КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 2 С. 211–241
DOI: 10.20537/2076-7633-2026-18-2-211-241

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 004.8

Объяснимый искусственный интеллект: принципы, методы и применение

П. Ю. Середа-Калинин[1,2,a], А. С. Власова[1,b]

[1]Институт психологии Российской академии наук,
Россия, 129366, г. Москва, ул. Ярославская, д. 13, корп. 1
[2]Институт системного программирования им. В. П. Иванникова Российской академии наук,
Россия, 109004, г. Москва, ул. Александра Солженицына, д. 25

E-mail: [a] seredapj@ipran.ru, [b] vlasovaas@ipran.ru

Получено 08.10.2025, после доработки — 26.01.2026.
Принято к публикации 13.03.2026.

Объяснимый искусственный интеллект (Explainable AI, XAI) представляет собой область искусственного интеллекта, направленную на создание методов и инструментов для генерации интерпретируемых и понятных для человека объяснений решений ИИ. Актуальность объяснимости моделей возрастает по мере внедрения искусственного интеллекта в критически важные сферы (медицина, финансы, юриспруденция), где непрозрачность алгоритмов может приводить к серьезным последствиям для пользователей и общества. В работе представлен аналитический обзор современного состояния области XAI, охватывающий теоретические основы, методологию и практические применения.

Рассматриваемые методы объяснимого ИИ были отобраны и систематизированы на основе многоуровневой классификации методов XAI по постановке задачи (цель, целевая аудитория, тип данных), методологии (стадия применения, модель-специфичность, методы, масштаб) и форме результата (представление, презентация, метрики оценки).

Проведен сравнительный анализ методов объяснимого ИИ для различных областей применения. Для классического машинного обучения детально рассмотрены SHAP и LIME с выявлением их теоретических оснований, вычислительных характеристик и ограничений. Для компьютерного зрения систематизированы градиентные методы (SmoothGrad, Integrated Gradients), методы визуализации активаций (Grad-CAM, Grad-CAM++), методы на основе возмущений (RISE, Occlusion) и концептуальные объяснения (TCAV, Network Dissection). Особое внимание уделено специфике применения XAI к обработке естественного языка и большим языковым моделям, включая анализ достоверности цепочек размышлений (Chain-of-Thought), естественно-языковых объяснений и методов на основе графов атрибуции. Выделены фундаментальные ограничения существующих подходов к объяснимости LLM и определены направления дальнейших исследований.

Результаты обзора демонстрируют, что методы XAI достигли значительной зрелости в области классического машинного обучения и компьютерного зрения, однако применение к большим языковым моделям остается открытой исследовательской проблемой, требующей разработки новых парадигм объяснения.

Ключевые слова: объяснимый искусственный интеллект, XAI, интерпретируемость, прозрачность моделей, машинное обучение, глубокое обучение, большие языковые модели

Работа выполнена в рамках госзадания в Институте психологии РАН (тема № 0138-2024-0020).

COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 2 P. 211–241
DOI: 10.20537/2076-7633-2026-18-2-211-241

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 004.8

Explainable artificial intelligence: principles, methods and applications

P. Yu. Sereda-Kalinin[1,2,a], A. S. Vlasova[1,b]

[1]Institute of Psychology of the Russian Academy of Sciences,
13/1 Yaroslavskaya st., Moscow, 129366, Russia
[2]Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25 Alexander Solzhenitsyn st., Moscow, 109004, Russia

E-mail: [a] seredapj@ipran.ru, [b] vlasovaas@ipran.ru

Received 08.10.2025, after completion — 26.01.2026.
Accepted for publication 13.03.2026.

Explainable Artificial Intelligence (XAI) is a field of artificial intelligence aimed at creating methods and tools for generating interpretable and human-understandable explanations of AI decisions. The relevance of model explainability increases with the deployment of artificial intelligence in critical domains (healthcare, finance, law), where algorithmic opacity can lead to serious consequences for users and society. This work presents an analytical review of the current state of the XAI field, covering theoretical foundations, methodology, and practical applications.

The examined explainable AI methods were selected and systematized based on a multi-level classification of XAI methods by problem formulation (goal, target audience, data type), methodology (application stage, model-specificity, methods, scale), and result form (representation, presentation, evaluation metrics).

A comparative analysis of explainable AI methods for various application domains is conducted. For classical machine learning, SHAP and LIME are examined in detail, revealing their theoretical foundations, computational characteristics, and limitations. For computer vision, gradient-based methods (SmoothGrad, Integrated Gradients), activation visualization methods (Grad-CAM, Grad-CAM++), perturbation-based methods (RISE, Occlusion), and conceptual explanations (TCAV, Network Dissection) are systematized. Special attention is paid to the specifics of applying XAI to natural language processing and large language models, including analysis of the faithfulness of Chain-of-Thought reasoning, natural language explanations, and attribution graph methods. Fundamental limitations of existing approaches to LLM explainability are identified and directions for future research are defined.

The review results demonstrate that XAI methods have reached significant maturity in classical machine learning and computer vision, however, their application to large language models remains an open research problem requiring the development of new explanation paradigms.

Keywords: explainable artificial intelligence, XAI, interpretability, model transparency, machine learning, deep learning, large language models

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 2, pp. 211–241 (Russian).

This work was supported by RF state assignment 0138-2024-0020.

Введение

Объяснимый искусственный интеллект (Explainable AI, XAI) — это область искусственного интеллекта, которая представляет собой набор инструментов, методов и алгоритмов, способных генерировать высококачественные, интерпретируемые и интуитивно понятные для человека объяснения решений ИИ [Das, Rad, 2020; Шевская, 2021]. Объяснимость ИИ приобретает особую значимость в критически важных областях применения искусственного интеллекта: здравоохранении, финансах, безопасности и других, где от решений моделей могут зависеть человеческая жизнь и ресурсы. В таких рискованных и строго регулируемых сферах, как медицина и финансы, ошибка алгоритма способна привести к катастрофическим последствиям: утрате жизни или капитала.

Принципы XAI

Представление о том, каким требованиям должны соответствовать методы XAI, чтобы получаемые объяснения могли считаться ясными и достоверными и соответствовать запросам практики, многокомпонентно, и в литературе отдельное внимание уделяется систематизации принципов объяснимого ИИ. Выделяемые разными авторами принципы объяснимого ИИ можно обобщить, обозначив три основных взаимосвязанных вектора: ориентация на конечного пользователя, социальная ответственность, а также техническая достоверность.

Ориентация на конечного пользователя

Ориентация на конечного пользователя заключается в таких конкретных положениях, как понимаемость [Arrieta et al., 2020], понятность [Arrieta et al., 2020; Belle, Papantonis, 2021; Phillips et al., 2021] и адаптация объяснений под знания, цели и контекст конечного пользователя [Phillips et al., 2021; Ronge et al., 2025].

Понимаемость (understandability) определяется как возможность функции объясняемой модели быть понятой без знания о ее внутренней структуре. В отличие от понимаемости термин «понятность» (comprehensibility) отражает требование относительно доступности представления именно выходных данных модели [Arrieta et al., 2020].

Необходимость адаптации объяснений под знания, цели и контекст конечного пользователя связана с тем, что в зависимости от того, кому и зачем понадобилось объяснение функционирования и предсказаний модели, степень детализации и формат, соответствующие удовлетворяющему пользователя объяснению, будут различными [Phillips et al., 2021; Ronge et al., 2025].

Социальная ответственность

Принципы, направленные на социально ответственное применение ИИ, связаны с такими этическими аспектами, как ограниченность [Phillips et al., 2021; Ronge et al., 2025], выявление предвзятости [Belle, Papantonis, 2021], а также обеспечение доверия к решениям модели [Arrieta et al., 2020; Phillips et al., 2021].

Согласно принципу ограниченности система ИИ должна работать только в пределах тех условий, для которых она была разработана, и только тогда, когда достигает достаточной уверенности в своем результате. Если запрос или операция выходят за пределы области, для которой система предназначена, или если уровень уверенности слишком низкий, система должна явно указать на это, тем самым определяя границы своих решений [Phillips et al., 2021].

Выявление предвзятости заключается в использовании методов объяснимого ИИ для обнаружения и оценки систематических ошибок и дискриминационных предубеждений в данных и модели, что позволяет понять, какие признаки влияют на решение, и выявить социально значимые предвзятости, повышая прозрачность и справедливость модели [Belle, Papantonis, 2021].

Принцип обеспечения доверия к решениям модели заключается в самом факте наличия объяснений, которые выступают связующим звеном между пользователем и моделью, принимающей решение, и отображают обоснованное принятие решений моделью с целью донесения логики работы, за счет чего и может быть достигнуто доверие к принятому решению [Arrieta et al., 2020; Phillips et al., 2021].

Техническая достоверность

Техническая достоверность — наиболее обширная группа принципов объяснимого ИИ, к которой относятся точность [Belle, Papantonis, 2021; Phillips et al., 2021], прозрачность [Arrieta et al., 2020], отслеживаемость [Ronge et al., 2025], надежность [Belle, Papantonis, 2021], экстраполируемость [Belle, Papantonis, 2021], снижение размерности [Ronge et al., 2025], возможность использования объяснений для улучшения качества предсказаний модели [Belle, Papantonis, 2021] и научная верификация [Ronge et al., 2025].

Один из ключевых принципов технической достоверности — точность. Объяснения должны корректно отображать причину получения тех или иных выходных данных или, в зависимости от типа объяснимого ИИ, точно отображать процесс работы модели. Верное объяснение не обязательно подразумевает корректную работу объясняемой модели: объяснения могут быть точными, даже если сама модель ошибочна, что важно для диагностики и улучшения модели [Belle, Papantonis, 2021; Phillips et al., 2021].

Прозрачность определяется как свойство модели быть понятной самой по себе без необходимости в дополнительных объяснениях. Ариетта с соавторами выделяют три вложенных уровня прозрачности: симулируемость, декомпозируемость и алгоритмическая прозрачность моделей. Под симулируемостью, как под наивысшей степенью прозрачности, понимается возможность построить математическую модель машинного обучения с помощью расчетов вручную. Декомпозируемость подразумевает возможность разложить модель МО на отдельные части, каждую из которых можно интерпретировать отдельно. Алгоритмическая прозрачность связана с возможностью понять устройство и логику модели, не прибегая к внешним инструментам. Вложенность этой классификации подразумевает, что симулируемая модель всегда также и декомпозируема, и алгоритмически прозрачна, а декомпозируемая модель может не быть симулируемой, но во всех случаях обладает алгоритмической прозрачностью. К таким простым и изначально прозрачным моделям относят логистическую и линейную регрессию, деревья решений малой глубины, метод k-ближайших соседей, генеральные аддитивные модели, а также байесовские методы машинного обучения [Arrieta et al., 2020].

Принцип отслеживаемости заключается в возможности проследить вклад определенных входных данных и отдельных функций модели [Ronge et al., 2025]. В сравнении с прозрачностью отслеживаемость ориентирована на конкретные взаимосвязи в данных и их следы в выводах модели, обеспечивая детальную обратную связь.

Согласно принципу надежности хорошее объяснение работы модели должно отвечать на вопрос о том, насколько предсказания модели зависят от зашумленности данных и неполноты информации, обеспечивая оценку доверия к результатам и позволяя выявлять нестабильные зоны модели [Belle, Papantonis, 2021].

Экстраполируемость подразумевает оценку того, каким конкретно образом модель прогнозирования, разработанная для одной прикладной области, может быть применена к другой прикладной области, а также того, какие характеристики данных и модели необходимо скорректировать, чтобы обеспечить их применимость. Этот принцип тесно связан с определением границ принятия решений, так как экстраполяция возможна лишь в пределах совместимых областей [Belle, Papantonis, 2021].

Снижение размерности способствует повышению прозрачности моделей за счет уменьшения числа признаков, что облегчает интерпретацию и анализ, снижая шум и повышая качество объяснений, а также повышая прозрачность [Ronge et al., 2025].

Возможность использования объяснений для улучшения качества предсказаний означает применение методов XAI не только для понимания модели, но и для ее доработки и оптимизации, что повышает точность и надежность результатов [Belle, Papantonis, 2021].

Научная верификация включает проверку объяснений и моделей с помощью внешних знаний и эмпирических данных, чтобы обеспечить обоснованность и повторяемость результатов, что критически важно для доверия и внедрения XAI [Ronge et al., 2025].

