https://doi.org/10.25559/SITITO.020.202402.309-315
УДК 004.8

Оригинальная статья

Использование текстов энциклопедий для искусственных нейронных сетей при обучении и обработке запросов

В. В. Медейко
АНО «Интернет-энциклопедия «Рувики», г. Москва, Российская Федерация
Адрес: 125040, Российская Федерация, Ленинградский пр., д. 15, стр. 14
vladimir.medeyko@gmail.com

Аннотация

В статье рассматривается использование текстов интернет-энциклопедий, таких как Википедия и Рувики, для обучения искусственных нейронных сетей (ИНС) класса больших языковых моделей (БЯМ), и обработки ими запросов. Основное внимание уделяется актуальности и качеству обучающих выборок, а также проблемам, связанным с достоверностью и предвзятостью генерируемых ответов.

ИНС, основанные на архитектуре «трансформер», демонстрируют исключительные возможности в различных задачах, связанных с обработкой естественного языка. Однако существует ряд ограничений, включая проблемы с галлюцинациями, когда модели генерируют несуществующие или ложные утверждения. Эти проблемы могут быть обусловлены качеством обучающих выборок, особенностями обучения моделей и обработки запросов.

Энциклопедии, особенно Википедия, широко используются для обучения ИНС благодаря их открытости и структурированности информации. Однако, несмотря на многоязычность и доступность, в статьях Википедии часто присутствует значительный разброс по качеству, что усложняет процесс обучения и повышает риск галлюцинаций.

В качестве дополнения существующих обучающих выборок предлагается использование Рувики — новой интернет-энциклопедии на языках народов России, создаваемой с участием экспертов и с фокусом на достоверность информации. Статьи Рувики проходят тщательную проверку и разметку, что способствует улучшению качества обучающих выборок и снижению риска галлюцинаций.

Также упоминаются другие проекты, такие как «Ковчег Знаний» и онлайн-энциклопедия Большой российской энциклопедии (БРЭ), которые направлены на создание точных и систематизированных информационных баз.

Подчёркивается важность создания региональных интернет-энциклопедий для повышения качества обучающих выборок и уменьшения юридических рисков при использовании БЯМ. Это позволит улучшить точность и релевантность ответов ИНС, что имеет особое значение для пользователей в различных регионах и на разных языках.

Ключевые слова: искусственные нейронные сети, большие языковые модели, машинное обучение, интернет-энциклопедии, Википедия, Рувики, «Ковчег знаний МГУ»

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Медейко В. В. Использование текстов энциклопедий для искусственных нейронных сетей при обучении и обработке запросов // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 2. С. 309-315. https://doi.org/10.25559/SITITO.020.202402.309-315

Original article

Using Encyclopedic Texts for Training and Inference of Artificial Neural Networks

V. V. Medeyko
Autonomous Nonprofit Organization "Internet-Encyclopedia "Ruwiki", Moscow, Russian Federation
Address: 15 Leningradsky Ave., build. 14, Moscow 125040, Russian Federation
vladimir.medeyko@gmail.com

Abstract

The article explores the use of texts from online encyclopedias, such as Wikipedia and Ruwiki, for training and query processing by artificial neural networks (ANNs) of large language models (LLMs). The focus is on the relevance and quality of training datasets, as well as issues related to the accuracy and bias of generated responses.

ANNs based on the "transformer" architecture exhibit exceptional capabilities in various natural language processing tasks. However, there are several limitations, including hallucination problems, where models generate nonexistent or false statements. These issues can be attributed to the quality of training datasets, model training features, and query processing.

Encyclopedias, especially Wikipedia, are widely used for training ANNs due to their openness and structured information. However, despite their multilingualism and accessibility, Wikipedia articles often exhibit significant quality variability, complicating the training process and increasing the risk of hallucinations.

As a supplement to the existing datasets, the use of Ruwiki, a new online encyclopedia in the languages of the peoples of Russia, is proposed. Ruwiki is being created with the participation of experts and focuses on information accuracy. Ruwiki articles undergo thorough verification and annotation, contributing to improved training dataset quality and reduced risk of hallucinations.

Other projects, such as the "Ark of Knowledge" and the online edition of the Great Russian Encyclopedia (GRE), are also mentioned, aimed at creating accurate and systematic information bases.

