим $1/n_1 = q \cdot 3^1/(3^1 + 1)$. Дробь $1/m_1$ представим в виде $1/m_1 = q/(3^1 + 1)$. Соответственно $1/n_2 = q \cdot 3^2/(3^2 + 1), 1/m_2 = q/(3^2 + 1)$. Тогда $1/n_i = q \cdot 3^i/(3^i + 1); 1/m_i = q/(3^i + 1)$.

Тогда получаем: $1/n_i + 1/m_i = q \cdot 3^i/(3^i + 1) + q/(3^i + 1) = q \cdot (3^i + 1)/(3^i + 1) = q$.
 $*nj! = mj : nj = q \cdot 3^j/(3^j + 1)! = q/(3^j + 1) = mj*$

Таким образом $n_i = (3^i + 1)/(q3^i), m_i = (3^i + 1)/q, q = 1/(3^1 + 3^2 + 3^3 + \dots + 3^{100})$

Ответ языковой модели:

Решение 9: Предоставляет сложную конструкцию числа и пар (n, m) , но не ясно, соответствует ли она требуемым критериям. Требуется более глубокий анализ. Оценка: неопределенно.

Текст решения ученика олимпиады (орфография и пунктуация сохранены):

$1/n + 1/m = (n + m)/(n \cdot m) = k$, можно разложить n и m на простые множители. В числителе можем за скобки вынести общие простые множители и сократить их со знаменателями. В итоге у нас получится, что мы получим новое решение для уравнения $k = (n + m)/(n \cdot m)$. Из этого следует, что мы можем таким же образом подобрать такие 100 пар натуральных n и m , что будут подходить под уравнение $k = (n + m)/(n \cdot m)$, следовательно, и под уравнение $k = 1/n + 1/m$.

Ответ языковой модели:

Решение 19: Предлагает метод подбора пар n и m , основываясь на разложении на простые множители. Это интересный подход, но без конкретных примеров и полного доказательства. Оценка: неопределенно.

5.2 Метод B

Сохраним запрос метода A , однако теперь попросим модель ставить решениям только оценки 0 или 2.

Текст запроса в языковую модель:

Тебе нужно проверить решения задач

Условие: Существует ли число, которое может быть представлено в виде $1/n + 1/m$, где m и n натуральные, не менее чем ста способами? Ответ объясните.

Пример верного решения: Пусть a, b такие что $a = 100 - v, b = 100 + v$. тогда для любого $1 \leq v \leq 100$: $x = (a+b)/200!$. x будет равен $1/199!$. $(a+b)/200! = a/200! + b/200!$. $n = 200!/a, m = 200!/b$. так-как $a, b < 200 \Rightarrow$ числа целые, всего 100 пар.

Критерии проверки: Максимум за задачу — 2 балла. Верными считаются решения, которые приводят пример явно или иным способом доказывают его существование. За задачу ставь только 0 или 2 балла. Если не знаешь, что поставить, пиши "Нужна подробная проверка".

Критерии:

- в решении пары (n, m) и (m, n) считаются различными — не влияет на оценку;
- только ответ «да, существует» или «нет, не существует» без доказательства — 0 баллов;
- приведено число, но не доказано, что его можно представить в нужном виде 100 способами — 0 баллов;
- в решении написано, что подходящее число найдено с помощью программы, но не приведён её код или не приведены 100 подходящих пар m и n — 0 баллов.

Готов ли ты проверять решения?

Выборка	B_{00}	B_{01}	B_{02}	B_{10}	B_{11}	B_{12}	B_{20}	B_{21}	B_{22}
Выборка неверных	46	-	-	-	-	-	4	-	-
Выборка верных	-	-	13	-	-	11	-	-	37
Выборка случайных	43	-	4	-	-	-	1	-	2
Выборка расхождений	7	8	9	-	-	-	10	5	11
Выборка ответов 1	-	23	-	-	-	-	-	8	-

Таблица 2: Результаты проверки метода B

5.3 Метод C

Сохраним запрос метода B , однако уточним, в каком случае мы хотим, чтобы решение получило 2 балла, а в каких - 0 баллов. А именно, напишем "Если решение имеет верные идеи, ставь 2. Если решение совсем неверное, ставь 0".

