УДК 004.855.5:336.648

Соловяненко Олег Юрьевич, магистрант, Казанский (Приволжский) федеральный университет, г. Казань

ИНТЕГРАЦИЯ ГРАДИЕНТНОГО БУСТИНГА, SHAP-ИНТЕРПРЕТАЦИИ И ЛОКАЛЬНЫХ ЯЗЫКОВЫХ МОДЕЛЕЙ В АГЕНТНОЙ АРХИТЕКТУРЕ REACT ДЛЯ ВЕРБАЛИЗАЦИИ РИСКОВ КОРПОРАТИВНОГО ДЕФОЛТА

Аннотация

Ансамблевые модели градиентного бустинга обеспечивают высокую точность прогнозирования корпоративного дефолта, однако их «непрозрачность» ограничивает применение в комплаенсе и риск-менеджменте. Существующие методы визуальной интерпретации (SHAP) малоприменимы в регламентированной деловой коммуникации, а использование облачных больших языковых моделей (LLM) для текстовой генерации недопустимо из-за строгих требований банковской тайны. В статье предложена архитектура изолированного вычислительного контура, объединяющая табличное ML-ядро (CatBoost), алгоритмы аддитивного объяснения (TreeExplainer) и дообученную локальную LLM (Qwen). Управление пайплайном реализовано на базе агентного подхода ReAct, где языковая модель в автоматическом цикле извлекает SHAP-значения и генерирует связный аналитический отчет с программным контролем фактической точности. Экспериментально на масштабном массиве российской финансовой отчетности подтверждено, что оптимизированное ML-ядро достигает ROC-AUC 0,84 и Brier Score 0,082, а предложенная агентная надстройка позволяет автоматизировать интерпретацию нелинейных зависимостей без передачи конфиденциальных данных во внешние сервисы.

Annotation

Ensemble gradient boosting models provide high accuracy in predicting corporate defaults, but their opacity limits their application in compliance and risk management. Existing visual interpretation methods (SHAP) are largely inapplicable in regulated business communication, while the use of cloud-based large language models (LLMs) for text generation is prohibited due to strict banking secrecy requirements. This paper proposes an isolated computational environment combining a tabular ML core (CatBoost), additive explanation algorithms (TreeExplainer), and a fine-tuned local LLM (Qwen). Pipeline orchestration is implemented using the ReAct agent framework, where the language model orchestrates tool calls to the tabular prediction service; SHAP contributions are computed by the gradient boosting backend (TreeExplainer), and the model then generates a coherent analytical report with programmatic validation of factual accuracy. Experiments conducted on a large-scale dataset of Russian corporate financial statements confirm that the optimized ML core achieves a ROC-AUC of 0.84 and a Brier Score of 0.082. Furthermore, the proposed agent-based framework automates the interpretation of non-linear dependencies without transmitting confidential data to external services.

Ключевые слова: градиентный бустинг; прогнозирование дефолта; объяснимый искусственный интеллект; SHAP; большие языковые модели; ReAct; агентная архитектура; корпоративные финансы; локальный вычислительный контур.

Keywords: Gradient boosting; default prediction; explainable AI (XAI); SHAP; large language models (LLMs); ReAct; agent-based architecture; corporate finance; isolated computational environment.

Введение. Задача раннего выявления признаков финансовой несостоятельности юридических лиц занимает центральное место в системе управления кредитным риском коммерческих банков, институтов развития и регуляторных органов. Рост объема и детализации корпоративной отчетности

создает предпосылки для применения методов машинного обучения, способных уловить нелинейные взаимодействия между финансовыми коэффициентами, недоступные классическим скоринговым картам. Современные ансамблевые модели на основе градиентного бустинга (CatBoost [8], XGBoost [4]) демонстрируют высокое качество ранжирования на задачах бинарной классификации с выраженным дисбалансом классов, однако носят характер «черного ящика». Даже при наличии формальных метрик качества, доверие со стороны бизнес-подразделений и комплаенса требует объяснений, привязанных к содержательным финансовым показателям (ликвидности, долговой нагрузке, рентабельности). Без них принятие решений по кредитному лимиту остается непрозрачным для внутреннего аудита.