Взаимосвязь принципов

Взаимосвязь трех групп принципов объяснимого ИИ — ориентации на конечного пользователя, социальной ответственности и технической достоверности — выражается в том, что технические принципы обеспечивают базовую достоверность и прозрачность работы модели, социальная ответственность гарантирует справедливость, отсутствие предвзятости и безопасность для общества, а ориентация на конечного пользователя направлена на понятность и формирование доверия со стороны пользователей. На стыке технической достоверности и социальной ответственности находятся принципы надежности и выявления предвзятости, между технической достоверностью и ориентацией на пользователя — прозрачность и объяснимость, а между социальной ответственностью и ориентацией на пользователя — обеспечение доверия к решениям.

Классификация методов XAI

Обобщая множество существующих таксономий XAI, систематизирующих его методы на различных основаниях, мы предлагаем классификацию по постановке задачи, методологии и результату.

Постановка задачи

На этапе постановки задачи модели XAI отличаются по цели [Ronge et al., 2025], целевой аудитории [Ronge et al., 2025], задаче машинного обучения [Schwalbe, Finzel, 2024] и типу входных данных [Schwalbe, Finzel, 2024].

По целям методы XAI можно разделить на направленные на укрепление доверия, улучшение удобства использования, извлечение знаний, регуляцию и оценку. При этом отмечается, что в литературе не очень четко и конкретно обозначены цели объяснимого ИИ по двум причинам. Во-первых, из-за сложности доказательства достижения цели. Во-вторых, из-за несоответствия между предполагаемым воплощением цели и фактическими результатами XAI [Ronge et al., 2025].

Целевые аудитории, использующие объяснения работы модели, можно различать по их знаниям об ИИ, знаниям области его применения и по тому, как объяснения используются этой целевой аудиторией. Субъект — тот, на кого непосредственно влияет решение ИИ. Представители данной целевой аудитории, не имея знаний ни в области ML, ни в конкретной предметной области, нуждаются в объяснениях для понимания, справедливы ли решения. Другой тип целевой аудитории — пользователи. Они получают подробные объяснения для формирования базы знаний, на основе которой можно решить, стоит ли использовать ИИ для решаемой ими задачи. Исполнитель, обладающий обширными знаниями об ML, использует объяснения для оценки модели. Разработчик, который также обладает знаниями об ML, применяет объяснения в целях улучшения модели. И наконец, регулятор является представителем внешней организации, тем, на кого решение ИИ не влияет. Его роль — контролировать справедливость решений ИИ по

отношению к субъектам [Ronge et al., 2025]. Классификация методов XAI по целевым аудиториям — это первый шаг к осуществлению принципа адаптации объяснений под того, для кого они предназначены.

Кроме того, важно учитывать тип задачи машинного обучения самой модели, поскольку разные задачи (например, классификация, регрессия, кластеризация) предъявляют специфичные требования к форме и содержанию объяснений, а также к методам их получения, что обеспечивает релевантность и полезность объяснений в конкретном контексте применения модели. Это позволяет создавать объяснения, которые наиболее эффективно помогают понять и оценить поведение модели для данного типа задачи, улучшая ее интерпретируемость и доверие к ее решениям [Schwalbe, Finzel, 2024].

Тип входных данных варьируется от числовых и текстовых до изображений, временных рядов, аудио и графов, что накладывает определенные требования к применяемым методам объяснения [Schwalbe, Finzel, 2024].

Методология

На методологических основаниях XAI классифицируются по стадии применения, модель-специфичности, методам и масштабу.

Стадия применения разделяет методы на анте-хок и пост-хок. Анте-хок-методы интегрируются непосредственно в архитектуру модели, обеспечивая ее изначальную прозрачность и интерпретируемость без последующих доработок. Пост-хок-методы применяются после обучения модели и направлены на создание объяснений с помощью внешних алгоритмов, не меняя внутреннюю структуру модели [Speith, 2022; Ali et al., 2023; Ronge et al., 2025]. В свою очередь, пост-хок-методы делятся по основанию модель-специфичности и по методу.

Модель-специфичность характеризует зависимость метода объяснения от структуры модели. Модель-специфичные методы используют внутренние характеристики и операции конкретной модели для генерации объяснений и обычно дают более точные и информативные результаты, обладая при этом узким диапазоном применения. К ним, например, относятся методы, основанные на принципе обратного распространения ошибки или механизме обратной свертки. Модель-неспецифичные методы не зависят от архитектуры модели, что обеспечивает их широкую применимость, но порой снижает степень точности объяснений. Представителями этой группы являются LIME и SHAP [Speith, 2022; Schwalbe, Finzel, 2024; Ronge et al., 2025].

Методы выделяются по принципу получения объяснений: использование структуры модели, извлечение примеров, изменение входных данных, изменение архитектуры модели, а также выделяют метаобъяснения. Под использованием структуры подразумеваются методы, основанные на интеграции внутренней структуры модели в процесс объяснения, но при этом на саму модель не оказывается дополнительное воздействие, изменяющее модель. Изменение входных данных, или метод локальных возмущений, заключается в точечном изменении входных данных, благодаря чему делаются выводы о влиянии этих частей входных данных на работу модели. Под изменением архитектуры модели понимается именно ее упрощение, что позволяет получить модель с более высоким уровнем прозрачности. Таким образом, объяснение исходной модели получается благодаря объяснению аналогичной более простой модели. Метод извлечения примеров представляет собой поиск репрезентативных примеров, которые отражают связи, обнаруживаемые анализируемой моделью. Наконец, метаобъяснения представляют собой обобщение объяснений, полученных разными методами [Speith, 2022; Ronge et al., 2025].

По масштабу объяснения разделяются на локальные, которые интерпретируют предсказание для отдельного или ограниченного числа примеров, и глобальные, раскрывающие общее поведение модели на всем пространстве данных [Speith, 2022; Ali et al., 2023; Ronge et al., 2025].

Результат

Основания классификации объяснимого ИИ по результату включают форму самого результата, форму презентации, а также метрику.

По форме результата объяснения могут представлять собой примеры входных данных, репрезентативно отражающие закономерности работы модели, контрфактические примеры, выявленные правила, суррогатные модели, значимость признаков и снижение размерности [Schwalbe, Finzel, 2024; Ronge et al., 2025].

Форма презентации — способ представления объяснений, который может быть текстовым, цифровым, визуальным, аудиальным и комбинированным [Schwalbe, Finzel, 2024; Ronge et al., 2025].

Классификация метрик объяснимого искусственного интеллекта основана на уровне вовлечения человека в их оценку и делится на три группы по следующим основаниям: функциональные, основанные на человеческих оценках и прикладные. Функциональные метрики не требуют участия человека и измеряют формальные свойства объяснения, соответствующие принципам технической достоверности. Метрики на основе человеческих оценок позволяют измерить соответствие принципам ориентации на конечного пользователя. Прикладные метрики ориентированы на оценку объяснений в реальных условиях работы системы с пользователем, включая удовлетворенность, воздействие на решения и улучшение общей производительности «человек – ИИ». Примером являются меры улучшения доверия к системе и повышение качества совместной диагностики у медицинских специалистов, а также оценка возможности автоматизации рутинных действий. Таким образом, прикладные метрики, выделяемые в работе Швальбе и Финцеля, не соответствуют однозначно определенной из предложенной нами групп принципов XAI, отражая преимущественно идею социальной ответственности, но в то же время затрагивая и другие группы принципов [Schwalbe, Finzel, 2024].

Методы XAI для классического машинного обучения

Существует ряд методов объяснимого искусственного интеллекта для классических алгоритмов машинного обучения. Одним из больших преимуществ этих алгоритмов является то, что признаки, на которых строится модель, являются исходно интерпретируемыми, так как они приходят чаще всего из реальных табличных данных. Таким образом, методы объяснимого ИИ должны на выходе выдавать нам ранжирование исходных признаков по уровню их влияния на вывод модели.

Простейшие алгоритмы машинного обучения, такие как линейная регрессия, допускают использование соответствующих весов для признаков для определения степени важности признаков на вывод модели. Наибольшие по модулю веса соответствуют самым важным нормированным признакам.

Большинство проблем классического машинного обучения не допускают приемлемого объяснения с помощью только линейных моделей и требуют применения более сложных нелинейных алгоритмов. Одним из таких алгоритмов являются решающие деревья. Несмотря на то что для решения большинства задач решающие деревья используются как составная часть более сложного алгоритма (бустинг, бэггинг и другие алгоритмы), они обладают одним интересным свойством. Логическая структура решающего дерева сама по себе является объяснением того, что повлияло на результат. Дополнительные статистические оценки могут быть использованы для оценки важности каждой вершины решающего дерева.

Как уже упоминалось ранее, бустинг-алгоритмы являются примером сложных составных алгоритмов машинного обучения. Несмотря на то что базовые модели, на которых строится

бустинг, могут обладать хорошей интерпретируемостью (решающие деревья, логистическая регрессия), алгоритм целиком теряет эти свойства и требует применения новых методов объяснимого искусственного интеллекта. Наиболее распространенными такими методами являются SHAP и LIME. Остановимся на каждом из них подробнее.

SHAP

Задачей объяснимого искусственного интеллекта назовем задачу ранжирования исходных признаков в порядке убывания по степени их важности для модели. SHAP (SHapley Additive exPlanations) [Lundberg, Lee, 2017], как теоретико-игровой алгоритм, предлагает оценивать важность признака, учитывая отклонение предсказания модели при замене значения признака для выбранного сэмпла на его среднее значение по датасету. Метод основан на концепции значений Шепли из кооперативной теории игр, где каждый признак рассматривается как игрок, а его вклад в предсказание оценивается через все возможные коалиции признаков. Формально значение Шепли для признака i определяется как

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)], \tag{1}$$

где N — множество всех признаков, $n = |N|$ — общее число признаков, S — подмножество признаков без i, $f(S)$ — предсказание модели при использовании только признаков из S (остальные заменяются на базовые значения). Тем самым SHAP дает информацию о важности признака для каждого сэмпла из датасета и впоследствии агрегирует эту важность по всему датасету для выбранного признака (наиболее распространены версии с усреднением и максимальным абсолютным отклонением).

SHAP является теоретически обоснованным методом, удовлетворяющим трем фундаментальным аксиомам Шепли.

- *Локальная точность*. Сумма всех значений SHAP равна разности между предсказанием модели и базовым значением: $f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$.
- *Симметрия*. Если два признака вносят одинаковый вклад во все коалиции, их значения SHAP равны.
- *Аддитивность* (*линейность*). Для комбинации моделей значения SHAP суммируются.

Метод является модель-агностичным, хотя существуют оптимизированные версии для деревьев (TreeSHAP). SHAP предоставляет как локальные объяснения для отдельных предсказаний, так и глобальные объяснения через агрегацию.

При этом критика данного метода указывает на то, что относительная важность признака, полученная с помощью точных Shapley-значений для моделей классификации, не отражает практического влияния признака. Это демонстрируется на простых логических, многозначных и дискретных классификаторах, где вычисленные точные значения SHAP явно вводят в заблуждение, поскольку они придают важность признакам, которые не влияют на прогнозируемый результат, и не придают значения признакам, которые на самом деле являются критическими для прогнозируемого класса [Huang, Marques-Silva, 2024].

Визуализация результатов SHAP включает несколько типов графиков: сводный график важности для отображения важности всех признаков на всем датасете, диаграмму вкладов для детального объяснения отдельного предсказания, а также график зависимости для анализа зависимости влияния признака от его значения [Lundberg, Lee, 2017]. Пример визуализации результатов SHAP для табличных данных представлен на рис. 1.

