

Новые подходы к решению прикладных задач геоэкологии и нелинейной геомеханики на основе больших языковых моделей

Ю. И. Шокин, В. П. Потапов, С. Е. Попов*

Федеральный исследовательский центр информационных и вычислительных технологий,
650025, Кемерово, Россия

*Контактный автор: Попов Семен Евгеньевич, e-mail: popov@ict.sbras.ru

Поступила 18 апреля 2025 г., доработана 26 июня 2025 г., принята в печать 01 июля 2025 г.

Рассматривается новый подход к решению задач в геоэкологических и геомеханических приложениях, основанный на использовании больших языковых моделей. Обсуждаются проблемы, возникающие в данной прикладной области, типовые модели для решения конкретных задач и сложности, возникающие при их реализации с использованием компьютерных технологий.

Ключевые слова: геоэкология, нелинейная геомеханика, большие языковые модели, методы искусственного интеллекта, нейроассистент, нейросоветник, нейроконсультант, трансфертное обучение.

Цитирование: 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; 30(4):26–40. DOI:10.25743/ICT.2025.30.4.004.

Введение

Сегодня большинство задач в области геоэкологии и нелинейной геомеханики решаются на основе хорошо разработанных вычислительных моделей, использующих такие известные методы, как метод граничных интегральных уравнений, метод конечных элементов и ряд других, а также экспериментальные данные, полученные из натурных измерений. В то же время такие подходы, сталкиваясь с большими потоками разнородных (мультимодальных) данных [1], чаще всего не срабатывают. Заметим, что имеющееся сегодня огромное количество текстового материала редко используется при анализе предметной области, чаще всего в виде ссылок на ту или иную работу, хотя он должен интенсивно использоваться в различных отраслях знаний, чего не происходит, особенно в прикладных задачах.

В предыдущей работе [2] мы рассмотрели возможности применения методов искусственного интеллекта (ИИ), в основном моделей глубокого обучения (deep learning) [3], к решению задач нелинейной геомеханики и экологии, определив конкретные классы задач. Однако бурное развитие больших языковых моделей позволяет не только расширять классы решаемых задач, но и формулировать новые подходы для получения новых знаний. Написание данной статьи стало возможным по результатам совместных обсуждений с Ю.И. Шокиным проблемных вопросов в области геоэкологии и геомеханики.

1. Генеративные модели искусственного интеллекта на основе больших языков

Генеративный ИИ — это тип искусственного интеллекта, который ориентирован на генерацию нового контента, данных или результатов, имитирующих человеческое творчество и воображение [4]. Генеративный ИИ может создавать новые и разнообразные результаты на основе изученных моделей и данных [5]. Как и для большинства моделей ИИ, при работе с ними требуется процесс обучения, который в зависимости от типа модели реализуется с помощью специальных программных архитектур [6]. В настоящей работе нас интересует достаточно широкий класс моделей, создаваемых на основе больших языков (large language models) [7].

Большая языковая модель (LLM) — это высокоточная компьютерная модель, предназначенная для обработки и понимания естественного языка. Она основана на глубоком обучении и может анализировать и генерировать текст, подобный тому, как пишут и говорят люди. Большие языковые модели основаны на нейронных сетях [8], которые используют большие объемы лингвистических данных для обучения, но это одновременно означает, что для них невозможно узнать что-либо из используемого набора информации (если вы не предоставите им это в подсказке!) [9]. Обучение LLM — дорогостоящий и трудоемкий процесс. Они обучаются на огромных объемах данных. Примером такой модели является GPT-3 (generative pre-trained transformer) от OpenAI [10]. Модель обучается на больших объемах текста из Интернета и других источников и может понимать и генерировать текст, имитирующий стиль и содержание речи реальных людей. Крупномасштабные языковые модели обучаются путем предоставления модели большого количества текстовых данных [11]. Затем модель автоматически настраивает веса и параметры, чтобы предсказать следующее слово в предложении или ответ на запрос. Такое обучение опирается на оптимизационные методы, такие как стохастический градиентный спуск, и требует значительных вычислительных ресурсов. Схожесть между крупномасштабными языковыми моделями и генеративным ИИ заключается в том, что LLM могут генерировать новый текст, ответы на запросы и завершения предложений. Они не только понимают язык, но и могут создавать свой собственный язык, что делает их мощными инструментами для автоматического создания контента [12]. Однако данные, на которых обучаются LLM, являются историческими (или датированными). Например, последняя модель GPT4 от OpenAI имеет сведения только до апреля 2023 г., и о любом событии, произошедшем после этой даты, информация недоступна [13]. Используя данную информацию, попытаемся определить классы задач нелинейной геомеханики и экологии, которые могут быть решены с использованием LLM.

