УДК 550.83:004.8

Возможности использования больших языковых моделей в геологии и геофизике

И.В. Геник
ГИ УрО РАН
614007, Пермь, ул. Сибирская, 78А. E-mail: IVGenik@rambler.ru
Пермский государственный национальный исследовательский университет
614068, Пермь, ул. Букирева, 15. E-mail: geophysics@psu.ru
(Статья поступила в редакцию 25.08.2025 г.)

Рассмотрены предпосылки возникновения больших языковых моделей и их главные особенности. Для геологии и геофизики выделены основные направления применения больших языковых моделей: анализ геологических и геофизических текстов; анализ данных и построение многофакторных связей; пространственное моделирование; генерация программного кода по его описанию. Выделены проблемы развития, во многом связанные с тем, что большие языковые модели находятся на ранней стадии развития. Дано описание перспектив, связанных как с повышением производительности электроники, так и с возможными другими направлениями в организации систем обработки данных.
Ключевые слова: *геофизика, геология, искусственный интеллект, большая языковая модель.*
DOI: 10.17072/psu.geol. 24.4.360

Введение

Развитие геофизики и геологии во многих случаях связано с заимствованием новаций из других наук и технологий. Так, решение обратных задач в геофизике методом подбора стало главным интерпретационным методом в связи с внедрением электронных вычислительных машин, а также использованием достижений математики: теории некорректных задач и математического программирования (оптимизации).

В настоящее время крупнейшей технологической инновацией являются большие языковые модели. Большая языковая модель (LLM) – разновидность с языковой модели (языковая модель – распределение вероятностей по последовательностям слов), построенная с использованием нейронной сети с архитектурой трансформера (архитектура, позволяющая эффективно обрабатывать логически связанные последовательности данных, в том числе текст) с большим числом параметров (от сотен млн в $BERT_{BASE}$ до более одного трлн в Qwen3-Max и Kimi K2), при обучении которой используется метод RLHF (Reinforcement Learning from Human Feedback – обучение с подкреплением на основе отзывов людей). LLM позволяют решать разнообразные задачи, включая генерацию текста, изображений, видео- и аудио-контента, перевод текста, рассуждения и др.

Также существуют малые языковые модели (SLM), предназначенные для обработки естественного языка, включая генерацию, и имеющие от нескольких млн до десятков млрд параметров. SLM нужны для решения узкоспециализированных задач (классификация коротких текстов, чат-боты, быстрый поиск, фильтрация спама и др.). Они требуют меньше ресурсов при обучении и анализе данных, чем большие модели.

История развития больших языковых моделей

Появление больших языковых моделей является одним из результатов работ в области искусственного интеллекта (ИИ), ведущихся более 80 лет, начиная с первого описания искусственной нейронной сети

(1943 г.) и появления в 1946 г. первой ЭВМ общего назначения, которую можно было программировать для решения широкого спектра задач. Работы Н. Винера «Кибернетика» (1948) и А. Тьюринга «Вычислительные машины и разум» («Может ли машина мыслить?») (1950) стали основополагающими для исследований в сфере ИИ.

Практически сразу с появлением ЭВМ, кроме чисто вычислительных задач, начались исследования по широкому кругу вопросов: компьютерная лингвистика (обработка естественного языка, включая перевод текстов, чат-боты (виртуальные собеседники)); нейронные сети; теория информации; игровой ИИ (шахматы, шашки и др.) и математическая теория игр; логические рассуждения. На Дартмутском семинаре (1956) было введено само понятие искусственного интеллекта. С конца 1950-х гг. велись исследования по теории распознавания образов. Позже, начиная с 1960-х гг., после первых космических полетов начались работы по цифровым обработкам изображений, а с 1980-х гг. после появления персональных компьютеров (ПК) наступила эпоха цифровой обработки видео.

На появление больших языковых моделей основное влияние оказали компьютерная лингвистика (главным образом обработка естественного языка), нейронные сети и теория распознавания образов.

Рассмотрим основные этапы развития этих направлений.

