

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЯЗЫКОВЫХ МОДЕЛЕЙ ИИ

А.В. Полуэктов¹, А.И. Заревич¹, Ф.В. Макаренко¹

¹ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

Аннотация. В работе рассматривается история появления языковых моделей, проводится оценка их возможностей. Выполняется сравнительный анализ каждой из моделей, оцениваются их индивидуальные особенности, проводится сравнительный анализ.

Ключевые слова: языковая модель, искусственный интеллект, BERT, GPT, Llama 2, YaLM, GigaChat, DeepSeek V3.

A COMPARATIVE ANALYSIS OF AI LANGUAGE MODELS

A.V. Poluektov¹, A.I. Zarevich¹, Ph.V. Makarenko¹

¹Voronezh State University of Forestry and Technologies named after G.F. Morozov

Abstract. This paper explores the history of the emergence of language models and evaluates their capabilities. A comparative analysis of each model is performed, assessing their individual characteristics, and a comparative analysis is conducted.

Key words: language model, artificial intelligence, BERT, GPT, Llama 2, YaLM, GigaChat, DeepSeek V3.

Развитие искусственного интеллекта (ИИ) – это не просто технологический прогресс, это трансформация человеческой цивилизации. Понимание истории и текущего состояния ИИ, а также его потенциала и ограничений, необходимо для эффективного использования этой технологии и построения будущего, где ИИ будет служить на благо общества. Развитие ИИ связано с решением следующих задач – экономический рост и повышение производительности; решение сложных проблем; улучшение качества жизни; научный прогресс; обеспечение безопасности; социальное воздействие.

Рассмотрение этих задач невозможно без привязки к истории, так ИИ в историческом контексте можно условно разделить на следующие этапы:

1950-е – 1970-е: начало исследований в области ИИ, попытки создать “мыслящие” машины, что привело к развитию алгоритмов поиска, логического программирования и представления знаний. ИИ был ограничен в решении неструктурированных задач и задач с большим объемом данных.

1980-е: практическое применение ИИ в бизнесе и промышленности, создание экспертных систем, основанных на правилах. Экспертные системы были хрупкими, требовали постоянной поддержки экспертов и не могли обучаться самостоятельно, что привело к спаду интереса.

1990-е – 2000-е: возрождение интереса к ИИ благодаря успехам в машинном обучении, особенно статистическим методам и алгоритмам, способным обучаться на больших данных.

2010-е – настоящее время: Революционный прорыв в области ИИ благодаря развитию глубокого обучения и нейронных сетей, позволяющих решать сложные задачи, которые ранее были не под силу. Этические вопросы, предвзятость алгоритмов, проблемы с объяснимостью решений, высокие вычислительные затраты.

Одним из аспектов развития ИИ стало создание языковых моделей (LLM) – от простых систем, основанных на правилах, до сложных нейронных сетей. Оценить их можно только сравнивая направления развития и их возможности.

До 2000-х: системы строились на основе правил. Они были ограничены и не могли обрабатывать сложные и неоднозначные предложения. Позволяли создавать модели, основанные на анализе частоты встречаемости слов и фраз в текстах.

Переход к нейронным сетям и глубокому обучению (2000-е – 2010-е) : характеризуется развитием в трех направлениях:

- рекуррентные нейронные сети (RNN) – позволило моделям обрабатывать последовательности слов и запоминать контекст. RNN значительно улучшили качество машинного перевода и генерации текста.
- эмбединги слов (Word2Vec и GloVe) – позволила представлять слова в виде векторных пространств, где семантически близкие слова расположены рядом. Это улучшило способность моделей к пониманию семантики.
- sequence-to-sequence модели – позволили создавать модели, способные преобразовывать текст из одной последовательности в другую.

2017 – настоящее время:

- трансформеры (Transformer) – модели могут обрабатывать параллельно длинные последовательности слов, значительно ускоряя процесс обучения и улучшая качество результатов.
- BERT (Bidirectional Encoder Representations from Transformers) – позволили точнее понимать текст используя двунаправленные представления слов для лучшего понимания контекста.
- GPT (Generative Pre-trained Transformer) – серия моделей GPT (GPT-1, GPT-2, GPT-3, GPT-4) продемонстрировала впечатляющие возможности в генерации текста, близкого к человеческому.
- большие языковые модели (LLM) – современные LLM, такие как GPT-3, GPT-4, LaMDA, PaLM, представляют собой огромные нейросети, обученные на гигантских наборах данных. Они способны выполнять широкий спектр задач, от генерации текста до написания кода, перевода, суммирования, и ответов на вопросы.

