НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2026 Том 26 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2026 Vol. 26 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-1-69-76
УДК 004.932

Подход к применению больших языковых моделей с дополненным поиском для повышения интерпретируемости моделей машинного обучения

Роман Дмитриевич Иванов[1], Артем Бакытжанович Менисов[2]✉, Олег Александрович Мишуков[3], Тимур Римович Сабиров[4]

[1,2,3,4] Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация

[1] vka@mil.ru, https://orcid.org/0009-0004-1664-1512
[2] men.arty@yandex.ru✉, https://orcid.org/0000-0002-9955-2694
[3] vka@mil.ru, https://orcid.org/0009-0004-4247-5810
[4] vka@mil.ru, https://orcid.org/0000-0002-6807-2954

Аннотация

Введение. Интерпретируемость моделей машинного обучения является ключевым условием доверия при внедрении технологий искусственного интеллекта. Традиционные методы объяснения обеспечивают формальную интерпретацию и представляют фрагментарные и трудно интерпретируемые результаты, что снижает прозрачность принятия решений. Для решения этой проблемы предлагается подход, основанный на использовании больших языковых моделей в сочетании с технологией дополненного поиска, обеспечивающей привлечение внешних знаний и повышающей связность объяснений. Отличительной особенностью подхода является ориентация на семантическую согласованность, стабильность интерпретаций и понятность для пользователя. **Метод.** В представленном подходе большая языковая модель выступает в качестве интерпретатора, формирующего объяснения на основе исходных данных модели и внешних знаний, получаемых через дополненный поиск. Подход включает механизм генерации человеческо-читаемых объяснений, верификацию интерпретируемых признаков (токенов) и оценку устойчивости выводов. В результате формируются выводы, согласованные с контекстом предметной области. **Основные результаты.** Эффективность подхода проверена на данных MITRE ATT&CK, представляющих стандартизированную информацию о киберугрозах. Сравнение с методом SHapley Additive exPlanations показало, что предложенный подход обеспечивает более высокую семантическую согласованность объяснений и большую устойчивость оценки важности признаков при сохранении сопоставимого уровня точности интерпретации. Экспериментальные результаты показали преимущество интерпретатора на основе больших языковых моделей в объяснимости и восприятии человеком. **Обсуждение.** Разработанный подход в сравнении с традиционными методами обеспечивает понятные и контекстно обоснованные объяснения. Полученные результаты делают предложенный метод перспективным для применения в системах, где критично объяснение причин принятия решений. Дальнейшие направления исследований включают интеграцию подхода в реальные системы принятия решений, исследование автоматического контроля достоверности объяснений и адаптацию подхода к различным архитектурам больших языковых моделей.

Ключевые слова

большая языковая модель, интерпретируемость, дополненный поиск, семантическая согласованность

Ссылка для цитирования: Иванов Р.Д., Менисов А.Б., Мишуков О.А., Сабиров Т.Р. Подход к применению больших языковых моделей с дополненным поиском для повышения интерпретируемости моделей машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 1. С. 69–76. doi: 10.17586/2226-1494-2026-26-1-69-76

An approach to using large language models with augmented search to improve the interpretability of machine learning models

Roman D. Ivanov[1], Artem B. Menisov[2]✉, Oleg A. Mishukov[3], Timur R. Sabirov[4]

[1,2,3,4] Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation

[1] vka@mil.ru, https://orcid.org/0009-0004-1664-1512
[2] men.arty@yandex.ru✉, https://orcid.org/0000-0002-9955-2694
[3] vka@mil.ru, https://orcid.org/0009-0004-4247-5810
[4] vka@mil.ru, https://orcid.org/0000-0002-6807-2954

Abstract

Interpretability of machine learning models is a key requirement for robustness in the implementation of artificial intelligence technologies. Traditional explanation methods provide formal interpretation and produce results that are fragmented and hard to interpret, reducing the transparency of decision making. To address this problem, a modern approach is proposed based on the use of large language models combined with augmented search, which ensures the involvement of external knowledge and increases the coherence of sequences. A distinctive feature of this approach is the focus on semantic consistency, the stability of interpretations, and user comprehensibility. In the presented approach, a large language model acts as an interpreter, generating explanations based on the model input data and external knowledge obtained through augmented search. The approach includes a mechanism for generating human-readable observations, verifying interpretable features (tokens), and assessing the robustness of inferences. As a result, inferences are generated that are consistent with the context of the subject domain. The effectiveness of the proposed approach is tested using MITRE ATT&CK data, which provides standardized information on cyber threats. A comparison with the SHapley Additive exPlanations method showed that the proposed approach provides higher semantic consistency of consequences and greater robustness of feature importance assessment while maintaining advanced level accuracy. Experimental results obtained by the interpreter based on large language models are in terms of human comprehensibility and perception. The developed approach, when considered with conservative methods, provides understandable and context-based explanations. The obtained results make the proposed method promising for application in economics where critical explanations of decision-making are essential. Future research includes integrating the solutions into real systems, investigating automatic validation of results, and adapting them to various large language model architectures.

