УДК 004.056 **DOI** 10.15622/ia.25.2.1

А.А. Вяткин, А.В. Попцов, В.Д. Олисеенко, М.В. Абрамов

ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ МЕТОДА МАТРИЧНОЙ ФАКТОРИЗАЦИИ ДЛЯ РАНЖИРОВАНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Вяткин А.А., Попцов А.В., Олисеенко В.Д., Абрамов М.В. **Исследование применимости метода матричной факторизации для ранжирования больших языковых моделей.**

Аннотация. В последние годы широкое применение в области финансов получили большие языковые модели (англ. Large Language Models, LLM). Прямое сравнение таких моделей может быть затруднено, так как наборы данных и сами LLM могут быть закрыты, а параметры при оценке могут отличаться. В работе для задачи заполнения неизвестных метрик предлагается использование метода матричной факторизации из рекомендательных систем, изначально созданного для прогнозирования предпочтений пользователей. Целью работы является оценка применимости матричной факторизации для предсказания метрик качества LLM на финансовых задачах, а также разработка метода ранжирования LLM на основе агрегации метрик качества. Проводится эксперимент по применению матричной факторизации на собранных из научных исследований данных о 34 LLM и 42 финансовых наборах данных. Усредненная MAE метода на всех запусках составляет 0.07 на тестовом наборе данных. Верхние позиции в рейтинге занимают модели DeepSeek R1, OpenAI GPT-4o, OpenAI o1-mini, Fin-R1, Claude 3.5 Sonnet. Двумя способами исследуется влияние ошибки прогнозирования на итоговые предсказания: при помощи MAE и метода Монте Карло. Анализируются полученные результаты, основными выводами которых являются: а) метод матричной факторизации может быть применен для прогнозирования неизвестных значений метрик моделей на наборах данных; б) ведущие большие языковые модели сблизились в оценке настолько, что невозможно выявить явного лидера; в) большие ошибки предсказания позволяют выявить специфические особенности моделей на конкретных задачах. Представленный метод ранжирования способен упростить выбор подходящей модели для финансовых задач.

Ключевые слова: большие языковые модели, оценка качества моделей, матричная факторизация, финансовая сфера.

1. Введение. По мере развития больших языковых моделей (*англ. Large Language Models*, LLM) расширяется охват научных и промышленных областей, в которых они все более успешно внедряются [1]. Одной из сфер, где большие языковые модели могут дать существенный социальный и экономический эффект, является финансовая сфера [2]. В данной сфере LLM могут выступать автономными чат-ботами первой линии технической поддержки или в качестве помощников для второй и остальных линий поддержки, значительно сокращая время отклика и рабочую нагрузку, положительно влияя на качество обслуживания клиентов и клиентоцентричность [3]. Кроме того, LLM могут использоваться в задачах, связанных с обработкой большого количества финансовой документации [4], выявлением мошеннических операций [5],

кредитным скорингом [6], прогнозированием котировок на финансовых рынках [7] и др.

За последние несколько лет разработано большое количество различных LLM [8], и предпринимаются попытки сравнить их как с точки зрения архитектуры, так и качества на разных задачах [9]. Для такого сравнения используются наборы данных и бенчмарки, которые проверяют способности LLM выполнять различные действия. Примерами таких задач в банковской сфере могут выступать: ответы на вопросы по финансовым документам [10], извлечение информации из финансовых текстов [11], анализ тональности финансовых новостей [12], суммаризация финансовой документации [13] и др.

При сравнении моделей классическим решением является использование единого набора данных, как предложено в [14]. В некоторых ситуациях такое сравнение затруднено, поскольку наборы данных и сами LLM могут быть закрыты, а методики оценки, параметры и условия тестирования могут значительно различаться [15]. Поэтому актуальным остается использование других способов сравнения моделей, таких как методы прогнозирования неизвестных значений метрик, например, из области рекомендательных систем. Одним из таких методов является матричная факторизация [16], постановка задачи которой может легко быть применена к LLM и наборам данных для их тестирования. Получающиеся же в результате табличные данные о производительности моделей, в том числе предсказанной, позволяют ранжировать модели и в этом смысле сравнить их.

Целью данной работы является оценка применимости матричной факторизации для предсказания метрик качества LLM на финансовых задачах, а также разработка метода ранжирования LLM на основе агрегации метрик качества. Теоретическая значимость работы заключается в применении известного метода матричной факторизации в новой задаче – оценке LLM в финансовой области для ранжирования моделей в условиях нехватки данных. Практическая значимость состоит в том, что используемый метод позволяет ускорить выбор подходящей для финансовых задач LLM, при отсутствии возможности проведения прямого сравнения на существующих наборах данных.

2. Обзор литературы. В первой части данного раздела рассмотрены работы, использующие методы машинного обучения, изначально разработанные для рекомендательных систем, но применяемые для задач прогнозирования производительности LLM.

Во второй части описаны наборы данных, используемые далее для проведения эксперимента.

Использование методов машинного обучения. Идея применения методов из области рекомендательных систем, таких как матричная факторизация или коллаборативная фильтрация, для прогнозирования оценки качества LLM находит отражение в современных исследованиях [17 – 19]. Это связано с необходимостью оценки качества моделей в условиях, когда прямое их тестирование на всех задачах становится вычислительно затратным, непрактичным или вовсе невозможным в силу закрытости некоторых моделей и бенчмарков [20]. При использовании указанных методов данные представляются в виде матрицы, где строки соответствуют LLM, а столбцы – задачам или наборам данных, а матричная факторизация или коллаборативная фильтрация используется для прогнозирования пропущенных значений в матрице.

В исследовании [21] изучается, насколько предсказуема оценка качества LLM в зависимости от масштаба обучения, который включает в себя увеличение вычислительных ресурсов, количества параметров модели и объема данных для тренировки. Авторы показывают, что для отдельных задач прогнозирование затруднено. Но следует обратить внимание, что общее качество, на больших и разнородных бенчмарках достаточно хорошо предсказывается с помощью гладкой функции потерь, основанной на законах масштабирования.

В работе [17] исследуется метод, позволяющий избежать необходимости проводить дорогостоящие тесты моделей на каждом наборе данных. Предлагается рассматривать задачу предсказания неизвестных значений метрик матрицы как задачу прогнозирования. В матрице строки соответствуют различным большим визуально-языковым моделям, а столбцы – наборам данных. Используя вероятностную матричную факторизацию с цепями Маркова Монте-Карло, авторы демонстрируют значительно лучшие результаты по сравнению с базовыми подходами использования средних значений метрик по моделям и наборам данных (1.5–2 кратное увеличение), в особенности, когда доля тестовой выборки меньше 90%. Таким образом, использование подходов, основанных на применении матричной факторизации, представляется целесообразным, по крайней мере относительно упомянутых выше базовых подходов.

Другой подход, решающий проблему прогнозирования неизвестных метрик, основанный на идеях коллаборативной фильтрации, представлен в исследовании [18]. Здесь для предсказания пропущенных значений также используется матрица с данными

о качестве различных моделей, но с добавлением внешних факторов, описывающих модели и задачи. Этот метод, названный «Collaborative Performance Prediction», позволяет анализировать важность ранее не учитываемых различных факторов и демонстрирует на собранной авторами матрице точность 45%.

В статье [19] для заполнения неизвестных метрик предлагается двухэтапный метод, в котором сначала выбираются наиболее репрезентативные примеры для оценки, а затем, на основе результатов этой выборки, прогнозируется общее качество. Авторы демонстрируют, что их подход превосходит 4 базовых метода, демонстрируя наименьшую среднюю абсолютную ошибку (MAE).

Существующие исследования показывают, что научное сообщество активно разрабатывает методы преодоления проблемы невозможности сравнения моделей при оценке LLM. Тем не менее, представленные выше работы опираются на наборы данных из самых разных областей. В отличие от них, данное исследование полностью сосредоточено на финансовом домене.

Использованные в эксперименте работы. Для применения матричной факторизации было необходимо собрать таблицу со значениями метрик моделей на наборах данных из разных работ. Агрегирование результатов оценки качества LLM из различных источников было выполнено посредством анализа научной литературы. Цель состояла в том, чтобы сформировать единую матрицу данных, где строки соответствуют исследуемым языковым моделям, а столбцы – наборам данных. Каждая ячейка такой матрицы содержит численное значение качества конкретной модели на конкретном задании (наборе данных), измеренное определенной метрикой. Полученная матрица по своей природе является разреженной, поскольку ни одна из публикаций не содержит исчерпывающей оценки всех моделей по всем возможным задачам.