Существующие подходы объяснимого искусственного интеллекта на основе алгоритмов SHAP [6] дают согласованную аддитивную декомпозицию вклада каждого признака. Тем не менее, визуализации SHAP (beeswarm-диаграммы, waterfall-графики) информативны преимущественно для специалистов по данным. Для аудитории риск-менеджмента и кредитных комитетов недостаточно статичных диаграмм; необходим связный текст, фиксирующий уровень риска, доминирующие факторы ухудшения и стабилизации, пригодный для включения в официальные протоколы.

В научной литературе задача автоматической вербализации структурированных данных все чаще решается с помощью больших языковых моделей. Исследования последних лет (в частности, архитектуры BloombergGPT [11] и FinGPT [13]) подтвердили применимость LLM в финансовом домене, однако выявили два фундаментальных ограничения: склонность к фактологическим искажениям (галлюцинациям) при прямой генерации текста и недопустимость передачи конфиденциальных данных (ИНН, балансовые строки) через публичные API на серверы третьих лиц.

В связи с этим возникает объективная необходимость перехода к локальному вычислительному контуру, в котором и табличное ядро, и языковая модель исполняются на контролируемой инфраструктуре заказчика. В качестве

базовой генеративной модели целесообразно использовать компактные открытые LLM семейства Qwen [12], пригодные для дообучения методами низкоранговой адаптации (LoRA/QLoRA) [5].

Как показывают исследования в области когнитивных архитектур [9, 10], проблему галлюцинаций эффективно решает переход к автономным агентам, где модель не генерирует численные значения «из весов», а извлекает их через вызов внешних детерминированных инструментов. Вместе с тем интеграция подобных агентных подходов с методами XAI для интерпретации табличного скоринга внутри on-premise-контура остается нерешенной научной задачей.

Цель настоящего исследования является разработка и валидация архитектуры замкнутого контура, обеспечивающей прозрачную интерпретацию рисков корпоративного дефолта. Научная новизна работы заключается в синтезе трёх компонентов: масштабируемого табличного моделирования на унифицированных признаках, алгоритмов SHAP-объяснений и агентной архитектуры (ReAct) [14] с дообученной локальной языковой моделью, автоматизирующей перевод численных метрик в регламентированный аналитический отчет.

Материалы и методы. Источником эмпирических данных послужили выгрузки Ресурса финансовых и статистических данных, содержащие бухгалтерскую отчетность юридических лиц РФ. Загрузка осуществлялась потоковым образом с использованием механизма условного обновления по бизнес-ключам, что обеспечило обработку многомиллионного массива записей без превышения лимитов оперативной памяти. Признаковое пространство формировалось из 20 финансовых коэффициентов, рассчитываемых единым доменным калькулятором. В их число вошли 5 классических интегральных моделей банкротства (Z-счет Альтмана [3], Z-счет Таффлера, H-счёт Фулмера, индексы Сайфуллина–Кадыкова и Зайцевой), 12 аналитических коэффициентов по методике В. В. Ковалёва [1], а также 3 структурных показателя. Использование идентичного алгоритма расчета

признаков как при формировании обучающей выборки, так и в контуре эксплуатации позволило исключить проблему рассинхронизации сред.