Keywords

large language model, interpretability, augmented search, semantic consistency

For citation: Ivanov R.D., Menisov A.B., Mishukov O.A., Sabirov T.R. An approach to using large language models with augmented search to improve the interpretability of machine learning models. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 1, pp. 69–76 (in Russian). doi: 10.17586/2226-1494-2026-26-1-69-76

Введение

Искусственный интеллект активно внедряется в разные сферы — от здравоохранения и финансов до транспорта и образования [1]. Современные системы искусственного интеллекта, особенно включающие нейросетевые модели, демонстрируют высокую производительность, что открывает возможность для автоматизации сложных задач [2]. Однако наряду с ростом возможностей систем искусственного интеллекта возрастает и потребность в доверии к ним [3]. Без понимания того, как нейронная сеть принимает решения, ее использование в критических задачах становится рискованным. Прежде всего, это касается высокоавтоматизированных систем, где непрозрачность работы модели может привести к непредсказуемым и потенциально опасным последствиям. По этой причине важнейшей задачей становится обеспечение интерпретируемости моделей машинного обучения — их способности объяснять свои действия и выводы понятным для человека образом.

В ответ на эти вызовы сформировалось направление Explainable Artificial Intelligence [4], целью которого является разработка методов и инструментов для интерпретации решений моделей машинного обучения, в том числе нейросетей. Тем не менее, традиционные подходы к интерпретируемости [5] зачастую ограничены, так как они либо оперируют абстрактными метриками важности признаков, либо предоставляют сложно интерпретируемые результаты. Эти ограничения особенно ощутимы при работе с высокоразмерными, мультимодальными или языковыми моделями машинного обучения [6].

Такие большие языковые модели (БЯМ) как GPT [7], LLaMA [8] и Gemma [9], продемонстрировали высокую способность к обработке естественного языка и генерации связных, осмысленных текстов. Данные особенности открывают перспективы их использования в качестве интерпретаторов — моделей, способных формировать человеко-понятные объяснения поведения других моделей машинного обучения. В связи с этим в настоящей работе выдвигается гипотеза о применимости БЯМ для интерпретируемости других моделей машинного обучения.

Анализ известных результатов в области интерпретации моделей машинного обучения

Современные подходы к интерпретации моделей машинного обучения можно классифицировать на типы, представленные на рис. 1.

Методы интерпретации моделей машинного обучения позволяют понять, как модели организуют знания и выявляют паттерны в данных. Структурные методы анализа [10] изучают внутренние компоненты модели:

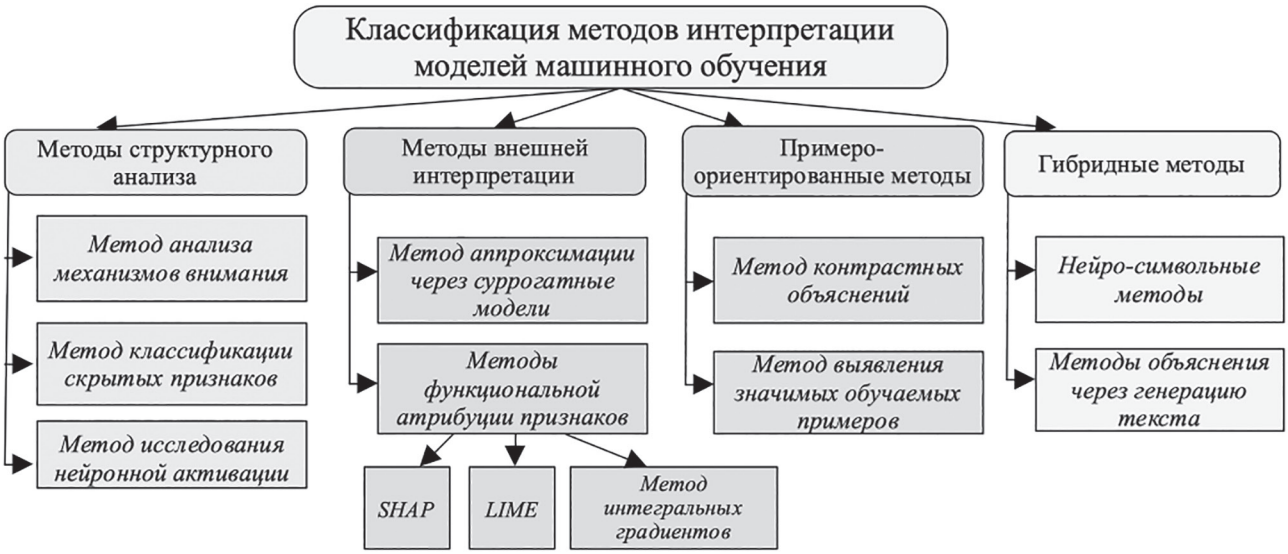

Рис. 1. Классификация подходов к интерпретации моделей машинного обучения
Fig. 1. Classification of approaches to interpreting machine learning models

анализ внимания показывает, какие входные элементы влияют на предсказания, вспомогательная классификация скрытых признаков оценивает информационное содержание слоев, а исследование нейронной активации выявляет функциональную специализацию отдельных нейронов.