Текст запроса в языковую модель:
Тебе нужно проверить решения задач
Условие: Существует ли число, которое может быть представлено в виде $1/n + 1/m$, где m и n натуральные, не менее чем ста способами? Ответ объясните.
Пример верного решения: Пусть a, b такие что $a = 100 - v$, $b = 100 + v$. тогда для любого $1 \leq v \leq 100$: $x = (a+b)/200!$. x будет равен $1/199!$. $(a+b)/200! = a/200! + b/200!$. $n = 200!/a$, $m = 200!/b$. так-как $a, b < 200 \Rightarrow$ числа целые, всего 100 пар.
Критерии проверки: Максимум за задачу — 2 балла. Верными считаются решения, которые приводят пример явно или иным способом доказывают его существование. За задачу ставь только 0 или 2 балла. Если решение имеет верные идеи, ставь 2. Если решение совсем неверное, ставь 0.
Критерии:
<ul style="list-style-type: none"> • в решении пары (n, m) и (m, n) считаются различными — не влияет на оценку; • только ответ «да, существует» или «нет, не существует» без доказательства — 0 баллов; • приведено число, но не доказано, что его можно представить в нужном виде 100 способами — 0 баллов; • в решении написано, что подходящее число найдено с помощью программы, но не приведён её код или не приведены 100 подходящих пар m и n — 0 баллов.
Готов ли ты проверять решения?

Выборка	C_{00}	C_{01}	C_{02}	C_{10}	C_{11}	C_{12}	C_{20}	C_{21}	C_{22}
Выборка неверных	49	-	-	-	-	-	1	-	-
Выборка верных	-	-	-	-	-	-	-	-	50
Выборка случайных	43	-	3	-	-	-	1	-	3
Выборка расхождений	11	10	9	-	-	-	6	3	11
Выборка ответов 1	-	14	-	-	-	-	-	17	-

Таблица 3: Результаты проверки метода C

5.4 Метод D

Во время одного из запросов нейросеть написала код для автоматической проверки. Код опирается исключительно на поиск ключевых слов "равно" и "пример" и "доказательство" в решении. Такой код, конечно, не может претендовать на качественную оценку решения. Однако такой метод мы тоже исследуем для сравнения с другими.

Код, который написала нейросеть:

```
1 def check_solution_updated(solution):
2     try:
3         solution_text = str(solution)
4
5         if "1/n + 1/m" in solution_text and "100" in solution_text:
6             if any(keyword in solution_text for keyword in ["=", "
7                 ↪ равно", "пример", "доказательство"]):
8                 return 2
9         return 0
10    except:
11        return 0
12
13 remaining_solutions['Оценка'] = remaining_solutions['Решение'].apply(
14     ↪ check_solution_updated)
15
16 all_results = pd.concat([first_15_results, next_15_results,
17     ↪ remaining_solutions[['Номер решения', 'Оценка']]])
18
19 all_results.reset_index(drop=True, inplace=True)
20 all_results
```

Листинг 1: Python example

Выборка	D_{00}	D_{01}	D_{02}	D_{10}	D_{11}	D_{12}	D_{20}	D_{21}	D_{22}
Выборка случайных	317	4	44	-	-	-	13	1	21

Таблица 4: Результаты проверки метода D

5.5 Проверки жюри

Для каждого из членов жюри мы также посчитаем процент ошибок. Процент ошибок найдём сразу на всей выборке.

Работой с расхождением в оценках жюри назовём работу, в которой итоговый балл проверки отличается от результата одной из проверок. На 3780 работ во всей выборке имеется 63 работы с расхождением жюри, что составляет 1,67%. На 692 работах с положительным баллом (1 или 2) имеется 43 работы с расхождением жюри, что составляет уже 6,2%.

Для каждого члена жюри (H_1 , H_2) найдём процент ошибок на выборке из всех решений, на выборке всех неверных решений, на выборке всех решений, получивших 1 балл, и на выборке всех решений, получивших 2 балла.