The article emphasizes the importance of creating regional online encyclopedias to improve training dataset quality and reduce legal risks when using LLMs. This will enhance the accuracy and relevance of ANN responses, which is especially important for users in various regions and languages.

Keywords: artificial neural nets, large language models, machine learning, online encyclopedias, Wikipedia, Ruwiki, Arc of Knowledge

Conflict of interests: The author declares no conflict of interests.

For citation: Medeyko V.V. Using Encyclopedic Texts for Training and Inference of Artificial Neural Networks. *Modern Information Technologies and IT-Education.* 2024;20(2):309-315. https://doi.org/10.25559/SITITO.020.202402.309-315

1. Большие языковые модели

Искусственные нейронные сети (ИНС) класса больших языковых моделей (БЯМ) являются одним из наиболее активно развивающихся и перспективных направлений в области искусственного интеллекта.

Современные БЯМ основаны преимущественно на архитектуре «трансформер»[1] и связанном с ними понятием самовнимания [1]; они содержат многие миллиарды (возможно, некоторые – до двух триллионов[2]) параметров (весов) [2].

Обучение БЯМ происходит в два этапа: предобучение[3], которое производится на огромных массивах практически не размеченных текстов (обучение без учителя) [2], и последующее дообучение[4] для решения конкретных классов задач (обычно с учителем – *instruct*[5] и др. схемы) [1].

Обученная БЯМ позволяет генерировать тексты в ответ на запросы. Этот процесс называется инференс[6].

БЯМ демонстрируют исключительные возможности в различных задачах, связанных с обработкой естественного языка[7], включая перевод, генерацию текстов, написание программ, ответы на вопросы [1], [3].

Для пользователя БЯМ преимущественно представлены чат-ботами – диалоговыми системами, позволяющими пользователю вести диалог с ними на естественном языке. Среди ведущих проектов есть как проприетарные (в первую очередь иностранные ChatGPT (GPT-4o), Bard (PaLM 2), Claude 2, Grok; российские YandexGPT 2, GigaChat [4], так и открытые (LLaMA 3.1, Qwen 2, Mistral (Mixtral), GPT 2, BLOOM, Falcon).

Несмотря на существенные успехи, достигнутые в области БЯМ за последние годы, остается ряд ограничений, которые необходимо учитывать и разрешать, в частности, проблемы достоверности результатов работы БЯМ (степень достоверности каждого конкретного утверждения, сделанного БЯМ, сложно оценить, так как возможности отслеживать, как именно БЯМ к нему пришла очень ограничены), предвзятости (системных отклонений в результатах из-за неравномерной представленности объектов в обучающей выборке), этические проблемы, связанные с эффективностью БЯМ (риски потери людьми работы и непонимания своего предназначения) [1].

2. Обучающие выборки

Для этого класса моделей необходимы большие объёмы текстовой информации. Хотя БЯМ и способны обрабатывать слабо структурированную информацию, важно, чтобы данные для обучения были организованны в систему знания о мире, обоснованную фактами и подкреплённую рациональными аргументами [2].

Обучающие выборки для предобучения можно разбить на восемь основных классов [5]:

1. веб-страницы: характеризуются огромными массивами, динамикой изменений, многоязычностью, широким охватом тем, частичной структурированностью; однако требуют тщательной очистки от мусора, нерелевантной и незаконной информации;
2. корпуса языков;
3. книги: характеризуются широтой охвата, высоким качеством, длиной;
4. академические материалы (самый широко используемый источник – arXiv);
5. исходные программные тексты;
6. параллельные корпуса текстов;
7. социальные сети (больше всего использованы StackExchange и Reddit);
8. энциклопедии (наиболее широко используется Википедия; для китайского – Bǎidù Bǎikē[8]).

Несмотря на то, что ведущие проекты в большинстве случаев выдают качественный результат, выходные тексты могут включать абсолютно не имеющие к реальности и не содержащиеся в обучающей выборке или запросе утверждения (т.н. «галлюцинации»), которые зачастую выглядят весьма убедительно.

Это может привести к чрезвычайно серьёзным рискам во многих сферах, в частности в юридической и медицинской практике [6].

«Галлюцинации» можно разбить на две большие группы [7]:

1. Утверждения, искажающие или фабрикующие факты, информацию о которых модель предположительно должна была получить при обучении.
2. Утверждения, рассогласованные с запросом, контекстом или с нарушением логики.