Методы внешней интерпретации оценивают модель «извне» [11]. Суррогатные модели аппроксимируют поведение сложных моделей с помощью простых, интерпретируемых алгоритмов, а функциональная атрибуция признаков (SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), интегральные градиенты) измеряет вклад каждого входного элемента в предсказание.

Эти подходы универсальны, но могут быть вычислительно затратными и нестабильными при локальных объяснениях.

Примеро-ориентированные методы анализируют поведение модели через входные данные [12]. Контрастные объяснения создают альтернативные примеры с изменением исходного результата, а выявление значимых обучающих элементов определяет данные, влияющие на конкретные предсказания. Такие методы обеспечивают наглядность, но требуют значительных вычислительных ресурсов и доступа к обучающим данным.

Гибридные методы объединяют различные подходы [10]. Нейро-символьные методы комбинируют нейросети и логический вывод, обеспечивая прозрачность и надежность решений, а генерация текстовых объяснений преобразует внутренние представления модели в понятные описания. Данные методы позволяют получать интерпретируемые объяснения даже для сложных моделей БЯМ.

Формализованная постановка задачи

В формализованном виде постановка задачи исследования может быть сформулирована следующим образом. Пусть заданы: $X = \{x_1, \ldots, x_m\}$ — множество основных признаков (токенов); $Y = \{y_1, \ldots, y_n\}$ — множество целевых признаков; F: $X \to Y$ — модель-классификатор на основе БЯМ, которая по входным данным $x_i \in X$ предсказывает метку класса $y_j \in Y$ с вероятностью распознавания класса $p(y_j)$; F_{base} — интерпретатор, принятый в качестве базового, который возвращает значения важности основных признаков (токенов) x_i.

Определим пространство объяснений E, где каждое объяснение $e_k \in E$ описывает логическую связь между x_i и классом y_j. Функция интерпретации $S(x_i, y_j) = [s_1, s_2, \ldots, s_k]$ определяет значения важности x_i для предсказанного класса y_j, на основе которого определяются значимые признаки (токены) $V = \{v_1, \ldots, v_n\}$.

Необходимо разработать такой способ интерпретации F^*, реализуемый с помощью БЯМ-интерпретатора $F_{\text{БЯМ}}$, при котором метрика качества объяснений $Q(F_{\text{БЯМ}})$ превосходит $Q(F_{\text{base}})$:

$$F^* = \underset{F \in \{F_{\text{base}}, F_{\text{БЯМ}}\}}{\operatorname{argmax}} Q(F),$$

где $Q(F) = \delta(e_F, e_{ref})$ — метрика качества объяснений; e_{ref} — эталонное объяснение на выходе базового интерпретатора; δ: $e_F \times e_{ref} \to [0, 1]$ — мера схожести между объяснениями.

Описание подхода

Предлагаемый подход к интерпретации решений моделей машинного обучения основан на применении БЯМ в качестве интерпретатора с использованием технологии дополненного поиска (ДП) [13].

Целью подхода является повышение интерпретируемости моделей («черных ящиков») за счет генерации человекопонятных, семантически обоснованных и контекстуализированных объяснений на естественном языке. Процесс интерпретации состоит в том, что входной текст, ранее обработанный моделью-классификатором, передается на вход интерпретатору, который формирует объяснение как в виде списка наиболее значимых

токенов, так и в виде текстовой интерпретации выбора предсказанного класса. В случае использования ДП извлекаются связанные фрагменты данных, расширяющие контекст генерации.

Основными компонентами подхода являются:

— интерпретируемая модель-классификатор, обученная на задаче классификации по входным текстовым данным из текстовых отчетов об инцидентах компьютерной безопасности;
— базовый интерпретатор, позволяющий оценить вклад каждого признака (токена) в результат модели машинного обучения, объясняя, насколько он изменен по сравнению с его базовым значением;
— БЯМ-интерпретатор, способный генерировать объяснения в форме текстов естественного языка;
— БЯМ-интерпретатор с интегрированной технологией ДП, которая добавляет информацию в запрос к модели машинного обучения.