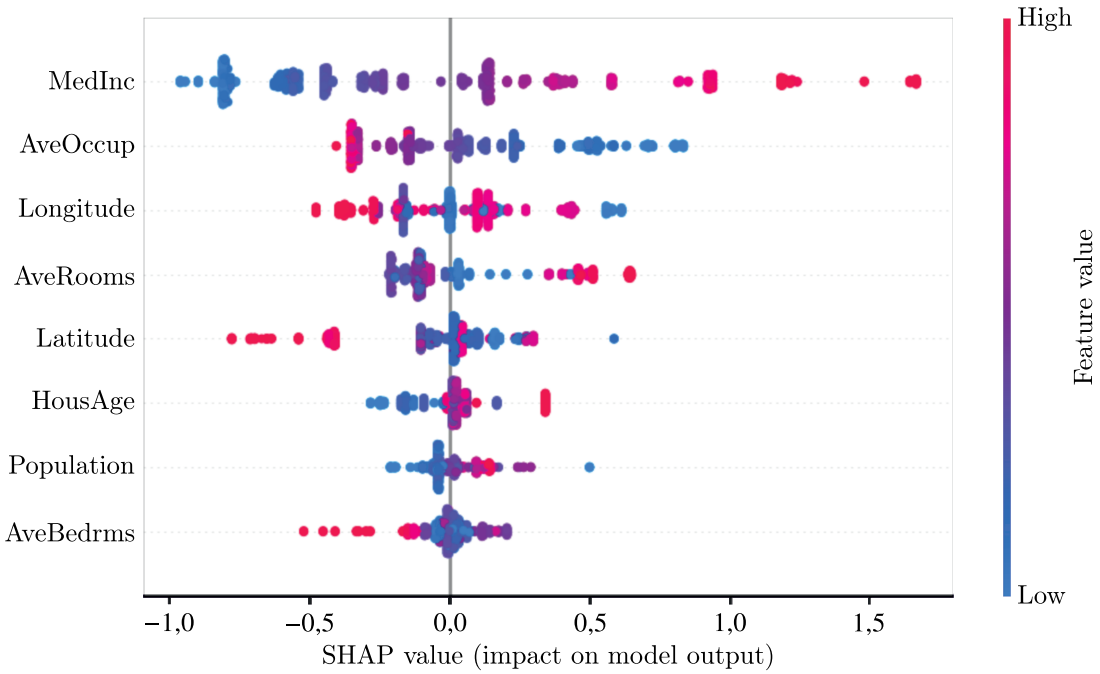


Рис. 1. Пример визуализации результатов SHAP для табличных данных (California Housing Dataset): сводный график важности признаков (summary plot), показывающий распределение значений SHAP для каждого признака по всему датасету. Каждая точка представляет значение SHAP для конкретного примера, цвет точки соответствует значению признака (красный — высокое, синий — низкое). Признаки: MedInc — медианный доход в районе, HouseAge — медианный возраст дома, AveRooms — среднее число комнат на домохозяйство, AveBedrms — среднее число спален на домохозяйство, Population — население района, AveOccup — среднее число членов домохозяйства, Latitude и Longitude — географические координаты района. Признаки ранжированы по средней абсолютной величине значений SHAP [Salih et al., 2024].

LIME

Другим популярным алгоритмом объяснимого ИИ является LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al., 2016]. В его основе лежит идея о том, что сложный глобальный алгоритм можно локально приблизить с помощью более простого интерпретируемого алгоритма. Поэтапно он формулируется следующим образом.

1. Выбирается некоторая точка из датасета.
2. Генерируются синтетические данные около этой точки и выход модели на этих точках.
3. Более интерпретируемая модель используется для предсказания этого локального датасета. Важно, что веса сэмплов в этом датасете прямо зависят от степени удаленности сгенерированной точки от исходной (определяется через ядро близости).
4. На основе построенной упрощенной модели строится интерпретация важности признаков.

Стоит заметить, что если на первом этапе выбирается точка, которая неадекватно представляет локальную границу решения, то объяснения могут быть недостаточно конкретными для объясняемого случая. Кроме того, шаг генерации синтетических данных неизбежно вносит степень нестабильности в различные запуски этого алгоритма. То есть разные запуски для одной и той же модели и датасета могут давать немного отличные интерпретации влияния признаков.

Третий этап связан с риском неточного отражения суррогатной моделью исходного сложного алгоритма. Наконец, отмечаются проблемы вычислительной эффективности, связанной с затратами на каждый из этапов, а также интерпретируемости конечного результата [Knab et al., 2025].

LIME является полностью модель-агностичным методом, что обеспечивает его широкую применимость к различным типам моделей и данных (табличные данные, тексты, изображения). Однако его локальная природа означает, что каждое объяснение относится только к одному предсказанию за раз. Визуализация результатов LIME включает барграфы важности признаков, а для текстовых данных и изображений — подсветку важных слов или выделение значимых областей [Ribeiro et al., 2016]. Пример визуализации результатов LIME для текстовых данных показан на рис. 2.

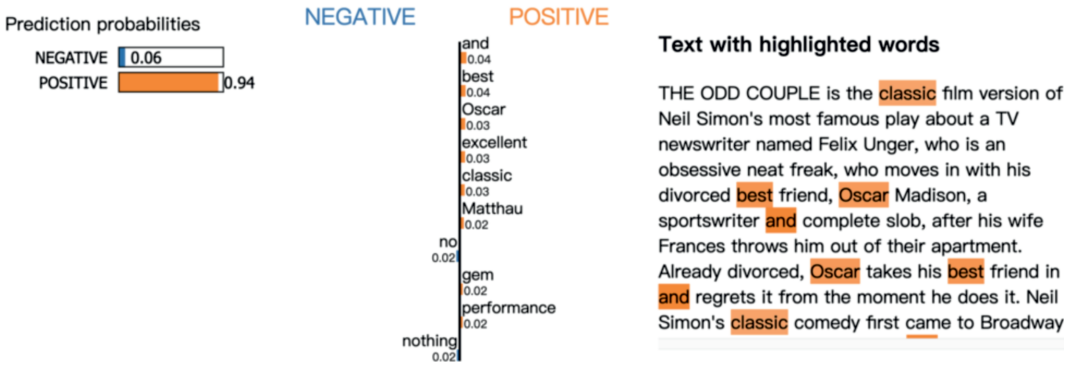


Рис. 2. Пример визуализации результатов LIME для анализа тональности текста: объяснение предсказания положительного sentiment для отзыва. Слова, выделенные оранжевым цветом (classic, best, excellent), вносят наибольший вклад в классификацию текста как положительного отзыва, что демонстрируется подсветкой и соответствующими весами важности каждого слова [Hsieh et al., 2024]. © 2024 Hsieh et al., воспроизведено с разрешения авторов

Сравнение SHAP и LIME

Систематическое сравнение SHAP и LIME по критериям предложенной классификации представлено в таблице 1. SHAP дает строгую теоретическую основу (значения Шепли), а LIME — больше практический инструмент для наглядных, локальных объяснений, но с меньшей стабильностью и строгой обоснованностью.

Оба метода находят широкое применение в различных областях. SHAP активно используется в медицине для объяснения предсказаний риска заболеваний, в финансовом секторе для обоснования решений о кредитовании, а также в промышленности для диагностики оборудования. LIME особенно эффективен в задачах, требующих быстрых объяснений отдельных решений, таких как модерация контента или объяснение рекомендательных систем конечным пользователям.

Методы XAI для компьютерного зрения

Задачи машинного зрения обладают своей спецификой в контексте возможных методов объяснимого искусственного интеллекта. Базовым объектом (признаком) в этих задачах являются отдельные пиксели, которые сами по себе не несут достаточной информации относительно их влияния на предсказание модели. Объяснение в контексте компьютерного зрения обычно представляет собой визуализацию областей или признаков изображения, наиболее значимых для

Таблица 1. Сравнение методов SHAP и LIME

Группа	Признак	SHAP	LIME
Постановка задачи	Цель	Извлечение знаний; укрепление доверия; обнаружение ошибок	Локальная интерпретация; укрепление доверия; обнаружение ошибок
	Целевая аудитория	Исполнители, разработчики; частично пользователи и регуляторы (через визуализации)	Пользователи, исполнители, разработчики; регуляторы (локальная проверка)
	Задача МО	Универсален (классификация, регрессия и др.)	Универсален (классификация, регрессия, тексты, изображения)
	Тип входных данных	Табличные (лучше всего), тексты и изображения (через адаптации)	Табличные, тексты, изображения (через адаптации)
Методология	Стадия применения	Post-hoc	Post-hoc
	Модель-специфичность	Модель-агностичный (есть оптимизации для деревьев: TreeSHAP)	Полностью модель-агностичный
	Методы	Аддитивные значения Шепли (теоретически обоснованные аксиомами)	Локальное приближение сложной модели простой интерпретируемой
	Масштаб	Локальный (основной), агрегируется до глобального	Локальный (одно предсказание за раз)
Результат	Форма результата	Вектор значений Шепли (разложение предсказания)	Набор важных признаков с весами
	Форма презентации	Сводный график важности, диаграмма вкладов, график зависимости, таблицы	Барграфы, подсветка слов, выделение областей изображений
	Метрика	Теоретическая строгость (аксиомы Шепли: локальная точность, симметрия, аддитивность)	Локальная точность аппроксимации (качество приближения интерпретируемой модели)

принятого моделью решения. Методы объяснимого ИИ для компьютерного зрения преимущественно являются модель-специфичными и основаны на анализе внутренней структуры нейронных сетей, в частности сверточных нейронных сетей (CNN).

Методы на основе градиентов

Gradient-based-методы используют информацию о градиентах функции потерь по отношению к входным данным для определения важности отдельных пикселей или областей изображения. Простейший метод — визуализация градиента (Gradient Visualization) — вычисляет градиент выходного класса по входному изображению, показывая, какие пиксели при небольшом изменении сильнее всего влияют на предсказание модели [Simonyan et al., 2014].

SmoothGrad [Smilkov et al., 2017] является улучшением базового градиентного метода и решает проблему визуального шума в картах важности. Метод добавляет небольшой гауссовский шум к входному изображению, вычисляет градиенты для множества зашумленных версий и усредняет полученные карты важности. Это позволяет получить более гладкие и визуально интерпретируемые результаты, снижая влияние локальных артефактов.

Integrated Gradients [Sundararajan et al., 2017] — теоретически обоснованный метод атрибуции признаков, который вычисляет интеграл градиентов вдоль прямолинейного пути от базового изображения (обычно черного или размытого) до исходного входного изображения. Формально атрибуция для i-го признака определяется как

$$IG_i(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha, \tag{2}$$

где x — входное изображение, x' — базовое изображение, F — функция модели. Метод удовлетворяет двум важным аксиомам: чувствительности (если признак влияет на выход, ему присваивается ненулевая важность) и полноты реализации $\left(\text{сумма атрибуций равна разности между предсказанием для входа и базового изображения: } \sum_{i=1}^{n} IG_i(x) = F(x) - F(x')\right)$. Эти свойства делают Integrated Gradients более надежным по сравнению с простыми градиентными методами.

Методы визуализации активаций

Class Activation Mapping (CAM) [Zhou et al., 2016] является методом, специфичным для архитектур со слоем Global Average Pooling (GAP) перед финальным полносвязным слоем. Метод использует веса финального классификационного слоя для взвешенной линейной комбинации карт признаков последнего сверточного слоя, что позволяет получить карту активаций размером с входное изображение, показывающую области, важные для предсказания конкретного класса. Главное ограничение CAM — требование определенной архитектуры сети с GAP.

Grad-CAM [Selvaraju et al., 2017] (Gradient-weighted Class Activation Mapping) обобщает идею CAM [Zhou et al., 2016] на произвольные архитектуры CNN без требования наличия GAP. Метод использует градиенты целевого класса, распространяющиеся обратно в выбранный сверточный слой, для получения весов важности каждой карты признаков. Взвешенная комбинация карт признаков с последующим применением функции усеченного линейного преобразования (ReLU) дает тепловую карту, показывающую области изображения, положительно влияющие на предсказание класса. Grad-CAM может применяться к любой дифференцируемой архитектуре и является одним из наиболее популярных методов визуализации в компьютерном зрении.

Grad-CAM++ [Chattopadhay et al., 2018] улучшает Grad-CAM за счет более точного взвешивания градиентов, особенно в ситуациях, когда в изображении присутствует несколько экземпляров объекта одного класса. Метод использует взвешенную комбинацию положительных частных производных второго и третьего порядка для лучшей локализации объектов.

Методы на основе возмущений

RISE (Randomized Input Sampling for Explanation) [Petsiuk et al., 2018] — модель-агностичный метод, основанный на случайном сэмплировании и маскировании входных данных. Метод генерирует большое количество случайных бинарных масок, применяет их к входному изображению и получает предсказания модели для каждого замаскированного изображения. Карта важности строится как взвешенная линейная комбинация масок, где веса соответствуют уверенности модели в предсказании целевого класса. RISE не требует доступа к градиентам и внутренней структуре модели, что делает его применимым к любым моделям, включая ансамбли.

Occlusion Sensitivity Analysis [Zeiler, Fergus, 2014] систематически закрывает различные области изображения (обычно скользящим окном) и измеряет падение уверенности модели в предсказании. Области, при закрытии которых уверенность падает сильнее всего, считаются наиболее важными для решения модели. Несмотря на простоту концепции, метод вычислительно затратен, так как требует множества прямых проходов через сеть.