Основными критериями отбора используемых публикаций являлись:

- наличие таблицы со значениями метрик современных моделей из различных финансовых наборов данных;
- наличие пересекающихся моделей или наборов данных с другими работами;
- совпадение значений метрик с другими рассматриваемыми работами (так, в силу невозможности сопоставления, не включались работы, в которых значения метрик моделей на наборах данных отличались от других работ).

После отбора публикаций, отвечающих озвученным критериям, основой для построения используемой далее сводной матрицы послужили следующие четыре работы для анализа.

В работе [22] представлен FLaME – комплексный набор для оценки языковых моделей на финансовых задачах, включающий 20 основных наборов данных для задач обработки естественного языка в финансовой сфере: FinQA, ConvFinQA, TAT-QA, ECTSum, EDTSum [14], FiNER-ORD, FinRED, REFinD, FNXL, FinEntity, SubjECTive-QA, FiQA, FPB, NumClaim, Banking77, FinBench, Headline, FOMC, FinCausal-SC. В рамках исследования используется таксономия на основе сценариев для классификации данных и приводятся значения метрик для 23 моделей (включая GPT-4, Llama-2, BloombergGPT, Gemini, Claude).

В работе [23] представлен набор открытых моделей Open-FinLLMs, проходящих оценку по 14 финансовым задачам из 30 наборов данных. Семейство включает модель FinLLaMA на архитектуре LLaMA3 (дообучена на 52 млрд. токенов), диалоговую модель FinLLaMA-Instruct (настроена на 573 тыс. инструкций) и мультимодальную FinLLaVA (1,43 млн. пар «изображение-текст»), для которых приведены значения метрик (FinLLaMA, FinLLaMA-Instruct, FinLLaVA, LLaMA3, LLaMA3.1).

В работе [24] представлена модель Fin-R1, специально созданная для финансовых рассуждений на основе Qwen2.5-7B-Instruct с использованием обучения с подкреплением. Вместе с моделью был разработан набор данных Fin-R1-Data, содержащий около 60 091 полных цепочек мыслей (Chain-of-Thought) для различных сценариев. Представлены метрики на 5 наборах данных для 10 моделей (основные: Fin-R1, Qwen2.5-7B-Instruct, DeepSeek-R1).

В работе [25] представлена модель KAI-GPT, позиционируемая как первая LLM для банковской отрасли, созданная на базе Pythia-Chat-Base-7B. Для оценки были созданы два новых набора данных: KAI-GPT Evaluation Set #1 (вопросы экспертов с ответами с сайта банка) и KAI-GPT Evaluation Set #2 (краудсорсинг вопросов с форума). Представлены метрики для двух моделей (KAI-GPT, Pythia-Chat-Base-7B). Данное исследование включено в рассмотрение, так как закрытость моделей и данных позволяет посмотреть на работу матричной факторизации на таком типе данных.

Некоторые работы (как, например, Fino1 [26]) не удалось включить в данное исследование, так как значения метрик моделей на наборах данных не совпадают со значениями ни одной другой статьи, что делает невозможным сопоставление. Дополнительно, добавление

новых данных в матрицу увеличило бы ее разреженность еще сильнее, превысив порог, при котором резко падает качество, как показано далее.

3. Постановка задачи. Метод матричной факторизации используется при построении рекомендательных систем [16] и позволяет прогнозировать неизвестные рейтинги товаров/предметов для пользователей. Если в процессе построения матричной факторизации рассматривать наборы данных в качестве «пользователей», а LLM – в качестве «товаров», то данный метод позволит определить неизвестные оценки пользователей-наборов данных по отношению к товарам-моделям. В данной постановке задачи, пользователи-наборы данных будут отдавать предпочтение лучшим товарам-моделям и поэтому ставить им большие оценки-метрики. Метод позволяет учесть скрытые факторы, соответствующие моделям и наборам данных, и с их помощью определять неизвестные рейтинги. Его основа – разложение матрицы метрик $R^{N_u \times N_i}$ на произведение двух матриц $H^{N_u \times N_f}$ и $W^{N_f \times N_i}$:

$$R = HW,$$

где N_u – количество наборов данных, N_i – количество LLM, и N_f – количество скрытых факторов. Прогнозируемая в таком случае оценка $r'_{u,i}$ для набора данных u и модели i рассчитывается следующим образом [16]:

$$r'_{u,i} = \sum_{f=0}^{N_f} H_{u,f} W_{f,i}.$$

Подобная оценка позволяет учесть явные числовые метрики, однако для дополнительного учета неявных взаимодействий используется усовершенствованная формула расчета прогнозируемых значений, которая и будет впоследствии использована для предсказания значений метрик моделей [16]:

$$r'_{u,i} = \mu + b_i + b_u + \sum_{f=0}^{N_f} H_{u,f} W_{f,i}, \quad (1)$$

где μ, b_i и b_u – соответственно общая средняя метрика, а также наблюдаемые отклонения, относящиеся к модели i или набору данных u. В результате, чтобы определить параметры H, W, все b_i и b_u (обозначенные как b_*), система минимизирует регуляризованную квадратичную ошибку:

$$min_{H,W,b_*}\left[\sum_{(u,i)\in K}\left(r_{u,i}-r'_{u,i}\right)^2+\lambda\left(b_u^2+b_i^2+\left|\left|H_{u,*}\right|\right|^2+\left|\left|W_{*,i}\right|\right|^2\right)\right], \quad (2)$$

где $r_{u,i}$ – действительная метрика для набора данных u и модели i, K – набор (u,i), для которых известна $r_{u,i}$ (обучающий набор), $H_{u,*}$ и $W_{*,i}$ – векторы, соответствующие набору данных u и LLM i соответственно, $||\cdot||$ – норма (например, Евклидова), λ – параметр регуляризации [16].

4. Эксперимент. В данном разделе представлено детальное описание эксперимента по предсказанию неизвестных значений метрик больших языковых моделей в финансовой области с использованием метода матричной факторизации. Первым шагом является сбор данных о метриках из различных научных работ, за которым следует их подготовка: фильтрация и формирование итоговой матрицы «модель-набор данных». Далее проводится предварительный анализ влияния разреженности данных на точность прогнозов. Основная часть посвящена непосредственному применению матричной факторизации, включая подбор гиперпараметров для минимизации ошибки прогнозирования. Вводится специальная агрегированная метрика для обеспечения возможности комплексного ранжирования моделей. Для оценки надежности полученных результатов и итогового рейтинга проводится анализ неопределенности прогнозов с использованием двух различных подходов, включая разброс по MAE и метод Монте-Карло.

Подготовка данных. На основе перечисленных в разделе **2** исследований были собраны метрики по 42 наборам данных. Их можно разбить на 9 категорий в зависимости от типа задачи. Предоставим краткое описание общих категорий использованных наборов данных.

Извлечение информации. Категория, включающая наборы данных, предназначенных для извлечения структурированной информации из неструктурированного текста. Наборы данных: FiNER-ORD [11], FinRED [27], REFinD [28], FNXL (Financial Numeric Extreme Labeling) [29], Headlines [30], Mergers and Acquisition (M&A) [30].

Ответы на вопросы. Категория наборов данных, проверяющих способности моделей предоставлять ответы на вопросы, иногда по заданному контексту. Наборы данных: SubjECTive-QA [32], FinQA [10], TAT-QA (Tabular and Textual Question Answering) [33], ConvFinQA (Conversational Finance Question Answering) [34], RegulationsQA [23], Evaluation Set №1 [25], Evaluation Set №2 [25].

Анализ тональности. Категория, содержащая наборы данных, оценивающих способности моделей определять эмоциональную окраску

текста. Наборы данных: FinEntity [12], Financial Phrase Bank (FPB) [35], Targeted Sentiment Analysis (TSA) [36], The Twitter Financial News (TFNS) [37].

Классификация. Группа наборов данных, предназначенных для классификации текстов по заранее заданным категориям. Наборы данных: Banking77 [38], FinCausal-SC [39], Numerical Claim Detection Dataset (NC) [40], FinArg-AUC [41].