1.1. Примеры задач для моделей больших языков

Моделирование нелинейного поведения грунта. В работе [14] успешно применена генеративная модель для прогнозирования деформации грунта при высоких технологических нагрузках. Аналогичные результаты могут быть получены при использовании модели типа GPT-3 для учета сложных нелинейных взаимосвязей, которые описываются на уровне предполагаемых качественных связей, получаемых при обработке конкретных корпусов текста, взятого из различных литературных источников.

Оценка воздействия процессов добычи на окружающую среду. В работе [15] проанализированы текстовые данные, полученные с помощью языковой модели, для

предсказания распространения загрязняющих веществ в окружающей среде. Крупномасштабные языковые модели, такие как ChatGPT [16], LLAMA [17], Mistral-7B [18], способны существенно улучшить эту методологию, привлекая дополнительные данные, обеспечивая тем самым комплексную оценку воздействия.

Оптимизация процессов горных работ. В работе [19] предложен метод оптимизации процесса добычи на основе предварительного обучения текстовых данных с использованием GPT-2 [20]. Этот подход может быть улучшен при использовании более мощных моделей, таких как GPT-3, GPT-4, с помощью которых можно исследовать большее количество методов оптимизации, а также оценивать возможности интерпретируемости применяемых моделей, используя известные методы ее оценки.

Трансферное обучение (перенос знаний, полученных в одной задаче, на решение другой задачи). В контексте больших языковых моделей трансферное обучение может быть применено для дообучения модели на специфических данных или задачах, чтобы улучшить ее производительность в конкретных областях. Джонсон и др. [21] представили метод трансферного обучения для адаптации крупномасштабных языковых моделей к специфическим особенностям геомеханических и экологических данных. Этот метод может быть эффективно применен в горнодобывающих отраслях и экологии.

Интеграция с географическими информационными системами (ГИС). Предлагается интегрировать данные языковых моделей с ГИС для более точного анализа пространственных параметров. Однако для успешного внедрения необходимо тщательно подготовить и структурировать данные. При этом потребуются разработка специализированных API (advanced program interface), связывающих модели и ГИС [22].

Интеллектуальное предсказание внезапных газогеохимических явлений в горнодобывающей промышленности. Работа [23] посвящена применению методов машинного обучения для прогнозирования сейсмических колебаний грунта, вызванных взрывными работами. В целом работа вносит вклад в растущий объем исследований, использующих методы машинного обучения и гибридные методы искусственного интеллекта для улучшения прогнозирования сейсмических колебаний грунта, вызванных взрывными работами в горной промышленности. Такие модели способствуют более эффективному планированию и контролю взрывных операций с целью снижения их негативного воздействия на близлежащие сооружения и окружающую среду и основываются на методах искусственного интеллекта.

Непрерывный мониторинг и обратная связь. Развертывание больших языковых моделей в системах мониторинга в реальном времени обеспечивает динамический подход к управлению геомеханическими и экологическими процессами. Такие модели могут непрерывно анализировать данные из различных источников, обеспечивая мгновенную информацию об экологических изменениях и состоянии недр, переведенные на естественный язык. Получаемая информация в режиме реального времени позволяет горнодобывающим компаниям оперативно корректировать свою деятельность, снижая вероятность непредвиденных нарушений окружающей среды и оптимизируя использование ресурсов [24].

