Компьютерные инструменты в образовании, 2024
№ 3: 33–47
УДК: 004.42
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-3-33-47

ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ ВЫЧИСЛИТЕЛЬНЫХ СПОСОБНОСТЕЙ СОВРЕМЕННЫХ ЧАТ-БОТОВ В РЕШЕНИИ ЗАДАЧ ПО МАТЕМАТИЧЕСКОМУ АНАЛИЗУ И ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

Винокурова Д. В.[1], аспирант, ✉ d.v.vinokurova@gmail.com

[1]Российский государственный педагогический университет им. А. И. Герцена,
набережная реки Мойки, д. 48, 191186, Санкт-Петербург, Россия

Аннотация

В статье описывается численный эксперимент по решению математических задач чат-ботами (YandexGPT 2, ChatGPT 3.5, Gemini, Copilot) по некоторым темам математического анализа (пределы, производные, интегралы), включающий 693 задачи, и вычислительной математики (решение нелинейных уравнений, решение систем линейных уравнений, интерполяция функций, численное интегрирование), состоящий из 45 задач. Рассматриваются основные характеристики современных виртуальных помощников. Представлен обзор исследований по применению искусственного интеллекта в решении математических задач на различных тестах и наборах данных. В работе рассматриваются недостатки, проявляющиеся в работе чат-ботов, анализируется их производительность на конкретных наборах данных. Проводится сравнительный анализ количества правильно решенных задач в рассматриваемых системах. Обсуждаются основные проблемы, с которыми можно столкнуться при подробном решении задач по вычислительной математике в каждом из чат-ботов. Данное исследование может представлять практический интерес для исследователей, разработчиков, преподавателей и пользователей, которые применяют данные виртуальные помощники в своей работе. Проведенный эксперимент позволит лучше оценить эффективность применения рассматриваемых систем в области математики.

Ключевые слова: *чат-бот, YandexGPT, ChatGPT, Gemini, Copilot, численный эксперимент, искусственный интеллект, математический анализ, вычислительная математика.*

Цитирование: Винокурова Д. В. Численный эксперимент вычислительных способностей современных чат-ботов в решении задач по математическому анализу и вычислительной математике // Компьютерные инструменты в образовании. 2024. № 3. С. 33–47. doi:10.32603/2071-2340-2024-3-33-47

1. ВВЕДЕНИЕ

Концепция чат-ботов начала своё развитие ещё в 1950-х годах, когда выдающийся ученый в области информатики Алан Тьюринг в своей статье предложил новый подход к пониманию мышления, переформулировав вопрос о том, могут ли машины думать,

на более практичный: «Может ли машина общаться таким образом, что ее поведение будет неотличимо от человеческого?». Для ответа на этот вопрос Тьюринг предложил критерий, который в настоящее время известен как тест Тьюринга. Он основывается на вариации игры «Имитация», где человек ведет общение с испытуемым без возможности видеть его. По окончании игры человек должен определить, кем являлся его собеседник — живым человеком или искусственным интеллектом. В этом контексте способность к мышлению заменяется способностью к общению на таком уровне, что участники разговора воспринимают машину как мыслящее существо [1].

Одним из старейших чат-ботов является программа Элиза, созданная в Массачусетском технологическом институте (1964–1966), которая имитировала роль психотерапевта, задавая открытые вопросы и отвлекая внимание от себя к пользователю. Люди доверяли ей свои конфиденциальные данные и секреты [1].

С начала 2000-х чат-боты значительно эволюционировали и стали использоваться в различных сферах — от клиентского обслуживания до образования. За последнее десятилетие появились голосовые помощники Cortana, Алиса, чат-боты ChatGPT, YandexGPT, Gemini, Copilot для генерации текстов на основе естественного языка, способных вести диалоги с пользователями. Нейронные сети удобно применять для распознавания речи и образов, генерации изображений по описанию, создания контента. Чат-боты эффективно обрабатывают естественный язык, однако в математических вопросах, требующих применения логических и абстрактных рассуждений, могут возникать трудности.

В данной статье приводится численный эксперимент по решению задач из некоторых разделов математического анализа и вычислительной математики, которые проводились в феврале и марте-апреле 2024 года с использованием чат-ботов YandexGPT 2 [2], ChatGPT 3.5 [3], Gemini [4], Copilot [5]. Выбор данных чат-ботов обусловлен возможностью бесплатного доступа, что делает их доступными для широкой аудитории, а также популярностью разработчиков, в связи с чем эти чат-боты вызывают у пользователей наибольшее доверие.

2. ХАРАКТЕРИСТИКИ СОВРЕМЕННЫХ ЧАТ-БОТОВ

В настоящее время к наиболее популярным чат-ботам можно отнести YandexGPT, ChatGPT, Gemini и Copilot. Данные системы основываются на большой языковой модели (Large Language Models (LLM)), которая содержит мощные алгоритмы обработки естественного языка. В них используются методы глубокого обучения для анализа, понимания и генерации текста. Они способны решать задачи в области обработки естественного языка, включая генерацию текста, машинный перевод, ответы на вопросы, суммаризацию текста. GPT модели (Generative Pre-trained Transformer models) — это класс нейронных сетей, входящих в состав чат-ботов, которые используют архитектуру трансформеров и обладают возможностью генерации текста и изображений.

YandexGPT (был известен под названием YaLM) — языковая модель, основанная на архитектуре GPT, разработанная компанией Яндекс. Впервые компания применила GPT-модели в 2021 году. Генеративные нейросети помогали быстро находить ответы на вопросы в поисковике. Трансформеры стали основной архитектурой нейросетей для задач NLP. Семейство языковых моделей YaLM включает младшую модель с 1 млрд параметров и старшую — с 13 млрд, которые используют этот же подход. Обучение и внедрение таких крупных моделей представляет собой сложную задачу [6]. База данных, на которой обучали YandexGPT, ограничена мартом 2023 года.

ChatGPT создан американской научно-исследовательской организацией OpenAI. В настоящее время в бесплатной версии представлена модель GPT-3.5. ChatGPT не имеет доступа в интернет. Информация, по которой происходило обучение нейронной сети, ограничена 2021 годом [7]. ChatGPT был оптимизирован для диалога с помощью обучения с подкреплением и обратной связью с человеком (RLHF) [8].