Для обеспечения временной корректности применялось темпоральное разбиение, при котором последний доступный год отводился под тестирование. Экстремальные значения признаков обрабатывались методом робастного клиппинга (винзоризация на уровнях 1% и 99%) с фиксацией границ для тестовой выборки. Естественный дисбаланс классов (2–5% дефолтов) компенсировался контролируемым undersampling-ом мажоритарного класса. В качестве базового алгоритма использовался градиентный бустинг CatBoost [8], гиперпараметры которого оптимизировались библиотекой Optuna [2] с целевой функцией average_precision_score (PR-AUC) для корректной работы с несбалансированными данными. В связи с неприменимостью фиксированного порога классификации (0,5) были реализованы два динамических операционных режима: сбалансированный (максимизация F_1 – меры с ограничением точности) и консервативный (нижняя граница полноты ≥ 0,85). В качестве baseline–моделей для сравнения выступали XGBoost [4] и логистическая регрессия.

Для интерпретации предсказаний обученной ансамблевой модели применялся алгоритм SHAP (TreeExplainer [7]). Для каждого оцениваемого наблюдения вычислялся вектор SHAP-значений, из которого извлекались доминирующие факторы риска, стабилизирующие факторы и базовое значение предсказания. С целью специализации языковой модели на задаче перевода полученных структурированных объяснений в аналитический текст был разработан конвейер формирования обучающего XAI-датасета. На репрезентативной подвыборке из действующих компаний и банкротов выполнялся полный цикл инференса, результат которого агрегировался во входной JSON-пакет. На

основании этого пакета LLM-учитель генерировала эталонный аналитический отчёт в деловом стиле. Полученные пары (instruction/input/output) использовались для дообучения компактной открытой модели семейства Qwen [12] методами низкоранговой адаптации (LoRA/QLoRA) [5].

Оркестрация всего пайплайна, включая управление логическим выводом и вызовом внешних инструментов, реализована в парадигме ReAct [14] (Reasoning + Acting). На каждом шаге языковая модель возвращает структурированный ответ: намерение вызвать инструмент либо финальный аналитический отчет. Агент ведет цепочку записей «Мысль — Действие — Наблюдение», передавая состояние через изолированный контекст запроса. Инструментарий агента инкапсулирует парсинг XBRL-отчетности, проверку качества данных, расчет метрик и вызов ML-ядра для получения оценки риска. Для предотвращения галлюцинаций LLM и преждевременной генерации ответов внедрен двухуровневый программный валидатор. Блокирующие правила прерывают выполнение при расхождении текстовой оценки с числовым предсказанием или отсутствии обязательных шагов. Перед подачей в промпт выполняется семантическая очистка: нормализация финансовых терминов и удаление избыточных технических полей. Взаимодействие с LLM осуществляется через stateless-маршрутизатор, абстрагирующий логику ReAct от конкретной среды исполнения (локальный инференс Ollama или облачное API), что обеспечивает возможность бесшовного A/B-тестирования моделей.

Результаты.

Эволюция и оценка предиктивной способности моделей. Итеративный процесс моделирования позволил проследить динамику качества ансамблевых методов при переходе от базовых конфигураций к целевой оптимизации. На ранних этапах применялся единый протокол сравнения CatBoost [8] и XGBoost [4] без глубокого поиска гиперпараметров. Результаты фиксировали паритет методов: XGBoost незначительно превосходил CatBoost по площади под ROC-кривой (0,7453 против 0,7408). Однако критическим ограничением выступал

чрезвычайно высокий показатель Brier Score (около 0,20 для обеих моделей), свидетельствующий о плохой калибровке вероятностей, что неприемлемо для риск-процессов, оперирующих абсолютными значениями вероятности дефолта (PD). Переход к целенаправленной оптимизации включал применение фреймворка Optuna [2] с целевой функцией максимизации площади под PR-кривой (PR-AUC), чувствительной к качеству ранжирования редкого класса. Был осуществлен отказ от встроенной балансировки классов для сохранения калибровки вероятностей, а также внедрен подход максимизации полноты при фиксированной нижней границе точности. Итоговая оценка на отложенной выборке представлена в Таблице 1.