Причины галлюцинаций могут быть связаны с обучающей выборкой, с особенностями обучения модели и с особенностями обработки запросов (инференса); наконец, галлюцинации могут быть связаны с запросом [7]. Понимание природы галлюцинаций улучшается, как и методы их устранения [8].

Для выявления галлюцинаций применяются различные методы: метрики риска [9], метки на уровне слов [10], безреференсная детекция на уровне токенов [11] и др. Разработаны ряд тестовых систем как общего назначения, так и для узких тем, в которых галлюцинации особенно опасны – таких как медицина [7].

Качество обучающей выборки очень важно для корректной работы систем искусственного интеллекта: увеличение ошибок и прочих проблем в обучающей выборке кардинально и нелинейно увеличивает количество ошибок в ответах нейронной сети.

[1] Англ. transformer. Понятие введено в 2017-м году исследователями из Google Brain в работе [12].

[2] Количество параметров в GPT 4 оценивается в 1,75 млрд [13] (оценка дана Джорджем Хоцем, подтверждена другими экспертами), а в GPT 4 Turbo – 1,96 млрд [14]. Официальных данных о размерах моделей, используемых в ChatGPT – нет.

[3] Англ. pre-training

[4] Англ. fine-tuning

[5] Метод Instruct был реализован в OpenAI, и описан в работе [15].

[6] Англ. inference

[7] Англ. natural language processing

[8] Кит. 百度百科

Проблемы обучающей выборки также распадаются на две большие группы – непосредственно вызванные недостоверностью вводимой при обучении информации (искажение фактов или предвзятость), и связанные с особенностями её обработки (выстраивание ложных корреляций в данных и пропуск важных деталей) – в этом случае повышение качества обучающей выборки также позволяет уменьшать количество «галлюцинаций» [7].

Качество информации важно и при генерации ответов, когда комбинируются нейронные сети и алгоритмические подходы. Наибольшую достоверность результатов дают системы с IR – Information Retrieval, в которых информация извлекается из доверенного корпуса текстов [16]. В частности, распространение получили системы на основе RAG[9], когда модели дополнительно подаётся информация, найденная в базах знания или подобных источниках. БЯМ хорошо умеют сжато излагать информацию, однако при этом могут упускаться важные факты и генерироваться ложные утверждения. Использование RAG позволяет существенно улучшить результат [17].

Существуют и другие подходы увеличения уменьшения галлюцинаций. В частности, метод дистилляции знаний, используемый для оптимизации используемых БЯМ ресурсов (в рамках этого метода на основе взаимодействия с исходной БЯМ-учителем, создаётся гораздо меньшая БЯМ-ученик) может применяться и для устранения галлюцинаций [18].

Предложен подход на основе дополненной массивной модели Mixture of Memory Experts, позволяющий устранять галлюцинации. В частности, такие модели способны запоминать большие массивы случайных чисел без порождения ошибок генерализации [8].

Дополнительной проблемой становится то, что в обучающие выборки попадает всё больше материалов, сформированных ИНС, что приводит к вымыванию важной информации – коллапсу модели. Тем самым возрастает ценность целенаправленно создаваемых пространных систематизированных наборов текстов [19].

3. Энциклопедии

Энциклопедии, особенно классические, хорошо соответствуют требованиям по достоверности материалов и широте и целостности охвата тем. Но классические печатные энциклопедии содержат слишком мало материалов, тогда как для обучения и последующей обработки запросов пользователей требуются большие объёмы информации.

Эта информация может быть использована различными способами (и в том числе и несколькими одновременно):

1. Обучение базовой модели на основе большой обучающей выборки.
2. Дообучение (донастройка) модели (в т.ч. LoRA[10], когда настраиваются только дополнительные веса, а изначальные веса не изменяются; при устранении погрешностей после квантизации весов и др.).
3. При генерации ответа (в т.ч. при использовании различных вариантов RAG, когда на вход модели подмешивается связанная с запросом пользователя информация, найденная во внешних по отношению к модели источниках и т.д.).