Обнаружение мошенничества. Категория, объединяющая наборы данных, направленных на выявление мошеннических операций в финансовых транзакциях и страховых исках. Наборы данных: PortoSeguro [6], travelinsurance [6], Credit Card Fraud (ccf) [6], ccFraud [6].

Суммаризация. Категория наборов данных, предназначенных для создания кратких изложений длинных финансовых текстов и документов. Наборы данных: ECTSum [13], EDTSum [14].

Работа с числами. Категория наборов данных, проверяющих способности моделей производить числовые рассуждения и решать математические задачи в финансовом контексте. Наборы данных: KnowledgeMath [42], DocMath-Eval [43], MC [23].

Прогнозирование, скоринг и оценка. Категория, содержащая наборы данных для проверки возможностей LLM предсказывать различные финансовые события, такие как дефолт, банкротство, финансовые затруднения. Наборы данных: FinBench [44], German (German Credit Data) [45], Australian (Australian Credit Approval) [46], LendingClub [6], polish (Polish Companies Bankruptcy) [6], taiwan (Taiwan Economic Journal dataset) [6].

Общая оценка моделей. Категория, объединяющая наборы данных для оценки общих знаний моделей и способностей в финансовой сфере. Наборы данных: Finance-Instruct-500K [47], Abbreviation [23], Ant_Finance [48], Federal Open Market Committee [49].

Из упомянутых работ были собраны данные значений метрик моделей на наборах данных. Не рассматривались метрики на обобщенных категориях задач (объединяющих только в рамках данной статьи несколько наборов данных по общей тематике), как, например, данные в таблице 7 статьи Open-FinLLMs. В работе использовалась min-max нормализация, и нормализованное значение метрики X составляло $\frac{X'-X_{min}}{X_{max}-X_{min}}$, где X' – первоначальное значение метрики, X_{max} и X_{min} – максимальное и минимальное значения метрики соответственно. Так как далее ко всем значениям применялась эта нормализация, то не рассматривались сеты, на которых считались метрики, не имеющие четкого диапазона возможных значений (так, например, не рассматривалась MSE). У наборов данных, имеющих

более одной подсчитанной метрики, брались те значения, которые соотносятся с метриками того же набора данных в других исследованиях. В отсутствие таковых использовались в первую очередь метрики с возможным диапазоном значений от 0 до 1.

В итоге была получена матрица, состоящая из 34 моделей и 42 наборов данных, с 531 известными значениями метрик (37.18% заполненности), представленная в таблицах, размещенных на внешнем ресурсе[1].

Анализ влияния разреженности. Сперва была протестирована устойчивость матричной факторизации к разреженности матрицы. Для этого на основе полностью заполненной таблицы из бенчмарка FLaME были сформированы 1000 урезанных таблиц (случайный выбор 80% столбцов и 80% строк). После чего из каждой новой матрицы случайным образом были выброшены от 10% до 90% значений (они использовались в качестве тестовых данных). В качестве гиперпараметров были взяты factors=35, learning rate = 0.05, regularization = 0.01, epochs = 200 (гиперпараметры, соотносящиеся с дальнейшими экспериментами). На рисунке 1 представлены 95% доверительные интервалы и средняя ошибка $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, где n – размер выборки, y_i и $\hat{y}_i$ – реальные и предсказанные метрики.

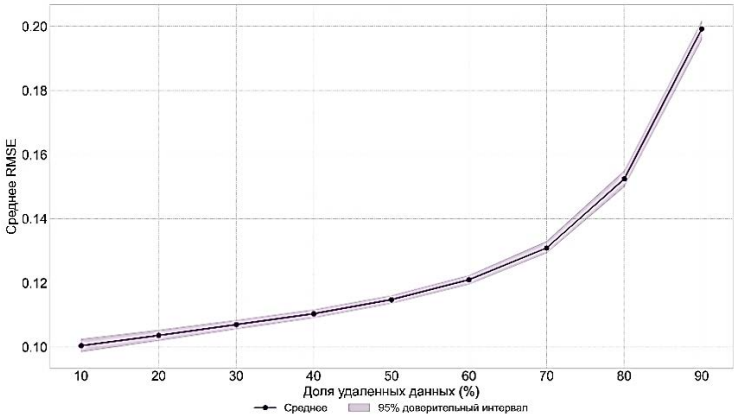

Рис. 1. RMSE в зависимости от процента пропущенных значений

[1] https://github.com/careepy/FinMF

Хорошо видно, что резкое ухудшение в качестве начинается в районе 70% пропущенных значений (30% заполненность матрицы). Разреженность составленной для экспериментов матрицы соотносится с данными результатами (37.18%).

Гиперпараметры. Эксперимент по заполнению пропущенных значений матрицы предсказанными метриками на основе матричной факторизации проводился с использованием пакета matrix-factorization[2]. Чтобы найти требуемые значения из уравнения 1, этот пакет использует метод стохастического градиентного спуска [50] и минимизирует уравнение 2 [16].

Изначально проводился поиск подходящих гиперпараметров, применяя поиск по сетке по широкому диапазону значений. Так, были рассмотрены factors: 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50; learning rate: 0.5, 0.1, 0.05, 0.01, 0.005, 0.001; regularization: 0.5, 0.1, 0.05, 0.01, 0.005, 0.001. Считалось среднее значение метрики RMSE по всем комбинациям с использованием пятикратной 5-блочной кросс-валидации (RepeatedKFold [51]). Результаты поиска показали, что наименьшая средняя ошибка достигается при значениях параметра learning rate 0.05, regularization 0.01 и числа скрытых факторов 20 (рисунок 2). Для данных параметров определялось оптимальное количество эпох — около 200, на основе графика динамики ошибки (рисунок 3). С использованием описанных выше гиперпараметров были получены итоговые предсказания R_{res}, которые представлены в таблицах, размещенных на внешнем ресурсе[3], в которых также приведены значения ошибок для показателей, которые были известны изначально.

Ошибки были рассчитаны как разница между фактическими и предсказанными измерениями. Помимо этого, также фиксируется значение $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$, (где n – размер выборки, y_i и $\widehat{y_i}$ – реальные и предсказанные метрики) итоговой модели (0.07), так как оно будет использовано при оценке агрегированной метрики. Стоит отметить, что, хотя в результирующей таблице этого не наблюдалось, такой подход может привести к тому, что предсказанные показатели превысят 1. Важно отметить, что этот подход позволяет найти только приблизительные показатели и показывает сложные, но относительные различия в результирующих оценках.

[2] https://github.com/Quang-Vinh/matrix-factorization
[3] https://github.com/careepy/FinMF

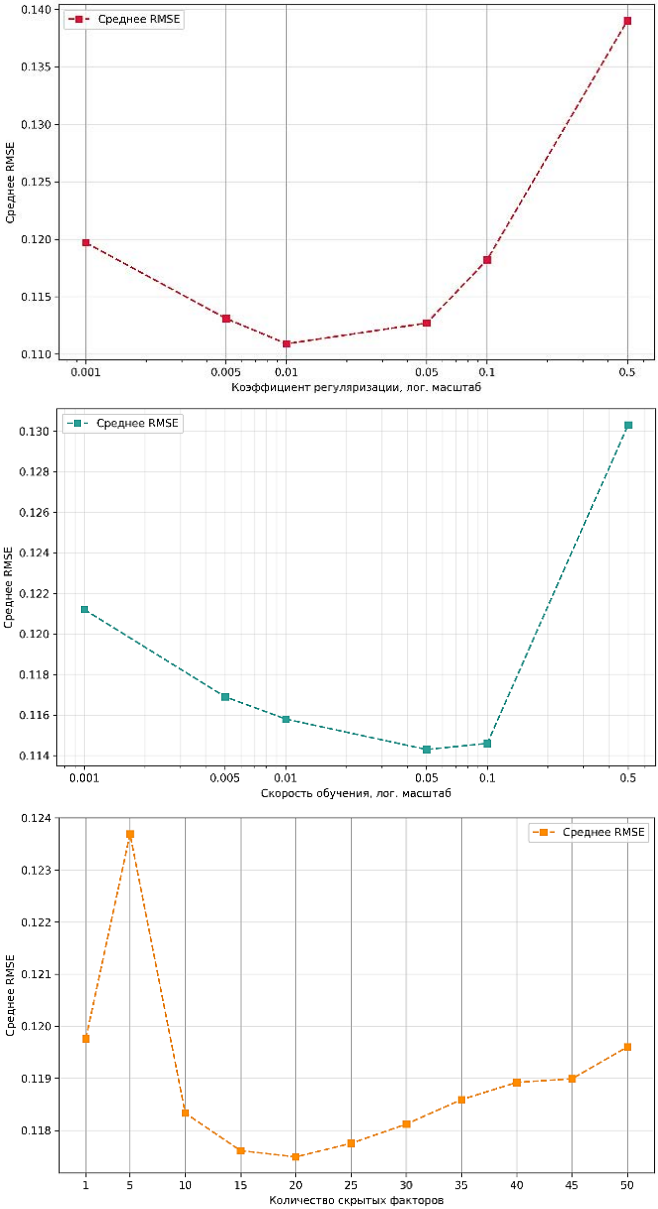

Рис. 2. Средняя RMSE для гиперпараметров регуляризации, скорости обучения и числа скрытых факторов