Модель	ROC-AUC	PR-AUC	Brier Score	F1 (сбаланс.)	Recall (сбаланс.)
CatBoost	0,8402	0,2817	0,0821	0,4476	0,7160
XGBoost	0,7619	0,2766	0,1926	0,3418	0,6347
Логистическая регрессия	0,6379	0,1751	0,2392	0,2403	0,4630

Таблица 1. Финальное сравнение моделей на отложенной выборке (2021 г.)

Оптимизированная модель CatBoost продемонстрировала абсолютный прирост ROC-AUC на 0,099 пунктов по сравнению с базовой версией, а также улучшение калибровки (Brier Score) в 2,5 раза. Преимущество бустинговых методов над линейным базисом (логистической регрессией) наглядно

подтверждается на ROC-кривых и PR-кривых (Рис. 1 и Рис. 2). Для CatBoost PR-кривая сохраняет точность выше 0,3 вплоть до уровня полноты 0,4, что является существенным фактором для операционных решений при выраженном дисбалансе классов.

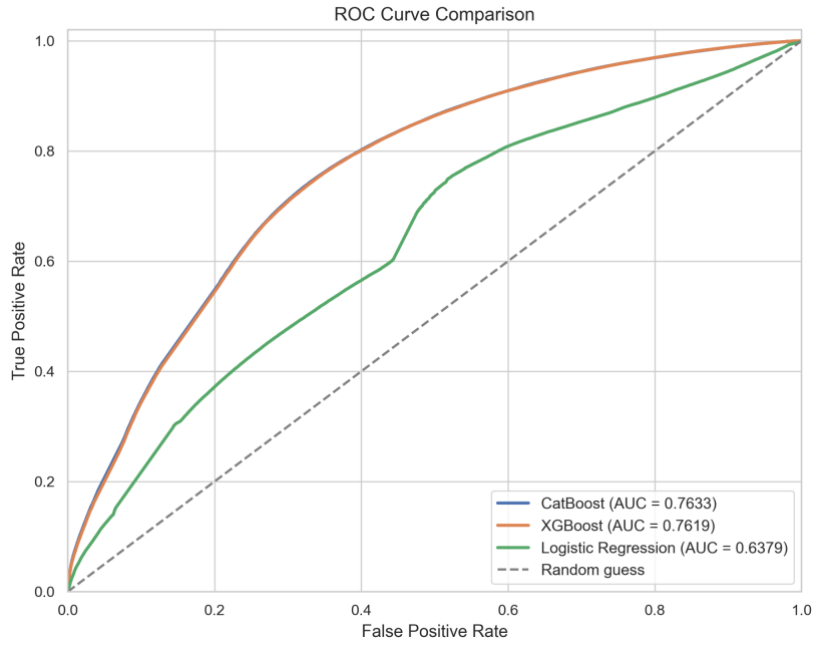

Рис. 1. Сравнение ROC-кривых моделей на отложенной выборке

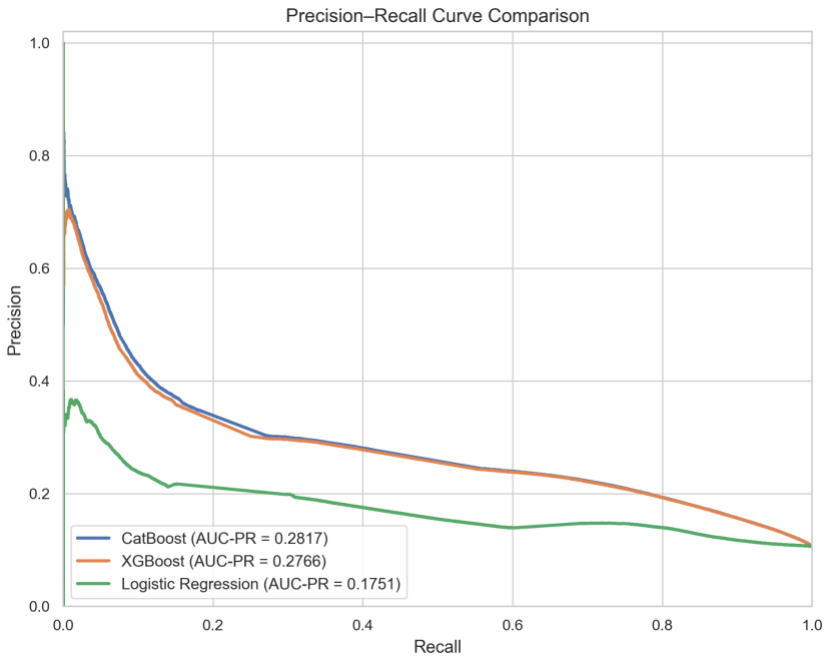

Рис. 2. Сравнение precision–recall кривых

Анализ матриц ошибок выявил ожидаемый компромисс между полнотой и точностью при переключении профиля риска. В сбалансированном режиме CatBoost распознаёт 71,6 % случаев дефолта. Переход в консервативный режим повышает этот показатель до 83,4 % ценой снижения общей точности, что