Концептуальные объяснения

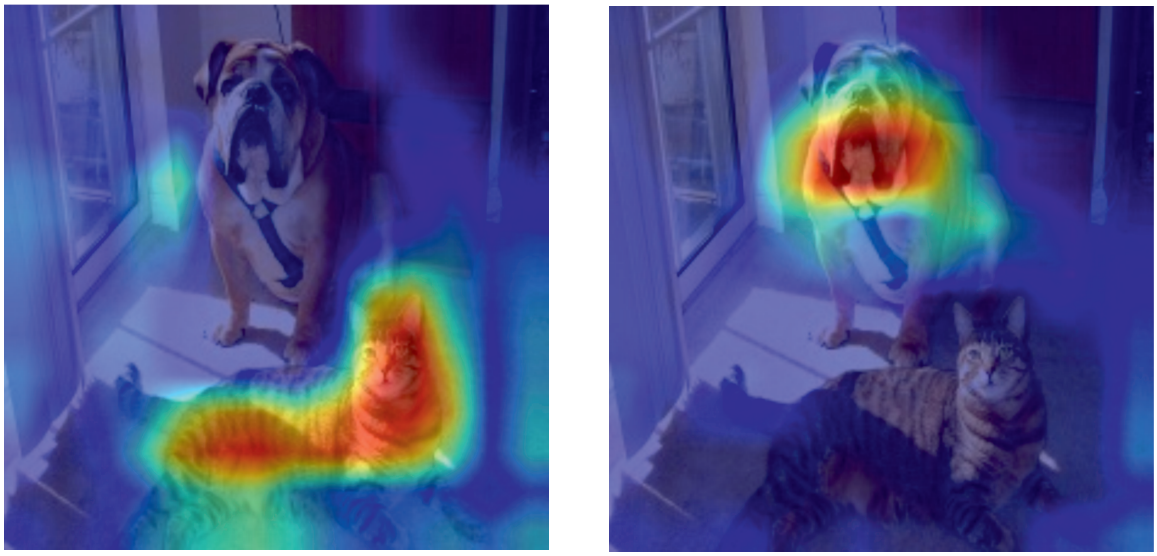
Testing with Concept Activation Vectors (TCAV) [Kim et al., 2018] представляет собой метод высокоуровневых концептуальных объяснений, выходящий за рамки пиксельной атрибуции. TCAV позволяет количественно оценить важность определенных человекопонятных концептов (например, «полоски», «текстура», «цвет») для предсказаний модели путем вычисления векторов активации концептов (CAV) во внутренних слоях сети. Пользователь предоставляет примеры, иллюстрирующие интересующий концепт, и TCAV обучает линейный классификатор в пространстве активаций, разделяющий примеры концепта от случайных примеров. Метод затем использует направленную производную выхода модели вдоль CAV для определения чувствительности предсказания к данному концепту.

Network Dissection [Bau et al., 2017] автоматически идентифицирует семантические концепты, закодированные в отдельных нейронах сверточных слоев. Метод сопоставляет активации нейронов с набором размеченных визуальных концептов (объекты, части, текстуры, цвета и т. д.) и количественно оценивает, насколько хорошо каждый нейрон детектирует определенный концепт. Это позволяет интерпретировать роль отдельных нейронов в терминах человекопонятных визуальных категорий.

Сравнение и применение методов

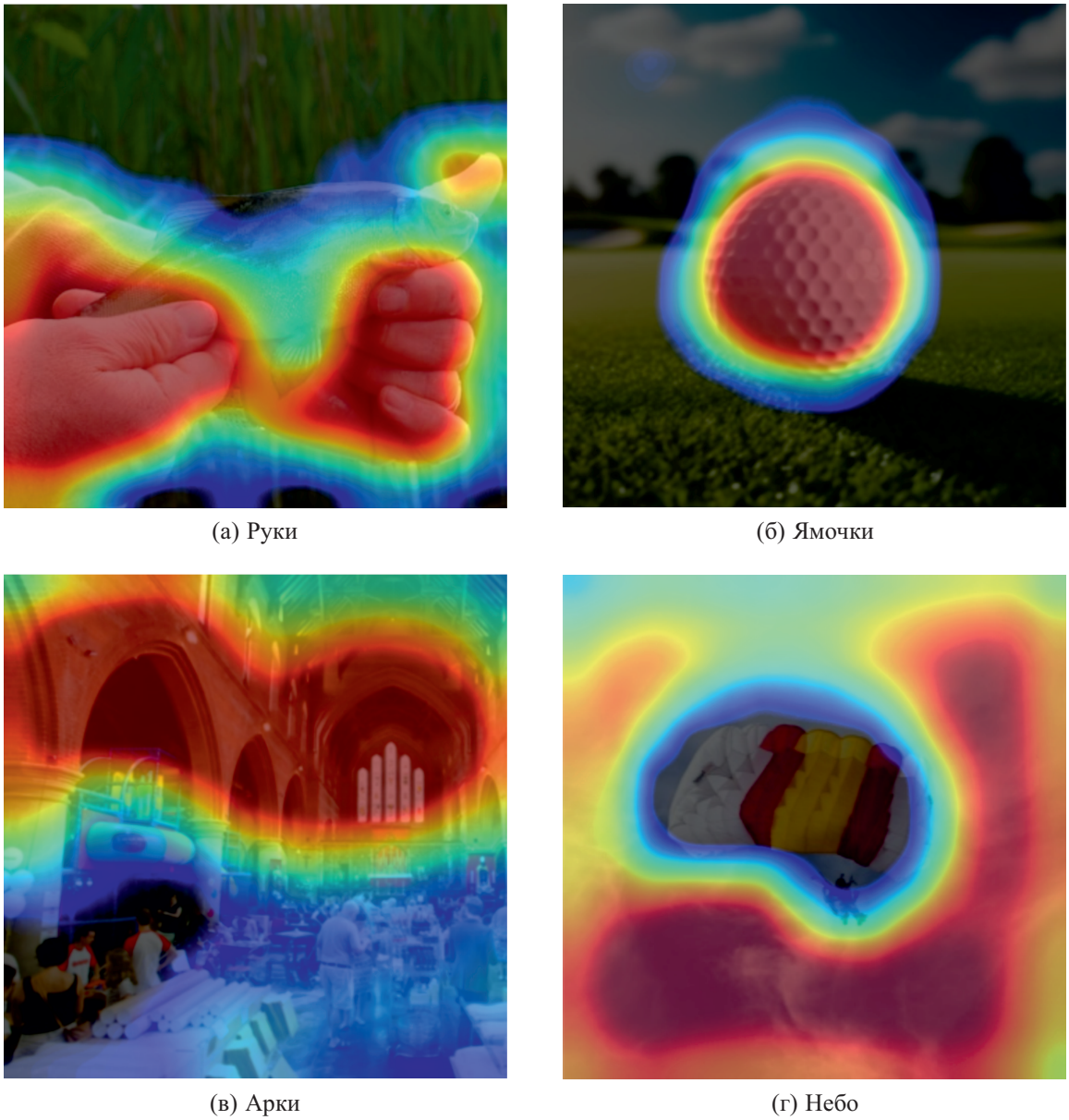
Методы XAI для компьютерного зрения различаются по вычислительной сложности, требованиям к доступу к модели и уровню детализации объяснений. Систематическое сравнение представителей основных групп методов представлено в таблице 2.

Градиентные методы (SmoothGrad, Integrated Gradients) обеспечивают пиксельный уровень точности, но требуют доступа к градиентам. Методы визуализации активаций (Grad-CAM, Grad-CAM++) дают более грубую локализацию на уровне областей, но являются визуально более интерпретируемыми и вычислительно эффективными (рис. 3). RISE обеспечивает баланс между



(а) (б)

Рис. 3. Примеры визуализации Grad-CAM для изображений животных: (а) тепловая карта для класса «кот»; (б) тепловая карта для класса «собака». Красные области показывают регионы изображения, наиболее важные для предсказания соответствующего класса нейронной сетью. Визуализация выполнена с использованием библиотеки pytorch-grad-cam pytorch-grad-cam (https://github.com/jacobgil/pytorch-grad-cam), лицензия MIT



(а) Руки (б) Ямочки

(в) Арки (г) Небо

Рис. 4. Примеры визуализации концептов для метода TCAV: (а) концепт «руки» (hands); (б) концепт «ямочки» (dimples); (в) концепт «арки» (arches); (г) концепт «небо» (sky). Каждое изображение демонстрирует примеры визуальных паттернов, используемых для обучения векторов активации концептов (CAV), которые позволяют количественно оценить влияние этих человекопонятных концептов на предсказания модели [De Santis et al., 2024].

модель-агностичностью и качеством объяснений, но требует значительных вычислительных ресурсов. Концептуальные методы (TCAV, Network Dissection) предоставляют объяснения на более высоком семантическом уровне, что особенно ценно для коммуникации с конечными пользователями, не обладающими техническими знаниями (рис. 4).

Эти методы широко применяются в медицинской диагностике для визуализации областей изображений, на основе которых модель делает диагностические заключения, в автономном вождении для понимания решений систем восприятия, в системах безопасности для объяснения детекции аномалий, а также в научных исследованиях для анализа того, какие признаки изображений использует модель для классификации.

Таблица 2. Сравнение основных групп методов XAI для компьютерного зрения

Признак	SmoothGrad (градиентные)	Grad-CAM (активации)	RISE (возмущения)	TCAV (концепты)
Уровень объяснения	Пиксельная атрибуция	Локализация областей	Пиксельная/региональная атрибуция	Концептуальный
Модель-специфичность	Модель-специфичный (требует градиенты)	Модель-специфичный (требует градиенты и активации)	Модель-агностичный (black-box)	Модель-специфичный (требует активации)
Вычислительная сложность	Средняя (множественные проходы с шумом)	Низкая (один обратный проход)	Высокая (тысячи масок)	Средняя (обучение CAV + + производные)
Визуальная интерпретируемость	Высокая (детальные карты)	Очень высокая (тепловые карты)	Высокая (карты важности)	Средняя (количественные оценки концептов)
Теоретическое обоснование	Эмпирическое улучшение базового метода	Градиентное взвешивание активаций	Вероятностное сэмплирование	Направленные производные в пространстве концептов
Стабильность результатов	Высокая (усреднение по шуму)	Высокая (детерминированный)	Средняя (стохастический)	Высокая (статистическая значимость)
Масштаб объяснения	Локальный (одно изображение)	Локальный (одно изображение, один слой)	Локальный (одно изображение)	Глобальный (класс/набор данных)
Требования к пользователю	Технические знания	Минимальные (визуальный выход)	Минимальные (визуальный выход)	Предоставление примеров концептов
Применимость	Любые дифференцируемые модели	CNN и подобные архитектуры	Любые модели (включая ансамбли)	CNN с доступом к активациям
Форма презентации	Карты важности пикселей	Тепловые карты поверх изображения	Карты важности (saliency maps)	Графики чувствительности к концептам

Методы XAI для обработки естественного языка

Объяснимая обработка естественного языка (XNLP) подразумевает применение методов XAI как к традиционным методам NLP, например BoW в различных вариациях, так и к методам, основанным на эмбедингах, включая трансформеры, в частности и большие языковые модели (LLM). С одной стороны, с усложнением архитектуры NLP-методов открываются и реализуются новые возможности решения практических задач, но с другой — стремительно снижается прозрачность моделей естественной обработки языка, вследствие чего мы сталкиваемся с большими вызовами в области объяснимости. Если BoW, TF, TF-IDF позволяли выявлять влияние отдельных слов на результат прогнозирования благодаря прозрачным классификаторам, то уже для объяснения первых моделей, использовавших эмбеддинги, разрабатывались такие методы, как визуализация векторных пространств, а также методы на основе градиента и механизма внимания. Дальнейший переход к трансформерам открыл новые возможности для решения практических задач, но еще больше снизил прозрачность моделей. Чтобы объяснить их работу, начали использовать визуализацию весов внимания, проверочные задания, визуализацию призна-

ков, атрибутивные методы, а также упрощение и дистилляцию моделей — обучение небольших интерпретируемых ученических сетей [Mohammadi et al., 2024].

Специфика применения объяснимого ИИ к большим языковым моделям заключается в нескольких ключевых аспектах. Помимо крайне высокой степени непрозрачности, модели сталкиваются с проблемой подверженности искажениям и предвзятости входных данных, обусловленной обучением на больших и разнородных интернет-ресурсах, что ведет к непредсказуемым и потенциально нежелательным последствиям вывода. Во-вторых, применение LLM широким кругом пользователей порождает риски неправильного понимания работы моделей и возможность злоупотреблений их генеративными способностями. В-третьих, обучение и использование LLM влечет за собой значительные вычислительные затраты, создающие барьеры для небольших исследовательских групп, что тормозит развитие XAI-методов применительно к LLM [Hsieh et al., 2024; Herrera, 2025]. В-четвертых, традиционные методы XAI зачастую оказываются малоэффективными или вовсе неприменимыми к LLM из-за их сложности и масштабности, что требует новых подходов и специализированных методик объяснения. На этой последней проблеме мы остановимся для подробного рассмотрения.