Ключевым компонентом разработанного подхода является использование БЯМ в роли интерпретатора, обладающего способностью к генерации текстов и выделению значимых признаков (токенов), влияющих на поведение целевой модели-классификатора. Подход сочетает возможности современных БЯМ и механизмов внешнего семантического поиска, обеспечивая интерпретацию, сочетающую точность, устойчивость и понятность результатов моделей машинного обучения. Для повышения качества объяснений и минимизации эффектов галлюцинации применяется ДП.

Экспериментальное исследование БЯМ-интерпретатора

Для подтверждения гипотезы о применимости БЯМ в качестве интерпретатора другой БЯМ был проведен эксперимент (рис. 2) на вычислительном комплексе со следующими характеристиками: процессор: Intel Core i9-10980XE (18 ядер, 36 потоков, частота 4,8 ГГц); оперативная память (ОЗУ) — 64 ГБ; графический процессор — NVIDIA GeForce TITAN RTX 24 ГБ.

Исходные данные:

— интерпретируемая модель: БЯМ-классификатор Distilgpt-2[1];
— базовый интерпретатор SHAP[2], используемый в качестве эталона для оценки значимости токенов;
— БЯМ-интерпретатор Saiga-2[3] в двух конфигурациях: без дополнительных модификаций и с применением ДП;
— поисковая модель для ДП All-MiniLM-L6-v2[4];
— векторная база данных, построенная на основании поисковой системы OpenSearch[5].

Для БЯМ-интерпретатора Saiga-2 настраивались следующие параметры: температура, ограничение пространства выбора (top-k) и порог вероятности (top-p). Так как для решения задачи интерпретации модели машинного обучения с помощью БЯМ требовалась наибольшая точность, поэтому опытным путем были определены рекомендуемые параметры модели БЯМ-интерпретатора Saiga-2: температура 0,1; значение top-p равно 0,5, значение top-k равно 50. Данные значения позволили уменьшить нестандартные ответы БЯМ-

[1] [Электронный ресурс]. Режим доступа: https://huggingface.co/distilbert/distilgpt2 (дата обращения: 25.12.2025).

[2] [Электронный ресурс]. Режим доступа: https://github.com/shap/shap (дата обращения: 25.12.2025).

[3] [Электронный ресурс]. Режим доступа: https://huggingface.co/IlyaGusev/saiga2_7b_gguf (дата обращения: 25.12.2025).

[4] [Электронный ресурс]. Режим доступа: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (дата обращения: 25.12.2025).

[5] [Электронный ресурс]. Режим доступа: https://github.com/opensearch-project/OpenSearch (дата обращения: 25.12.2025).

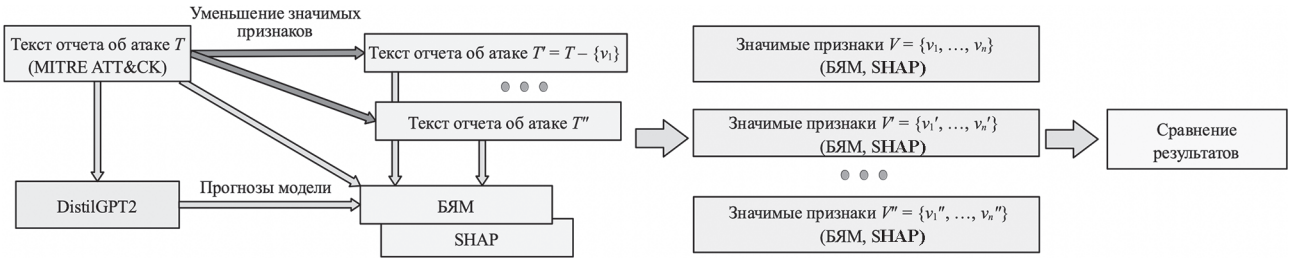

Рис. 2. Схема эксперимента по интерпретации большой языковой модели.

T — исходный текст описания инцидента в терминах MITRE ATT&CK; T' — текст отчета T после этапа уменьшения значимых признаков; T'' — альтернативная или дополнительно обработанная версия текста отчета, полученная после применения иной стратегии уменьшения признаков; $v_1, \ldots, v_n$ — отдельные признаки (токены), извлеченные из соответствующего текста отчета; n — общее количество используемых токенов.

Вектор признаков: $V = (v_1, \ldots, v_n)$, сформированный на основе исходного текста T; $V' = (v_1', \ldots, v_n')$, соответствующий преобразованному тексту T'; $V'' = (v_1'', \ldots, v_n'')$, соответствующий тексту T''

Fig. 2. Scheme of the experiment on interpretation of a large language model.

T is the original text of the incident description in terms of MITRE ATT&CK; T' is the text of the report T after the stage of significant features reduction; T'' is an alternative or additionally processed version of the report text obtained after applying a different feature reduction strategy; $v_1, \ldots, v_n$ are individual features (tokens) extracted from the corresponding report text; $V = (v_1, \ldots, v_n)$ is the feature vector formed on the basis of the original text T; $V' = (v_1', \ldots, v_n')$ is the feature vector corresponding to the transformed text T'; $V'' = (v_1'', \ldots, v_n'')$ is the feature vector corresponding to the text T''; n is the total number of tokens used

интерпретатора Saiga-2 при интерпретации модели БЯМ-классификатора Distilgpt-2.