Gemini (ранее известный под названием Google Bard) создан компанией Google. В основе чат-бота лежит языковая модель для диалоговых приложений LaMDA, разработанная Google на базе архитектуры нейронной сети Transformer [9]. LaMDA имеет до 137 млрд параметров и использует одну модель для выполнения нескольких задач: генерирует потенциальные ответы, которые основываются на внешнем источнике знаний и повторно ранжируются для поиска ответа высочайшего качества [10]. В бесплатной версии в качестве входных данных Gemini, помимо текстовых данных, принимает изображения.

Copilot (ранее известный под названием Bing Chat) разработан компанией Microsoft на базе модели ChatGPT следующего поколения. Для работы с моделью OpenAI использует модель Microsoft Prometheus. Доступен через поисковую систему Bing и как встроенная функция веб-браузера Microsoft Edge. Чат-бот обладает возможностью принимать на вход не только текстовые данные, но и изображения, позволяет осуществлять поиск необходимой информации в интернете и представлять обработанные результаты со ссылками на первоисточники [11]. В Copilot имеется функция создания изображений на основе текстовых описаний с использованием модели DALL-E-3. Существует возможность выбора стиля разговора в трех вариантах: творческий, сбалансированный и точный [12].

3. ОБЗОР ИССЛЕДОВАНИЙ

Применение искусственного интеллекта (ИИ) в решении математических задач обсуждается во многих научных работах. В исследовании А. И. Дроздова [13] была оценена способность чат-ботов ChatGPT 4, Gemini и Copilot решать задачи из области математического анализа по введенной системе оценок. При сравнительном анализе было отмечено, что ChatGPT показал лучшие результаты по сравнению с Copilot и Gemini. ChatGPT допускал ошибки в задачах на индукцию, наблюдались проблемы в рациональности представленных решений. Результаты исследования выявили, что у каждого из чат-ботов имелись некоторые недостатки: у Gemini наблюдалась тенденция подстраивать свои ответы под заданные вопросы, в то время как у Copilot возникали проблемы с форматированием текста и применением нерациональных формул для решения простых задач. Все чатботы испытывали затруднения при вычислении интегралов.

Исследование производительности ChatGPT на математических словесных задачах из набора данных DRAW-1K, содержащего 1000 задач с ответами и алгебраическими уравнениями, выявило зависимость производительности модели от вида представления расчетов: подробного описания или ответа без пояснений. Установлено, что вероятность ошибок возрастает с увеличением числа операций сложения и вычитания [14]. В другой работе [15] было выявлено, что GPT-3 задачи на деление и вычитание даются тяжелее, чем задачи на сложение и умножение. Исследование показало, что современные системы ИИ не всегда справляются с многими случайно сгенерированными текстовыми задачами, требующими выполнения нескольких этапов, и сталкиваются с трудностями при оперировании абстрактными понятиями.

В работе [16] сравнивались чат-боты ChatGPT-3.5, ChatGPT-4 и Google Bard по их способности предоставлять корректные результаты на математические и логические задачи

в объеме 30 вопросов. Данные вопросы были разделены на два набора. Первый состоял из задач, которые не содержатся в интернете, второй — из задач, которые можно было найти в интернете с решениями. В результате эксперимента Google Bard (Gemini) лучше справился с вопросами, которые уже были опубликованы в интернете. Чат-боты не всегда предоставляли точные ответы на арифметические, алгебраические выражения и простые логические головоломки. Сложные математические задачи, в которых было необходимо осуществить логические рассуждения, вызвали трудности. Наблюдалась неправильная интерпретация вводимых данных. Предоставляемые решения были подробными, длинными и в большинстве случаев не имели смысла. Некоторые ответы, которые были правильными при первой попытке, приводили к неудачам в последующих.

В статье [17] анализировалась возможность генерации вопросов ChatGPT для каждого урока по математике для начальной, средней и старшей школ на основе тестов, включающих 121 тему и 428 уроков. Исследователи установили, что сгенерированные вопросы были менее сложными, чем предполагалось. В некоторых случаях ChatGPT генерировал вопросы, которые не соответствовали заявленным темам. При генерации сложных вопросов получались задачи, лишенные смысла.

Авторы работы [18] анализировали математические возможности ChatGPT на общедоступных наборах данных и данных GHOSTS и miniGOSTS, созданных вручную. Оценивалась способность ChatGPT решать задачи, связанные с вычислениями, завершать математические доказательства, решать сложные вопросы, требующие оригинальных решений, анализировать литературу и мыслить в различных областях. В результате проведенной работы было выявлено, что ChatGPT не справляется с задачами курса математики университетского уровня, испытывает затруднения при выполнении сложных символьных вычислений, плохо справляется с математическими олимпиадами и доказательствами упражнений, предназначенных для выпускников.

В исследовании [19] проводился анализ математических способностей ChatGPT на вьетнамском национальном выпускном экзамене средней школы (VNHSGE) с 250 вопросами и несколькими вариантами ответов, разделенными на четыре уровня сложности. Успешность чат-бота варьировалась от 52 % до 66 %, что является более низким показателем по сравнению с показателями успеха вьетнамских студентов. ChatGPT имел проблемы с решением сложных математических задач и интерпретацией графических данных.

В научном исследовании [20] производилось тестирование GPT-4 с плагинами Wolfram Alpha и Code Interpreter на 105 оригинальных задачах по естествознанию и математике на уровне средней школы и колледжа. Плагины расширяли способности модели, но GPT-4 не смог полностью использовать их возможности, выполняя бессмысленные вызовы плагинов и неточно осуществляя интерпретацию результатов.