К указанным выше содержательным требованиям к обучающей выборке следует добавить потребность в правовой доступности материалов. Использование классических энциклопедий обычно ограничено авторами, либо правовой статус материалов не определён, а потому их использование влечёт определённые риски (в отношении других охраняемых авторским правом текстов уже есть прецеденты правовых споров).

Одной из наиболее пригодных обучающих выборок на настоящий момент является интернет-энциклопедия «Википедия» и производные от неё сетевые энциклопедии. В совокупности они содержат большие массивы достаточно объёмных текстов, включающих основные знания человечества и, что важно, эти тексты распространяются на основе свободных лицензий, исключающих конфликт с правообладателями.

Благодаря тому, что «Википедия» имеет открытый характер и обеспечивает определённый уровень достоверности и структурности, превышающий большинство более доступных источников, её содержимое используются в обучающих выборках практически всех ведущих языковых моделей. Например, одной из первых моделей на основе трансформеров была BERT[11] – в основу её обучающей выборки легла Википедия на английском языке, размером более 2,5 млрд слов [20].

С учётом многоязычности Википедии, её использование в обучающей выборке позволяет модели работать сразу с большим количеством языков. Для некоторых языков – а иногда даже для таких значимых языков, как русский – в обучающей выборке присутствуют только тексты Википедии.

Проблемой является то, что Википедии присущ очень большой разброс по качеству от статьи к статье. Часть статей написана и выверена специалистами в соответствующей, смежной или близкой области знаний; большинство же статей представляют собой результат взаимодействия интересующихся дилетантов; а некоторая часть материалов сфальсифицирована. Эффективного способа различать статьи Википедии по степени достоверности не существует. Интересно, что эти проблемы Википедии схожи с проблемами систем искусственного интеллекта на основе БЯМ.

Многообещающим выглядит эксперимент с системой WikiChat (*Stanford Open Virtual Assistant Lab*), представляющей из себя комплекс подсистем, выполняющих следующие действия:

1. поиск в Википедии релевантной запросу информацией;
2. суммаризация и фильтрация найденной информации;
3. генерация текста при помощи БЯМ;
4. извлечение утверждений из полученного текста и проверка фактов на соответствие Википедии;
5. подготовка содержательного ответа и формулирование его в удобном для восприятия виде.

Эта система достигает достоверности фактов в 97,9% в диалоге с человеком, что примерно на 55% превосходит результат, показываемый GPT-4 [22].

[9] Англ. аббр. от Retrieval Augmented Generation

[10] Англ. аббр. от Low-Rank Adaptation; понятие введено в 2021-м году в работе [21].

[11] Англ. аббр. от Bidirectional Encoder Representaions from Transformers; модель разработана Google в 2018 г.

Возможен и обратный процесс – увеличение достоверности текстов Википедии благодаря применению ИНС [23].

Наконец, материалы Википедии используются при создании метрик, по которым оценивается достоверность информации, генерируемая ИНС [24].

Кроме касающихся достоверности недостатков Википедии, указанных выше, проблема с использованием обучающей выборки или информационной базы на основе Википедии также усугубляется тем, что в этой энциклопедии не учитывается вариация запретов между юрисдикциями. Это приводит к юридическим рискам при применении организациями больших языковых моделей в своих сервисах. Кроме того, Википедия зачастую обладает системным смещением в сторону политических и мировоззренческих позиций, находящихся в России в меньшинстве, а значит, обученные на её текстах ИНС менее эффективно взаимодействуют с большинством российских пользователей.

Сейчас в России реализуется ряд проектов, нацеленных на уменьшение указанных проблем в выборках, связанных с русским языком и другими языками народов России.

Наиболее масштабным проектом на настоящий момент является новая интернет-энциклопедия «Рувики[12]». Она создаётся на языках народов России (уже открыты 20 языковых разделов[13]) [25]. Изначальное содержимое Рувики было легально скопировано из Википедии, что обусловило большой стартовый массив текстов. Заимствованные тексты подвергаются анализу и редактированию на основе заключений экспертов, как являющихся сотрудниками проекта Рувики, так и внешних – в том числе, статьи поступают на экспертизу в Российскую академию наук.

Статьи, прошедшие проверку как внутри редакции, так и внешнюю проверку, соответствующим образом помечаются (они называются «выверенными» и «рецензированными» статьями). Тщательно и подробно помечаются также и статьи, содержание которых может быть неприемлемым в каком-то контексте – например, содержащие информацию, не подходящую для несовершеннолетних.