Передача знаний в горнодобывающем секторе и экологии. Для повышения квалификации работников можно создавать учебные материалы, учебные модули и соответствующую документацию, предоставляя всеобъемлющие и доступные ресурсы, которые способствуют формированию грамотной и квалифицированной рабочей силы, знакомой с традиционными методами добычи полезных ископаемых [25]. Это,

в свою очередь, будет способствовать повышению способности горнодобывающего сектора сбалансировать добычу ресурсов и защиту окружающей среды.

Даже такое достаточно небольшое перечисление возможных задач, решение которых можно осуществлять с использованием LLM, говорит о их универсальности и широких возможностях. В дальнейшем мы планируем расширить список таких задач, однако необходимо оценить и сложности их применения для исследуемых областей.

2. Сложности применения больших языковых моделей

При использовании больших языковых моделей в нелинейной геомеханике и экологии могут возникать проблемы и ограничения. Рассмотрим некоторые из них.

Вычислительные ресурсы. Большие языковые модели требуют значительных вычислительных ресурсов для обучения и инференса, включая высокопроизводительные графические процессоры (GPU) или специализированные аппаратные средства, такие как tensor processing units (TPU) [26].

Объем данных. Для обучения больших языковых моделей требуется значительное количество данных. В области нелинейной геомеханики и экологии может оказаться сложным найти достаточное количество качественных данных, особенно в случае редких или слабо исследованных явлений.

Интерпретируемость. Большие языковые модели, такие как Transformer [27], могут быть сложными для интерпретации. Они воспринимаются как “черный ящик”, поскольку внутренние механизмы и процессы моделирования окажутся сложны для понимания или объяснения [28].

Переносимость. Языковые модели могут быть обучены на доступных данных, но их применимость к новым геоматериалам или условиям может быть ограничена. Модели могут не всегда обладать достаточной робастностью при обработке новых или нетипичных данных.

Исследование влияния размера языковой модели на вычислительные ресурсы и время, необходимое для анализа. Размер языковой модели имеет значительное влияние на вычислительные ресурсы и время, необходимое для проведения анализа. Большие модели с большим количеством параметров требуют больше памяти и вычислительных мощностей для их обучения и использования. Увеличение размера модели может привести к улучшению качества генерации, но это будет требовать больших вычислительных ресурсов. Поэтому исследователи должны учитывать свои вычислительные ограничения, сбалансировать размер модели и качество результатов с точки зрения доступных ресурсов.

Ограничения в обучающих данных. В контексте оценки состояния окружающей среды в исследовании [29] подчеркивается, что ограниченность данных может стать препятствием для эффективного прогнозирования на основе больших языковых моделей. Применение методов расширения (augmentation) данных может преодолеть эту проблему.

Проблема галлюцинаций. Это одна из сложнейших проблем применения больших языковых моделей, особенно в реализации любого приложения, поддерживаемого ими. В отличие от традиционных систем искусственного интеллекта, ориентированных на ограниченные задачи, LLM во время обучения связаны с обработкой огромных объемов текстовых онлайн-данных. При этом в системе проявляются отрицательные особенности, такие как экстраполяция информации на основе систематических ошибок

в обучающих данных, неправильная интерпретация, двусмысленные подсказки или изменение информации для поверхностного соответствия входным данным. Это вызывает значительные опасения, когда возможности генерации языка используются для критических приложений, таких как безопасность горных работ, оценка экологического воздействия и др. В дальнейшем изложении мы приведем один из методов устранения галлюцинаций, но отметим как многогранность, так и сложность проблемы, которая может привести к неэффективному использованию больших языковых моделей, особенно обученных на больших текстовых корпусах [30, 31].