Метод	<i>НМО</i>	<i>НМ2</i>	<i>НМ1</i>	<i>НМ*</i>
<i>Н1</i>	0,2%	1,5%	35,5%	0,7%
<i>Н2</i>	0,4%	2%	29%	0,9%

Таблица 5: Процентом ошибок жюри на выборках

6 Анализ результатов

6.1 Анализ результатов

Для каждой проверки посчитаем следующие данные:

$XM0 = 1 - \frac{X00}{|X|}$ – процент ошибок на выборке неверных

$XM2 = 1 - \frac{X22}{|X|}$ – процент ошибок на выборке верных

$XM1 = 1 - \frac{X11}{|X|}$ – процент ошибок на выборке решений с результатом 1

$XM* = 1 - \frac{X00+X11+X22}{|X|}$ – процент ошибок на выборке случайных

$XMD = 1 - \frac{X00+X11+X22}{|X|}$ – процент ошибок на выборке с расхождением жюри

Где $|X|$ – размер выборки в каждом конкретном случае

Метод	<i>ХМО</i>	<i>ХМ2</i>	<i>ХМ1</i>	<i>ХМ*</i>	<i>ХМD</i>
<i>A</i>	6%	34%	38,7%	6%	70%
<i>B</i>	8%	26%	100%	10%	36%
<i>C</i>	2%	0%	100%	8%	56%
<i>D</i>	-	-	-	15,5%	-
<i>Н1</i>	0,2%	1,5%	35,5%	0,7%	-
<i>Н2</i>	0,4%	2%	29%	0,9%	-

Таблица 6: Процентом ошибок при каждом методе проверки

Можно заметить, что доля ошибок при проверке методами *A, B, C* отличаются. Это показывает, что точность запроса в языковую модель сильно влияет на качество проверки.

Методы на случайной выборке ошибаются с вероятностью от 6% до 10%.

Если же проверить метод на "выборке расхождений" то процент ошибок сильно возрастет и составляет от 36% до 70%. Этот результат можно интерпретировать как "спорность" решений из этой выборки. Получается, что их сложно оценить как людям, так и нейросети.

Качество проверки все еще хуже человеческой (6% – 10% против 0,7% – 0,9%), но этот результат уже позволяет значительно упростить процесс проверки олимпиад.

6.2 Комментарии языковой модели

Другим бонусом от проверки при помощи языковой модели является наличие комментариев. Модель кратко поясняет, почему ставит ту или иную оценку. Сейчас не все комментарии модели являются разумными, однако такие комментарии позволяют в будущем упростить вторую проверку, а также применить результаты исследования в других направлениях, таких как учебные ассистенты для учеников и преподавателей.

Решения, отправленные в модель:

31 " $1/n + 1/m = (m+n)/nm$ $xm = m+n$ $m(xn-1)=n$ из $m(xn-1)=n$ следует, что n и $m > 1$, отсюда следует, что x может принимать более 100 чисел"

32 переведем выражение $1/n + 1/m = x$ (где x равен сумме $1/n + 1/m$) в вид $(n+m)/(n*m) = x$, продолжим преобразовывать $\Rightarrow n+m = n*m*x \Rightarrow n*x*m-m = n \Rightarrow (n*x-1)*m = n \Rightarrow m=n/(n*x-1) \Rightarrow$ получаем функцию гипербола с $x - \text{const}$ и аргументом n ; функция гипербола может принимать любое значение отличное от нуля, следовательно, больше 100 значений. Следовательно, такое число существует

33 Откуда мне знать

34 ркцыр

35 Так как знаменатели данных обыкновенных дробей содержатся в переменных m и n , то эти переменные могут быть любыми натуральными.