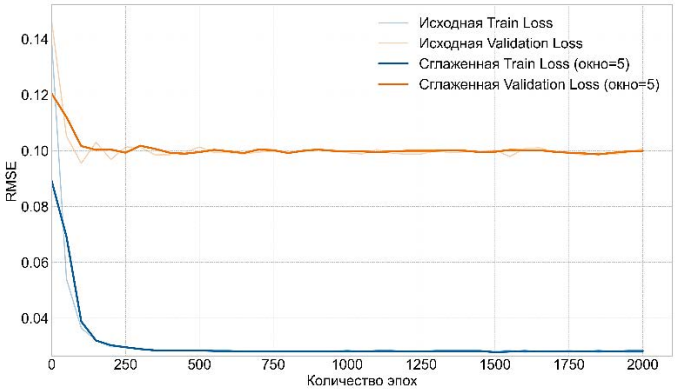

Рис. 3. Зависимость RMSE от количества эпох

Агрегированная метрика LLM. Составленная таблица с оценками значений метрик не позволит напрямую сравнивать и ранжировать модели. Для этого предлагается ввести агрегирующей метрику A, которая для каждой модели m суммирует нормализованные значения метрик по всем наборам данных:

$$A_m = \sum_{d \in D} A_m^d = \sum_{d \in D} \frac{d_m - d_{min}}{d_{max} - d_{min}},$$

где D – наборы данных, d – один выбранный набор данных, d_m – значение метрики модели на наборе данных d, d_{min} и d_{max} – минимальные и максимальные полученные по всем моделям значения метрик на наборе данных d. Итоговые отранжированные значения A_m, полученные по R_{res}, представлены в таблице 1.

Важно также рассмотреть, насколько значение метрики A зависит от ошибок предсказаний и оценить возможные значения метрики A при различных предположениях. В данной работе будет предложено два метода: с использованием вариации на ошибку MAE и с использованием метода Монте-Карло.

Первый метод заключается в том, что рассматриваются максимальные и минимальные значения A_m при вариации предсказаний на MAE, вычисленное при кросс-валидации (0.07). MAE было выбрано, потому что оно дает косвенную оценку потенциальной ошибки. Для этого:

– для каждого набора данных d вычисляются все возможные d_{min} и d_{max} при различных вариациях предсказаний;

– вычисляются максимальные и минимальные возможные A_m^d при различных d_{min} и d_{max} и различных вариациях d_m;

– верхняя граница A_m – сумма максимальных A_m^d, нижняя – сумма минимальных.

Часть полученных таким образом интервалов указана на рисунке 4, на котором представлены оценки для некоторых моделей, полученные при использовании метода Монте-Карло и MAE. По вертикали отложена предсказанная оценка A_m, по горизонтали – LLM (числовые значения указывают на положение в итоговом рейтинге).

Для подтверждения полученных выше результатов предложим еще один способ, использующий метод Монте-Карло. Он заключается в вычислении различных A_m при вариациях d_m на случайные значения из распределения ошибок:

– Оценивается распределение ошибок. Для этого генерируется N^{sp} разбиений на обучающие (80%) и тестовые (20%) выборки и на тестовых выборках вычисляются ошибки предсказаний.

– Генерируется N^{tb} таблиц, где каждая ячейка — сумма соответствующей ячейки R_{res} и случайной ошибки из распределения, определенного на шаге 1.

– Для каждой таблицы t из шага 2 и модели m вычисляется A_m^t.

– Для каждой m вычисляется среднее по A_M^* и определяется доверительный интервал (5 и 95 перцентили).

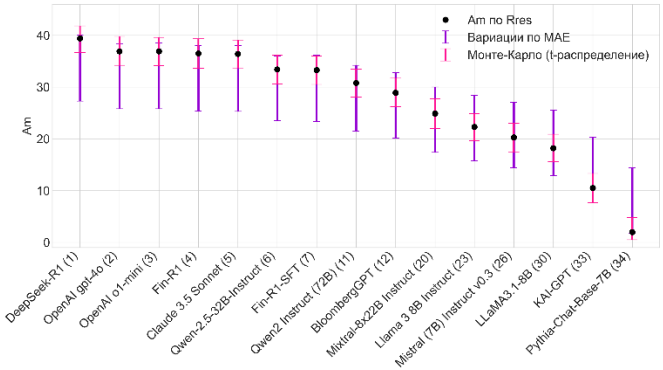

Рис. 4. Интервалы, построенные с помощью MAE и метода Монте Карло для некоторых моделей

Для расчета было взято $N^{tb} = N^{sp} = 1000$. На рисунке 5 представлено полученное эмпирическое распределение ошибок и аппроксимирующее его распределение Стьюдента со степенями свободы = 4.71 и матожиданием = –0.002. По оси X отложены значения ошибок, а по оси Y – плотность вероятности. Без учета выбросов стандартное отклонение составляет 0.063. Таким образом, случайные ошибки были взяты из распределения Стьюдента с озвученными параметрами. Итоговые A_m и доверительные интервалы для части моделей показаны на рисунке 4. Полные полученные значения можно найти в таблице 1.

Таблица 1. Значения метрики A и границ интервалов MAE и Монте-Карло

№	Модель	A	A_{min}^{MAE}	A_{max}^{MAE}	A_{min}^{MC}	A_{max}^{MC}
1	DeepSeek R1	39,9	27,6	40,5	37,1	42,7
2	OpenAI gpt-4o	37,0	25,6	38,5	34,3	39,8
3	OpenAI o1-mini	36,9	25,7	38,5	34,1	39,6
4	Fin-R1	36,6	25,4	38,2	33,7	39,5
5	Claude 3,5 Sonnet	36,5	25,2	38,1	33,7	39,1
6	Qwen-2,5-32B-Instruct	33,3	23,2	36,1	30,5	36,1
7	Fin-R1-SFT	33,2	23,2	36,0	30,3	36,0
8	DeepSeek-V3	33,1	22,9	35,7	30,5	35,7
9	Qwen-2,5-14B-Instruct	32,1	22,3	35,2	29,3	35,1
10	Llama 3 70B Instruct	31,3	21,8	34,6	28,5	34,1
11	Qwen 2 Instruct (72B)	30,9	21,4	34,3	28,1	33,5
12	Google Gemini 1,5 Pro	29,1	20,1	33,0	26,3	31,8
13	BloombergGPT	29,0	20,0	32,8	26,3	31,7
14	Gemma 2 27B	28,6	19,8	32,6	25,8	31,3
15	Gemma 2 9B	26,6	18,5	31,4	23,8	29,5
16	Jamba 1,5 Large	26,4	18,4	31,2	23,6	29,1
17	Claude 3 Haiku	26,3	18,4	31,2	23,5	29,0
18	Cohere Command R +	26,1	18,0	30,8	23,3	28,9
19	WizardLM-2 8x22B	25,1	17,4	30,2	22,3	27,8
20	Mixtral-8x22B Instruct	24,4	16,9	29,8	21,5	27,2
21	Qwen-2,5-7B-Instruct	24,1	17,0	29,8	21,3	27,0
22	Cohere Command R 7B	24,0	16,7	29,5	21,1	26,6
23	QwQ-32B-Preview	22,2	15,4	28,3	19,6	24,8
24	Llama 3 8B Instruct	21,9	15,3	28,2	19,2	24,5
25	Mixtral-8x7B Instruct	21,8	15,3	28,1	19,2	24,6
26	Mistral (7B) Instruct v0,3	20,0	14,0	26,8	17,0	22,7
27	DeepSeek LLM (67B)	19,6	13,7	26,5	16,7	22,3
28	FinLLaMA	18,7	13,2	26,0	15,7	21,4
29	Jamba 1,5 Mini	18,1	12,6	25,5	15,2	20,7
30	LLaMA3,1-8B	18,0	12,6	25,5	15,5	20,7
31	DBRX Instruct	12,9	9,0	21,9	10,2	15,6
32	LLaMA3-8B	11,5	8,1	21,0	8,8	14,3
33	KAI-GPT	10,2	7,3	20,1	7,4	12,9
34	Pythia-Chat-Base-7B	1,5	1,2	14,0	0,0	4,3