4. МЕТОДОЛОГИЯ ЭКСПЕРИМЕНТА

В вышеперечисленных исследованиях основное внимание уделялось анализу математических способностей ChatGPT [14, 15, 18, 19], сравнению вычислительных возможностей двух чат-ботов — ChatGPT и Gemini [16], а также трех — ChatGPT, Gemini и Copilot [13]. В основном в работах оценивались возможности искусственного интеллекта на экзаменационных задачах в начальной, средней и старшей школах [17, 19], на алгебраических и логических задачах [14–16]. Однако стоит отметить, что не было выполнено численного эксперимента производительности математических способностей над линейкой чат-

ботов YandexGPT, ChatGPT, Gemini, Copilot. В рамках данной работы был проведен анализ возможностей чат-ботов по решению задач, связанных с вычислительной математикой и математическим анализом.

Общая численность задач математического анализа (МА) составила 693 задачи для трех тем (пределы, производные, интегралы):

- пределы числовых последовательностей (предел отношения двух многочленов — задача 2 (31 задача) [21, с. 7–8]; предел от иррациональностей — задача 3 (31 задача) [21, с. 8–9], задача 4 (31 задача) [21, с. 9–10]; второй замечательный предел — задача 6 (31 задача) [21, с. 12]);
- пределы функций (предел отношения двух многочленов — задача 9 (31 задача) [21, с. 14]; первый замечательный предел — задача 11 (31 задача) [21, с. 15–16]; нахождение пределов с использованием замены переменной — задача 12 (31 задача) [21, с. 16]; пределы от эквивалентных бесконечно малых — задача 14 (31 задача) [21, с. 7–8], задача 17 (31 задача) [21, с. 17–18]; второй замечательный предел — задача 16 (31 задача) [21, с. 19]);
- производные (производные от частного, разности, суммы, сложной функции — задача 5 (31 задача) [21, с. 30–31]; производные от показательных, тригонометрических и степенных функций — задача 6 (28 задач) [21, с. 31–32]; производные от логарифмов — задача 7 (29 задач) [21, с. 32]; производные от тригонометрических функций — задача 8 (31 задача) [21, с. 33]; производные от обратных тригонометрических функций и степенных функций — задача 9 (20 задач) [21, с. 33–34]; производные от сложных функций, гиперболических функций, обратных тригонометрических функций — задача 10 (22 задачи) [21, с. 34–35]);
- интегралы (метод интегрирования по частям — задача 1 (31 задача) [21, с. 58], задача 2 (30 задач) [21, с. 59–60]; интегрирование заменой переменной — задача 3 (26 задач) [21, с. 61–62], задача 4 (27 задач) [21, с. 61–62]; интегрирование рациональных дробей — задача 5 (21 задача) [21, с. 62–64]; универсальная тригонометрическая подстановка — задача 8 (27 задач) [21, с. 66–67]; $\int \sin^m x \cdot \cos^n x \, dx$ — задача 10 (31 задача) [21, с. 70–71]; интегрирование иррациональных функций с помощью тригонометрической подстановки — задача 12 (29 задач) [21, с. 74–75]).

В задачах по вычислительной математике (ВМ) для рассмотрения были взяты следующие темы: решение нелинейных уравнений (метод дихотомии (метод половинного деления), метод простой итерации, метод Ньютона, метод хорд) [22, с. 60], решение систем линейных уравнений (СЛУ) (метод Гаусса, метод LU-Разложения, метод правой прогонки) [22, с. 70], интерполирование функций (интерполяционные полиномы Лагранжа и Ньютона) [23, с. 113], численное интегрирование (метод прямоугольников (формулы для левых, правых и средних прямоугольников), метод трапеции, метод Симпсона (метод парабол), метод Ньютона (правило 3/8)) [23, с. 93]. В каждом методе для решения было предложено по три варианта задач.

В исследовании по математическому анализу оценивалась точность конечных результатов и для произвольных задач проводилась проверка промежуточных этапов вычислений. В эксперименте по вычислительной математике анализировались способности чат-ботов выполнять правильные промежуточные вычисления и давать корректные итоговые ответы.

Набор задач для ввода в системы подготавливался в Latex формате. В первоначальном виде данные были представлены в виде изображений и для их распознавания применялся Gemini. В процессе идентификации условий задач с изображений возникали ошибки,

и приходилось несколько раз объяснять чат-боту, в каком формате необходимо предоставлять результаты. Возникали ситуации, когда нейронная сеть могла придумывать задачи или отказываться распознавать изображения. Задачи с изображений приходилось неоднократно проверять на правильность их представления в LaTeX и в некоторых случаях полностью набирать вручную. В Copilot попытки распознавания изображений были хуже, чем у Gemini. Числа плохо распознавались, и результаты часто представлялись в неправильном формате. Диалог многократно прерывался, что затрудняло продолжение работы.

После получения набора задач в LaTeX формате на втором этапе (рис. 1) для каждого чат-бота [2–5] формировался запрос о необходимых действиях. Системы предоставляли результаты в зависимости от контекста запроса в подробном виде или кратком. Как правило, это занимало некоторое время в зависимости от объема задач и количества отправляемых запросов от пользователей в единицу времени.

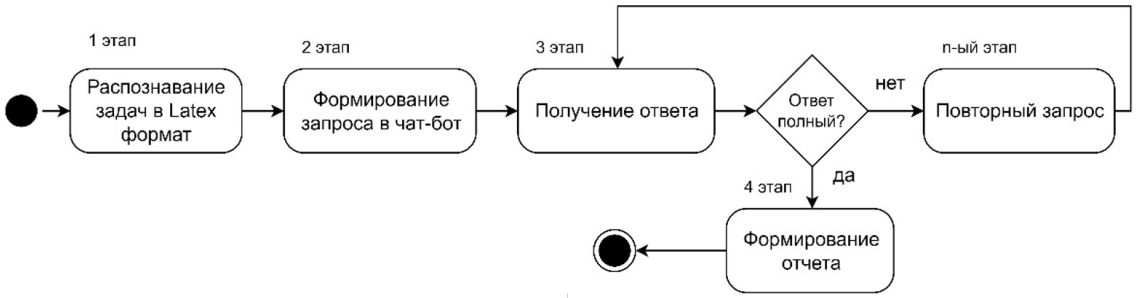

Рис. 1. Диаграмма состояний, отражающая последовательность этапов

После получения ответа проводилась проверка его полноты. В случае удовлетворительного результата на следующем этапе формировался отчет о предоставленных ответах. В случае возникновения проблем запрос уточнялся и осуществлялся повторный запрос. Данный процесс мог повторяться n раз в зависимости от получаемых результатов.