36 Пусть $x[1], x[2], \dots, x[99], x[100]$ - различные простые числа. Число вида $1/(x[1]*x[2]*\dots*x[99]*x[100])$ можно представить в виде $(x[k]+1)/(x[1]*x[2]*\dots*x[99]*x[100]*(x[k]+1))$, где k - натуральное число от 1 до 100. Эту дробь можно представить в виде суммы дробей $1/(x[1]*x[2]*\dots*x[99]*x[100]*(x[k]+1))$ и $1/(x[1]*\dots*x[k-1]*x[k+1]*\dots*x[100]*(x[k]+1))$. Тогда, в силу того, что $x[1], x[2], \dots, x[100]$ - различные простые числа, представить дробь в виде суммы двух дробей так, как показано выше, можно ста способами (т.к. существует 100 вариантов выбора $x[k]$). То есть, существует число, которое можно представить в виде $1/m + 1/n$ не менее, чем 100 способами (например число вида $1/(x[1]*x[2]*\dots*x[99]*x[100])$, где $x[1], x[2], \dots, x[100]$ - первые 100 простых чисел)

37 Рау

38 " $k=(m+n)/mn$

k будет больше или равно 100"

39 нужно взять число равное k

40 " $1/n + 1/m = C$, какое-то число, которое возможно существует

$(m+n)/(m*n) = C$, очевидно, что для натуральных чисел C может быть только 1 или 2, а мы хотим более 100 способов $\Rightarrow C$ дробь в виде $p/q \Rightarrow (m+n)/(m*n) = p/q$ Заметим, что в квадратном уравнении $(x**2 - px + q)$ по теореме Виета корнями будут как раз m и n : $m+n = p$ и $m*n = q$ (из соотношения выше) \Rightarrow пар m и n не больше чем 2 (т.к. у квадратного уравнения не более 2 корней) \Rightarrow не существует."

Комментарий нейросети:

31 Рассуждение о том, что x может принимать более 100 чисел, не демонстрирует, как именно число может быть представлено в виде $1/n + 1/m$ ста раз. Оценка: 0 баллов.

32 Преобразование выражения в функцию гиперболы и утверждение о том, что она может принимать более 100 значений, не является достаточным доказательством существования числа, удовлетворяющего условию задачи. Оценка: 0 баллов.

33 Отсутствует попытка решения. Оценка: 0 баллов.

34 Неясное утверждение, не связанное с задачей. Оценка: 0 баллов.

35 Утверждение о возможности выбора любых натуральных чисел для переменных m и n не доказывает, что число может быть представлено в виде $1/n + 1/m$ ста раз. Оценка: 0 баллов.

36 Интересный подход с использованием различных простых чисел. Это решение демонстрирует, как можно представить число в виде $1/n + 1/m$ ста раз, используя различные комбинации простых чисел. Оценка: 2 балла.

37 Неясное утверждение, не связанное с задачей. Оценка: 0 баллов.

38 Утверждение о том, что k будет больше или равно 100, не демонстрирует, как число может быть представлено в виде $1/n + 1/m$ ста раз. Оценка: 0 баллов.

39 Простое утверждение о выборе числа k без доказательства или объяснения, как это число может быть представлено в виде $1/n + 1/m$ ста раз. Оценка: 0 баллов.

40 Рассуждение о возможности представления C в виде дроби p/q и анализ квадратного уравнения интересно, но не демонстрирует, как число может быть представлено в виде $1/n + 1/m$ ста раз. Оценка: 0 баллов.

7 Возможности применения результатов

7.1 Сокращение количества проверок олимпиады на 75%

Как можно увидеть на данном примере, на проверку олимпиады понадобилось $3780 \cdot 2 + 63 = 7623$ проверок. Следующими способами мы можем уменьшить количество человеческих проверок:

1) Использовать один из наших методов в качестве первой проверки, на вторую и третью оставить человека. Таким образом первая проверка становится автоматизированной. Расхождение 1 и 2 проверки может увеличиться до 10% работ. Таким образом, нужно $3780 + 3780 \cdot 10\% = 4158$ проверок. Такой метод сокращает количество необходимых проверок на 45%.

2) Использовать метод C в качестве первой проверки и после этого два раза проверить решения с положительным баллом. При этом может увеличиться количество апелляций в связи с ошибками при определении неверных решений (хотя на нашей выборке ложно отрицательных оценок не было). При выборе такого способа не может оказаться решений, которые по ошибке получают положительные оценки. В таком случае понадобится $2 \cdot (8\% \cdot 3088 + 692) \approx 1878$. Это означает сокращение

количества проверок почти на 75%.