Таким образом, указанные языковые модели позволили обрабатывать информацию в следующих областях: генерация текста, машинный перевод, понимание текста, образование, научные исследования, разработка программного обеспечения, медицина, финансы, юриспруденция, искусство и креативность.

Рассмотрим и проведем анализ использования языковых моделей, разработанных на данный момент.

Языковая модель BERT (Bidirectional Encoder Representations from Transformers). Страна разработки – США, компания-разработчик – OpenAI AI, модель является открытой. Архитектура построения BERT основана на архитектуре Transformer, в частности, на его энкодерной части, использует многослойный двунаправленный механизм внимания (self-attention), что позволяет учитывать контекст как слева, так и справа от слова в предложении. Не использует декодер, так как BERT предназначен для понимания текста, а не для его генерации. Особенностью обработки данных то, что BERT обучается предсказывать как предыдущее, так и последующее слово в предложении, что дает более глубокое понимание контекста. Модель учиться понимать отношения между словами, а также обучается предсказывать, является ли одно предложение продолжением другого. Для обучения используются текстовые данные, такие как Википедия и книги, а затем может быть до обучен для конкретных задач. В качестве заявленных характеристик и возможностей, озвученных разработчиком – улучшенное понимание контекста и семантики текста, высокая точность в задачах классификации

текста, ответов на вопросы, извлечения информации и т.д. При этом модели BERT для различных языков могут давать различные по качеству результаты, хотя и существуют мультиязычные версии BERT (mBERT). BERT хорошо подходит для задач классификации текста, ответов на вопросы, извлечения информации, распознавания именованных сущностей, анализа тональности. Применяется в поисковых системах, чат-ботах, анализе отзывов, медицинской диагностике, анализе юридических документов, анализе социальных сетей.

Следующая языковая модель – GPT (Generative Pre-trained Transformer), причем это целый класс моделей. Страна разработки – США, Компания-разработчик – OpenAI. Архитектура построения GPT построена на использовании одностороннего механизма внимания, что позволяет генерировать текст последовательно, слово за словом. В качестве уникальных особенностей и инновации можно выделить то, что GPT специализируется на генерации связного и релевантного текста, обучается на большом объеме текстов из интернета, а также может быть дообучен на специфических наборах данных для конкретных задач. В качестве основных задач GPT может быть использован для создания чат-ботов, написания статей, создания сценариев, перевода текста, генерации программного кода и как виртуальный ассистент.

Llama 2 представляет собой открытую модель, доступную как для исследовательских, так и коммерческих целей. Страна разработки – США, компания-разработчик – Meta AI (ранее Facebook AI). Архитектура построения близка GPT и оптимизирована для эффективной работы с длинными последовательностями текста. Особенность Llama 2 заключается не только в том, что она предоставляет открытый доступ к исходному коду и весам модели, но и открыта к исследованиям и развитию. Подходит для генерации текста, ответов на вопросы, чат-ботов, диалоговых систем, создания контента.

DeepSeek V3 разработана в Китае. Архитектура построения основана на усовершенствованной архитектуре Transformer. В качестве заявленных характеристик и возможностей должна обладать более высокой точностью и качеством генерации текста по сравнению с существующими моделями, может предоставлять возможности для работы с мультимодальными данными. Должна быть способна выполнять широкий спектр задач: от генерации текста до перевода, суммирования, ответа на вопросы, решения логических задач. Области применения – чат-боты, виртуальные ассистенты, создание контента, научные исследования, образование, разработка, финансы.

И последние две языковые модели разработаны в Российской Федерации – это YaLM от Yandex и GigaChat от Сбербанка.

YaLM (Yet Another Language Model). Архитектура YaLM основана на архитектуре Transformer, аналогично GPT и использует механизм внимания для обработки контекста. Предназначена для решения широкого спектра задач, связанных с русским языком. Обучалась на больших объемах текстовых данных, собранных из интернета, в основном русскоязычных. Также используется англоязычные и другие языковые данные для улучшения обобщения. В качестве заявленных возможностей – генерация текста, перевод, суммаризация, ответы на вопросы, диалоговые системы.