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

При проведении исследования были получены следующие результаты. В таблице 1 представлено количество правильных ответов на задачи по математическому анализу для производных и интегралов. Оценка правильности решений осуществлялась при помощи Wolfram Mathematica. Для некоторых задач с подробным объяснением решения получались громоздкими, промежуточные шаги которых не имели смысла, что приводило к неверным итоговым результатам. Нейросети лучше справились с решением задач на пределы (рис. 2). ChatGPT значительно опережал другие GPT модели в задачах на нахождение пределов числовой последовательности двух многочленов (задача 2), что составляет 61 % корректно решенных задач. Пределы от эквивалентных бесконечно малых функций лучше решили ChatGPT (39 %), Gemini (39 %) и Copilot (42 %). Хуже всего были вычислены пределы от иррациональностей (задача 3), YandexGPT (0 %) и Gemini (0 %) не справились ни с одной задачей. Общий итог корректно решенных задач по пределам из 310 задач в процентном соотношении составляет: YandexGPT — 7 %, ChatGPT — 23 %, Gemini — 16 %, Copilot — 21 %.

Таблица 1. Количественная информация по проведенному численному эксперименту по математическому анализу

Тип задачи \ Чат-бот	YandexGPT	ChatGPT	Gemini	Copilot
Производные				
Задача 5	0/31	0/31	0/31	0/31
Задача 6	0/28	0/28	0/28	0/28
Задача 7	0/29	0/29	0/29	0/29
Задача 8	0/31	0/31	0/31	0/31
Задача 9	0/20	0/20	0/20	0/20
Задача 10	0/22	0/22	0/22	1/22
Всего задач	0/161	0/161	0/161	1/161
Интегралы				
Задача 1	0/31	5/31	0/31	0/31
Задача 2	0/30	0/30	0/30	0/30
Задача 3	0/26	3/26	0/26	1/26
Задача 4	0/27	1/27	1/27	0/27
Задача 5	0/21	0/21	0/21	0/21
Задача 8	0/27	0/27	0/27	0/27
Задача 10	0/31	0/31	0/31	2/31
Задача 12	1/29	2/29	0/29	3/29
Всего задач	1/222	11/222	1/222	6/222

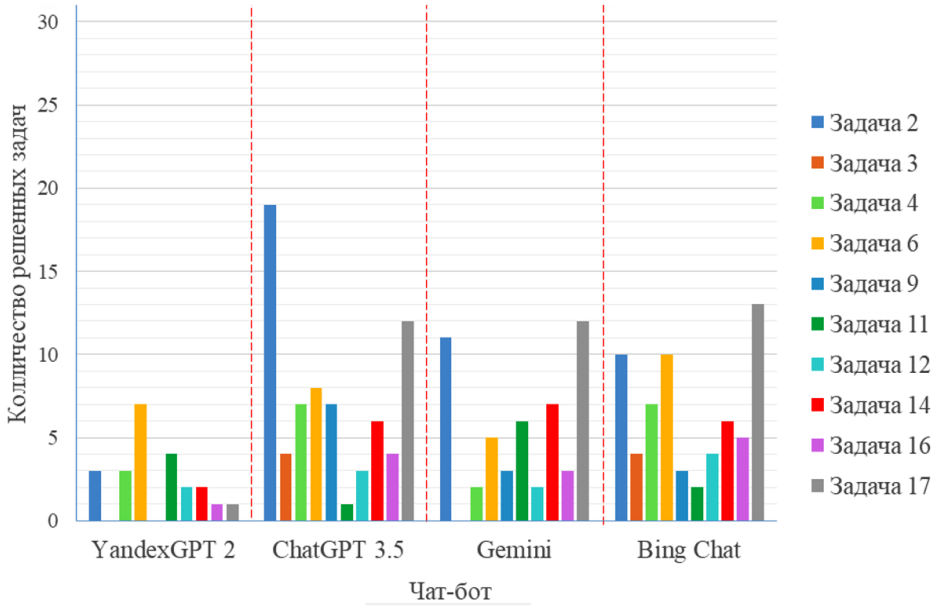

Рис. 2. Результаты решения задач по пределам

Значительные проблемы возникли у нейросетей при решении задач с вычислением производных. В задаче 10 Copilot решил всего 1 задачу, что составляет 5 % от количества

предлагаемых к решению задач. Количество правильно решенных задач по нахождению производных из 161 задачи в процентном соотношении составляет: YandexGPT, ChatGPT и Gemini — 0 %, Copilot — 1 %. У диалоговых систем возникли затруднения при вычислении интегралов. Лучше всех справился ChatGPT (в задаче 1 — 16 % правильно решенных задач, в задаче 3 — 12 %). Copilot в задаче 12 предоставил корректные ответы на 10 % задач. Количество корректно решенных задач из 222 задач составляет: YandexGPT и Gemini — 0 %, ChatGPT — 5 %, Copilot — 3 %.

Результаты эксперимента с задачами по вычислительной математике представлены в таблице 2, где ИО обозначает итоговый ответ, ПВ — промежуточные вычисления. Нецелочисленные значения в столбце ИО означают, что чат-ботом было осуществлено приближение путем «подгона» до правильного результата. Дробные значения в столбце ПВ указывают на несколько ситуаций: решение было представлено наполовину правильно, решение полностью не расписано или ответ обобщен до выражений без конкретных значений. Встречались подробные решения, в которых итоговый ответ был правильным,

Таблица 2. Количественная информация по проведенному численному эксперименту по вычислительной математике