Подобная разметка может быть использована при обучении и использовании нейронных сетей, что позволяет достичь нужного баланса между достоверностью и уместностью ответов с одной стороны, и широты охвата тем и глубины выдачи – с другой.

Также следует упомянуть онлайн-проект Большой российской энциклопедии (БРЭ), а также информационная система «Ковчег Знаний», создаваемый МГУ. Эти проекты тоже направлены на создание максимально точных онлайн-энциклопедий. Если энциклопедия «Рувики» нацелена на широкую аудиторию, то «Ковчег Знаний», в первую очередь, ориентирован на создание точных статей, адресованных учёным и студентам; кроме того, в этом проекте используется онтологический подход, позволяющий эффективно структурировать информацию [26].

Поскольку технологически «Рувики» и «Ковчег знаний» близки к «Википедии», то разработанные для неё технологии могут быть использованы и для них.

БРЭ продолжает классические традиции Большой советской энциклопедии (БСЭ), технологически обособлена, и очень широко известна.

Сочетание таких различных источников при обучении систем искусственного интеллекта должно обеспечить качественную выдачу любого уровня сложности и подходящую для самой различной русскоязычной аудитории, а многоязычный характер «Рувики» позволит получать высокие результаты и на других языках народов России.

4. Заключение

Интернет-энциклопедии будут продолжать играть существенную роль в обучении и использовании систем искусственного интеллекта. Целесообразно создавать региональные интернет-энциклопедии, ориентированные на потребности конкретных регионов. В частности, создание качественного свободного корпуса текстов на русском языке и других языках народов России должно послужить расширению использования русского языка в обучающих выборках для ИНС; уменьшить количество «галлюцинаций» в их работе; позволить российским организациям использовать такие обучающие выборки без юридических рисков.

References

[1] Kasneci E., et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*. 2023;103:102274. https://doi.org/10.1016/j.lindif.2023.102274

[2] Shen Y., et al. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*. 2023;307(2):e230163. https://doi.org/10.1148/radiol.230163

[3] Raiaan M.A.K., et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*. 2024;12:26839-26874. https://doi.org/10.1109/ACCESS.2024.3365742

[4] Kononov I.A., Petrov V.E. Automation in the tasks of streaming configuration validation. *Science Bulletin.* 2024;1(6):1438-1443. (In Russ., abstract in Eng.) EDN: QIAHIC

[5] Liu Y., et al. Datasets for large language models: A comprehensive survey. *arXiv:*2402.18041. 2024. https://doi.org/10.48550/arXiv.2402.18041

[12] Онтологический подход в информационной системе «Ковчег знаний»: синтез знаний и формализация энциклопедической деятельности / А. Л. Семенов, И. Ю. Гришин, Н. И. Галлини, Р. Р. Тимиргалеева // Международная конференция по мягким вычислениям и измерениям. СПб.: СПбГЭТУ «ЛЭТИ», 2024. Т. 1. С. 243-245. EDN: ACRMRY

[13] На алтайском, башкирском, бурятском, вепсском, ингушском, калмыцком, коми, коми-пермяцком, карельском (ливвиковское наречие), марийском, мокшанском, русском, саха (якутском), татарском, тувинском, удмуртском, хакасском, чеченском, чувашском, эрзянском языках.