Для проведения экспериментов по интерпретации БЯМ использовались данные о тактиках и техниках, представленные в базе знаний MITRE ATT&CK [14]. Этот фреймворк является общепризнанным стандартом для описания методов, которые злоумышленники применяют на различных этапах компьютерных атак MITRE ATT&CK включает подробные сведения о тактиках (стратегических целях атакующего) и техниках (конкретных действиях, реализующих эти цели).

Перед подачей на вход БЯМ данные были приведены к унифицированному текстовому формату: удалены дублирующиеся или устаревшие записи; объединены описание тактики и соответствующих техник в связные тексты; добавлены контекстные метаданные (категория тактики, идентификатор техники) для дальнейшей интерпретации модели.

Подготовленные данные использовались для нескольких типов задач интерпретации БЯМ: генерации описаний техник на основе тактик; выявления, какие элементы текста модели считают ключевыми для классификации или генерации; проверки способности интерпретаторов выявлять скрытые связи между тактиками и техниками (рис. 3).

Этап 1. Ввод входных данных. На вход экспериментальной системы подавались текстовые сообщения, содержащие описания компьютерных инцидентов. Эти данные формировали множество входных отчетов $X = \{x_1, \ldots, x_m\}$.

Этап 2. Классификация техник MITRE ATT&CK с помощью БЯМ. Для каждого отчета выполнялась классификация с использованием БЯМ-классификатора Distilgpt-2, обученного распознавать тактики и техники MITRE ATT&CK. Результатом распознавания являлась метка класса , определяющая технику нарушителя.

Этап 3. Настройка параметров базового интерпретатора SHAP. Производилось конфигурирование параметров интерпретатора, включая количество сэмплов для аппроксимации nsamples, коэффициент L_1-регуляризации для снижения избыточной детализации и шума l1_reg и количество наиболее важных признаков (токенов) для отбора num_features.

Этап 4. Интерпретация базовым интерпретатором SHAP. Интерпретатор SHAP применялся к входным данным и предсказанию модели F: $X \to Y$ для получения базового списка значимых токенов с оценками важности:

$$V_i^{SHAP} = \{(\omega_k, s_k)\}_{k=1}^{10},$$

где ω_k — значимый k-й токен текстового отчета x_i; s_k — оценка важности k-го токена.

Этап 5. Вывод 10 токенов в порядке убывания степени важности интерпретатором SHAP. Интерпретатор SHAP возвращал список из 10 наиболее значимых токенов. Эти данные формировали опорную точку для дальнейшего сравнения результатов интерпретации.

Этап 6. Настройка параметров БЯМ-интерпретатора Saiga-2. Выполнялась настройка параметров генерации БЯМ-интерпретатора Saiga-2, включая температуру, top-k, top-p, и способ форматирования промпта. Эти настройки обеспечивали стабильность и достоверность интерпретаций.

Этап 7. Интерпретация с помощью БЯМ-интерпретатора Saiga-2, которая генерировала список наиболее значимых токенов, влияющих на решение БЯМ-классификатора:

$$V_i^{Saiga2} = \{(\omega_k, s_k)\}_{k=1}^{10}.$$

Этап 8. Вывод 10 токенов в порядке убывания степени важности БЯМ-интерпретатором Saiga-2.

Этап 9. Настройка параметров БЯМ-интерпретатора Saiga-2 с ДП RAG. Для повышения контекстуальности интерпретации БЯМ-интерпретатор Saiga-2 использовал внешнюю базу знаний и семантический поиск. В процессе настройки конфигурировались параметры поиска: модель All-MiniLM, количество возвращаемых документов $n_{results}$, и агрегация результатов в промпт.

Этап 10. Интерпретация с помощью БЯМ-интерпретатора Saiga-2 с ДП RAG, которая генерировала список наиболее значимых токенов:

$$V_i^{Saiga2\_RAG} = \{(\omega_k, s_k)\}_{k=1}^{10}.$$

Рис. 3. Этапы эксперимента по интерпретации большой языковой модели

Fig. 3. Stages of the experiment on interpretation of a large language model

Этап 11. Вывод 10 токенов в порядке убывания степени важности БЯМ-интерпретатором Saiga-2 с ДП.

Этап 12. Сравнение результатов интерпретации.