Наименование задачи	**Корректно решеные задачи**							
	YandexGPT		**ChatGPT**		**Gemini**		**Copilot**	
	ИО	ПВ	ИО	ПВ	ИО	ПВ	ИО	ПВ
Решение нелинейных уравнений								
Метод дихотомии	0/3	0/3	0/3	0/3	0/3	0/3	1/3	0.5/3
Метод простой итерации	0/3	0/3	1/3	1/3	0/3	0/3	0/3	0/3
Метод Ньютона	0/3	0/3	0.5/3	0.5/3	0/3	0/3	0/3	0/3
Метод хорд	0/3	0/3	0/3	0/3	1/3	1/3	0/3	0/3
Решение СЛУ								
Метод Гаусса	0/3	0/3	0/3	0/3	0/3	0/3	0.5/3	0/3
Метод LU-Разложения	0/3	0/3	0/3	0/3	0/3	0/3	0.5/3	0/3
Метод правой прогонки	0/3	0/3	0/3	0/3	0/3	0/3	0/3	0/3
Интерполирование функций								
Интерполяционный полином Лагранжа	0/3	1.5/3	0/3	1.5/3	0/3	0/3	3/3	1.5/3
Интерполяционный полином Ньютона	0/3	0/3	0/3	1/3	0/3	0/3	1/3	0.5/3
Численное интегрирование								
Метод прямоугольников (формула левых прямоугольников)	0/3	0.5/3	0/3	1/3	0.5/3	1/3	3/3	1.5/3
Метод прямоугольников (формула правых прямоугольников)	0/3	0/3	0/3	0.5/3	0/3	0.5/3	3/3	1/3
Метод прямоугольников (формула средних прямоугольников)	0/3	0/3	1/3	1/3	0.5/3	0.5/3	3/3	1.5/3
Метод трапеции	0/3	1/3	0.5/3	0.5/3	0/3	0.5/3	3/3	1.5/3
Метод Симпсона	0/3	0/3	0/3	1/3	0/3	0/3	3/3	1.5/3
Метод Ньютона (правило 3/8)	0/3	0/3	1/3	1.5/3	0/3	0/3	3/3	1.5/3

а в промежуточных вычислениях встречались ошибки, в таких решениях нельзя было проследить за логикой рассуждений, которые приводили к такому результату.

Рассмотрим количество правильно решенных задач в таблице 2 в процентном соотношении по применяемым методам в каждой из рассмотренных тем: решение нелинейных уравнений: YandexGPT — 0 %, ChatGPT — 13 %, Gemini — 8 %, Copilot — 8 %, решение СЛУ: YandexGPT — 0 %, ChatGPT — 0 %, Gemini — 0 %, Copilot — 11 %, интерполирование функций: решение нелинейных уравнений: YandexGPT — 0 %, ChatGPT — 0 %, Gemini — 0 %, Copilot — 67 %, численное интегрирование: YandexGPT — 0 %, ChatGPT — 14 %, Gemini — 6 %, Copilot — 100 %.

Из анализа полученных результатов видно, что YandexGPT плохо справился с задачами по всем темам, у ChatGPT иногда встречались правильные решения, Gemini справлялся с задачами лучше YandexGPT, но его результаты ниже, чем у ChatGPT. Copilot предоставлял ответы на задачи, которые вызывали затруднения у других чат-ботов. Вероятно, чат-бот осуществлял запрос к вспомогательным математическим пакетам. Опишем основные ситуации, возникшие в каждом из чат-ботов при решении задач по вычислительной математике.

В YandexGPT в методе дихотомии присутствовали произвольные интервалы, которые не соответствовали условию задач. При применении метода простой итерации прерывалось пошаговое описание. При нажатии кнопки «Продолжи» появлялось сообщение о невозможности найти ответ. В методе Ньютона YandexGPT в процессе решения начинал искать производные, часто диалог прерывался, при повторном запросе выдавалась часть ответа. Нахождение производных и проверка выполнения условий осуществлялись правильно, в процессе нахождения приближенных значений происходило зацикливание на определенном значении. Многократно рассказывалось о принципах решения, конкретных действий не производилось. В методе хорд аналогично предыдущему методу описание этапов нахождения корня внезапно прекращалось, при продолжении терялась связь с предыдущим сообщением. В решениях СЛУ различными методами получаемые матрицы отображались в нечитаемом виде.

В методе дихотомии ChatGPT осуществлял неверное вычисление значений функции. Нейросеть «подгоняла» вычисления под значения, чтобы они были больше или меньше, и в результате выбирался неверный интервал. В методе простой итерации и методе Ньютона ChatGPT аналогичным образом происходил подбор приблизительных значений и получались корни близкие к эталонным. В методе хорд в одном из случаев появлялось сообщение о том, что ни одно из условий не выполнено и нужны другие методы для решения, что являлось ошибочным. При решении СЛУ наблюдались ошибки в промежуточных вычислениях. При нахождении интерполяционного полинома в форме Лагранжа в подробных объяснениях с конкретными значениями возникали ошибки. В решениях задач на нахождение интерполяционного многочлена в форме Ньютона описывались алгоритмы действий для общего случая без конкретных значений. В задачах численного интегрирования правильные решения и окончательный результат по одному разу встречались в методе прямоугольников для формулы средних прямоугольников и методе Ньютона, в остальных случаях рассуждения не приводили к правильным результатам.

В Gemini в задачах на применение метода дихотомии выбор интервалов происходил неверно. В методе Ньютона чат-бот выполнял проверку условий и сообщал, что условия не выполняются (в случае, когда они выполнялись), и переходил к вычислениям значений, не соответствовавших корректным значениям, в результате получались неправильные значения корней нелинейных уравнений. В решениях СЛУ пошаговые решения об-

рывались, для дальнейшего получения решения после соответствующих запросов чат-боту часть информации терялась, и ответ искажался. В задачах численного интегрирования встречались недопустимые числа в ответах. В некоторых случаях начало решений сопровождалось правильными рассуждениями, однако в дальнейшем происходили ошибки в рассуждениях, которые приводили к некорректным результатам (рис. 3).