Архитектура GigaChat основана на архитектуре Transformer, использует комбинацию энкодера и декодера.

Ориентирована на русский язык, но поддерживает и другие языки и направлена на бизнес-применения в сервисах и продуктах Sber. Обучается на больших наборах русскоязычных и англоязычных текстов. Также используются специфические наборы данных для обучения моделям конкретным задачам. Выполняет генерацию текста, перевод, ответы на вопросы, диалоговые системы, создание контента. GigaChat доступен через API Sber, а также используется в различных продуктах экосистемы компании.

Рассмотрев ряд языковых моделей можно выделить некоторые характеристики присущие только определенным языковым моделям, рассмотрим их.

Особенности BERT.

- Двухнаправленное обучение (Bidirectional Training) – позволяет модели одновременно анализировать контекст слова с обеих сторон, что обеспечивает глубокое понимание семантики и контекста.
- Маскирование слов (Masked Language Modeling) – обучение строится на предсказании пропущенных слов в предложении, что помогает модели лучше понимать контекстные связи между словами.
- Предобучение для понимания текста – отлично подходит для классификации текста, ответов на вопросы, извлечения информации и других задач, где требуется понимание контекста.
- BERT лучше приспособлен для задач, требующих анализа и понимания текста, а не его генерации.

Особенности GPT.

- Генерация текста, близкого к человеческому (High-quality text generation) – специализация на генерации связного, последовательного и релевантного текста. GPT в больших версиях (GPT-3, GPT-4) достигает очень высокого качества в генерации текста, который трудно отличить от написанного человеком.
- Масштабируемость (Scalability) – GPT-3, GPT-4 демонстрируют способность улучшать качество генерации с увеличением размера модели и объема данных для обучения.
- Способность следовать инструкциям (Instruction Following) – модели хорошо умеют выполнять задачи на основе текстовых инструкций.
- Few-Shot Learning – способность решать новые задачи, используя лишь несколько примеров (few-shot learning).
- Способность к креативности: GPT способен генерировать креативный текст, стихи, сценарии, код и другие виды контента.
- GPT лучше подходит для задач, где требуется генерация текста.

Особенности Llama 2.

- Выделяется тем, что предоставляет открытый доступ к исходному коду и весам модели, что делает ее более доступной для исследований и коммерческого использования.
- Доступна в разных размерах (7B, 13B, 70B), что позволяет выбирать подходящий баланс между производительностью и качеством для конкретных задач.
- Оптимизирована для работы в диалоговых системах, что делает ее хорошим выбором для чат-ботов и виртуальных ассистентов.
- Имеет открытый исходный код и ориентирована больше на генерацию текста.

Особенности YaLM.

- Разработана специально для русского языка и других языков, включая тюркские. Она обладает высоким качеством понимания и генерации текста на русском, что делает ее хорошим выбором для задач, связанных с этим языком.
- Интегрирована в различные сервисы Yandex, что обеспечивает ее эффективное использование в экосистеме компании.
- Yandex предоставляет доступ к некоторым моделям и библиотекам YaLM, что способствует исследованиям.

- Лучше подходит для русскоязычных задач и хорошо интегрирована в сервисы Яндекса.

Особенности GigaChat.

- Разработан с акцентом на бизнес-применения, что делает его подходящим для использования в различных сервисах и продуктах Sber.

- Обеспечивает хорошую поддержку разных NLP задач.

- Разрабатывался для применения в бизнесе.

Особенности DeepSeek.

- Обладает увеличенной длиной контекста, позволяя модели понимать более длинные последовательности текста.

- Может использовать новые методы обучения или оптимизации, которые могут приводить к более высокой точности и качеству генерации текста.

- Может быть оптимизирована для конкретных задач или языков.