Этические риски. В работе [32] поднимается вопрос об этических рисках при использовании ИИ в экологически чувствительных областях. Для снижения таких рисков необходимо разработать стандарты исследований и их прозрачность. Существуют этические и потенциальные риски, связанные с использованием больших языковых моделей, такие как возможность создания и распространения фейковых новостей, создание вводящего в заблуждение контента и проблемы конфиденциальности данных из-за необходимости хранения больших объемов информации для обучения. Эту достаточно большую область необходимо в дальнейшем исследовать для конкретных приложений.

3. Особенности создания больших языковых моделей для решения задач нелинейной геомеханики и экологии

Реализации любого приложения, ориентированного на использование LLM, можно разделить на два больших класса:

- Приложения, использующие универсальные генеративные большие языковые модели типа ChatGPT, Mistral-7B, Llama-2 и др. [33]. При этом LLM являются основными компонентами современных приложений искусственного интеллекта, применяемых для естественного языка. У них есть потенциал для эффективной обработки и понимания человеческого языка с помощью самых разных приложений: от виртуальных помощников и машинного перевода до обобщения текста и ответов на вопросы.
- Традиционные подходы к LLM, которые опираются в основном на большие, предварительно обученные модели и общедоступные наборы данных и предлагают ограниченные возможности для тонкой настройки и адаптации. Технологически это означает, что пользователь получает ответы на свои вопросы, просто задавая их в диалоговом режиме. Однако в силу универсальности больших моделей ответ может не соответствовать действительности. Именно для решения этой проблемы существует технология подсказок (prompts) [34], которая детализирует и уточняет запрос пользователя за счет дополнительного контента. Затем этот запрос уже с подсказками поступает в универсальную модель, которая дает уточненный ответ на запрос пользователя. И чем подробнее составлены подсказка и ее дополнительный контент, тем точнее может быть ответ.

Общая схема работы с большой универсальной моделью выглядит следующим образом (рис. 1).

Существуют конкретные правила составления подсказок для произвольных запросов, которые можно найти в сети Интернет. В наиболее развитых генеративных моделях больших языков, например ChatGPT-3.5, ChatGPT-4.0, существует специальный сервис, обеспечивающий создание подробных подсказок [35].

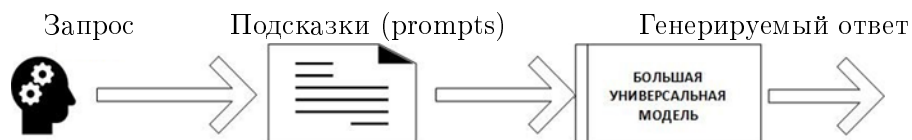


Рис. 1. Схема работы с универсальной моделью

Fig. 1. Scheme of work with the universal model

Приведем конкретный пример подсказки и соответствующего контента для вопроса “Что такое нелинейная геомеханика?” Сгенерированная системой ChatGPT-3.5 подсказка “Расскажите о принципах и основных концепциях нелинейной геомеханики, и какие аспекты она включает в себя в контексте изучения механического поведения геологических материалов под воздействием различных нагрузок и условий.”

Приведем ответ системы. “Нелинейная геомеханика — это раздел науки о горных породах и грунтах, который изучает нелинейное механическое поведение геологических материалов под воздействием различных нагрузок и условий. В отличие от линейной геомеханики, которая предполагает линейные зависимости между напряжением и деформацией, нелинейная геомеханика рассматривает нелинейные, часто неоднородные и неоднородно деформируемые материалы. Принципы нелинейной геомеханики включают в себя учет таких явлений, как пластичность, разрушение, изменение свойств материалов в зависимости от условий нагружения, деформаций и времени. Этот подход позволяет более точно моделировать и предсказывать поведение геологических материалов в различных геотехнических условиях, таких как строительство подземных сооружений, горные работы, а также природные геологические процессы, например сейсмическая активность. Основные концепции нелинейной геомеханики включают в себя исследование гиперпластичности, деформации под воздействием больших напряжений, моделирование разрушения материалов, а также анализ влияния геотермических и гидрогеологических факторов на механическое поведение геологических формаций. Этот подход позволяет инженерам и геологам более эффективно управлять рисками и разрабатывать безопасные и устойчивые геотехнические проекты.”