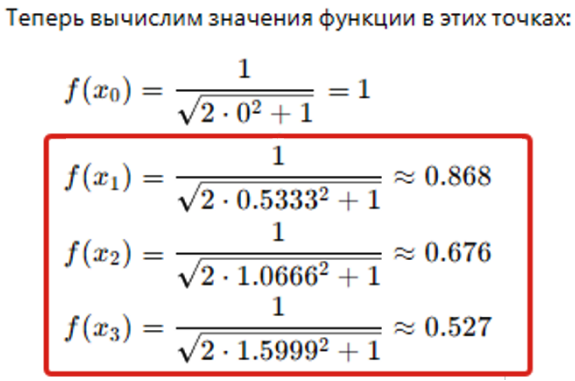

Рис. 3. Численное интегрирование: метод трапеции в Gemini

В Copilot в задачах на применение метода дихотомии в ходе проведения итераций вручную чат-ботом до требуемой точности наблюдались неверные вычисления промежуточных результатов, что способствовало выбору неправильного интервала и нахождению неверных корней уравнения (рис. 4). В методе простой итерации присутствовали подробные объяснения о применяемом методе, чему равны соответствующие параметры в общем виде, как выглядит итерационная формула. Численные решения в формате чата Copilot не смог осуществить. В качестве решений были представлены функции и итерационные формулы с меняющимися индексами. При использовании метода Ньютона без указания необходимых условий в запросе решение задач было некорректным. После уточнения запроса путем объяснений необходимых условий проверки для метода Ньютона и правил оценки ошибки нейросеть начинала выполнять вычисления, но игнорируя данные требования. Это приводило к получению неверных значений и, как следствие, к неправильным корням уравнений. В некоторых случаях Copilot дублировал информацию, изначально введенную в запрос, не предпринимая никаких действий.

Решение СЛУ в Copilot в кратком виде было более корректно, но не для всех корней. При объяснении хода решения возникали проблемы. В процессе вычисления интерполя-

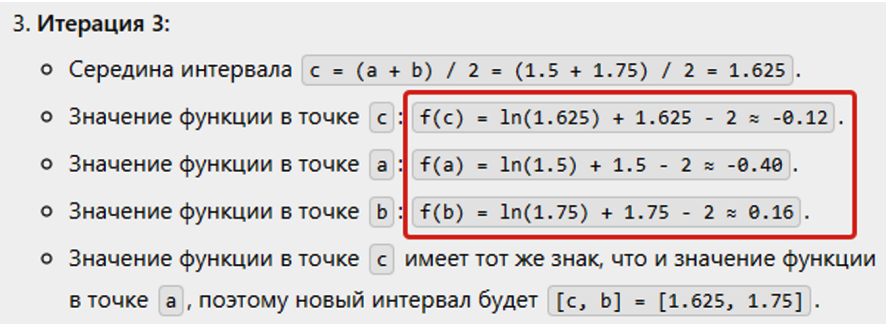

Рис. 4. Метод дихотомии, решение в Copilot

ционного полинома Лагранжа кратко в виде ответа ошибок не возникало. Однако при проведении подробного решения задач многократно дублировались значения, которые необходимо вычислять без упрощения. В бесплатной версии в беседе с чат-ботом быстро достигалось ограничение, в случае создания новой беседы терялась цепочка рассуждений. Базисные полиномы не упрощались. Интерполяционный полином Ньютона в виде конечного ответа был вычислен только для одной задачи из трех, для остальных задач Copilot не смог выполнить вычисления (рис. 5). Полное решение для единственной задачи, которую чат-бот смог решить, содержало ошибки и в результате приводило к неверным результатам.

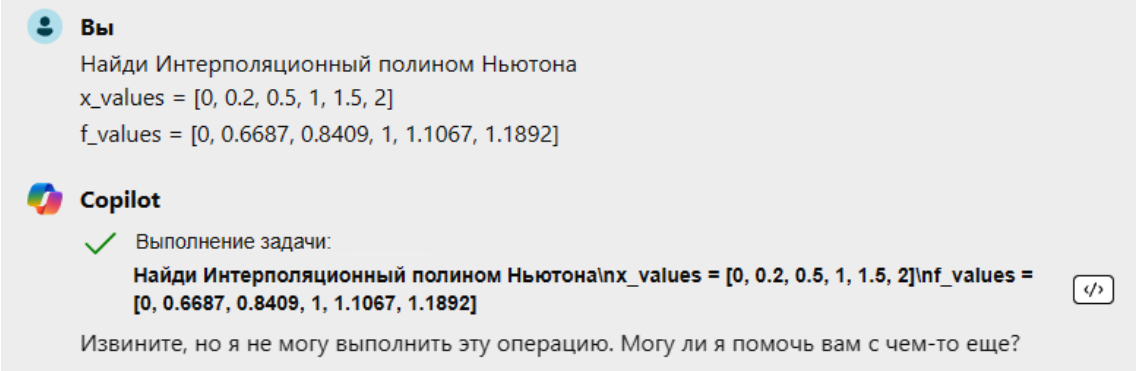

Рис. 5. Вычисление интерполяционного полинома Ньютона в Copilot

В «Записках Гейтса» [24] было отмечено, что искусственный интеллект не в полной мере улавливает контекст запроса человека, что может приводить к появлению странных результатов, и не всегда способен выдавать верные решения математических задач из-за ограничений в абстрактном мышлении. К сожалению, нейронным сетям трудно достичь уровня мышления, сравнимого с человеческим мозгом. Прогноз авторов статьи [25] о том, что запас высококачественных данных для обучения нейронных сетей может исчерпаться к 2026 году, вызывает опасения. Учитывая, что доля математического контента в обучающей выборке невелика, существует риск приостановки развития нейронных сетей и ограничения возможностей предоставления качественных решений.

6. ЗАКЛЮЧЕНИЕ

Данное исследование показало, что математические задачи для чат-ботов являются трудоемкими и предлагаемые решения не всегда корректны. Применение рассмотренных автоматизированных помощников для решения математических задач оказывается малоэффективным из-за необходимости использования дополнительных затрат времени на выявление и исправление ошибок в предоставляемых решениях. В ходе эксперимента и исследования, проведенного коллективом авторов [16], было замечено, что предлагаемые чат-ботами ответы очень длинные и большинство из них могут оказаться абсурдными, поэтому человеку сложнее оценить достоверность получаемых ответов. Надеемся, что результаты численного эксперимента окажутся полезными для исследователей, разработчиков, преподавателей и пользователей, которые применяют данные системы в рамках своей деятельности. Представленное исследование поможет глубже понять возможности виртуальных помощников в области математики на сегодняшний день.