экономически оправдано для сценариев кредитования, где ошибка второго рода (пропуск дефолта) несет критические убытки. Оценка временной устойчивости подтвердила стабильность CatBoost на историческом горизонте: ROC-AUC составил 0,87–0,89 в период 2017–2019 гг., с закономерным снижением до 0,84 в 2021 г., что отражает нестационарность экономических процессов на фоне макроэкономических шоков.

Интерпретация признаков и агентная вербализация. Сопоставление глобальной важности признаков выявило, что различные архитектуры выделяют пересекающиеся, но не идентичные группы ведущих показателей. Сводная оценка SHAP [6, 7] по обучающей совокупности (Рис. 3) иллюстрирует устойчивый вклад ключевых финансовых метрик. Доминирующим признаком выступает абсолютная ликвидность: низкие значения сосредоточены в зоне положительных SHAP-вкладов и повышают прогнозируемую вероятность дефолта; высокие значения смещены к отрицательным вкладам и снижают оценку риска. Высокая оборачиваемость активов, напротив, формирует наибольший положительный вклад в предсказание дефолта. Второстепенными драйверами выступают рентабельность продаж и интегральные показатели (Z-счет).

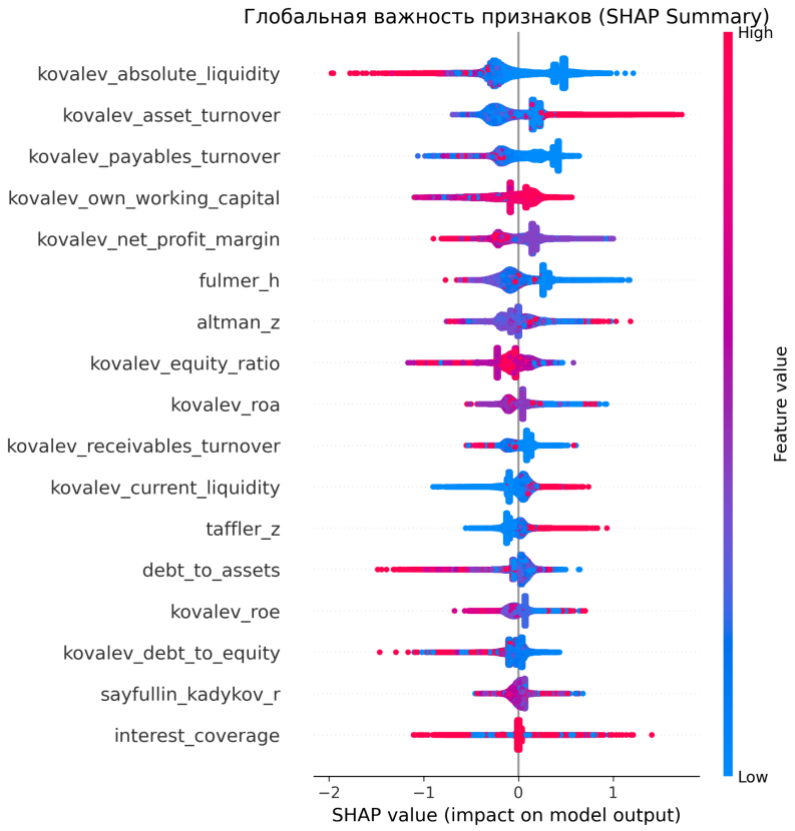

Рис. 3. Глобальная важность признаков (сводная диаграмма SHAP)