Таким образом, этот простой пример указывает на определенную гибкость предлагаемых моделей, которые в определенной степени могут достаточно подробно отвечать на более сложные вопросы. Гибкость проявляется в способности LLM:

- адаптироваться к специализированным терминам (на примере “нелинейной геомеханики”);
- генерировать структурированные объяснения сложных концепций без дополнительного обучения;
- обрабатывать мультимодальные данные (как указано во введении и заключении).

Дополнение: по сравнению с традиционными методами (например, МКЭ), где такие объяснения требуют ручного моделирования, предлагаемый подход имеет понятные преимущества.

Однако большие универсальные языковые модели (LLM) неизбежно имеют галлюцинации, поскольку точность сгенерированных текстов не может быть обеспечена исключительно за счет параметрических знаний, которые они инкапсулируют [36]. Поэтому при создании проблемно-ориентированных моделей на основе LLM требуется иной подход, заключающийся в использовании LLM, которые могут постоянно обучаться и адаптироваться на основе новых данных и опыта. Такая необходимость в частом повторном обучении всей модели делает ее более актуальной и современной. Метод

ды непрерывного обучения и адаптации, такие как активное обучение, трансферное обучение, метаобучение и интеграция данных в реальном времени, позволяют LLM постоянно корректировать свои внутренние параметры на основе новой информации. Это повышает производительность модели с течением времени. База знаний LLM постоянно развивается. По мере добавления новых данных и взаимодействия с пользователем [37] система на основе технологий RAG-генерации (retrieval augmented generation) с расширенным поиском обновляет свое понимание и улучшает возможности рассуждения, тем самым отдавая предпочтение релевантности и точности.

В настоящей работе мы не имеем возможности описать большинство новых технологий создания моделей такого типа, но в дальнейшем постараемся более подробно исследовать вопросы их адекватности. Ниже рассмотрим одну из наиболее эффективных технологий (RAG), которая позволяет дообучать модели, созданные на основе универсальных, и частично решать проблемы, связанные с галлюцинациями, тем самым предоставляя более релевантную информацию для конкретных задач. Иными словами, RAG подпадает под категорию методов контекстного обучения, которые направлены на предоставление дополнительных знаний предварительно обученным моделям с целью повышения качества их ответов.

Одна из возможных концептуальных схем RAG показана на рис. 2. Опишем работу схемы:

- пользователь формирует запрос, который преобразуется через механизм поиска;
- преобразованный запрос через семантический векторный поиск отправляется в (векторную) базу дополнительных знаний;
- наиболее подходящие (релевантные) данные с соответствующим контекстом в форме подсказки (prompts) передаются в предобученную модель (соответствующую прикладной области);
- модель отвечает сгенерированным текстом, преобразуемым в наиболее удобную для пользователя форму.

Рассмотренная схема достаточно проста, но в реальности для устранения галлюцинаций могут применяться более сложные архитектуры, которые позволяют улучшить качество ответа на специальные запросы. В настоящей работе мы рассматриваем наиболее общую технологию для реализации больших языковых моделей на основе универсальных. Таким образом, для решения конкретных задач с использованием LLM можно применять два рассмотренных выше подхода:

- Решать задачу, используя расширенные подсказки для универсальных моделей типа ChatGPT, MISTRAL-7B, LLAMA-2 и др., количество которых стремительно растет. Отметим, что не все модели успешно работают с русским языком.



Рис. 2. Концептуальная схема RAG

Fig. 2. RAG Conceptual Diagram



Рис. 3. Концептуальная модель нейросоветника

Fig. 3. Conceptual model of a neuroadvisor

- Создавать прикладные инструменты, применяя RAG или другие способы конкретного обучения с использованием дополнительных обучающих материалов.