7.2 Способы применения исследования на практике

1. Проверка олимпиад. Предыдущий параграф показывает, что уже сейчас применение языковых моделей позволяет без потери качества ускорить проверку олимпиады в 4 раза.

2. Помощник для преподавателей. Можно создать ассистента на основе языковой модели, который будет помогать преподавателю проверять решения учеников и писать комментарии по их решениям.

3. Ассистент для самообучения. Ученик может задавать вопросы в языковую модель и спрашивать, верно ли решена задача. Также модель умеет объяснять различные темы олимпиадной математики.

4. Кружок с автоматической проверкой задач. Одной из проблем создания качественных кружков для учеников из регионов является нехватка квалифицированных преподавателей. Уже существует множество онлайн курсов (например, Сириус-курсы). Однако, на этих курсах далеко не всегда предусмотрена проверка решений, что снижает мотивацию учеников и качество занятий. Языковые модели позволяют создать кружок с автоматической проверкой решений.

5. Помощь в оценке развернутых ответов. Различные статьи предлагают методы оценки компетенций учащихся [10]. Часто оценка компетенций требует оценки развернутого ответа учащегося. Можно применить языковые модели в оценке этих развернутых ответов. Такая оценка может быть даже более объективной, ведь все ученики будут проверяться в равных условиях, одной программой.

6. Проверка заданий по другим предметам, в том числе ЕГЭ. Другие статьи показывают важность и сложность оценки развернутого ответа на экзаменах [11]. Использование языковых моделей может помочь в проверке ответов и упростить проверку экзаменов. В том числе по обществознанию, биологии, русскому языку и литературе.

8 Заключение

Исходя из результатов исследования, стоит отметить, что языковые модели уже способны без существенной потери качества ускорить процесс проверки олимпиад по математике.

При этом видны и другие способы применения моделей. Например, создание ассистентов для учеников и преподавателей, а также создание кружков с автоматической проверкой решений.

Это подчеркивает важность темы и указывает на необходимость продолжения исследований в этой области, а также важность разработки практических приложений и применения результатов исследования на практике.

Отдельно хочу поблагодарить Ивана Валериевича Яценко за постановку задачи и продуктивное обсуждение поставленной проблемы.

Список литературы

1. Solving Quantitative Reasoning Problems with Language Models / A. Lewkowycz [и др.]. — 2022.

2. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter With Code-Based Self-Verification / A. Zhou [и др.]. — 2023.
3. Solving olympiad geometry without human demonstrations / Т. Н. Trinh [и др.] // Nature. — 2024. — Т. 625. — С. 476—482. — DOI: <https://doi.org/10.1038/s41586-023-06747-5>.
4. Математический праздник. — 2024. — URL: <https://olympiads.mcsme.ru/matprazdnik/>.
5. *Яценко И. В.* Приглашение на математический праздник. — 3-е изд., испр. и доп. — МЦНМО, 2009.
6. Турнир Ломоносова. — 2024. — URL: <https://turlom.olimpiada.ru/news/>.
7. Сайт ОЦ "Сириус". — 2024. — URL: <https://siriusolymp.ru/school2023/2/about>.
8. *Гениевич Д. О.* Вероятностный подход к определению педагогической оценки по результатам компьютерного тестирования // Педагогические измерения. — 2023. — № 1. — С. 8—19.
9. Scalable Extraction of Training Data from (Production) Language Models / M. Nasr [и др.]. — 2023.
10. *Демидова Марина Юрьевна Добротин Дмитрий Юрьевич Р. В. С.* Подходы к разработке заданий по оценке естественнонаучной грамотности обучающихся // Педагогические измерения. — 2020. — № 2. — С. 8—19.
11. *Ленвладовна Е. Е.* Анализ развёрнутого ответа на едином государственном экзамене по обществознанию: речеведческий аспект // Педагогические измерения. — 2022. — № 2. — С. 17—22.