Локальные SHAP-значения, вычисляемые для индивидуального запроса, формируют профиль риска конкретной компании. Агентная архитектура ReAct [14] автоматизирует перевод данного профиля в аналитический текст через последовательный конвейер: извлечение метрик, семантическая группировка факторов по знаку вклада с нейтрализацией аббревиатур, компактификация контекста (удаление избыточных числовых массивов), генерация структурированного ответа (вводная часть, факторы снижения и роста риска, рекомендации).

Программная валидация гарантирует сверку итоговой зоны риска в сгенерированном тексте с первичным математическим предсказанием. Результатом работы контура является связный аналитический текст, пригодный для включения в протоколы кредитного комитета, исключающий вымышленные числовые данные и раскрытие внутренних алгоритмических структур.

Предложенная архитектура решает методологическую проблему, которую невозможно устранить использованием ее компонентов по отдельности. Табличная модель обеспечивает формальную предсказательную силу, но лишена объяснимости; алгоритм SHAP [6] предоставляет точную декомпозицию, формат которой затруднителен для интерпретации бизнес-подразделениями; стандартные LLM способны к вербализации, но подвержены генерации недостоверных фактов (галлюцинациям) при отсутствии жесткой структурной привязки. Синтез данных методов в контуре ReAct [14] с программной JSON-валидацией компенсирует указанные недостатки. Проведенные эксперименты позволяют выделить ряд ключевых наблюдений. Во-первых, целенаправленная оптимизация по метрике PR-AUC оказалась решающим фактором превосходства: при отказе от искусственного взвешивания классов удалось не только повысить качество ранжирования, но и обеспечить корректную калибровку вероятностей. Во-вторых, механизм маршрутизации запросов к LLM (stateless-роутинг) подтвердил свою эффективность для проведения сравнительного тестирования локальных и облачных моделей без изменения логики агента. В-третьих, встроенная система многоуровневой обратной связи критически снижает частоту ошибок логического вывода: блокирующие проверки отсекают нарушение математической консистентности, а рекомендательные — повышают информативность. К ограничениям текущего исследования следует отнести отсутствие автоматизированных метрик оценки качества генерации естественного языка, что затрудняет потоковое количественное сравнение моделей. Кроме того, размер обучающего датасета может оказаться недостаточным для глубокой доменной специализации малых языковых моделей.

Заключение. В рамках исследования разработана и валидирована архитектура локального вычислительного контура, объединяющего высокоточное табличное моделирование [8] корпоративных дефолтов, алгоритмы аддитивного объяснения SHAP [6, 7] и дообученную локальную языковую

модель [12]. Эмпирически на массиве российской корпоративной отчётности подтверждено, что оптимизированная модель CatBoost достигает ROC-AUC 0,8402 и Brier Score 0,0821, значительно превосходя альтернативные подходы и демонстрируя высокую временную устойчивость. Интеграция агентного подхода ReAct [14] с жесткой JSON-валидацией позволяет автоматически транслировать многомерные SHAP-профили в регламентированные аналитические отчеты. Предложенный подход разрешает фундаментальное противоречие между потребностью финансового сектора в объяснимом ИИ и строгими ограничениями на передачу конфиденциальных данных в публичные генеративные сервисы. Перспективными направлениями развития работы являются расширение специализированного обучающего датасета, внедрение автоматических NLP-метрик для оценки качества вербализации и масштабирование системы на задачи мультипериодного мониторинга финансового состояния контрагентов.