Подводя итог, мы можем указать сводную таблицу уникальных возможностей:

Таблица 1 – Сводная таблица уникальных возможностей

Модель	Уникальные возможности
BERT	Двунаправленное обучение, маскирование слов, оптимизация для понимания текста
GPT	Высококачественная генерация текста, креативность, Few-shot learning, масштабируемость, инструктаж
Llama 2	Открытый исходный код, разные размеры, оптимизация для диалоговых систем
YaLM	Русскоязычная ориентация, интеграция в сервисы Yandex, доступность для сообщества
GigaChat	Направленность на бизнес-применения, интеграция в экосистему Sber, поддержка разных NLP задач
DeepSeek V3	Увеличенный контекст, улучшенные алгоритмы обучения, специализация на конкретных задачах или языках, мультимодальность

Таким образом, ни одна модель не является уникальной по всем параметрам изучения, у каждой языковой модели имеются свои уникальные возможности и своя специализация. Поэтому, при выборе модели для конкретной задачи необходимо учитывать ее сильные стороны и возможности. BERT отлично подходит для понимания текста, GPT – для генерации, Llama 2 – открытая и гибкая, YaLM – для русскоязычных задач, GigaChat – для бизнес-задач, а DeepSeek V3, по предположениям, возможно, станет лидером по скорости и универсальности.

Список литературы

1. Васильев, Г. В. Программный-аппаратный комплекс для сбора данных, которые применяются в обучении искусственного интеллекта в области энергетики / Г. В. Васильев, А. В. Васильев, В. Д. Бердоносков // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований: Материалы IV Всероссийской национальной научной конференции студентов, аспирантов и молодых ученых. В 4-х частях, Комсомольск-на-Амуре, 12–16 апреля 2021 года / Редколлегия: Э.А. Дмитриев (отв. ред.), А.В. Космынин (зам. отв. ред.). Том Часть 3. – Комсомольск-на-Амуре: Комсомольский-на-Амуре государственный университет, 2021. – С. 163-165. – DOI 10.17084/978-5-7765-1480-7-2021-163. – EDN FJMMLZ.
2. Савин, И. М. Некоторые особенности эффективности применения искусственного интеллекта в сметах (на примере оптимизации календарных графиков) / И. М. Савин // OpenScience. – 2022. – Т. 4, № 1. – С. 75-82. – DOI 10.51632/2658-7939_2022_4_1_75. – EDN RHEIMR.
3. Косенкова, Е. Ю. Применение искусственного интеллекта в медицине / Е. Ю. Косенкова, Л. В. Гаев // Интеграция, эволюция, модернизация: пути развития науки и образования : сборник статей по итогам Международной научно-практической конференции, Саратов, 18 декабря 2023 года. – Стерлитамак: Общество с ограниченной ответственностью "Агентство международных исследований", 2023. – С. 191-194. – EDN SETDCZ.
4. Мананников, А. О. Применение искусственного интеллекта в образовании / А. О. Мананников, Л. В. Гаев // Инновационные исследования: проблемы внедрения результатов и тенденции развития: Сборник статей по итогам Международной научно-практической конференции, Пермь, 24 декабря 2023 года. – Стерлитамак: ООО "Агентство международных исследований", 2023. – С. 180-182. – EDN HWFVFY.
5. Смольянинова, Е. Е. Применение искусственного интеллекта в психологии / Е. Е. Смольянинова, Л. В. Гаев // Инновационные исследования: проблемы внедрения результатов и тенденции развития: Сборник статей по итогам Международной научно-практической конференции, Пермь, 24 декабря 2023 года. – Стерлитамак: ООО "Агентство международных исследований", 2023. – С. 184-186. – EDN SVJXOI.
6. Визгина, Е. О. Искусственный интеллект как средство перевода / Е. О. Визгина // Проблемы языка и перевода в трудах молодых ученых. – 2024. – № 23. – С. 58-68. – EDN ACWZDT.

7. Мещерякова, А. А. Разработка автоматизированной системы управления мобильным роботом / А. А. Мещерякова // Моделирование систем и процессов. – 2024. – Т. 17, № 1. – С. 73-84. – DOI 10.12737/2219-0767-2024-17-1-73-84. – EDN GCHUTB.

8. Мочалов, В. П. Алгоритм динамического распределения и балансировки нагрузки в распределенных облачных вычислениях / В. П. Мочалов, Н. Ю. Братченко, Д. В. Гостева // Моделирование систем и процессов. – 2024. – Т. 17, № 1. – С. 92-102. – DOI 10.12737/2219-0767-2024-17-1-92-102. – EDN EWMPYU.

9. Минакова, О. В. Повышение эффективности работы в проектах open source на основе архитектурного анализа (на примере проекта Сахана) / О. В. Минакова, И. В. Поцбнева, П. Ю. Гусев // Моделирование систем и процессов. – 2024. – Т. 17, № 1. – С. 84-92. – DOI 10.12737/2219-0767-2024-17-1-84-92. – EDN NZNSKM.