Для каждого интерпретатора выполнялись четыре итерации преобразования текста входного отчета об атаке путем последовательного удаления токена с наибольшей степенью важности из текста на каждой итерации. Преобразованный текст отчета и предсказанная моделью Distilgpt-2 метка класса подавались на вход интерпретаторов SHAP, БЯМ Saiga-2, БЯМ Saiga-2 с ДП. На выходе каждого интерпретатора формировался список токенов и степени важности токенов, которые повлияли на решение модели Distilgpt-2. Далее токены, полученные на выходе интерпретатора SHAP, сравнивались с токенами на выходе интерпретаторов БЯМ Saiga-2, БЯМ Saiga-2 с ДП.

Обсуждение

Сравнение результатов по определению важных токенов с помощью интерпретаторов SHAP и БЯМ Saiga-2 представлено в табл. 1.

Анализ полученных результатов показал, что применение БЯМ-интерпретатора Saiga-2 позволяет семантически верно выделять наиболее важные токены на каждой итерации, повышая важность оставшихся токенов в ответе. Интерпретатор SHAP показал, напротив, неудовлетворительные результаты, в каждой итерации хаотично присваивая значения важности различным токенам. Отметим, что БЯМ-интерпретатор Saiga-2 токенами считает слова или отдельные фразы, тогда как интерпретатор SHAP — слоги и, в редком случае, небольшие слова.

В табл. 2 представлены ключевые параметры интерпретаторов SHAP и БЯМ Saiga-2 и даны их сравнительные оценки, полученные в результате эксперимента.

Результаты экспериментальных исследований демонстрируют, что интерпретатор SHAP превосходит по точности интерпретации, но уступает БЯМ-интерпретатору Saiga-2 в семантической точности, скорости обработки и удобочитаемости. Интерпретатор SHAP демонстрирует более высокую точность в связи с тем, что он основан на строгих математических методах (теория игр Шепли), тогда как БЯМ-интерпретатор Saiga-2 может допускать «галлюцинации» при генерации объяснений. По стабильности объяснений БЯМ-интерпретатор Saiga-2 (коэффициент вариации 0,12) дает более устойчивые результаты по сравнению с SHAP (0,18) в связи с тем, что при повторных запусках БЯМ сохраняет более согласованные объяснения.

Таблица 1. Пример результатов выделения наиболее важных токенов с помощью интерпретаторов SHAP и БЯМ
Table 1. Example of results of extracting the most important tokens using the SHAP and BJM interpreters

Интерпретатор	Важные токены	
SHAP	1. sign — 0.046 2. Azure — 0.035 3. considers — 0.028 4. Active — 0.021 5. user — 0.019	6. query — 0.017 7. each — 0.016 8. «002D»[1] — 0.013 9. ins — 0.009 10. This — 0.007
БЯМ Saiga-2	1. Azure Active Directory — 0.219 2. sign-in — 0.191 3. application — 0.164 4. anomalous — 0.137 5. change — 0.109	6. profile — 0.082 7. user — 0.055 8. within — 0.027 9. individual — 0.014 10. possibly — 0.003

[1] Unicode.

Таблица 2. Сравнительная оценка параметров интерпретаторов SHAP и БЯМ
Table 2. Comparative evaluation of parameters of SHAP and LLM interpreters

Параметр		SHAP	БЯМ Saiga-2 ДП
Точность интерпретации	Метрика F1-score согласованности с эталонными объяснениями	0,92	0,87
Семантическая согласованность	Метрика BERTScore (сходство с контекстом)	0,65	0,86
Скорость обработки	Время на запрос, с	9,70	4,20
Стабильность объяснений	Коэффициент вариации важности токенов между итерациями	0,18	0,12
Понятность для пользователей	Оценка удобочитаемости (1–5)	2,50	4,60
Вычислительные затраты	Потребление VRAM, ГБ	6,10	14,30

Заключение

В работе продемонстрирована возможность применения больших языковых моделей в качестве интерпретаторов решений различных моделей машинного обучения, в том числе и других языковых моделей. Разработанный подход позволяет обеспечивать генерацию человеко-понятных, семантически обоснованных и устойчивых к вариациям объяснений. Большие языковые модели, применяемые в качестве интерпретаторов, продемонстрировали способность к выделению ключевых признаков (токенов), значимо влияющих на предсказание модели-классификатора, а также к формированию связанных объяснений на естественном языке. Дополненный поиск повышает контекстную релевантность и обеспечивает включение внешних знаний в интерпретацию, что способствовало минимизации семантических ошибок и усилению обоснованности объяснений.

Результаты экспериментов подтверждают достижение поставленной в исследовании гипотезы о том, что большие языковые модели могут быть использованы не только как основа предсказательных моделей, но и как универсальные генераторы объяснений, способные существенно повысить интерпретируемость современных нейросетевых решений.

Таким образом, применение БЯМ-интерпретаторов с компонентом дополненного поиска является перспективным направлением для построения интерпретируемых и доверенных систем искусственного интеллекта. На основе результатов экспериментальных исследований реализации подхода с использованием БЯМ-интерпретатора Saiga-2 и технологии дополненного поиска получены практические результаты, свидетельствующие о высокой эффективности предложенного подхода.