Ву с соавторами выделяет 7 групп методов XAI, применимых к LLM: атрибутивные методы; интерпретация параметров модели; объяснение на основе выборок; объяснимость как проверка надежности; объяснимость в промтинг-парадигме; объяснимость, основанная на промтинге, дополняющем знания LLM; объяснения, основанные на аугментации обучающей выборки. Методы каждой из этих групп обладают своей проблематикой, связанной с определенными вызовами.

Применительно к LLM атрибутивные методы позволяют не только повысить прозрачность моделей, но и улучшить их качество, обнаруживая потенциальные проблемы и уязвимости. Среди основных типов атрибутивных методов можно выделить несколько групп. Методы на основе пертурбаций оценивают важность входных признаков путем их изменения и анализа влияния на достоверность прогноза. Однако у них есть существенные ограничения: предположение о независимости признаков, которому зачастую не соответствуют текстовые данные, а также высокие требования к вычислительным ресурсам. Объяснение на основе градиентов позволяет вычислить релевантность признаков прогнозируемого текста относительно заданного входного слова. Проблема здесь заключается в дискретности вектора признаков, из-за чего разработанные для изображения градиентные методы XAI требуют адаптации к NLP, включая LLM. Суррогатные методы, предполагающие создание упрощенной модели для интерпретации работы целевой модели, сталкиваются с проблемами многократного взаимодействия с целевой моделью и трудностями в поиске адекватной простой модели для сложных языковых систем. Методы декомпозиции присваивают входным данным линейно-аддитивные оценки релевантности и эффективно разбивают прогноз на составные части, но требуют индивидуальных стратегий для различных архитектур моделей, что ограничивает их универсальность [Mumuni et al., 2025].

Применение атрибутивных методов к LLM охватывает два ключевых направления: диагностику модели, которая заключается в сравнении выходных данных с результатами атрибуции для выявления необоснованной или неподтвержденной информации, и улучшение модели: использование результатов атрибуции как дополнительной информации для уточнения запросов и повышения общей эффективности LLM. Тем не менее использование атрибутивных методов в работе с LLM сопряжено с рядом проблем: высокой вычислительной сложностью, достоверностью атрибуций и необходимостью разработки новых парадигм объяснения — динамичный и разнообразный характер генерации LLM требует выхода за рамки существующих методов атрибуции и создания принципиально новых подходов.

Применение SHAP и LIME к текстовым данным

Методы SHAP и LIME, первоначально разработанные для табличных данных и рассмотренные в контексте классического машинного обучения, получили широкое распространение в области обработки естественного языка благодаря своей модель-агностичности [Salih et al., 2024]. Однако их применение к текстовым данным требует специфической адаптации, учитывающей дискретную природу языка и последовательную структуру текста.

Адаптация SHAP для NLP заключается в трактовке токенов (слов, подслов или символов) как признаков, для которых вычисляются значения Шепли. В отличие от табличных данных, где признаки независимы, в тексте существуют сложные зависимости между токенами: порядок слов, синтаксические конструкции и контекстуальные связи. Для работы с текстом используются специализированные версии SHAP. Partition SHAP группирует токены в более крупные смысловые единицы (иерархические партиции) для снижения вычислительной сложности, вычисляя значения Оуэна вместо полных значений Шепли, что делает расчеты более tractable за счет сокращения числа рассматриваемых коалиций [Chen et al., 2020]. При объяснении предсказаний модели SHAP вычисляет вклад каждого токена в итоговый результат, создавая распределение важности по всей входной последовательности. Значения Шепли для токенов могут быть как положительными (токен увеличивает вероятность предсказанного класса), так и отрицательными (токен уменьшает эту вероятность).

LIME для текстовых данных работает через пертурбации входной последовательности: метод создает множество вариантов исходного текста, удаляя или заменяя отдельные слова, и анализирует, как эти изменения влияют на предсказание модели. Генерируются синтетические примеры путем случайного исключения токенов из исходного текста, формируя окрестность вокруг объясняемого примера. Для каждого синтетического примера получается предсказание исходной модели, и вес примера определяется его близостью к исходному тексту (обычно через косинусное сходство эмбеддингов или простое совпадение токенов). На этих взвешенных синтетических данных обучается простая интерпретируемая модель (например, линейная регрессия), веса которой интерпретируются как важность соответствующих токенов. Особенностью LIME для текста является необходимость определения единицы пертурбации: это могут быть отдельные слова, n-граммы или даже предложения (в зависимости от задачи и уровня желаемой детализации объяснения).

Оба метода нашли применение в разнообразных задачах NLP. В анализе тональности SHAP и LIME используются для объяснения классификации отзывов, выявляя конкретные слова и фразы, которые определяют позитивную или негативную окраску текста. Это особенно ценно для бизнес-аналитики, где важно понимать не только общую тональность отзывов о продукте, но и конкретные аспекты, вызывающие недовольство или удовлетворение клиентов. В задачах детекции токсичности и модерации контента методы помогают идентифицировать оскорбительные выражения и паттерны речи ненависти, что критично для обеспечения безопасности онлайн-платформ. При классификации текстов по тематике (например, категоризация новостных статей или научных публикаций) SHAP и LIME выявляют ключевые термины и концепты, характерные для каждой категории [Chen et al., 2020]. В системах детекции дезинформации методы позволяют понять, какие лингвистические признаки и семантические паттерны указывают на потенциально ложную информацию.

Сравнение производительности SHAP и LIME на текстовых данных показывает компромисс между стабильностью и вычислительной эффективностью [Salih et al., 2024]. SHAP обеспечивает более стабильные и теоретически обоснованные объяснения благодаря аксиоматическому подходу, основанному на теории кооперативных игр, однако требует значительных вычислительных ресурсов, особенно для длинных текстов с большим количеством токенов. Вычислительная

сложность SHAP растет экспоненциально с числом признаков, что делает его применение к текстам из сотен или тысяч слов практически невозможным без аппроксимаций. LIME, в свою очередь, работает значительно быстрее благодаря локальной аппроксимации, но может демонстрировать нестабильность: повторные запуски на одном и том же примере могут давать различающиеся объяснения из-за случайной природы генерации пертурбаций. Для критических приложений, требующих надежных и воспроизводимых объяснений (например, медицинская диагностика по клиническим текстам), предпочтителен SHAP, тогда как для быстрого прототипирования и интерактивной отладки моделей LIME может быть более практичным выбором.

Также было показано, что объяснения LIME не только менее надежные, но и менее точные в обычных условиях. Однако было показано, что при злонамеренном вмешательстве в функцию активации NLP-классификатора точность LIME, как правило, превышает этот показатель для SHAP [Ali et al., 2022].

Кроме того, внутренняя ненадежность LIME применительно к текстовым данным усиливается зашумленностью естественного языка, в котором один и тот же смысл передается множеством разных способов. В работе Бюргера и коллег были смоделированы условия для демонстрации этих уязвимостей. В первом случае для определения выраженности внутренней нестабильности LIME варьировался размер датасета при неизменном документе. При этом среднее падение сходства объяснений, полученных с помощью трех разных моделей, для выборок размером 1000–7000 сравнительно с исходной выборкой объемом 5000 оказалось равным от 18 до 28,5. Во втором случае решалась задача оптимизации, которая заключалась в том, чтобы при данном документе d_b, целевой модели f, пороговом значении семантической близости δ, числе изменений в документе ϵ и k наиболее важных признаков необходимо найти такой документ d_p, что

$$d_p = \arg\min_{d_p} \mathrm{Sim}_e\left(e_{d_b}, e_{d_p}\right), \tag{3}$$

при условиях $f(d_b) = f(d_p)$, $\mathrm{Sim}_s(d_b, d_p) \geqslant \delta$, $i \leqslant \epsilon \cdot |f|$, $e_{d_p} \cap c \neq \varnothing \; \forall c \in e_{d_b}[:k]$.

При семантической близости между исходным и атакующим документами на уровне 0,89–0,91 метод XAIFOOLER показал следующие результаты манипуляции объяснениями LIME относительно базового уровня внутренней нестабильности: абсолютное изменение рангов значимых признаков выросло на 128 %, метрика деградации ранговой согласованности (ΔRC) увеличилась на 101 %, а пересечение между k наиболее важными признаками сократилось на 14,6 % [Burger et al., 2023].

Другая сложность, возникающая в задаче обработки естественного языка, заключается в том, что минимальное изменение смысла текста может привести к необоснованному изменению предсказания модели. В таком случае необходимое требование к точности метода объяснимого ИИ — адверсариальная чувствительность, которая определяется тем, что метод XAI для схожих текстов с разными метками дает также разные объяснения, тем самым отражая внутреннюю логику работы модели. Оказалось, что к такого рода воздействиям относительно градиентных методов как SHAP, так и LIME, оба основанные на пертурбации, кратно более устойчивы. Более того, Gradient, Integrated Gradient, Gradient × Input уступают LIME и SHAP при сравнении по соотношению чувствительности к изменению модели при минимальной разнице во входных данных и достоверности объяснений, определенной через удаление признаков, определенных XAI как наиболее важные. Однако если смотреть на абсолютные метрики адверсариальной точности (то есть точности предсказаний при целенаправленных минимальных возмущениях входных данных), то в зависимости от уровня воздействия и датасета значения составили от 0,64 до 0,83 для LIME и от 0,61 до 0,84 для SHAP, то есть в отдельных случаях уязвимость была значительной [Manna, Sett, 2024].

Интерпретация параметров модели заключается в анализе внутренних компонентов LLM, таких как слои внимания и нейроны сетей прямого распространения (feed-forward neurons), а также их взаимодействий. Этот подход позволяет локализовать, какие части модели ответственны за выполнение определенных задач или кодируют конкретные знания, что дает возможность для редактирования знаний и упрощения модели при сохранении производительности.

Методы на основе механизма внимания

Особое место среди методов интерпретации параметров занимают подходы, основанные на анализе механизма внимания (attention mechanism). В трансформерных архитектурах механизм внимания с множеством голов (multi-head attention) позволяет модели взвешивать важность различных токенов при обработке последовательности. Подходы к объяснению моделей обработки естественного языка, основанные на механизме внимания, предполагают анализ матриц весов внимания для выявления того, какие входные токены модель считает релевантными при генерации каждого выходного токена [Li et al., 2016].

Существует несколько стратегий использования механизма внимания для объяснения работы модели. Анализ весов отдельных голов внимания позволяет обнаружить специализацию: разные головы в multi-head attention могут фокусироваться на различных аспектах входной последовательности, таких как синтаксические зависимости, семантические связи или позиционные отношения между токенами. Агрегация весов внимания по слоям помогает отследить эволюцию представлений от низкоуровневых признаков к высокоуровневым концептам. Визуализация attention flow демонстрирует, как информация распространяется через последовательные слои трансформера, создавая наглядное представление о вычислительных путях в модели.

Некоторые исследования демонстрируют успешное применение анализа внимания для обнаружения лингвистических структур: определенные головы внимания в BERT и подобных моделях специализируются на выявлении синтаксических зависимостей, таких как отношения между подлежащим и сказуемым, или на разрешении кореференций. Это указывает на то, что механизм внимания действительно кодирует некоторую структурированную информацию о языке.

Однако эксперименты показывают существенные ограничения использования весов внимания как объяснения работы модели. Стандартные модули внимания NLP-моделей не всегда дают содержательных объяснений, поскольку веса внимания не всегда связаны с градиентными методами оценки важности признаков, то есть не отражают истинного вклада входов в предсказание. Кроме того, можно найти разные распределения весов внимания, которые при этом дают одинаковые предсказания модели, что указывает на отсутствие уникальности или исключительности объяснения через внимание [Jain, Wallace, 2019]. Это ставит под сомнение интерпретацию весов внимания как прямого объяснения того, почему модель делает определенный вывод.