10. Полуэктов А.В., Макаренко Ф.В., Ягодкин А.С. Использование сторонних библиотек при написании программ для обработки статистических данных // Моделирование систем и процессов. – 2022. – Т. 15, № 2. – С. 33-41.

References

1. Vasiliev, G. V. Hardware-Software Complex for Data Collection Used in Training Artificial Intelligence in the Field of Energy / G. V. Vasiliev, A. V. Vasiliev, V. D. Berdonosov // Youth and Science: Current Issues of Fundamental and Applied Research: Proceedings of the IV All-Russian National Scientific Conference of Students, Postgraduates, and Young Scientists. In 4 Parts, Komsomolsk-on-Amur, April 12–16, 2021 / Editorial Board: E. A. Dmitriev (Ed.), A. V. Kosmynin (Deputy Ed.). Part 3. – Komsomolsk-on-Amur: Komsomolsk-on-Amur State University, 2021. – Pp. 163-165. – DOI 10.17084/978-5-7765-1480-7-2021-163. – EDN FJMMMLZ.

2. Savin, I. M. Some Features of the Effectiveness of Artificial Intelligence Application in Estimates (Using the Example of Calendar Schedule Optimization) / I. M. Savin // OpenScience. – 2022. – Vol. 4, No. 1. – Pp. 75-82. – DOI 10.51632/2658-7939_2022_4_1_75. – EDN RHEIMR.

3. Kosenkova, E. Yu. Application of Artificial Intelligence in Medicine / E. Yu. Kosenkova, L. V. Gaev // Integration, Evolution, Modernization: Paths of Development of Science and Education: Collection of Articles Based on the Results of the International Scientific and Practical Conference, Saratov, December 18, 2023. – Sterlitamak: Limited Liability Company "Agency for International Research", 2023. – Pp. 191-194. – EDN SETDCZ.

4. Manannikov, A. O. Application of Artificial Intelligence in Education / A. O. Manannikov, L. V. Gaev // Innovative Research: Problems of Implementation and Development Trends: Collection of Articles Based on the Results of the International Scientific and Practical Conference, Perm, December 24, 2023. – Sterlitamak: LLC "Agency for International Research", 2023. – Pp. 180-182. – EDN HWFVIFY.
5. Smolyaninova, E. E. Application of Artificial Intelligence in Psychology / E. E. Smolyaninova, L. V. Gaev // Innovative Research: Problems of Implementation and Development Trends: Collection of Articles Based on the Results of the International Scientific and Practical Conference, Perm, December 24, 2023. – Sterlitamak: LLC "Agency for International Research", 2023. – Pp. 184-186. – EDN SVJXOI.
6. Vizgina, E. O. Artificial Intelligence as a Translation Tool / E. O. Vizgina // Problems of Language and Translation in the Works of Young Scientists. – 2024. – No. 23. – Pp. 58-68. – EDN ACWZDT.
7. Meshcheryakova, A. A. Development of an Automated Control System for a Mobile Robot / A. A. Meshcheryakova // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 1. – Pp. 73-84. – DOI 10.12737/2219-0767-2024-17-1-73-84. – EDN GCHUTB.
8. Mochalov, V. P. Algorithm for Dynamic Load Distribution and Balancing in Distributed Cloud Computing / V. P. Mochalov, N. Yu. Bratchenko, D. V. Gosteva // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 1. – Pp. 92-102. – DOI 10.12737/2219-0767-2024-17-1-92-102. – EDN EWMPYM.
9. Minakova, O. V. Improving the Efficiency of Work in Open Source Projects Based on Architectural Analysis (Using the Example of the Sahana Project) / O. V. Minakova, I. V. Potsebneva, P. Yu. Gusev // Modeling of Systems and Processes. – 2024. – Vol. 17, No. 1. – Pp. 84-92. – DOI 10.12737/2219-0767-2024-17-1-84-92. – EDN NZNSKM.
10. Poluektov A.V., Makarenko Ph.V., Yagodkin A.S. The use of third-party libraries when writing programs for processing statistical data // Modeling systems and processes. – 2022. – Vol. 15, No. 2. – pp. 33-41.