Компьютерная лингвистика зарождалась как машинный перевод (МП), который уже на ранних этапах (1954) продемонстрировал определенные успехи. Во второй половине 1950-х гг. появилась теории Ноама Хомского, где основное внимание уделялось моделированию синтаксиса и формальной логике. К 1966 г. выявились ограничения существующих подходов к МП и снизилось его финансирование. 1960–1970-е гг. характеризовались решением частных вопросов, связанных с МП: разрабатывались программы-«вопросники» и диалоговые системы.

В 1980–1990-е годы рост производительности ЭВМ и доступность больших объемов текстовых данных (корпусов) инициировали переход от формальных грамматик, основанных на правилах, к статистическим методам. Применение статистики ведет к проблеме цифровой бесконечности, когда надо задавать вероятности последовательностям слов, которых нет в обучающих данных. Для решения этой проблемы стали применять марковские цепи или нейронные сети, которые значительно улучшили качество систем перевода и распознавания речи.

Период с 2000-х гг. стал эпохой машинного обучения и нейронных сетей. Стало широко применяться глубокое обучение – методы машинного обучения (с учителем, без учителя, с подкреплением), основанные на обучении представлениям, а не на алгоритмах для конкретных задач. Хорошие результаты связаны с развитием теории нейронных сетей (предобучение) и ростом вычислительных возможностей, включая использование графических процессоров (GPU) и технологий распараллеливания задач; можно отметить кратный рост производительности GPU с начала 2000-х гг. Появление в 2017 году новой архитектуры глубоких нейронных сетей (трансформеров) с возможностью распараллеливания стало основой для создания LLM после предобучения на больших объемах данных. Таким образом, нейронные сети оказали решающее влияние на развитие компьютерной лингвистики вследствие отказа от жестких шаблонов и правил.

Теория распознавания образов до 1980-х гг. преимущественно базировалась на методах теории вероятностей и математической статистики. В 1980–2000-е годы наступил период использования нейросетей и машинного обучения: в 1980-е гг. предложен алгоритм обратного распространения ошибки, что позволило обучать многослойные нейронные сети; были созданы сети Хопфилда и Хэмминга. В 1990-е годы исследования смещаются в сторону машинного обучения, где распознавание образов рассматривается как частный случай задачи обучения. Развиваются такие методы, как метод опорных векторов и другие алгоритмы обучения с учителем и без. С 2000-х годов происходит рост интереса к глубокому обучению, появляются сверточные нейронные сети (CNN), которые стали главными в решении задач компьютерного зрения (классификация изображений, обнаружение объектов и др.).

Таким образом, во всех направлениях, которые привели к созданию LLM, важнейшим фактором успеха стало использование искусственных нейронных сетей. Современный этап развития LLM начинается с 2017 г. (появление архитектуры трансформера), когда одна за другой выпускаются большие языковые модели разных фирм: 2018 – модели BERT (Google) и GPT-1 (OpenAI); 2019 – GPT-2 (OpenAI); 2020 – GPT-3 (OpenAI); 2021 – Copilot (GitHub), BERT 2 (Google); 2022 – ChatGPT (OpenAI); 2023 – LLaMa (Meta AI), GigaChat (Сбер), YandexGPT («Яндекс»), Deepseek (High-Flyer); 2024 – Gemini (Google), Llama 3 (Meta AI); 2025 – Alice AI («Яндекс»).

Большие языковые модели в науках о Земле и проблемы использования

Основные направления развития больших языковых моделей позволяют сформулировать варианты их применения в геологии и геофизике: а) обработка геологических описаний и анализ геологических текстов: извлечение информации из текста; реферирование публикаций; сравнение и классификация описаний; генерация пространственных данных и др. (Патук, Наумова, 2023); б) анализ разнородных данных – бурения, геофизических, геохимических, геомеханических и дистанционных измерений; выявление сложных многофакторных связей; обеспечение непрерывного мониторинга (Шокин и др., 2025); в) пространственное моделирование с построением моделей рудных тел, полностью учитывая геологическую обстановку (Колесников, 2024); г) кодогенерация: создание отдельных программ и скриптов для обрабатывающих систем, допускающих автоматизацию (Медведев и др., 2024). Основной объем русскоязычных публикаций по применению больших языковых моделей в геологии и геофизике начинается с 2023 г.

Использование LLM в науках о Земле находится все еще на ранней стадии, а сами языковые модели имеют много проблем.