Особое внимание следует уделить универсальности предложенного подхода, который может быть адаптирован для различных архитектур БЯМ и применен в широком спектре задач — от кибербезопасности до аналитики и интеллектуальных систем поддержки принятия решений.

Направления дальнейших исследований включают разработку метрик достоверности объяснений, интеграцию гибридных архитектур объяснения и адаптацию подхода к мультимодальным данным.

Литература

1. Yang Y., Zhang Y., Sun D., He W., Wei Y. Navigating the landscape of AI literacy education: insights from a decade of research (2014–2024) // Humanities and Social Sciences Communications. 2025. V. 12. N 1. P. 374. https://doi.org/10.1057/s41599-025-04583-8
2. Mon-Williams R., Li G., Long R., Du W., Lucas C.G. Embodied large language models enable robots to complete complex tasks in unpredictable environments // Nature Machine Intelligence. 2025. V. 7. N 4. P. 592–601. https://doi.org/10.1038/s42256-025-01005-x
3. Brandenburg J.M., Müller-Stich B.P., Wagner M., van der Schaar M. Can surgeons trust AI? Perspectives on machine learning in surgery and the importance of eXplainable Artificial Intelligence (XAI) // Langenbeck's Archives of Surgery. 2025. V. 410. N 1. P. 53. https://doi.org/10.1007/s00423-025-03626-7
4. Kalasampath K., Spoorthi K.N., Sajeev S., Kuppa S.S., Ajay K., Maruthamuthu A. A Literature review on applications of explainable artificial intelligence (XAI) // IEEE Access. 2025. V. 13. P. 41111–41140. https://doi.org/10.1109/access.2025.3546681
5. Kalmykov V.L., Kalmykov L.V. Towards eXplicitly eXplainable Artificial Intelligence // Information Fusion. 2025. V. 123. P. 103352. https://doi.org/10.1016/j.inffus.2025.103352
6. Mohanty P.K., Francis S.A.J., Barik R.K., Reddy K.H.K., Roy D.S., Saikia M.J. IMPACT: an interactive multi-disease prevention and counterfactual treatment system using explainable AI and a multimodal LLM // PeerJ Computer Science. 2025. V. 11. P. e2839. https://doi.org/10.7717/peerj-cs.2839
7. Gadekallu T.R., Yenduri G., Kaluri R., Rajput D.S., Lakshmanna K., Fang K., Chen J.X., Wang W. The role of GPT in promoting inclusive higher education for people with various learning disabilities: a review // PeerJ Computer Science. 2025. V. 11. P. e2400. https://doi.org/10.7717/peerj-cs.2400
8. Aydin O., Karaarslan E., Erenay F.S., Bacaninet N. Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma // arXiv. 2025. arXiv:2503.04765. https://doi.org/10.48550/arXiv.2503.04765
9. Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., et al. Gemma 3 technical report // arXiv. 2025. arXiv:2503.19786. https://doi.org/10.48550/arXiv.2503.19786
10. Dwivedi R., Dave D., Naik H., Singhal S., Omer R., Patel P., et al. Explainable AI (XAI): core ideas, techniques, and solutions // ACM Computing Surveys. 2023. V. 55. N 9. P. 194. https://doi.org/10.1145/3561048

References

1. Yang Y., Zhang Y., Sun D., He W., Wei Y. Navigating the landscape of AI literacy education: insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications*, 2025, vol. 12, no. 1, pp. 374. https://doi.org/10.1057/s41599-025-04583-8
2. Mon-Williams R., Li G., Long R., Du W., Lucas C.G. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 2025, vol. 7, no. 4, pp. 592–601. https://doi.org/10.1038/s42256-025-01005-x
3. Brandenburg J.M., Müller-Stich B.P., Wagner M., van der Schaar M. Can surgeons trust AI? Perspectives on machine learning in surgery and the importance of eXplainable Artificial Intelligence (XAI). *Langenbeck's Archives of Surgery*, 2025, vol. 410, no. 1, pp. 53. https://doi.org/10.1007/s00423-025-03626-7
4. Kalasampath K., Spoorthi K.N., Sajeev S., Kuppa S.S., Ajay K., Maruthamuthu A. A Literature review on applications of explainable artificial intelligence (XAI). *IEEE Access*, 2025, vol. 13, pp. 41111–41140. https://doi.org/10.1109/access.2025.3546681
5. Kalmykov V.L., Kalmykov L.V. Towards eXplicitly eXplainable Artificial Intelligence. *Information Fusion*, 2025, vol. 123, pp. 103352. https://doi.org/10.1016/j.inffus.2025.103352
6. Mohanty P.K., Francis S.A.J., Barik R.K., Reddy K.H.K., Roy D.S., Saikia M.J. IMPACT: an interactive multi-disease prevention and counterfactual treatment system using explainable AI and a multimodal LLM. *PeerJ Computer Science*, 2025, vol. 11, pp. e2839. https://doi.org/10.7717/peerj-cs.2839
7. Gadekallu T.R., Yenduri G., Kaluri R., Rajput D.S., Lakshmanna K., Fang K., Chen J.X., Wang W. The role of GPT in promoting inclusive higher education for people with various learning disabilities: a review. *PeerJ Computer Science*, 2025, vol. 11, pp. e2400. https://doi.org/10.7717/peerj-cs.2400
8. Aydin O., Karaarslan E., Erenay F.S., Bacaninet N. Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma. *arXiv*, 2025. arXiv:2503.04765. https://doi.org/10.48550/arXiv.2503.04765
9. Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., et al. Gemma 3 technical report. *arXiv*, 2025. arXiv:2503.19786. https://doi.org/10.48550/arXiv.2503.19786
10. Dwivedi R., Dave D., Naik H., Singhal S., Omer R., Patel P., et al. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Computing Surveys*, 2023, vol. 55, no. 9, pp. 194. https://doi.org/10.1145/3561048