Список литературы

1. Ковалёв В.В. Финансовый анализ: методы и процедуры. М.: Финансы и статистика, 2002. 560 с.
2. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. P. 2623–2631.
3. Altman E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy // The Journal of Finance. 1968. Vol. 23, No. 4. P. 589–609.
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794.
5. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. LoRA: Low-Rank Adaptation of Large Language Models // Proceedings of the 10th International Conference on Learning Representations (ICLR). 2022.

6. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30 (NeurIPS). 2017. P. 4765–4774.
7. Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees // Nature Machine Intelligence. 2020. Vol. 2, No. 1. P. 56–67.
8. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased Boosting with Categorical Features // Advances in Neural Information Processing Systems 31 (NeurIPS). 2018. P. 6639–6649.
9. Schick T., Dwivedi-Yu J., Dessì R., Raileanu R., Lomeli M., Hambro E., Zettlemoyer L., Cancedda N., Scialom T. Toolformer: Language Models Can Teach Themselves to Use Tools // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 68539–68551.
10. Wu Q., Bansal G., Zhang J., Wu Y., Li B., Zhu E., Jiang L., Zhang X., Wang C. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation // arXiv preprint arXiv:2308.08155. 2024.
11. Wu S., Irsoy O., Lu S., Dabravolski V., Dredze M., Gehrmann S., Kambadur P., Rosenberg D., Mann G. BloombergGPT: A Large Language Model for Finance // arXiv preprint arXiv:2303.17564. 2023.
12. Yang A., Yang B., Hui B., Zheng B., Yu B., Zhou C. et al. Qwen2 Technical Report // arXiv preprint arXiv:2407.10671. 2024.
13. Yang Y., Liu X.-Y., Zhong R., Wang J., Walid A. FinGPT: Open-Source Financial Large Language Models // arXiv preprint arXiv:2306.06031. 2023.
14. Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models // Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.

References

1. Kovalev V.V. Financial Analysis: Methods and Procedures. Moscow: Finance and Statistics, 2002. 560 p. (In Russian).
2. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. P. 2623–2631.
3. Altman E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy // The Journal of Finance. 1968. Vol. 23, No. 4. P. 589–609.
4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794.
5. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. LoRA: Low-Rank Adaptation of Large Language Models // Proceedings of the 10th International Conference on Learning Representations (ICLR). 2022.
6. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30 (NeurIPS). 2017. P. 4765–4774.
7. Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees // Nature Machine Intelligence. 2020. Vol. 2, No. 1. P. 56–67.
8. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased Boosting with Categorical Features // Advances in Neural Information Processing Systems 31 (NeurIPS). 2018. P. 6639–6649.
9. Schick T., Dwivedi-Yu J., Dessì R., Raileanu R., Lomeli M., Hambro E., Zettlemoyer L., Cancedda N., Scialom T. Toolformer: Language Models Can Teach Themselves to Use Tools // Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 68539–68551.

10. Wu Q., Bansal G., Zhang J., Wu Y., Li B., Zhu E., Jiang L., Zhang X., Wang C. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation // arXiv preprint arXiv:2308.08155. 2024.
11. Wu S., Irsoy O., Lu S., Dabravolski V., Dredze M., Gehrmann S., Kambadur P., Rosenberg D., Mann G. BloombergGPT: A Large Language Model for Finance // arXiv preprint arXiv:2303.17564. 2023.
12. Yang A., Yang B., Hui B., Zheng B., Yu B., Zhou C. et al. Qwen2 Technical Report // arXiv preprint arXiv:2407.10671. 2024.
13. Yang Y., Liu X.-Y., Zhong R., Wang J., Walid A. FinGPT: Open-Source Financial Large Language Models // arXiv preprint arXiv:2306.06031. 2023.
14. Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models // Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.