Чтобы упростить использование и понимание LLM, мы предлагаем методику разработки прикладного программного обеспечения с применением нейросоветников. Общая структура нейросоветника состоит из трех блоков (рис. 3). Каждый из них выполняет характерный набор функций, реализующих элементы обработки текстовой информации, используемые для больших языковых моделей.

Блок “Нейроассистент” осуществляет первоначальную подготовку информации, получая ее из различных источников с разнообразными текстовыми форматами. Загрузка документов из списка источников может оказаться сложным процессом. Поэтому надо обязательно заранее спланировать все источники и загрузчики. Чаще всего требуются преобразования, очистка загруженных данных, такие как удаление дублированного контента, анализ его разделения на главы и разделы и т. д. Для этого существуют специализированные программные средства. Следующим шагом в конвейере нейроассистента является разделение документов на управляемые фрагменты. Для чего необходим такой шаг? На это имеются две причины.

- Простой доступ к большим объемам данных довольно трудно получить. Поэтому разбиение данных на более мелкие фрагменты повышает качество данных.
- Существенен размер контекстного окна. В LLM количество лексем, доступных для подсказки и завершения, ограничено, так как влечет за собой резкое увеличение вычислительных ресурсов. Поэтому контекст не может быть больше, чем позволяет контекстное окно.

На следующем этапе конвейера обработки необходимо разделение корпуса текста на части. Оно может показаться простым процессом, однако здесь существуют определенные методы, от которых может зависеть успешность создаваемой LLM. Таким образом, нейроассистент должен ответить на два вопроса, на которых следует сосредоточиться.

1. Как разбивается текст?
2. Каков размер текстового фрагмента?

Наиболее распространенный подход заключается в том, что заранее определяются размер текстовых фрагментов и возможность их перекрытия. Такой подробно рассмотренный процесс называется токенизацией [38], и для него имеется хорошо отработанные алгоритмы [39].

Все модели машинного обучения ИИ работают с числовыми данными. Перед выполнением какой-либо операции все данные (текст, изображения, аудио, видео) должны быть преобразованы в числовое представление. При этом создаются так называемые эмбединги (embeddings — встраивание) — это векторные представления данных, которые фиксируют значимые отношения между сущностями [40]. По общему определению встраивания это данные, которые были преобразованы в n -мерные матрицы для

использования в вычислениях глубокого обучения. Встраивание слов фактически означает векторное представление слов. С момента выпуска ChatGPT и появления других LLM создание моделей встраивания резко ускорилось. Сегодня развивается множество стандартов для оценки LLM и моделей встраивания. При создании приложений LLM с использованием RAG возникают вопросы: какую модель встраивания использовать для моделей и какую модель встраивания нужно использовать для работы с RAG? Обычно рассматриваются алгоритмы встраивания в семи вариантах использования — классификация, кластеризация, парная классификация, переранжирование, поиск, семантическое текстовое сходство и суммаризация. Однако несмотря на многочисленные практические рекомендации [41], выбор алгоритма встраивания, как и токенизации, сегодня представляет определенную сложность. В данной работе мы уделяем большое внимание схеме работы нейроассистента только потому, что, как и в нейронных сетях глубокого обучения, подготовка данных имеет решающее значение и может занимать до 60–70 % в общем времени разработки модели.

Блок “Нейроконсультант” воспринимает уже подготовленные данные и обучает большую модель, которая формируется либо при помощи расширенных подсказок, либо с использованием RAG или других технологий. Этот блок имеет отличительную особенность, так как дообучение модели на конкретных данных требует значительных вычислительных ресурсов на уровне мини-кластеров, в чем мы убедились на собственном опыте.