Во-первых, дорогое оборудование как для сетевого, так и для локального использование, поскольку практически не было конкурентов для чипов американских фирм (Nvidia, AMD, Intel). Только с 2023 г. начались крупные заказы китайских фирм у китайских производителей, и только в 2024 году появилась первая большая языковая модель, обученная с использованием только китайских чипов. Современный уровень развития GPU можно сравнить с уровнем развития персональных компьютеров в 1985–1989 гг.: интересные возможности, но пока еще большие цены и ранняя стадия развития с многими инфраструктурными вопросами.

Во-вторых, большие затраты на обучение моделей как чисто энергетические, так и финансовые. Если цены на первые LLM составляли десятки тысяч долларов, то самые последние языковые модели требуют уже сотен млн долларов, т.е. цена обучения с 2017 года растет экспоненциально.

В-третьих, галлюцинации (ответы, не имеющие отношения к действительности) – общая проблема моделей на основе нейросетей, т.к. при обучении LLM происходит сжатие с потерей информации. Одним из выходов является увеличение параметров, например 1,5 трлн в языковой модели Gemini (Google) в 2025 г.

В-четвертых, в большинстве случаев закрытый характер наборов для обучения, неизвестны использованные фильтры, потенциальная неполнота и противоречивость данных.

В-пятых, проблемы как с сетевыми, так и с локальными LLM. Сетевые модели могут привести к утечке конфиденциальных данных при обучении модели. Локальные модели в условиях быстрой эволюции LLM устаревают и, следовательно, требуют постоянного дообучения и разнообразия обучающих данных для избегания переобучения, т.е. ситуации, когда модель хорошо аппроксимирует обучающую выборку, но плохо работает на других задачах.

В-шестых, большие языковые модели потенциально снижают требования к численности персонала, но значительно повышают требования к квалификации пользователей, а также к надежности оборудования. Микронеисправности GPU и даже неоднородности в работе видеокарт одной модели могут приводить к появлению ошибок. Необходимо применять LLM как инструмент поддержки принятия решений, но не как единственного компетентного специалиста. Квалифициро-

ванное использование больших языковых моделей может приходить в противоречие с желаниями менеджмента по все большей оптимизации рабочих процессов.

Перспективы применения больших языковых моделей

Пути решения указанных выше проблем видятся в следующем.

Во-первых, часть проблем разрешится по мере удешевления графических процессоров (ситуация представляется аналогичной с производительностью ПК в 1981–2006 гг. – быстрый рост производительности и падение цен при массовом использовании).

Во-вторых, преимущественное использование LLM для работ, для которых языковые модели наиболее хорошо обучены: работа с текстами и программирование. Тестирование возможностей новых моделей для все более сложных задач, связанных с выводами и заключениями по представленным материалам. Использование RAG (Retrieval-Augmented Generation) – генерации ответа с использованием результатов поиска (поиск по документам пользователя).

В-третьих, по мере роста числа параметров в больших языковых моделях и роста затрат на их обучение вполне вероятно появление специализированных языковых моделей, включая геолого-геофизические.

В-четвертых, исследование больших языковых моделей может включать элементы обратного инжиниринга (изучение и модификация внутренней структуры) LLM. В настоящее время уже появились инструменты и фреймворки для этого (ChainForge, LangSmith и др.).

В-пятых, возможен переход от чисто нейросетевого подхода к анализу данных к комбинированию нейросетей с другими методами. LBS (Logic-Based Systems) – системы на основе логики, использующие формальную логику для принятия решений или вывода заключений (экспертные системы, логическое программирование, базы данных на основе вывода). CESP (Cognitive Event-Driven Symbolic Processing) – системы, в которых когнитивные процессы запускаются в ответ на значимые события в среде и связываются с символьным представлением и планированием действий (системы управления бизнес-процессами). Neuro-symbolic AI – методы машинного обучения и символические системы (управление складскими роботами; генерация программного кода с помощью Google AlphaCode, Microsoft CodePilot, Facebook DeepCode). Возможно создание гибридов на основе указанных выше методов.