В попытке преодолеть эти ограничения были разработаны гибридные методы, такие как взвешивание матриц внимания на основе градиентов (gradient-weighted attention), которые комбинируют информацию о весах внимания с градиентами функции потерь, чтобы получить более точную оценку важности токенов. Тем не менее вопрос о том, можно ли использовать механизм внимания как методологическую основу для XAI-методов, породил большую дискуссию в научном сообществе. Обобщение этой дискуссии представлено в работе Андирена Бибаля и соавторов, в которой заключено, что внимание может быть частью объяснения, но не единственным и не всегда основным элементом. Это связано с тем, что, с одной стороны, внимание связывает входные данные с остальной частью модели и обычно используется для локальных объяснений отдельных решений, однако внимание не всегда бывает достоверным объяснением внутреннего механизма модели, а иногда лишь правдоподобным для пользователя [Bibal et al., 2022]. Таким образом, веса внимания следует рассматривать как один из компонентов комплексного анализа работы модели, а не как самодостаточное объяснение.

Применение методов интерпретации параметров модели в целом позволяет не только глубже понять работу LLM, но и использовать полученные знания для редактирования знаний, удаления избыточных или нерелевантных компонентов для ускорения вывода и управления моделью на этапе вывода, направленного изменения скрытых представлений для достижения желаемых свойств ответов. Несмотря на значительный потенциал, методы интерпретации внутренних параметров сталкиваются с рядом вызовов, включая высокую вычислительную сложность, проблемы с достоверностью интерпретаций и необходимость разработки новых подходов для комплексного понимания разнообразных механизмов работы.

Объяснимость на основе частей входных данных использует влияние отдельных обучающих примеров для объяснения поведения модели. Анализируется, как изменение или удаление конкретных примеров из обучающего набора влияет на предсказания модели. Основные методы включают использование функций влияния, методов, основанных на эмбеддингах, и различные техники оценки вклада обучающих данных. Это помогает выявлять, какие данные действительно определяют поведение модели, служит инструментом для отладки и диагностики, а также повышает прозрачность моделей на уровне данных.

В случае группы методов объяснимости как проверки надежности для оценки качества и безопасности выводов LLM методы направлены на выявление проблем с подлинностью, честностью, отсутствием токсичности, нейтральностью и справедливостью ответов. Сюда входят проверки на галлюцинации, дублирование, предвзятость и другие проблемы, которые могут повлиять на надежность модели в реальных приложениях. Такой анализ помогает создавать более доверительные и этически безопасные AI-системы.

Естественно-языковые объяснения и цепочки размышлений

Поскольку большие языковые модели работают с естественным языком, появилась идея объяснять получаемые с помощью них решения также на естественным языке. Такой подход получил название промтинг-парадигмы и реализуется в методах chain-of-thought (CoT) — цепочках размышлений, а также в естественно-языковых пост-хок-объяснениях самой LLM. Цепочки размышлений представляют собой пошаговые рассуждения, которые используются LLM при генерации ответа и, как правило, повышают его продуктивность [Turpin et al., 2023]. А естественно-языковые пост-хок-объяснения, как можно понять из названия, создаются LLM по запросу пользователя для обоснования предыдущего ответа модели [Matton et al., 2025]. Таким образом, реализация принципов ориентации на конечного пользователя в промтинг-парадигме, как правило, оказывается на высоте, однако возникает проблема с технической достоверностью, в особенности с точностью таких объяснений.

Экспериментально показано, что в сконструированном предвзятом контексте возрастает частота неверных ответов, однако при этом как цепочки рассуждений [Turpin et al., 2023], так и естественно-языковые пост-хок-объяснения [Matton et al., 2025], во-первых, выглядят правдоподобно, а во-вторых, не содержат указание на внесенные искажения, которые и являются причиной ошибочных ответов. Причины неточности CoT, с одной стороны, связаны с неполнотой, искаженностью и противоречивостью хода мыслей авторов текстов, на которых обучена модель. С другой стороны, сам алгоритм обучения с подкреплением с человеческой оценкой RLHF (Reinforcement Learning with Human Feedback) заточен на те ответы, которые нравятся оценщикам, то есть на правдоподобные, а не на истинно верные [Turpin et al., 2023]. Эти же причины можно считать релевантными и для естественно-языковых пост-хок-объяснений.

Мэттон с коллегами предложили метод, который сравнивает концепции, указанные в объяснениях как важные для ответа, с теми, что действительно влияют на результат модели. Этот подход использует дополнительную LLM для генерации правдоподобных контрфактических вопросов, в которых изменяются значения определенных концепций во входных данных. Далее

с помощью иерархической байесовской модели количественно оценивается влияние этих концепций на ответы модели [Matton et al., 2025].

На материале простых арифметических задач, требующих пошаговые решения, был сделан вывод о том, что неявные поэтапные рассуждения, представленные через состояния модели, связанные с последним токеном в условии задачи, являются менее надежными, чем CoT, хотя и более быстрыми. Было показано, что неявные рассуждения имеют относительный успех только у LLM, обученных на скрытое вычисление последовательности. Из этого следует, что необученные неявному рассуждению модели могут полагаться на накопленные знания, а не на пошаговые рассуждения. И даже модели, обученные неявному рассуждению, демонстрируют снижение эффективности этих рассуждений при изменении формата задачи [Yu, 2024].

Однако мы не можем уверенно говорить о том, что CoT позволяет получить объяснения, релевантные работе модели. В исследовании Пфау и коллег утверждается, что повышение эффективности ответов LLM с использованием CoT связано не с тем, что CoT является механизмом, аналогичным рассуждениям вслух у человека, а с тем, что это ее использование позволяет задействовать дополнительные токены для вычисления на скрытом слое. Было продемонстрировано, что при решении некоторых задач, допускающих параллельные вычисления, замена цепочки размышления набором слов-филлеров, которые сами по себе не несут семантическую нагрузку, приводит к повышению производительности LLM. Из этого следует, что при наличии достаточного количества обучающих данных промежуточные токены между входными данными и ответом могут использоваться исключительно для повышения вычислительной мощности, а не для последовательного мышления [Pfau et al., 2024].

Таким образом, CoT, судя по всему, являются скорее лишь правдоподобными, но недостоверными пародиями на то, как рассуждает человек, приводя причинные обоснования своих решений, чем методом XAI для LLM, в котором воплощены все выделенные нами группы принципов объяснимого ИИ.

Выявленные проблемы с достоверностью явных цепочек размышлений стимулировали разработку методов более глубокого анализа внутренних вычислительных процессов моделей. Более глубокий анализ цепочек размышлений стал возможен благодаря развитию методов атрибуции на уровне внутренних представлений модели. Графы атрибуции (attribution graphs) представляют собой метод трассировки вычислительных шагов, которые модель использует для генерации ответа. В отличие от явных CoT, которые видит пользователь, графы атрибуции строят графы причинно-следственных взаимодействий между интерпретируемыми вычислительными единицами (features) внутри модели. Узлы в таких графах представляют концепты или вычислительные операции — интерпретируемые признаки, которые часто соответствуют семантическим понятиям, а ребра показывают каузальные взаимодействия между этими элементами в процессе генерации ответа.

Методология построения графов атрибуции включает несколько этапов. Сначала создается локальная модель замещения (local replacement model), аппроксимирующая вычисления исходной модели. Затем строятся графы атрибуции путем трассировки взаимодействий между признаками для получения выходных данных. Графы подвергаются прунингу для выделения наиболее важных вычислительных путей. Наконец, сформулированные гипотезы о работе модели валидируются через интервенционные эксперименты, в которых целенаправленно изменяются активации определенных признаков для проверки их каузального влияния на выход модели.

Критически важным открытием, полученным с помощью графов атрибуции, является то, что шаги рассуждения модели, выявленные через анализ внутренних представлений, часто существенно отличаются от явных цепочек размышлений в CoT. Модели могут использовать обратные рассуждения от целевого состояния (backward reasoning), задействовать метакогнитивные схемы для мониторинга собственного процесса решения задачи или применять множественные

параллельные вычислительные пути, которые не отражены в линейной структуре CoT. Это указывает на фундаментальное ограничение CoT как метода XAI: цепочки размышлений могут быть скорее постфактум рационализацией, приемлемой для пользователя, чем точным отображением внутренних вычислительных процессов модели.

Тем не менее графы атрибуции также имеют свои ограничения. Построенные графы являются несовершенными репрезентациями фактических вычислений модели и не могут захватить все механизмы, особенно в слоях внимания. Успешность метода зависит от конкретной модели и промпта. Несмотря на эти ограничения, графы атрибуции представляют важный шаг к пониманию того, как модели действительно выполняют сложные задачи рассуждения, предоставляя подход «снизу вверх» к анализу внутренних механизмов и помогая выявлять потенциальные смещения или неожиданные вычислительные стратегии.

В случае объяснимости, основанной на промтинге, дополняющем знания LLM, добавляется явное внешнее знание, чтобы помочь модели генерировать более точные, проверяемые и обоснованные ответы. Чаще всего это достигается через генерацию с дополнением извлечением (retrieval-augmented generation, RAG) — модель сначала получает релевантную информацию из внешних источников (корпусов, баз данных), а потом использует ее для ответа. Это повышает качество и релевантность объяснений, снижая вероятность галлюцинаций.

Объяснения, основанные на аугментации обучающей выборки, используют пояснения и разъяснения для расширения и улучшения обучающих данных. Например, к исходным примерам добавляются метки с дополнительными обоснованиями или контекстом, усиливая способность модели учиться на более информативных данных. Это повышает качество обучения и позволяет моделям не только точнее предсказывать, но и создавать более обоснованные объяснения своих решений, улучшая доверие к системе.

Практические применения XAI в обработке естественного языка

Методы объяснимого искусственного интеллекта находят широкое применение в различных прикладных областях обработки естественного языка, где понимание логики принятия решений моделью критично для доверия пользователей, соответствия регуляторным требованиям и улучшения качества систем.

В области анализа тональности и модерации контента методы XAI играют ключевую роль в обеспечении прозрачности систем автоматической обработки пользовательского контента. Применение SHAP и LIME к моделям классификации отзывов позволяет не только определить общую тональность (позитивная, негативная, нейтральная), но и выявить конкретные слова и фразы, определяющие эту оценку. Это особенно ценно для бизнес-аналитики: компании могут идентифицировать специфические аспекты продукта или сервиса, вызывающие недовольство клиентов, что позволяет принимать обоснованные решения по улучшению предложения. В задачах модерации контента объяснения помогают понять, почему определенные сообщения были классифицированы как токсичные, оскорбительные или содержащие речь ненависти, что важно как для пользователей (возможность оспорить решение модерации), так и для операторов платформ (совершенствование правил модерации и обучающих данных).

Медицинские приложения представляют собой критически важную область применения XAI для NLP, где от объяснимости моделей могут зависеть здоровье и жизнь пациентов. Системы автоматического анализа клинических записей, истории болезней и медицинской литературы используют методы XAI для интерпретации диагностических заключений. Например, модели, предсказывающие риск заболеваний на основе симптомов, описанных в свободной текстовой форме, должны объяснять, какие конкретные медицинские термины и их комбинации привели к определенному заключению. Это позволяет врачам оценить обоснованность предсказаний модели, выявить потенциальные ошибки и принять информированное решение о дальнейшей

диагностике и лечении. Графы атрибуции и методы на основе механизма внимания помогают отследить, как модель связывает симптомы с диагнозами, выявляя как корректные медицинские ассоциации, так и потенциальные ложные корреляции в данных.

В юридических технологиях (Legal Tech) объяснимость моделей NLP необходима для автоматизированного анализа контрактов, классификации юридических документов и систем поддержки принятия решений. Модели, классифицирующие документы по типам (договоры купли-продажи, трудовые соглашения, лицензионные договоры) или извлекающие ключевые условия из контрактов, должны объяснять свои решения юристам, которые несут ответственность за правовые последствия этих решений. SHAP и LIME позволяют выявить, какие юридические термины, формулировки и структурные элементы документа были определяющими для классификации. В системах предсказательной юриспруденции, анализирующих судебные решения для прогнозирования исходов дел, методы XAI помогают понять, какие прецеденты и аргументы модель считает наиболее релевантными, что может служить дополнительной информацией для юристов при подготовке дел.

Выявление дезинформации и фейковых новостей представляет собой актуальную проблему информационной безопасности, где XAI-методы помогают понять, какие лингвистические и семантические признаки указывают на недостоверную информацию. Модели детекции фейков анализируют не только содержание текста, но и стилистические особенности, эмоциональную окраску, наличие сенсационных заголовков и других маркеров манипулятивного контента. Применение SHAP и attribution-based-методов позволяет идентифицировать конкретные слова, фразы и риторические приемы, характерные для дезинформации. Это важно как для пользователей медиаплатформ (понимание, почему контент был помечен как потенциально ложный), так и для исследователей и разработчиков систем факт-чекинга (выявление новых паттернов распространения дезинформации).