Список литературы

1. *Zemčík M.T.* A Brief History of Chatbots // DEStech Transactions on Computer Science and Engineering, 2019. P. 1–19. doi: 10.12783/dtcse/aicae2019/31439
2. *YandexGPT2.* Alisa: Intelligent personal assistant. 2024 [Электронный ресурс]. URL: https://ya.ru/alisa_davay_pridumaem?utm_source=landing (дата обращения: 15.10.2024).
3. *OpenAI, Inc.* ChatGPT: Generative artificial intelligence chatbot. 2024 [Электронный ресурс]. URL: https://chat.openai.com/ (дата обращения: 15.10.2024).
4. *Google LLC.* Gemini: Generative artificial intelligence chatbot. 2024 [Электронный ресурс]. URL: https://gemini.google.com/ (дата обращения: 15.10.2024).
5. *Microsoft Corp.* Copilot in Microsoft Bing. 2024 [Электронный ресурс]. URL: https://www.bing.com/chat/ (дата обращения: 15.10.2024).
6. *Yandex LLC.* Как Яндекс применил генеративные нейросети для поиска ответов. 2024 [Электронный ресурс]. URL: https://habr.com/ru/companies/yandex/articles/561924/ (дата обращения: 15.10.2024).
7. *OpenAI, Inc.* What is ChatGPT? 2024 [Электронный ресурс]. URL: https://help.openai.com/en/articles/6783457-what-is-chatgpt (дата обращения: 15.10.2024).
8. *OpenAI, Inc.* How ChatGPT and our language models are developed. [Электронный ресурс]. URL: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed (дата обращения: 15.10.2024).
9. *Collins E., Ghahramani Z.* LaMDA: our breakthrough conversation technology. 2021. [Электронный ресурс]. URL: https://blog.google/technology/ai/lamda/ (дата обращения: 15.10.2024).
10. *Thoppilan R. et al.* Lamda: Language models for dialog applications [Электронный ресурс]. URL: https://arxiv.org/abs/2201.08239 (дата обращения: 15.10.2024).
11. *Mehdi Y.* Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. 2023 [Электронный ресурс]. URL: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ (дата обращения: 15.10.2024).
12. *Microsoft Corp.* What is Bing Chat, and How Can You Use It? 2023 [Электронный ресурс]. URL: https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it?form=MA13KP (дата обращения: 15.10.2024).
13. *Дроздов А. И.* Применение нейронных сетей в задачах математического анализа // Компьютерные системы и сети : сборник статей 59-й научной конференции аспирантов, магистрантов и студентов. Минск, 2023. С. 473–479.
14. *Shakarian P. et al.* An independent evaluation of ChatGPT on mathematical word problems (MWP). 2023 [Электронный ресурс]. URL: https://arxiv.org/abs/2302.13814 (дата обращения: 15.10.2024).
15. *Novak D.* Analyzing the GPT-3 AI's Ability to Predict the Answer to Algebraical Questions // Journal of Student Research, 2023. Т. 1, №. 1. doi:10.47611/jsrhs.v12i1.3998
16. *Plevris V., Papazafeiropoulos G., Jiménez Rios A.* Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard // AI. 2023. Т. 4, №. 4. С. 949–969. doi:10.3390/ai4040048
17. *Van Long P. P. et al.* ChatGPT as a Math Questioner? Evaluating ChatGPT on Generating Pre-university Math Questions2023 [Электронный ресурс]. URL: https://arxiv.org/abs/2312.01661 (дата обращения: 15.10.2024).
18. *Frieder S. et al.* Mathematical capabilities of chatgpt. 2023. [Электронный ресурс]. URL: https://arxiv.org/abs/2301.13867 (дата обращения: 15.10.2024).
19. *Dao X. Q., Le N. B.* Investigating the effectiveness of chatgpt in mathematical reasoning and problem solving: Evidence from the vietnamese national high school graduation examination. 2023 [Электронный ресурс]. URL: https://arxiv.org/abs/2306.06331 (дата обращения: 15.10.2024).
20. *Davis E., Aaronson S.* Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on math and science problems. 2023 [Электронный ресурс]. URL: https://arxiv.org/abs/2308.05713 (дата обращения: 15.10.2024).
21. *Кузнецов Л. А.* Сборник заданий по высшей математике (типовые расчеты). М.: «Высшая школа», 1994.

22. *Зенков А. В.* Вычислительная математика для IT-специальностей : учебное пособие. Москва; Вологда: Инфра-Инженерия, 2022.
23. *Зализняк В. Е.* Теория и практика по вычислительной математике : учеб. пособие. Красноярск: Сиб. федер. ун-т, 2012.
24. *Gates B.* The Age of AI has begun. 2023 [Электронный ресурс]. URL: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun (дата обращения: 15.10.2024).
25. *Villalobos P. et al.* Will we run out of data? An analysis of the limits of scaling datasets in machine learning. 2022 [Электронный ресурс]. URL: https://arxiv.org/abs/2211.04325 (дата обращения: 15.10.2024).

Поступила в редакцию 19.09.2024, окончательный вариант — 15.10.2024.

Винокурова Дарья Валентиновна, аспирант, Институт информационных технологий и технологического образования, РГПУ им. А. И. Герцена, ✉ d.v.vinokurova@gmail.com

Computer tools in education, 2024
№ 3: 33–47
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-3-33-47

Numerical Experiment of Computational Capabilities of Modern Chat-Bots in Solving Problems in Mathematical Analysis and Computational Mathematics

Vinokurova D. V.[1], Postgraduate, ✉ d.v.vinokurova@gmail.com

[1]Herzen University, 48 Moika river embankment, 191186, Saint Petersburg, Russia

Abstract

The paper describes a numerical experiment on calculation of mathematical problems by chatbots (Yandex GPT 2, ChatGPT 3.5, Gemini, Copilot) on some topics of mathematical analysis (limits, derivatives, integrals), including 693 problems, and computational mathematics (solution of nonlinear equations, solution of systems of linear equations, interpolation of functions, numerical integration), consisting of 45 problems. The main characteristics of modern virtual assistants are considered. A review of research on the application of artificial intelligence in solving mathematical problems on various tests and data sets is presented. The paper considers the shortcomings manifested in the work of chatbots, analyzes their performance on specific data sets. A comparative analysis of the number of correctly solved problems in the considered systems is carried out. The main problems that can be encountered when solving computational mathematics problems in detail in each of the chatbots are discussed. This study may be of practical interest for researchers, developers, teachers and users who use these virtual assistants in their work. The conducted experiment will allow to better evaluate the effectiveness of the application of the considered systems in the field of mathematics.