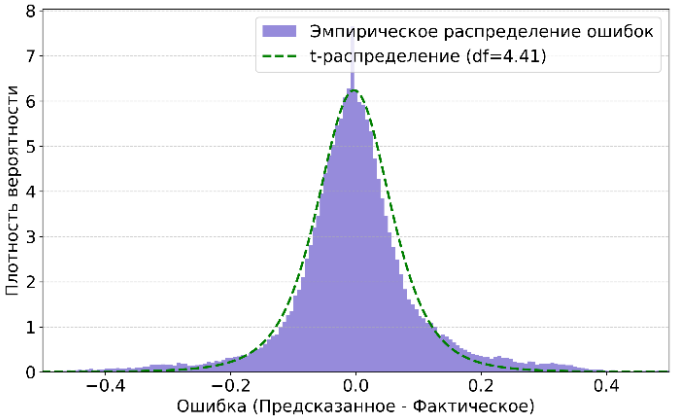

Рис. 5. Эмпирическое распределение ошибок и распределение Стьюдента

5. Обсуждение. Данный раздел посвящен обсуждению и интерпретации результатов, полученных в ходе эксперимента по применению матричной факторизации. Обсуждение разбито на две части: анализ итогового рейтинга моделей и анализ ошибок.

Анализ результатов. В данном подразделе представлен анализ результатов, полученных с использованием метода матричной факторизации для набора больших языковых моделей. Оценка качества на основе агрегированной метрики A, а также её оценок, полученных с помощью вариации по MAE и метода Монте-Карло.

Итоговый рейтинг моделей (таблица 1 и рисунок 4) показывает следующую тенденцию: оценка качества ведущих больших языковых моделей сблизилась до такой степени, что объявить однозначного лидера становится затруднительно. Разница в итоговых метриках между моделями из верхней части рейтинга, представленного в таблице 1 (DeepSeek R1, OpenAI gpt-4o, OpenAI o1-mini, Fin-R1, Claude 3.5 Sonnet), оказывается в пределах погрешности. Хоть DeepSeek R1 имеет небольшой отрыв от всех остальных моделей, существующий отрыв является минимальным. Так, если посмотреть на группу моделей, включающую OpenAI gpt-4o (36.96), OpenAI o1-mini (36.85), Fin-R1 (36.63) и Claude 3.5 Sonnet (36.47), то видно, что их агрегированные показатели A отличаются на десятые доли балла.

Ключевым аспектом является пересечение их доверительных интервалов, рассчитанных методом Монте-Карло (A^{MC}_{min} и A^{MC}_{max} в таблице 1). Например, диапазон оценки качества для gpt-4o

(34.27-39.85), o1-mini (34.10-39.59), Fin-R1 (33.74-39.48) и Claude 3.5 Sonnet (33.73-39.09) в значительной степени совпадает. Это говорит о том, что при повторных тестах или на несколько ином наборе данных их порядок в рейтинге мог бы легко измениться.

Однако здесь стоит особенно выделить модель Fin-R1, имеющую в своей архитектуре всего 7 миллиардов параметров, и находящуюся на первых строчках рейтинга, в то время как другие модели того же диапазона оценки качества имеют сотни миллионов параметров.

Невысокие результаты получили модели KAI-GPT и Pythia-Chat-Base-7B. В исходной таблице эти модели представляют собой особый случай в рамках данного исследования. Исходные метрики для них были предоставлены исключительно для двух наборов данных (KAI-GPT Eval. Set №1 и KAI-GPT Eval. Set №2), в которых не оценивалась ни одна другая модель из выборки. Несмотря на это ограничение, итоговый рейтинг, сгенерированный методом, последовательно размещает KAI-GPT и Pythia-Chat-Base-7B на низких позициях в общем зачете. Однако, предсказанные значения на остальных 30+ наборах данных скорее являются результатом обобщения модели на усредненных данных всей матрицы, нежели реальной оценкой их метрик.

Анализ ошибок предсказания матричной факторизации. Средняя абсолютная ошибка предсказания по всем известным значениям составляет 0.024 – данный показатель свидетельствует о том, что модель успешно справилась с описанием исходных данных через латентные представления. Средняя абсолютная ошибка для каждой отдельной модели также не превышает 0.033, по наборам данных – 0.04.

Дополнительно метод помогает выявить специфические особенности моделей на конкретных задачах, когда ошибки предсказания аномально высоки или низки. Так, ошибок со значением по модулю, большим 0.1, всего 5: Google Gemini 1,5 Pro – Banking77 (0.158), DeepSeek LLM – TAT-QA (0.151), Gemma 2 9B – ECTSum (0.129), DBRX Instruct – Federal Open Market Committee (0.113), QwQ-32B-Preview – REFinD (0.1).

Самая высокая полученная ошибка (0.158) принадлежит модели Google Gemini 1,5 Pro на наборе данных Banking77. Такой результат может объясняться тем, что в итоговом рейтинге Gemini 1,5 Pro находится на 12 месте, но в исходной матрице значение метрики для данной модели на наборе данных Banking77 ниже всех остальных на 0.09 пунктов. Это выявляет специфическую особенность модели (возможно, в многоклассовой классификации намерений в узкой

банковской тематике), которую не удалось уловить обобщенным латентным представлениям.

DeepSeek LLM продемонстрировал самые низкие метрики на TAT-QA (разница от 0.19), задаче, связанной с анализом табличных данных в финансовой области. Gemma 2 9B показала невысокие результаты на наборе данных ECTSum, связанном с суммаризацией длинных стенограмм, отчетов об убытках и прибылях компаний. Высокая метрика ошибки на DBRX Instruct, возможно, связана с тем, что модель в целом демонстрирует невысокие показатели на финансовых задачах, в особенности на наборе данных FOMC (на 0.2 ниже ближайшей модели). QwQ-32B-Preview демонстрирует результаты сильно ниже среднего на REFinD (0.2 пункта) – задаче, связанной с извлечением связей в финансовых и юридических документах.

Все полученные ошибки представлены в таблицах, размещенных на внешнем ресурсе[4].

6. Заключение. В данной работе была оценена применимость матричной факторизации для предсказания неизвестных метрик качества LLM на финансовых наборах данных. Для предсказанных значений была представлена агрегированная метрика, позволяющая ранжировать модели. Исследование, охватившее 42 набора данных и 34 модели, продемонстрировало потенциальную применимость использования матричной факторизации в качестве инструмента для предсказания оценки качества LLM. Предложенный метод может способствовать выбору моделей для решения финансовых задач.

Проведенный эксперимент показал, что оценка качества ведущих LLM, таких как DeepSeek R1, OpenAI gpt-4o, OpenAI o1-mini, Fin-R1 и Claude 3.5 Sonnet, находится приблизительно на одном уровне. Различия их итоговых показателей настолько малы, что попадают в пределы погрешности, и не позволяют выделить явного лидера. Данный вывод был подкреплен анализом доверительных интервалов, рассчитанных с помощью метода Монте-Карло.

Однако стоит подчеркнуть, что данный эксперимент показывает только общую эмпирическую картину. Некоторые значения метрик могут не отображать реальные метрики, полученные на новых наборах данных. Полученные ошибки могут быть вызваны разными использованными промптами, нюансами в тестовых данных, использованием разных языков, погрешностями факторизации. Анализ ошибок позволяет выявить чрезмерно хороший/плохой результат

[4] https://github.com/careepy/FinMF

на известных данных, но не позволяет предсказать подобное на новых данных.

Также, хоть матричная факторизация в целом подходит для решения задачи заполнения неизвестных значений метрик, следует проверять ее применимость при каждом новом эксперименте.

Дальнейшие исследования могут быть направлены на использование более сложных методов предсказания неизвестных значений метрик, в том числе нелинейных.