Разработка и отладка чат-ботов и виртуальных ассистентов существенно выигрывают от применения методов XAI. При анализе диалогов методы объяснимости помогают понять, почему модель генерирует определенные ответы, какие части пользовательского запроса были наиболее важны для формирования ответа и какие знания из обучающих данных или базы знаний были задействованы. Графы атрибуции и CoT-анализ позволяют отследить процесс рассуждения модели при формировании ответа, что критично для выявления и исправления ошибок в логике диалоговой системы. Это особенно важно в доменах, требующих высокой точности ответов, таких как банковские виртуальные ассистенты, медицинские консультационные боты или образовательные системы.

Инструменты и библиотеки XAI

Существует множество инструментов и библиотек, которые позволяют применять существующие методы объяснимого искусственного интеллекта. В этом разделе мы представим их в соответствии с предложенной классификацией.

Для анте-хок-методов объяснимого ИИ применимо к линейным моделям при использовании библиотеки scikit-learn (https://scikit-learn.org/) весовые коэффициенты признаков, представляющие собой глобальное объяснение, можно получить через атрибут `coef_`, а значение свободного члена — через `intercept_`. По форме результата данный инструмент представляет значимость признака, который может быть представлен как в числовом, так и в графическом или в табличном виде. Функциональные метрики, такие как устойчивость оценок и проверка мультиколлинеарности, обеспечивают техническую достоверность результата. Прикладные метрики отражают удобство использования этих моделей при анализе решений, тогда как человеческие оценки касаются понятности весовых коэффициентов для специалистов предметной области.

В случае применения анте-хок-объяснимого ИИ к деревьям решений форма результата может быть представлена и в виде правила, и как значимость признаков. Для извлечения правил вида «если . . . , то . . . » существует функция `sklearn.tree.export_text`, а для визуального представления правил — `sklearn.tree.plot_tree`. Получить оценку значимости признаков можно через атрибут `feature_importances_`.

В ансамблевых методах, таких как Random Forest, Gradient Boosting, прозрачность уменьшается, однако сохраняется через глобальные меры важности признаков (`feature_importances_`). Здесь функциональные метрики включают стабильность структуры дерева, прикладные — воспроизводимость решений, а метрики на основе человеческих оценок отражают доступность логических правил для пользователей. Ансамблевые методы сохраняют прозрачность благодаря оценке важностей признаков, которую также сопровождают функциональные метрики согласованности между моделями.

Аддитивные модели, реализуемые в pyGAM (https://pygam.readthedocs.io/) и Explainable Boosting Machine (https://github.com/interpretml/interpret), обеспечивают глобальную объяснимость через функции отклика для каждого признака. Формы результата представлены визуальными кривыми и профилями зависимости, а метрики оценивают гладкость и устойчивость этих профилей, а также воспринимаемость графической информации аналитиками.

Модель-неспецифичные пост-хок-методы XAI представлены в библиотеках SHAP, LIME, Anchors, DALEX, ELI5, Occlusion, Alibi и DiCE.

Библиотека SHAP (https://shap.readthedocs.io/en/latest/) предоставляет унифицированный интерфейс для получения локальных и глобальных объяснений на основе значений Шепли. Поскольку метод формально обоснован аксиоматически, он обеспечивает строго определенную значимость признаков. Библиотека включает широкий набор визуальных представлений: диаграммы вклада признаков, агрегированные распределения вкладов, графики последовательного разложения предсказания. Результат представлен как значимости признаков в числовом виде, а также может быть представлен графиками через инструменты и в графическом виде посредством `force_plot`, `waterfall_plot`, `summary_plot`.

Форма результата LIME — список локальных весовых коэффициентов, определяющих направление и силу влияния отдельных признаков на предсказание. В библиотеке LIME он представлен функциями `as_list()` или `as_map()`. В модуле `lime_text` объяснения могут быть представлены как подсветка наиболее значимых токенов, что делает метод удобным для интерпретации текстовых данных. Модальность объяснений преимущественно текстовая и графическая. Форма результата — значимость признаков, представленная в числовой, табличной и графической форме их рангами и весами.

Anchors (https://github.com/marcotcr/anchor) генерирует локальные логические правила, которые остаются устойчивыми к вариациям признаков; их метрики — точность и охват — относятся к функциональным.

Фреймворк DALEX (https://dalex.drwhy.ai/) предлагает систематизированный набор методов для анализа моделей, включая глобальные и локальные профили зависимостей, оценки важности признаков, разложения предсказаний и методы выявления аномального поведения. Форма результата варьируется от глобальных профилей зависимости: частичных диаграмм зависимости (Partial Dependence Plots, PDP) и накопленных локальных эффектов (Accumulated Local Effects, ALE), отражающих влияние признаков на предсказание в среднем, до локальных разложений предсказаний через диаграммы разбивки (break-down diagrams). Презентация преимущественно графическая: прогностические кривые, диаграммы вкладов, значимости признаков.

Библиотека ELI5 (https://eli5.readthedocs.io/en/latest/overview.html) известна прежде всего реализацией метода перестановочной важности признаков, который позволяет оценить влияние отдельных признаков путем случайной перестановки их значений и анализа изменения точности

модели. Помимо этого, ELI5 предоставляет визуализации весов линейных моделей и структуры деревьев решений. Форма результата — числовые значения важности признаков и их графическое отображение в виде столбчатых диаграмм. Функциональные метрики основываются на изменении качества модели при перестановке признака, в качестве прикладных метрик может быть рассмотрена пригодность этих оценок для анализа рисков, а в качестве человеческих — понятность диаграмм.

Метод Occlusion, реализованный в библиотеке Captum (https://captum.ai/api/occlusion.html), основан на систематическом маскировании частей входа и наблюдении за изменением предсказания модели. Он представляет собой наиболее непосредственный пертурбационный подход: влияние входного элемента определяется путем его исключения из входа без использования аппроксимаций или внутренних параметров модели. Результат выводится в виде тепловых карт и числовых профилей влияния отдельных элементов входа. Функциональные метрики представлены устойчивостью к изменению размера маски.

Фреймворк Alibi (https://github.com/SeldonIO/alibi) предоставляет инструменты для построения контрфактических объяснений, выявления прототипов и анализа аномалий. Контрфактические примеры представляются в числовом, текстовом и графическом виде. Метрики функциональны и представлены через проверку минимальности изменения, семантической близости, правдоподобия и устойчивости к вариациям модели.

Библиотека DiCE (https://github.com/microsoft/DiCE) развивает контрфактический подход, предоставляя не одно, а множество альтернативных сценариев изменения входа. Результат представляется контрфактическими примерами в табличной или текстовой форме. Метрики разнообразия, достижимости (actionability), близости к исходному примеру, правдоподобия и разреженности относятся к функциональным.

Теперь рассмотрим инструменты модель-специфичных пост-хок-методов применительно к градиентным методам, трансформерам и, в частности, LLM.

В Captum (PyTorch) и TF-Explain реализованы такие градиентные методы, как IntegratedGradients, GradientShap, Saliency, InputXGradient. Они опираются на вычисление производных предсказания по входу и отражают чувствительность модели к изменению каждого элемента ввода. Результат представлен в виде тепловых карт или численных карт атрибуций. Для исследователей и разработчиков такие методы особенно важны, поскольку позволяют выявлять слепые зоны модели, артефакты обучения и нестабильные признаки. Метрики включают sensitivity-n, completeness и проверку знаковой согласованности.

Интерпретация внимания трансформеров в HuggingFace Transformers представлена включенным в них механизмом вывода матриц внимания (`outputs.attentions`), а BertViz предоставляет удобные визуализации на уровне голов внимания и слоев. Форма презентации результата — графическая: матрицы вниманий, диаграммы, связи между токенами. Хотя внимание не всегда является причинным механизмом, оно отражает распределение зависимостей, что важно для разработчиков и исследователей, анализирующих поведение модели на уровне межтокенных отношений.

Для объяснимости LLM TransformerLens предоставляет доступ к активациям уровней трансформера, поддерживает механизмы патчинга (`activation_patch`), логит-линзы (`logit_lens`) и отслеживание влияния отдельных компонент модели. CircuitsVis визуализирует зависимости между слоями и подсетями модели, а GTraces позволяет прослеживать причинные пути прохождения информации. Презентация результата может быть текстовой, графовой или интерактивной. Метрики включают оценку вклада активаций, проверку причинной значимости и устойчивость цепочек. Эти инструменты предназначены преимущественно для исследователей и разработчиков моделей, которым необходимо понять, какие внутренние механизмы приводят к тем или иным рассуждениям LLM.

Практические рекомендации по выбору методов XAI

Выбор метода объяснимого искусственного интеллекта зависит от множества факторов: типа данных и модели, целевой аудитории, доступных вычислительных ресурсов и требований к теоретической обоснованности. Для облегчения выбора подходящего метода в таблице 3 представлено систематическое сравнение основных методов XAI по ключевым критериям.

Таблица 3. Сравнительная характеристика методов XAI

Метод	Область применения	Модель-специфичность	Масштаб	Вычислительная сложность	Теоретическая обоснованность	Стабильность	Целевая аудитория
Классическое машинное обучение							
SHAP	Табличные, тексты, изображения	Агностичный	Локальный + + глобальный	Высокая	Строгая (аксиомы Шепли)	Высокая	Исследователи, разработчики
LIME	Табличные, тексты, изображения	Агностичный	Локальный	Средняя	Частичная	Средняя (стохастический)	Разработчики, пользователи
Компьютерное зрение							
Grad-CAM	Изображения	CNN с GAP или произвольные	Локальный	Низкая	Эмпирическая	Высокая	Разработчики, пользователи
Integrated Gradients	Изображения	Дифференцируемые модели	Локальный	Средняя	Строгая (аксиомы)	Высокая	Исследователи, разработчики
SmoothGrad	Изображения	Дифференцируемые модели	Локальный	Средняя	Эмпирическое улучшение	Высокая	Исследователи, разработчики
RISE	Изображения	Агностичный (black-box)	Локальный	Высокая	Вероятностное сэмплирование	Средняя (стохастический)	Разработчики
TCAV	Изображения	CNN с доступом к активациям	Глобальный	Средняя	Направленные производные	Высокая	Исследователи
Обработка естественного языка							
Attention-визуализация	Тексты	Трансформеры	Локальный	Низкая	Частичная	Высокая	Разработчики, исследователи
Attribution graphs	Тексты (LLM)	Трансформеры с доступом к активациям	Локальный	Высокая	Каузальная трассировка	Высокая	Исследователи
Chain-of-Thought	Тексты (LLM)	LLM	Локальный	Низкая	Эмпирическая	Средняя	Пользователи, разработчики

Область применения определяет совместимость метода с типом входных данных. Модель-специфичность отражает требования к архитектуре модели: модель-агностичные методы (LIME, RISE) применимы к любым моделям, включая black-box-системы, тогда как модель-специфичные методы (Grad-CAM, Integrated Gradients) требуют доступа к внутренней структуре модели и градиентам. Масштаб объяснения показывает, предоставляет ли метод локальные объяснения для отдельных предсказаний, глобальные объяснения поведения модели в целом или оба типа объяснений.

Вычислительная сложность является критическим фактором для практического применения: методы с низкой сложностью (Grad-CAM) подходят для интерактивных систем и анализа

больших объемов данных, тогда как методы с высокой сложностью (SHAP для больших моделей, RISE) требуют значительных вычислительных ресурсов. Теоретическая обоснованность отражает наличие математических гарантий корректности объяснений: SHAP и Integrated Gradients обладают строгой теоретической базой (аксиомы Шепли, аксиомы чувствительности и полноты), LIME имеет частичное обоснование через локальную аппроксимацию, а некоторые методы (например, простая визуализация весов внимания) носят преимущественно эмпирический характер.

Стабильность результатов характеризует воспроизводимость объяснений при повторных запусках: детерминированные методы (Grad-CAM, Integrated Gradients) всегда дают одинаковые результаты для одного и того же входа, тогда как стохастические методы (LIME, RISE) могут давать различающиеся объяснения из-за случайной природы алгоритма. Наконец, целевая аудитория указывает, для каких пользователей метод наиболее подходит: исследователи нуждаются в теоретически обоснованных и детальных методах, разработчики — в быстрых и практичных инструментах, а конечные пользователи — в визуально интерпретируемых объяснениях.