Keywords: *chatbot, YandexGPT, ChatGPT, Gemini, Copilot, numerical experiment, artificial intelligence, mathematical analysis, computational mathematics.*

Citation: D. V. Vinokurova, “Numerical Experiment of Computational Capabilities of Modern Chat-Bots in Solving Problems in Mathematical Analysis and Computational Mathematics,” *Computer tools in education*, no. 3, pp. 33–47, 2024 (in Russian); doi:10.32603/2071-2340-2024-3-33-47

References

1. M. T. Zemčík, “A Brief History of Chatbots,” *DEStech Transactions on Computer Science and Engineering*, pp. 1–19, 2019; doi: 10.12783/dtcse/aicae2019/31439
2. Yandex LLC, “Alisa: Intelligent personal assistant,” in *yandex.ru*, 2024. [Online] (in Russian). Available: https://yandex.ru/yandexapp/ru/voiceassistant/yagpt/davay_pridumayem/
3. OpenAI, Inc., “ChatGPT: Generative artificial intelligence chatbot,” in *chat.openai.com*, 2024. [Online] (in Russian). Available: https://chat.openai.com/
4. Google LLC, “Gemini: Generative artificial intelligence chatbot,” in *gemini.google.com*, 2024. [Online] (in Russian). Available: https://gemini.google.com/
5. Microsoft Corp., “Copilot in Microsoft Bing,” in *copilot.microsoft.com*, 2024. [Online]. Available: https://copilot.microsoft.com/chats/
6. Yandex LLC, “How Yandex applied generative neural networks to search for answers,” in *habr.com*, 2024. [Online] (in Russian). Available: https://habr.com/ru/companies/yandex/articles/561924/
7. OpenAI, Inc., “What is ChatGPT?,” in *help.openai.com*, 2024. [Online]. Available: https://help.openai.com/en/articles/6783457-what-is-chatgpt
8. OpenAI, Inc., “How ChatGPT and our language models are developed,” in *help.openai.com*. [Online]. Available: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed
9. E. Collins and Z. Ghahramani, “LaMDA: our breakthrough conversation technology”, in *blog.google*, 18 May 2021. [Online]. Available: https://blog.google/technology/ai/lamda/
10. R. Thoppilan et al., “Lamda: Language models for dialog applications,” in *blog.google*, 2022. [Online]. Available: https://arxiv.org/abs/2201.08239
11. Y. Mehdi, “Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web,” *The Official Microsoft Blog*, 07 Feb. 2023. [Online]. Available: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/
12. Microsoft Corp., “What is Bing Chat, and How Can You Use It?,” in *microsoft.com*, 29 Sep. 2023. [Online]. Available: https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it?form=MA13KP
13. A. I. Drozdov, “Primenenie nejronnyh setej v zadachah matematicheskogo analiza” [Application of neural networks in calculus], in *Komp’yuternye sistemy i seti : sbornik statej 59-j nauchnoj konferencii aspirantov, magistrantov i studentov, Minsk, Belarus*, pp. 473–479, 2023 (in Russian).
14. P. Shakarian, et al. “An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP),” in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13814
15. D. Novak, “Analyzing the GPT-3 AI’s Ability to Predict the Answer to Algebraical Questions,” *Journal of Student Research*, vol. 12, no. 1, pp. 1–8, 2023; doi:10.47611/jsrhs.v12i1.3998
16. V. Plevris, G. Papazafeiropoulos, and A. Jiménez Rios, “Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard,” *AI*, vol. 4, no. 4, pp. 949–969, 2023; doi:10.3390/ai4040048
17. P. P. Van Long, D. A. Vu, N. M. Hoang, X. L. Do, and A. T. Luu, “ChatGPT as a Math Questioner? Evaluating ChatGPT on Generating Pre-university Math Questions,” in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2312.01661
18. S. Frieder et al., “Mathematical Capabilities of ChatGPT,” in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2301.13867
19. X.Q. Dao and N.B. Le, “Investigating the Effectiveness of ChatGPT in Mathematical Reasoning and Problem Solving: Evidence from the Vietnamese National High School Graduation Examination,” in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2306.06331
20. E. Davis and S. Aaronson, “Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on math and science problems,” in *arXiv*, 2023. [Online]. Available: https://arxiv.org/abs/2308.05713
21. L. A. Kuznetsov, *Sbornik zadaniy po vysshey matematike (tipovye raschety)* [Collection of tasks in higher mathematics (typical calculations)], Moscow, Russia: Vysshaya Shkola, 1994 (in Russian).
22. A. V. Zenkov, *Vychislitel’naya matematika dlya IT-spetsial’nostey: uchebnoe posobie* [Computational mathematics for IT specialties: a textbook], Moscow, Vologda, Russia: Infra-Inzheneriya, 2022 (in Russian).

23. V. E. Zaliznyak, G. I. Shchepanovskaya, *Teoriya i praktika po vychislitel'noy matematike: ucheb. posobie* [Theory and practice in computational mathematics: a textbook], Krasnoyarsk, Russia: Siberian Federal University, 2012 (in Russian).
24. B. Gates, "The Age of AI has begun," in *www.gatesnotes.com*, 21 Mar. 2023. [Online]. Available: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun
25. P. Villalobos, J. Sevilla, L. Heim, et al., "Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning," in *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2211.04325

Received 19-09-2024, the final version — 15-10-2024.

Daria Vinokurova, Postgraduate, Institute of Computer Science and Technology Education, Herzen University, ✉ **d.v.vinokurova@gmail.com**