11. Salih A.M., Raisi-Estabragh Z., Galazzo I.B., Radeva P., Petersen S.E., Lekadir K., Menegaz G. A perspective on explainable artificial intelligence methods: SHAP and LIME // Advanced Intelligent Systems. 2025. V. 7. N 1. P. 2400304. https://doi.org/10.1002/aisy.202400304
12. Mersha M.A., Yigezu M.G., Tonja A.L., Shakil H., Iskander S., Kolesnikova O., Kalita J. Explainable AI: XAI-guided context-aware data augmentation // Expert Systems with Applications. 2025. V. 289. P. 128364. https://doi.org/10.1016/j.eswa.2025.128364
13. Ng K.K.Y., Matsuba I., Zhang P.C. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations // NEJM AI. 2025. V. 2. N 1. P. 2400380. https://doi.org/10.1056/AIra2400380
14. Strom B.E., Applebaum A., Miller D.P., Nickels K.C., Pennington A.G., Thomas C.B. MITRE ATT&CK: Design and Philosophy. The MITRE Corporation, 2020. 36 p.

11. Salih A.M., Raisi-Estabragh Z., Galazzo I.B., Radeva P., Petersen S.E., Lekadir K., Menegaz G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 2025, vol. 7, no. 1, pp. 2400304. https://doi.org/10.1002/aisy.202400304
12. Mersha M.A., Yigezu M.G., Tonja A.L., Shakil H., Iskander S., Kolesnikova O., Kalita J. Explainable AI: XAI-guided context-aware data augmentation. *Expert Systems with Applications*, 2025, vol. 289, pp. 128364. https://doi.org/10.1016/j.eswa.2025.128364
13. Ng K.K.Y., Matsuba I., Zhang P.C. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI*, 2025, vol. 2, no. 1, pp. 2400380. https://doi.org/10.1056/AIra2400380
14. Strom B.E., Applebaum A., Miller D.P., Nickels K.C., Pennington A.G., Thomas C.B. *MITRE ATT&CK: Design and Philosophy*. The MITRE Corporation, 2020, 36 p.

Авторы

Иванов Роман Дмитриевич — старший помощник начальника отдела, младший научный сотрудник, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, https://orcid.org/0009-0004-1664-1512, vka@mil.ru

Менисов Артем Бакытжанович — доктор технических наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57220815185, https://orcid.org/0000-0002-9955-2694, men.arty@yandex.ru

Мишуков Олег Александрович — кандидат технических наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, https://orcid.org/0009-0004-4247-5810, vka@mil.ru

Сабиров Тимур Римович — кандидат технических наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург, 197198, Российская Федерация, SC 57188236500, https://orcid.org/0000-0002-6807-2954, vka@mil.ru

Authors

Roman D. Ivanov — Senior Assistant to the Head of the Department, Junior Research, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, https://orcid.org/0009-0004-1664-1512, vka@mil.ru

Artem B. Menisov — D.Sc., Senior Lecturer, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57220815185, https://orcid.org/0000-0002-9955-2694, men.arty@yandex.ru

Oleg A. Mishukov — PhD, Senior Lecturer, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, https://orcid.org/0009-0004-4247-5810, vka@mil.ru

Timur R. Sabirov — PhD, Senior Lecturer, Mozhaisky Military Aerospace Academy, Saint Petersburg, 197198, Russian Federation, SC 57188236500, https://orcid.org/0000-0002-6807-2954, vka@mil.ru

Статья поступила в редакцию 30.08.2025
Одобрена после рецензирования 05.11.2025
Принята к печати 19.01.2026

Received 30.08.2025
Approved after reviewing 05.11.2025
Accepted 19.01.2026