Литература

1. Ali C.-S.-M., Mahmood I. A Comprehensive Survey on Large Language Models: Architectures, Applications, and Ethical Considerations // Engineering and Technology Journal. 2025. vol. 10. no. 04. pp. 4578–4593. DOI: 10.47191/etj/v10i04.26.
2. Zhao H., Liu Z., Wu Z., Li Y., Yang T., Shu P., Xu S., Dai H., Zhao L., Mai G., Liu N., Liu T. Revolutionizing finance with LLMs: An overview of applications and insights // arXiv preprint arXiv:2401.11641. 2024.
3. Dhake S.-P., Lassi L., Hippalgaonkar A., Gaidhani R.-A., Jyothi N.-M. Impacts and Implications of Generative AI and Large Language Models: Redefining Banking Sector // Journal of Informatics Education and Research. 2024. vol. 4. no. 2. pp. 248–257. DOI: 10.52783/jier.v4i2.767.
4. Zhao W.-X., Liu J., Ren R., Wen J.-R. Dense text retrieval based on pretrained language models: A survey // ACM Transactions on Information Systems. 2024. vol. 42. no. 4. pp. 1–60. DOI: 10.1145/3637870.
5. Luo B., Zhang Z., Wang Q., Ke A., Lu S., He B. AI-powered fraud detection in decentralized finance: A project life cycle perspective // arXiv preprint arXiv:2308.15992. 2023.
6. Feng D., Dai Y., Huang J., Zhang Y., Xie Q., Han W., Lopez-Lira A., Wang H. Empowering many, biasing a few: Generalist credit scoring through large language models // arXiv preprint arXiv:2310.00566. 2023.
7. Dong Y., Yan D., Almudaifer A.-I., Yan S., Jiang Z., Zhou Y. BELT: A pipeline for stock price prediction using news. IEEE International Conference on Big Data // IEEE. 2020. pp. 1137–1146. DOI: 10.1109/BigData50022.2020.9378345.
8. Zhao W.-X., Zhou K., Li J., et al. A Survey of Large Language Models // arXiv preprint arXiv:2303.18223. 2023.
9. Li Y., Wang S., Ding H., Chen H. Large language models in finance: A survey // Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF '23). 2023. pp. 374–382. DOI: 10.1145/3604237.3626869.
10. Chen Z., Chen W., Smiley C., et al. FinQA: A Dataset of Numerical Reasoning over Financial Data // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2021. pp. 3697–3711. DOI: 10.18653/v1/2021.emnlp-main.300.
11. Shah A., Gullapalli A., Vithani R., Galarnyk M., Chava S. FiNER-ORD: Financial Named Entity Recognition Open Research Dataset // arXiv preprint arXiv:2302.11157. 2023.
12. Tang Y., Yang Y., Huang A., Tam A., Tang J. FinEntity: Entity-level Sentiment Classification for Financial Texts // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2023. pp. 15465–15471. DOI: 10.18653/v1/2023.emnlp-main.956.
13. Mukherjee R., Bohra A., Banerjee A., et al. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts // Proceedings of the

Conference on Empirical Methods in Natural Language Processing. 2022. pp. 10893–10906. DOI: 10.18653/v1/2022.emnlp-main.748.
14. Xie Q., Han W., Chen Z., et al. The FinBen: An Holistic Financial Benchmark for Large Language Models // arXiv preprint arXiv:2402.12659. 2024.
15. Chang Y., Wang X., Wang J., et al. A Survey on Evaluation of Large Language Models // ACM Transactions on Intelligent Systems and Technology. 2024. vol. 15. no. 3. pp. 1–45. DOI: 10.1145/3641289.
16. Koren Y., Bell R., Volinsky C. Matrix Factorization Techniques for Recommender Systems // Computer. 2009. vol. 42. no. 8. pp. 30–37. DOI: 10.1109/MC.2009.263.
17. Zhao Q., Xu M., Gupta K., et al. Can We Predict Performance of Large Models across Vision-Language Tasks // arXiv preprint arXiv:2410.10112. 2024.
18. Zhang Q., Lyu F., Liu X., Ma C. Collaborative Performance Prediction for Large Language Models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024. pp. 2576–2596. DOI: 10.18653/v1/2024.emnlp-main.150.
19. Zhong X.-X., Yi C., Ye H.-J. Efficient Evaluation of Large Language Models via Collaborative Filtering // arXiv preprint arXiv:2504.08781. 2025.
20. Laskar M.-T.-R., Alqahtani S., Bari M.-S., et al. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024. pp. 13785–13816. DOI: 10.18653/v1/2024.emnlp-main.764.
21. Owen D. How predictable is language model benchmark performance // arXiv preprint arXiv:2401.04757. 2024.
22. Matlin G., Okamoto M., Pardawala H., Yang Y., Chava S. Finance Language Model Evaluation (FLaME) // Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics. 2025. pp. 880–926. DOI: 10.48550/arXiv.2506.15846.
23. Huang J., Xiao M., Li D., et al. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications // arXiv preprint arXiv:2408.11878. 2024.
24. Liu Z., Guo X., Lou F., et al. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning // arXiv preprint arXiv:2503.16252. 2025.
25. KAI-GPT: The First Large Language Model Purpose-Built for Banking // Kasisto. URL: https//kasisto.com/blog/kai-gpt-the-first-large-language-model-purpose-built-for-banking/ (дата обращения: 25.11.2025).
26. Qian L., Zhou W., Wang Y., Peng X., Huang J., Xie Q. Fino1: On the Transferability of Reasoning-Enhanced LLMs and Reinforcement Learning to Finance // arXiv preprint arXiv:2502.08127. 2025.
27. Sharma S., Nayak T., Bose A., et al. FinRED: A Dataset for Relation Extraction in Financial Domain // Companion Proceedings of the Web Conference (WWW '22). 2022. pp. 595–597. DOI: 10.1145/3487553.3524637.
28. Kaur S., Smiley C., Gupta A., et al. REFinD: Relation Extraction Financial Dataset // Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023. pp. 3054–3063. DOI: 10.1145/3539618.3591911.
29. Sharma S., Khatuya S., Hegde M., et al. Financial Numeric Extreme Labelling: A dataset and benchmarking // Findings of the Association for Computational Linguistics: ACL 2023. 2023. pp. 2933–2946. DOI: 10.18653/v1/2023.findings-acl.219.
30. Sinha A., Khandait T. Impact of News on the Commodity Market: Dataset and Results // Advances in Information Retrieval (ECIR 2021). 2021. pp. 589–601. DOI: 10.1007/978-3-030-73103-8_41.
31. Yang L., Kenny E., Ng T.-L., Yang Y., Smyth B., Dong R. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification // Proceedings of the 28th International Conference on Computational Linguistics. 2020. pp. 6150–6160. DOI: 10.18653/v1/2020.coling-main.541.

32. Pardawala H., Sukhani S., Shah A., et al. SubjECTive-QA: Measuring Subjectivity in Earnings Call Transcripts' QA Through Six-Dimensional Feature Analysis // Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24). 2024. pp. 59342–59372.
33. Zhu F., Lei W., Chao Y., et al. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021. pp. 3277–3287. DOI: 10.18653/v1/2021.acl-long.254.
34. Chen Z., Li S., Smiley C., et al. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering // Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2022. pp. 6279–6290. DOI: 10.18653/v1/2022.emnlp-main.421.
35. Malo P., Sinha A., Korhonen P., Wallenius J., Takala P. Good debt or bad debt: Detecting semantic orientations in economic texts // Journal of the Association for Information Science and Technology. 2014. vol. 65. no. 4. pp. 782–796. DOI: 10.1002/asi.23062.
36. Cortis K., Freitas A., Daudert T., et al. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News Headlines // Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017. pp. 519–535. DOI: 10.18653/v1/S17-2089.
37. Twitter Financial News Sentiment // Hugging Face: Zeroshot. 2024. URL: https//huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment (дата обращения: 25.11.2025).
38. Casanueva I., Temcinas T., Gerz D., Henderson M., Vulic I. Efficient Intent Detection with Dual Sentence Encoders // Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. 2020. pp. 38–45. DOI: 10.18653/v1/2020.nlp4convai-1.5.
39. Mariko D., Abi-Akl H., Labidurie E., et al. The Financial Document Causality Detection Shared Task (FinCausal 2020) // Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020. pp. 23–32. DOI:10.48550/arXiv.2012.02505
40. Shah A., Hiray A., Shah P., et al. Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis // Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). 2024. pp. 170–185. DOI: 10.18653/v1/2024.fever-1.21.
41. Chen C.-C., Lin C.-Y., Chiu C.-J., et al. Overview of the NTCIR-17 FinArg-1 Task: Fine-grained argument understanding in financial analysis // Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies. 2023. pp. 16–20. DOI: 10.20736/0002001323.
42. Zhao Y., Liu H., Long Y., et al. FinanceMATH: Knowledge-Intensive Math Reasoning in Finance Domains // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. vol. 1. pp. 12841–12858. DOI: 10.18653/v1/2024.acl-long.693.
43. Zhao Y., Long Y., Liu H., et al. DocMath-Eval: Evaluating Math Reasoning Capabilities of LLMs in Understanding Long and Specialized Documents // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. vol. 1. pp. 16103–16120. DOI: 10.18653/v1/2024.acl-long.852.
44. Yin Y., Yang Y., Yang J., Liu Q. FinPT: Financial Risk Prediction with Profile Tuning on Pretrained Foundation Models // arXiv preprint arXiv:2308.00065. 2024.
45. Hofmann H. Statlog (German Credit Data) // UCI Machine Learning Repository. 1994. URL: https//archive.ics.uci.edu/dataset/144/statlog+german+credit+data (дата обращения: 25.11.2025).