Блок “Нейроэксперт” является одним из самых проблемных. По нему можно дать лишь только общие рекомендации, так как его структура определяется той задачей, для которой должны быть получены новые знания. Например, для задач нелинейной геомеханики этот блок может быть связан с выбором оптимальных технологий, которые снижали бы техногенное воздействие. В таком случае создается узкоспециализированный RAG, в который загружаются данные по технологиям добычи, а также определенные методы их оценки, которые можно описать на уровне текстов или таблиц. Фактически данный блок может использовать специфические модели для дообучения по рассмотренной выше схеме, а также получать новые знания средствами общих генеративных моделей. Кроме того, возможно объединение отдельных частей эксперта и консультанта посредством разработки так называемых гибридных моделей. Данная работа в настоящее время находится на начальном этапе, и мы проводим вычислительные эксперименты в этом направлении. Итак, структура прикладной LLM имеет вид:

- Нейроассистент:
 - 1) загрузка данных из разноформатных источников (PDF, ГИС, текстовые файлы);
 - 2) очистка и токенизация (алгоритмы SentenceTransformers);
 - 3) создание векторных эмбеддингов (модель all-MiniLM-L6-v2).
- Нейроконсультант:
 - 1) тонкая настройка (fine-tuning) универсальной LLM (например, Mistral-7B) на доменных данных;
 - 2) интеграция RAG для доступа к актуальным источникам;
 - 3) оптимизация ресурсов (мини-кластеры GPU).
- Нейроэксперт:
 - 1) специализированные агенты для задач (например, прогноз выбросов газа на основе [23]);
 - 2) гибридные модели (комбинация LLM с численными методами).

В предлагаемой нами технологии многие процессы, которые необходимо детально рассматривать, инкапсулируются в соответствующие блоки. При изменении алгоритмов, входящих в любой из них, остальные блоки не изменяются, так как связаны только через обмен данными с конкретной структурой. Само количество алгоритмов в блоке также может изменяться, например при изменении структуры RAG в нейроконсультанте. Предлагаемая структура организации прикладной LLM достаточно удобна и гибко адаптируется под любой класс задач.

Заключение

Рассмотрен новый подход к созданию моделей нелинейной геомеханики и экологии на основе использования больших языковых моделей (LLM). Обозначен возможный круг задач, для которых такие модели могут найти применение, он значительно увеличен. Предложена технологическая структура, называемая нейросоветником, которая позволяет инкапсулировать определенные этапы построения моделей и упростить пользование ими. В целом предложенный подход может быть использован для решения широкого круга задач нелинейной геомеханики и экологии. Преимуществами предлагаемого подхода являются возможность работы с мультимодальной информацией и его независимость от традиционных методов математического моделирования, так как извлечение знаний происходит на закрытом для пользователя уровне, но при этом любые методы математического моделирования достаточно просто могут быть включены в систему на уровне агентов. Создан интерфейс для работы нейросоветника с системой экологического мониторинга угольных предприятий и выполняется его обучение на экспериментальных данных по загрязнению воздушной среды. На этапе обучения модель анализирует данные датчиков (загрязнение воздуха, вибрации) и техническую документацию, генерируя прогнозы рисков на основе RAG с векторной базой данных нормативных актов (ГОСТ Р 58901-2020). В последующих работах планируется более подробно рассмотреть каждый из этапов создания нейросоветника с учетом всех ограничений, указанных выше.

Благодарности. Работа выполнена при финансовой поддержке Российского научного фонда (грант № 23-17-00148, <https://rscf.ru/project/23-17-00148/>).

Список литературы

- [1] **Philippe S., Souchet A.D., Lameris P., Petridis P., Caporal J., Coldeboeuf G., Duzan H.** Image processing and pattern recognition: fundamentals and techniques. N.Y.: IEEE Press; 2010: 537.
- [2] **Опарин В.Н., Киряева Т.А., Потапов В.П., Попов С.Е.** Особенности применения методов искусственного интеллекта в задачах нелинейной геомеханики. Безопасность и мониторинг природных и техногенных систем: материалы и доклады, Красноярск, 16–20 октября 2023 года. Красноярск: ФИЦ ИВТ; 2023: 65–67. DOI:10.25743/SSTS.2023.70.41.014.
- [3] **Ugail H.** Deep learning in visual computing: explanations and examples. CRC Press; 2023: 140.
- [4] **Solanki A., Nayyar A., Naved M.** Generative adversarial networks for image-to-image translation. Academic Press; 2021: 427.
- [5] **Ghayoumi M.** Generative adversarial networks in practice. Boca Raton London: CRC Press; 2024: 671.