[6] Yao J.Y., et al. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples. In: The Twelfth International Conference on Learning Representations (ICLR 2024). Vienna, Austria; 2024. p. 1-13. Available at: https://openreview.net/forum?id=Rh1aThKliu (accessed 29.04.2024).
[7] Huang L., et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems.* 2025;43(2):42. https://doi.org/10.1145/3703155
[8] Li J., et al. Banishing LLM Hallucinations Requires Rethinking Generalization. *arXiv:*2406.17642. 2024. https://doi.org/10.48550/arXiv.2406.17642
[9] Akani E., Favre B., Bechet F., Gemignani R. Reducing named entity hallucination risk to ensure faithful summary generation. In: Proceedings of the 16th International Natural Language Generation Conference. Prague, Czechia: Association for Computational Linguistics; 2023. p. 437-442. https://doi.org/10.18653/v1/2023.inlg-main.33
[10] Rebuffel C., Roberti M., Soulier L., et al. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*. 2022;36:318-354. https://doi.org/10.1007/s10618-021-00801-4
[11] Liu T., et al. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Vol. 1: Long Papers. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 6723-6737. https://doi.org/10.18653/v1/2022.acl-long.464
[12] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Polosukhin I., Kaiser Ł. Attention Is All You Need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA; 2017. p. 1-11. Available at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed 29.04.2024).
[13] Yin J., et al. Evaluation of pre-training large language models on leadership-class supercomputers. *The Journal of Supercomputing*. 2023;79(18):20747-20768. https://doi.org/10.1007/s11227-023-05479-7
[14] Rasheed Z., Waseem M., Systä K., Abrahamsson P. Large Language Model Evaluation Via Multi AI Agents: Preliminary results. *arXiv:*2404.01023. 2024. https://doi.org/10.48550/arXiv.2404.01023
[15] Ouyang L., et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Red Hook, NY, USA: Curran Associates Inc.; 2022. Article number: 2011. p. 27730-27744. Available at: https://openreview.net/forum?id=TG8KACxEON (accessed 29.04.2024).
[16] Zhang Z., Fang M., et al. How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 8289-8311. https://doi.org/10.18653/v1/2023.emnlp-main.516
[17] Gao F., et al. Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts. *arXiv:*2308.10410. 2024. https://doi.org/10.48550/arXiv.2308.10410
[18] McDonald D., Papadopoulos R., Benningfield L. Reducing LLM Hallucination Using Knowledge Distillation: A Case Study with Mistral Large and MMLU Benchmark. *TechRxiv*. 2024. p. 1-19. https://doi.org/10.36227/techrxiv.171665607.76504195/v1
[19] Shumailov I., et al. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv:*2305.17493. 2024. https://doi.org/10.48550/arXiv.2305.17493
[20] Kuznetsov A.V. Digital History and Artificial Intelligence: Perspectives and Risks of Pretrained Language Models. *New Information Technologies In Education and Science.* 2022;(5):53-57. (In Russ., abstract in Eng.) https://doi.org/10.17853/2587-6910-2022-05-53-57
[21] Hu E.J., et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:*2106.09685. 2021. https://doi.org/10.48550/arXiv.2106.09685
[22] Semnani S.J., et al. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. *arXiv:*2305.14292. 2023. https://doi.org/10.48550/arXiv.2305.14292
[23] Petroni F., et al. Improving Wikipedia verifiability with AI. *Nature Machine Intelligence*. 2023;5(10):1142-1148. https://doi.org/10.1038/s42256-023-00726-1
[24] Kuo T.-S., et al. Wikibench: Community-Driven Data Curation for AI Evaluation on Wikipedia. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). New York, NY, USA: Association for Computing Machinery; 2024. Article number: 193. https://doi.org/10.1145/3613904.3642278
[25] Bakeikin S. *Jazykovoe i kul'turnoe raznoobrazie javljaetsja chrezvychajno znachimym aktivom* [Linguistic and cultural diversity is an extremely important asset]. *Universitas*. 2024;(3):68-73. (In Russ., abstract in Eng.) EDN: HVBWKV
[26] Goryachko V.V., Bubnov A.S., Rayevskii E.V., et al. Digital Ark of Knowledge. *Doklady Mathematics*. 2022;106(Suppl 1):S113-S117. https://doi.org/10.1134/S106456242206009

Поступила 29.04.2024; одобрена после рецензирования 27.05.2024; принята к публикации 11.06.2024.
Submitted 29.04.2024; approved after reviewing 27.05.2024; accepted for publication 11.06.2024.

Об авторе:

Медейко Владимир Владимирович, генеральный директор, АНО «Интернет-энциклопедия «Рувики» (125040, Российская Федерация, Ленинградский пр., д. 15, стр. 14), **ORCID: https://orcid.org/0009-0002-2339-2783,** vladimir.medeyko@gmail.com

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Vladimir V. Medeyko, General Director of Autonomous Nonprofit Organization “Internet-Encyclopedia “Ruwiki” (15 Leningradsky Ave., build. 14, Moscow 125040, Russian Federation), **ORCID: https://orcid.org/0009-0002-2339-2783,** vladimir.medeyko@gmail.com

The author has read and approved the final manuscript.