46. Quinlan R. Statlog (Australian Credit Approval) // UCI Machine Learning Repository. URL: https//archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval (дата обращения: 25.11.2025).
47. Flowers J.G. Finance Instruct 500k // Hugging Face. 2025. URL: https//huggingface.co/datasets/Josephgflowers/Finance-Instruct-500k (дата обращения: 25.11.2025).
48. Financial Evaluation Dataset // GitHub: Alipay Team. 2023. URL: https//github.com/alipay/financial_evaluation_dataset (дата обращения: 25.11.2025).
49. Shah A., Paturi S., Chava S. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. vol. 1. pp. 6664–6679. DOI: 10.18653/v1/2023.acl-long.368.
50. Bottou L. Large-Scale Machine Learning with Stochastic Gradient Descent // Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT '2010). Physica-Verlag HD. 2010. pp. 177–186. DOI: 10.1007/978-3-7908-2604-3_16.
51. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95). 1995. vol. 2. pp. 1137–1143.

Вяткин Артем Андреевич — младший научный сотрудник, лаборатория прикладного искусственного интеллекта, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: вероятностные графические модели, алгебраические байесовские сети, нечеткие вычисления, большие языковые модели. Число научных публикаций — 24. aav@dscs.pro; 14-линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

Попцов Александр Владимирович — стажер-исследователь, лаборатория прикладного искусственного интеллекта, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: обработка естественного языка, большие языковые модели, ИИ-агенты. Число научных публикаций — 2. avp@dscs.pro; 14-линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(927)887-1617.

Олисеенко Валерий Дмитриевич — научный сотрудник, лаборатория прикладного искусственного интеллекта, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: обработка естественных языков, глубокое обучение, анализ социальных сетей, информационная безопасность, социоинженерные атаки. Число научных публикаций — 60. vdo@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

Абрамов Максим Викторович — канд. техн. наук, доцент, руководитель лаборатории, лаборатория прикладного искусственного интеллекта, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: генеративный искусственный интеллект, большие языковые модели, машинное обучение, анализ данных, социоинженерные атаки. Число научных публикаций — 170. mva@dscs.pro; 14-линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

Поддержка исследований. Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № FFZF-2024-0003.

DOI 10.15622/ia.25.2.1

A. VYATKIN, A. POPTSOV, V. OLISEENKO, M. ABRAMOV

RESEARCH ON THE APPLICABILITY OF MATRIX FACTORIZATION FOR RANKING LARGE LANGUAGE MODELS

Vyatkin A., Poptsov A., Oliseenko V., Abramov M. **Research on the Applicability of Matrix Factorization for Ranking Large Language Models.**

Abstract. In recent years, Large Language Models (LLMs) have gained widespread adoption in the financial domain. Direct comparison of models can be challenging, as datasets and LLMs may be closed, and evaluation parameters may vary. This paper proposes using the matrix factorization method from recommender systems, originally designed to predict user preferences, to address the task of predicting unknown metrics. The aim is to evaluate the applicability of matrix factorization for predicting LLM performance metrics on financial tasks, as well as to develop an LLM ranking method based on metric aggregation. An experiment involving the application of matrix factorization is conducted using data collected from academic research, covering 34 LLMs and 42 financial datasets. The average Mean Absolute Error (MAE) of the method across all runs is 0.07 on the test dataset. The top positions in the ranking are held by DeepSeek-R1, OpenAI GPT-4o, OpenAI o1-mini, Fin-R1, and Claude 3.5 Sonnet. The impact of prediction error on the final results is investigated using two approaches: analysis of MAE and the Monte Carlo method. The results are analyzed, yielding the following main conclusions: a) matrix factorization can be applied to predict missing model metric values on datasets; b) leading large language models have converged in performance to such an extent that identifying a clear leader is difficult; c) large prediction errors allow for the identification of specific model features on particular tasks. The proposed method can simplify the selection of a suitable model for financial tasks.

Keywords: large language models, performance evaluation, matrix factorization, financial domain.

References

1. Ali C.-S.-M., Mahmood I. A Comprehensive Survey on Large Language Models: Architectures, Applications, and Ethical Considerations. Engineering and Technology Journal. 2025. vol. 10. no. 04. pp. 4578–4593. DOI: 10.47191/etj/v10i04.26.
2. Zhao H., Liu Z., Wu Z., Li Y., Yang T., Shu P., Xu S., Dai H., Zhao L., Mai G., Liu N., Liu T. Revolutionizing finance with LLMs: An overview of applications and insights. arXiv preprint arXiv:2401.11641. 2024.
3. Dhake S.-P., Lassi L., Hippalgaonkar A., Gaidhani R.-A., Jyothi N.-M. Impacts and Implications of Generative AI and Large Language Models: Redefining Banking Sector. Journal of Informatics Education and Research. 2024. vol. 4. no. 2. pp. 248–257. DOI: 10.52783/jier.v4i2.767.
4. Zhao W.-X., Liu J., Ren R., Wen J.-R. Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems. 2024. vol. 42. no. 4. pp. 1–60. DOI: 10.1145/3637870.
5. Luo B., Zhang Z., Wang Q., Ke A., Lu S., He B. AI-powered fraud detection in decentralized finance: A project life cycle perspective. arXiv preprint arXiv:2308.15992. 2023.
6. Feng D., Dai Y., Huang J., Zhang Y., Xie Q., Han W., Lopez-Lira A., Wang H. Empowering many, biasing a few: Generalist credit scoring through large language models. arXiv preprint arXiv:2310.00566. 2023.