При выборе метода рекомендуется учитывать следующие практические соображения. Для быстрого прототипирования и отладки моделей компьютерного зрения наиболее эффективен Grad-CAM благодаря низкой вычислительной сложности и наглядной визуализации. Для критически важных приложений, требующих обоснованных объяснений (медицинская диагностика, финансовые решения), предпочтительны методы со строгой теоретической базой — SHAP или Integrated Gradients. При работе с black-box-моделями, к которым нет доступа к внутренней структуре, единственными опциями являются модель-агностичные методы: LIME для быстрого анализа или RISE для более точных, но вычислительно затратных объяснений.

Для больших языковых моделей выбор метода зависит от цели анализа. Визуализация весов внимания подходит для быстрого понимания того, какие токены модель считает релевантными, однако не гарантирует точного отражения причинных механизмов. Графы атрибуции обеспечивают более глубокий анализ внутренних вычислений, но требуют значительных ресурсов и специализированных инструментов. Естественно-языковые объяснения (Chain-of-Thought) наиболее понятны конечным пользователям, но их достоверность должна быть верифицирована другими методами.

При ограниченных вычислительных ресурсах следует отдавать предпочтение методам с низкой или средней сложностью: Grad-CAM для изображений, визуализация внимания для текстов, LIME для табличных данных. Если требуются высокая стабильность и воспроизводимость результатов, необходимо избегать стохастических методов (LIME, RISE) или использовать их с фиксированным seed и множественными запусками для усреднения результатов. Для мультимодальных моделей может потребоваться комбинация методов из разных областей, адаптированных для совместного анализа различных типов входных данных.

Заключение

Объяснимый искусственный интеллект находится на пересечении фундаментальных вызовов современного машинного обучения: необходимости создания все более мощных и точных моделей и одновременного обеспечения их прозрачности, понятности и надежности. Настоящая работа представила аналитический обзор принципов, методов и практических применений XAI, демонстрирующий как достижения этой области, так и ее текущие ограничения.

Проведенный анализ показывает, что методы XAI достигли значительной зрелости в области классического машинного обучения и компьютерного зрения, где теоретически обоснованные подходы (SHAP, Integrated Gradients) сочетаются с практически эффективными техниками (LIME, Grad-CAM). Однако применение XAI к большим языковым моделям и системам на

основе трансформеров остается открытой исследовательской проблемой, требующей разработки новых парадигм объяснения, учитывающих специфику архитектур с механизмом внимания и emergent capabilities.

Фундаментальная проблема компромисса между точностью модели и интерпретируемостью объяснений продолжает определять развитие области. Появление архитектур вроде смеси экспертов (Mixture-of-Experts), моделей с внешней памятью и мультимодальных систем требует пересмотра существующих методов объяснения. Особенно остро стоит вопрос верификации достоверности естественно-языковых объяснений, генерируемых самими языковыми моделями: методы графов атрибуции выявили существенные расхождения между явными цепочками рассуждений и фактическими вычислительными процессами.

Практическое внедрение методов XAI в критически важные области применения — медицину, юриспруденцию, финансовый сектор — сталкивается не только с техническими ограничениями, но и с необходимостью адаптации объяснений к различным категориям пользователей с разным уровнем технической подготовки. Разработка персонализированных объяснений, балансирующих простоту восприятия и техническую точность, остается важнейшей задачей.

Дальнейшее развитие области XAI видится в нескольких направлениях. Во-первых, создание эффективных методов объяснения для моделей с миллиардами и триллионами параметров, требующих новых подходов к масштабированию существующих техник. Во-вторых, разработка унифицированных метрик оценки качества объяснений, позволяющих количественно сравнивать различные подходы. В-третьих, интеграция объяснимости в процесс обучения моделей (ante-hoc-подходы) как альтернатива пост-хок-анализу. В-четвертых, формирование регуляторных стандартов объяснимости для высокорисковых областей применения ИИ.

Особого внимания заслуживают новые вызовы, связанные с мультимодальными моделями, объединяющими обработку текста, изображений, аудио и видео [Rodis et al., 2024]. Объяснение решений таких систем требует методов, способных интегрировать атрибуции из разных модальностей и выявлять кросс-модальные взаимодействия. Аналогичные проблемы возникают в контексте федеративного обучения, где модели обучаются на распределенных данных без их централизации: здесь необходимы методы XAI, сохраняющие конфиденциальность данных при обеспечении прозрачности принятия решений [Bárcena et al., 2022; Daole et al., 2023]. Модели с внешней памятью и агентные системы с доступом к инструментам также порождают новые требования к объяснимости — необходимость отслеживать не только внутренние вычисления модели, но и ее взаимодействие с внешними ресурсами.

Объяснимый искусственный интеллект является не просто набором технических методов, но необходимым условием ответственного развития и широкого внедрения технологий машинного обучения в общество. Дальнейший прогресс в этой области критически важен для построения доверия к системам ИИ, обеспечения их справедливости и безопасности, а также для реализации потенциала искусственного интеллекта в решении сложных научных и прикладных задач.

Список литературы (References)

Шевская Н. В. Объяснимый искусственный интеллект и методы интерпретации результатов // Моделирование, оптимизация и информационные технологии. — 2021. — Т. 9, № 2. — DOI: 10.26102/2310-6018/2021.33.2.024

Shevskaya N. V. Ob"yasnimyj iskusstvennyj intellekt i metody interpretacii rezul'tatov [Explainable artificial intelligence and methods for interpreting results] // Modelirovanie, optimizaciya i informacionnye tekhnologii [Modeling, Optimization and Information Technology]. — 2021. — Vol. 9, No. 2. — https://moitvivt.ru/ru/journal/pdf?id=1005 (in Russian).

Ali H., Khan M. S., Al-Fuqaha A., Qadir J. Tamp-X: Attacking explainable natural language classifiers through tampered activations // Computers & Security. — 2022. — Vol. 120. — P. 102791.

Ali S., Abuhmed T., El-Sappagh S., Muhammad K., Alonso-Moral J. M., Confalonieri R., Guidotti R., Del Ser J., Díaz-Rodríguez N., Herrera F. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence // Information Fusion. — 2023. — Vol. 99. — P. 101805. — DOI: 10.1016/j.inffus.2023.101805

Arrieta A. B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. — 2020. — Vol. 58. — P. 82–115.

Bárcena J. L. C., Daole M., Ducange P., Marcelloni F., Renda A., Ruffini F., Schiavo A. Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models // XAI.it@AI*IA. — 2022. — P. 104–117.

Bau D., Zhou B., Khosla A., Oliva A., Torralba A. Network dissection: Quantifying interpretability of deep visual representations // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2017. — P. 6541–6549.

Belle V., Papantonis I. Principles and practice of explainable machine learning // Frontiers in Big Data. — 2021. — Vol. 4. — P. 688969.

Bibal A., Cardon R., Alfter D., Wilkens R., Wang X., François T., Watrin P. Is attention explanation? An introduction to the debate // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. — 2022. — P. 3889–3900.

Burger C., Chen L., Le T. Are your explanations reliable? Investigating the stability of LIME in explaining text classifiers by marrying XAI and adversarial attack // arXiv preprint. — 2023. — arXiv:2305.12351

Chattopadhay A., Sarkar A., Howlader P., Balasubramanian V. N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks // Proceedings of the IEEE Winter Conference on Applications of Computer Vision. — 2018. — P. 839–847.

Chen H., Lundberg S., Lee S. I. SHAP values for explaining CNN-based text classification models // arXiv preprint. — 2020. — arXiv:2008.11825

Daole M., Schiavo A., Bárcena J. L. C., Ducange P., Marcelloni F., Renda A. OpenFL-XAI: Federated learning of explainable artificial intelligence models in Python // SoftwareX. — 2023. — Vol. 23. — P. 101505. — DOI: 10.1016/j.softx.2023.101505

Das A., Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey // arXiv preprint. — 2020. — arXiv:2006.11371

De Santis A., Campi R., Bianchi M., Brambilla M. Visual-TCAV: Concept-based attribution and saliency maps for post-hoc explainability in image classification // arXiv preprint. — 2024. — arXiv:2411.05698

Herrera F. Making sense of the unsensible: Reflection, survey, and challenges for XAI in large language models toward human-centered AI // arXiv preprint. — 2025. — arXiv:2505.20305

Hsieh W., Bi Z., Jiang C., Liu J., Peng B., Zhang S., Pan X., Xu J., Wang J., Chen K., Yin C., Feng P., Wen Y., Song X., Wang T., Yang J., Li M., Jing B., Ren J., Liu M. A comprehensive guide to explainable AI: From classical models to LLMs // arXiv preprint. — 2024. — arXiv:2412.00800

Huang X., Marques-Silva J. On the failings of Shapley values for explainability // International Journal of Approximate Reasoning. — 2024. — Vol. 171. — P. 109112.

Jain S., Wallace B. C. Attention is not explanation // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. — 2019. — P. 3543–3556.

Kim B., Wattenberg M., Gilmer J., Cai C., Wexler J., Viegas F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV) // Proceedings of the 35th International Conference on Machine Learning. — 2018. — Vol. 80. — P. 2668–2677.

Knab P., Marton S., Schlegel U., Bartelt C. Which LIME should I trust? Concepts, challenges, and solutions // World Conference on Explainable Artificial Intelligence. — Cham: Springer Nature Switzerland, 2025. — P. 28–52.

Li J., Monroe W., Jurafsky D. Understanding neural networks through representation erasure // arXiv preprint. — 2016. — arXiv:1612.08220

Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. — 2017. — Vol. 30. — P. 4765–4774.

Manna S., Sett N. Faithfulness and the notion of adversarial sensitivity in NLP explanations // arXiv preprint. — 2024. — arXiv:2409.17774

Matton K., Ness R. O., Guttag J., Kıcıman E. Walk the talk? Measuring the faithfulness of large language model explanations // arXiv preprint. — 2025. — arXiv:2501.14150

Mohammadi H., Mersha M., Bitewa M., Abay T., Kalita J. Explainability in neural networks for natural language processing tasks // arXiv preprint. — 2024. — arXiv:2412.18036

Mumuni A., Mumuni F., Gerrar N. K. A survey of synthetic data augmentation methods in computer vision // arXiv preprint. — 2024. — arXiv:2403.10075

Petsiuk V., Das A., Saenko K. RISE: Randomized input sampling for explanation of black-box models // Proceedings of the British Machine Vision Conference. — 2018. — P. 151.

Pfau J., Merrill W., Bowman S. R. Let's think dot by dot: Hidden computation in transformer language models // arXiv preprint. — 2024. — arXiv:2404.15758

Phillips P. J., Hahn C. A., Fontana P. C., Yates A. N., Greene K., Przybocki M. A. Four principles of explainable artificial intelligence // NIST Interagency Report. — 2021. — P. 8312.

Ribeiro M. T., Singh S., Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 1135–1144.

Rodis N., Sardianos C., Radoglou-Grammatikis P., Sarigiannidis P., Varlamis I., Papadopoulos G. T. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions // IEEE Access. — 2024. — Vol. 12. — P. 76686–76728. — DOI: 10.1109/ACCESS.2024.3408890

Ronge R., Bauer B., Rathgeber B. Approaching principles of XAI: A systematization // IEEE Transactions on Artificial Intelligence. — 2025. — Vol. 6, No. 5. — P. 1067–1079.

Salih A. M., Raisi-Estabragh Z., Boscolo Galazzo I., Radeva P., Petersen S. E., Lekadir K., Menegaz G. A perspective on explainable artificial intelligence methods: SHAP and LIME // Advanced Intelligent Systems. — 2024. — DOI: 10.1002/aisy.202400304

Schwalbe G., Finzel B. A. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts // Data Mining and Knowledge Discovery. — 2024. — Vol. 38. — P. 3043–3101.

Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision. — 2017. — P. 618–626.

Simonyan K., Vedaldi A., Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps // arXiv preprint. — 2014. — arXiv:1312.6034

Smilkov D., Thorat N., Kim B., Viégas F., Wattenberg M. SmoothGrad: removing noise by adding noise // arXiv preprint. — 2017. — arXiv:1706.03825

Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods // Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. — 2022. — P. 2239–2250.

Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning. — 2017. — Vol. 70. — P. 3319–3328.

Turpin M., Michael J., Perez E., Bowman S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting // Advances in Neural Information Processing Systems. — 2023. — Vol. 36.

Yu Y. Do LLMs really think step-by-step in implicit reasoning? // arXiv preprint. — 2024. — arXiv:2411.15862

Zeiler M. D., Fergus R. Visualizing and understanding convolutional networks // European Conference on Computer Vision. — Springer, 2014. — P. 818–833.

Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. Learning deep features for discriminative localization // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2016. — P. 2921–2929.