Заключение

Таким образом, большие языковые модели имеют длительную предысторию, когда тестировались различные подходы к работе с информацией. К началу 1990-х годов стали преобладать нейросетевые подходы, к которым впоследствии добавилось глубокое обучение. Создание архитектуры трансформера дало старт бурному росту LLM. Наибольший вклад в развитие языковых моделей вносят компании США и КНР. В России также имеются большие языковые модели собственной разработки: GigaChat (Сбер), YandexGPT/Alice AI («Яндекс»), Cotype (МТС), T-Pro (Т-Банк). Использование LLM открывает большие возможности в обработке и генерации текстов (включая программы), изображений, видео и аудио. Современный этап развития больших языковых моделей характеризуется быстрым ростом их возможностей, можно предположить такое же развитие в ближайшие годы. Большие языковые модели позволяют (как ранее персональные компьютеры) существенно повысить производительность труда в различных областях деятельности, включая геологию и геофизику.

Библиографический список

Колесников А.А. Использование больших языковых моделей в геоинформационных технологиях // Известия высших учебных заведений. Геодезия и аэрофотосъемка. 2024. Т. 68, № 1. С. 33–43. DOI: 10.30533/GiA-2024-003 EDN: JLCPFE

Медведев Д.Н., Сабашный В.Е., Немешев М.Х., Смирнов М.Н. Разработка и применение языковых моделей на основе глубокого обучения в геологии // Санкт-Петербург 2024. Геонауки: современные вызовы и пути решений: сборник материалов 11-й международной геолого-геофизической конференции. М., 2024. С. 178–181. EDN: NLADBZ

Патук М.И., Наумова В.В. Методы искусственного интеллекта для научных исследований в геологии // Электронные библиотеки. 2023. Т. 26, № 5. С. 673–696. DOI: 10.26907/1562-5419-2023-26-5-673-696 EDN: KKZOFT

Шокин Ю.И., Потапов В.П., Попов С.Е. Новые подходы к решению прикладных задач геоэкологии и нелинейной геомеханики на основе больших языковых моделей // Вычислительные технологии. 2025. Т. 30, № 4. С. 26–40. DOI: 10.25743/ICT.2025.30.4.004 EDN: HFXSKR

The Possibilities of Using the Large Language Models in Geology and Geophysics

I.V. Genik
Mining Institute UB RAS, 78a Sibirskaya Str., Perm 614007, Russia.
Perm State University, 15 Bukireva Str., Perm 614068, Russia.
E-mail: geophysics@psu.ru; IVGenik@rambler.ru

The prerequisites for the emergence of the Large Language Models and their main features are considered. The highlighted main areas of application of large language models for geography and geophysics are as follows: analysis of geological and geophysical texts; data analysis and construction of multifactorial relationships; spatial modeling; generation of program code based on its description. The problems are largely related to the fact that large language models are at an early stage of development. The paper describes the prospects related to both improving the performance of electronics and finding the possible other directions in organization of data processing systems.
Keywords: *geophysics; geology; artificial intelligence; large language model*

References

Kolesnikov A.A. 2024. Ispolzovanie bolshikh yazykovykh modeley v geoinformatsionnykh technologiyakh [The use of large language models in geoinformation technologies]. Izvestiya vysshikh uchebnykh zavedeniy. Geodeziya i aerofotosyomka. 68(1):33-43. doi: 10.30533/GiA-2024-003. (in Russian)

Medvedev D.N., Sabashny V.E., Nemeshev M.Kh., Smirnov M.N. 2024. Razrabotka i primenenie yazykovykh modeley na osnove glubokogo obucheniya v geologii [Development and application of language models based on deep learning in geology]. *In:* Geonauki: sovremennye vizovy i puti reshenii. Moskva, pp. 178-181. (in Russian)

Patuk M.I., Naumova V.V. 2023. Metody iskusstvennogo intellekta dlya nauchnykh issledovaniy v geologii [Artificial intelligence methods for scientific research in geology]. Elektronnye biblioteki. 26(5):673-696. doi: 10.26907/1562-5419-2023-26-5-673-696. (in Russian)

Shokin Yu.I., Potapov V.P., Popov S.E. 2025. Novye podkhody k resheniyu prikladnykh zadach geoekologii i nelineynoy geomekhaniki na osnove bolshikh yazykovykh modeley [New approaches to solving applied problems of geoecology and nonlinear geomechanics based on large language models] Vychislitelnye tekhnologii. 30(4):26-40. doi: 10.25743/ICT.2025.30.4.004. (in Russian)