7. Dong Y., Yan D., Almudaifer A.-I., Yan S., Jiang Z., Zhou Y. BELT: A pipeline for stock price prediction using news. IEEE International Conference on Big Data. IEEE. 2020. pp. 1137–1146. DOI: 10.1109/BigData50022.2020.9378345.
8. Zhao W.-X., Zhou K., Li J., et al. A Survey of Large Language Models. arXiv preprint arXiv:2303.18223. 2023.
9. Li Y., Wang S., Ding H., Chen H. Large language models in finance: A survey. Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF '23). 2023. pp. 374–382. DOI: 10.1145/3604237.3626869.
10. Chen Z., Chen W., Smiley C., et al. FinQA: A Dataset of Numerical Reasoning over Financial Data. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2021. pp. 3697–3711. DOI: 10.18653/v1/2021.emnlp-main.300.
11. Shah A., Gullapalli A., Vithani R., Galarnyk M., Chava S. FiNER-ORD: Financial Named Entity Recognition Open Research Dataset. arXiv preprint arXiv:2302.11157. 2023.
12. Tang Y., Yang Y., Huang A., Tam A., Tang J. FinEntity: Entity-level Sentiment Classification for Financial Texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2023. pp. 15465–15471. DOI: 10.18653/v1/2023.emnlp-main.956.
13. Mukherjee R., Bohra A., Banerjee A., et al. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2022. pp. 10893–10906. DOI: 10.18653/v1/2022.emnlp-main.748.
14. Xie Q., Han W., Chen Z., et al. The FinBen: An Holistic Financial Benchmark for Large Language Models. arXiv preprint arXiv:2402.12659. 2024.
15. Chang Y., Wang X., Wang J., et al. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 2024. vol. 15. no. 3. pp. 1–45. DOI: 10.1145/3641289.
16. Koren Y., Bell R., Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer. 2009. vol. 42. no. 8. pp. 30–37. DOI: 10.1109/MC.2009.263.
17. Zhao Q., Xu M., Gupta K., et al. Can We Predict Performance of Large Models across Vision-Language Tasks. arXiv preprint arXiv:2410.10112. 2024.
18. Zhang Q., Lyu F., Liu X., Ma C. Collaborative Performance Prediction for Large Language Models. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024. pp. 2576–2596. DOI: 10.18653/v1/2024.emnlp-main.150.
19. Zhong X.-X., Yi C., Ye H.-J. Efficient Evaluation of Large Language Models via Collaborative Filtering. arXiv preprint arXiv:2504.08781. 2025.
20. Laskar M.-T.-R., Alqahtani S., Bari M.-S., et al. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2024. pp. 13785–13816. DOI: 10.18653/v1/2024.emnlp-main.764.
21. Owen D. How predictable is language model benchmark performance. arXiv preprint arXiv:2401.04757. 2024.
22. Matlin G., Okamoto M., Pardawala H., Yang Y., Chava S. Finance Language Model Evaluation (FLaME). Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics. 2025. pp. 880–926. DOI: 10.48550/arXiv.2506.15846.
23. Huang J., Xiao M., Li D., et al. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications. arXiv preprint arXiv:2408.11878. 2024.
24. Liu Z., Guo X., Lou F., et al. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. arXiv preprint arXiv:2503.16252. 2025.

25. KAI-GPT: The First Large Language Model Purpose-Built for Banking. Kasisto. Available at: https.kasisto.com/blog/kai-gpt-the-first-large-language-model-purpose-built-for-banking/ (accessed 25.11.2025).
26. Qian L., Zhou W., Wang Y., Peng X., Huang J., Xie Q. Fino1: On the Transferability of Reasoning-Enhanced LLMs and Reinforcement Learning to Finance. arXiv preprint arXiv:2502.08127. 2025.
27. Sharma S., Nayak T., Bose A., et al. FinRED: A Dataset for Relation Extraction in Financial Domain. Companion Proceedings of the Web Conference (WWW '22). 2022. pp. 595–597. DOI: 10.1145/3487553.3524637.
28. Kaur S., Smiley C., Gupta A., et al. REFinD: Relation Extraction Financial Dataset. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023. pp. 3054–3063. DOI: 10.1145/3539618.3591911.
29. Sharma S., Khatuya S., Hegde M., et al. Financial Numeric Extreme Labelling: A dataset and benchmarking. Findings of the Association for Computational Linguistics: ACL 2023. 2023. pp. 2933–2946. DOI: 10.18653/v1/2023.findings-acl.219.
30. Sinha A., Khandait T. Impact of News on the Commodity Market: Dataset and Results. Advances in Information Retrieval (ECIR 2021). 2021. pp. 589–601. DOI: 10.1007/978-3-030-73103-8_41.
31. Yang L., Kenny E., Ng T.-L., Yang Y., Smyth B., Dong R. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. Proceedings of the 28th International Conference on Computational Linguistics. 2020. pp. 6150–6160. DOI: 10.18653/v1/2020.coling-main.541.
32. Pardawala H., Sukhani S., Shah A., et al. SubjECTive-QA: Measuring Subjectivity in Earnings Call Transcripts' QA Through Six-Dimensional Feature Analysis. Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24). 2024. pp. 59342–59372.
33. Zhu F., Lei W., Chao Y., et al. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021. pp. 3277–3287. DOI: 10.18653/v1/2021.acl-long.254.
34. Chen Z., Li S., Smiley C., et al. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2022. pp. 6279–6290. DOI: 10.18653/v1/2022.emnlp-main.421.
35. Malo P., Sinha A., Korhonen P., Wallenius J., Takala P. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology. 2014. vol. 65. no. 4. pp. 782–796. DOI: 10.1002/asi.23062.
36. Cortis K., Freitas A., Daudert T., et al. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News Headlines. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017. pp. 519–535. DOI: 10.18653/v1/S17-2089.
37. Twitter Financial News Sentiment. Hugging Face: Zeroshot. 2024. Available at: https.huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment (accessed 25.11.2025).
38. Casanueva I., Temcinas T., Gerz D., Henderson M., Vulic I. Efficient Intent Detection with Dual Sentence Encoders. Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. 2020. pp. 38–45. DOI: 10.18653/v1/2020.nlp4convai-1.5.

39. Mariko D., Abi-Akl H., Labidurie E., et al. The Financial Document Causality Detection Shared Task (FinCausal 2020). Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. 2020. pp. 23–32. DOI:10.48550/arXiv.2012.02505
40. Shah A., Hiray A., Shah P., et al. Numerical Claim Detection in Finance: A New Financial Dataset, Weak-Supervision Model, and Market Analysis. Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). 2024. pp. 170–185. DOI: 10.18653/v1/2024.fever-1.21.
41. Chen C.-C., Lin C.-Y., Chiu C.-J., et al. Overview of the NTCIR-17 FinArg-1 Task: Fine-grained argument understanding in financial analysis. Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies. 2023. pp. 16–20. DOI: 10.20736/0002001323.
42. Zhao Y., Liu H., Long Y., et al. FinanceMATH: Knowledge-Intensive Math Reasoning in Finance Domains. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. vol. 1. pp. 12841–12858. DOI: 10.18653/v1/2024.acl-long.693.
43. Zhao Y., Long Y., Liu H., et al. DocMath-Eval: Evaluating Math Reasoning Capabilities of LLMs in Understanding Long and Specialized Documents. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. vol. 1. pp. 16103–16120. DOI: 10.18653/v1/2024.acl-long.852.
44. Yin Y., Yang Y., Yang J., Liu Q. FinPT: Financial Risk Prediction with Profile Tuning on Pretrained Foundation Models. arXiv preprint arXiv:2308.00065. 2024.
45. Hofmann H. Statlog (German Credit Data). UCI Machine Learning Repository. 1994. Available at: https.archive.ics.uci.edu/dataset/144/statlog+german+credit+data (accessed 25.11.2025).
46. Quinlan R. Statlog (Australian Credit Approval). UCI Machine Learning Repository. Available at: https.archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval (accessed 25.11.2025).
47. Flowers J.G. Finance Instruct 500k. Hugging Face. 2025. Available at: https.huggingface.co/datasets/Josephgflowers/Finance-Instruct-500k (accessed 25.11.2025).
48. Financial Evaluation Dataset. GitHub: Alipay Team. 2023. Available at: https.github.com/alipay/financial_evaluation_dataset (accessed 25.11.2025).
49. Shah A., Paturi S., Chava S. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. vol. 1. pp. 6664–6679. DOI: 10.18653/v1/2023.acl-long.368.
50. Bottou L. Large-Scale Machine Learning with Stochastic Gradient Descent. Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT '2010). Physica-Verlag HD. 2010. pp. 177–186. DOI: 10.1007/978-3-7908-2604-3_16.
51. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95). 1995. vol. 2. pp. 1137–1143.

Vyatkin Artyom — Junior researcher, Laboratory of applied artificial intelligence, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: probabilistic graph models, algebraic Bayesian networks, fuzzy computing, large language models, multi-agent systems. The number of publications — 24. aav@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)508-3311.

Poptsov Alexander — Intern researcher, Laboratory of applied artificial intelligence, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: natural language processing, large language models, AI agents. The number of publications — 2. avp@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(927)887-1617.

Oliseenko Valerii — Researcher, Laboratory of applied artificial intelligence, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: natural language processing, deep learning, social network analysis, information security, social engineering attacks. The number of publications — 60. vdo@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)508-3311.

Abramov Maxim — Ph.D., Associate Professor, Head of laboratory, Laboratory of applied artificial intelligence, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: generative artificial intelligence, large language models, machine learning, data analysis, social engineering attacks. The number of publications — 170. mva@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)508-3311.

Acknowledgements. This work was performed under the State Assignment of the St. Petersburg Federal Research Center of the Russian Academy of Sciences № FFZF-2024-0003.