- [6] **Joe R., Matt H.** Fundamentals of data engineering: plan and build robust data systems. O'Reilly Media, Inc.; 2022: 446.
- [7] **Sinan O.** Quick start guide to large language models: strategies and best practices for using ChatGPT and other LLMs. Addison-Wesley: Pearson Education; 2023: 132.
- [8] **Vasilev I.** Python deep learning: understand how deep neural networks work and apply them to real-world tasks. 3rd ed. Birmingham: Packt Publishing; 2023: 361.
- [9] **Kulkarni A., Shivananda A., Kulkarni A., Gudivada D.** Applied generative AI for beginners: practical knowledge on diffusion models, ChatGPT, and other LLMs. Apress Berkeley; 2023: 212.
- [10] **Kubik S., Shubham S.** GPT-3: building innovative NLP products using large language models. O'Reilly Media; 2022: 151.
- [11] **He W., Liu P., Qian Q.** Machine learning contests: a guidebook. Springer: Posts & Telecom Press; 2023: 398.
- [12] **Kaplan J.** Generative artificial intelligence: what everyone needs to know. Oxford University Press; 2024: 241.
- [13] **IRMA.** Natural language processing: concepts, methodologies, tools, and applications. IGI Global; 2020: 1704.
- [14] **Baghbani A., Choudhury T., Costa S., Reiner J.** Application of artificial intelligence in geotechnical engineering: a state-of-the-art review. Earth-Science Reviews. 2022; (228):10.
- [15] **Jones B.E., Grantz K.H., Lindgren J.F.** Using natural language processing to extract environmental footprint data from text sources. Environmental Science & Technology. 2019; 53(13):7397–7406.
- [16] **Alto V.** Modern generative AI with ChatGPT and OpenAI models: leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4. Birmingham: Packt Publishing; 2023: 285.
- [17] **Bsharat S.M., Myrzakhan A., Shen Z.** Llama2 principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. Available at: <https://arxiv.org/abs/2312.16171> (accessed March 19, 2025).
- [18] **Jiang A.Q., Sablayrolles A., Mensch A., Bamford C., Albert Q., Chaplot D.S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L.R., Lachaux M.-A., Stock P., Le Scao T., Lavril T., Wang T., Lacroix T., El Sayed W.** Mistral 7B. Available at: <https://arxiv.org/pdf/2310.06825> (accessed March 19, 2025).
- [19] **Wang S., Sun Y., Zhao L., Li L.** Application of transfer learning in intelligent optimization of coal mining technology. 2020 IEEE 17th International Conference on Networking, Sensing and Control (ICNSC'2020). 2020: 1–6.
- [20] **Rothman D.** Transformers for natural language processing: build innovative deep neural network architectures for NLP with Python, PyTorch. BERT, RoBERTa, T5, GPT-2, architecture of GPT-3, and much more. Packt Publishing. 2021: 385.
- [21] **Johnson C., Zhang Y., Cai H.** A transfer learning framework for geotechnical engineering applications. Available at: <https://arxiv.org/pdf/2104.05704> (accessed March 17, 2025).
- [22] **Li Z., Guo W., Yuan Y., Wang H.** A spatiotemporal ensemble model for predicting urban land use changes. Computers, Environment and Urban Systems. 2022; (91):101688.
- [23] **Gupta A., Verma A.K.** Machine learning techniques for predicting ground vibration due to blasting in mines. Journal of Rock Mechanics and Geotechnical Engineering. 2022; 14(3):616–631.

-
- [24] **Santos B.F., Nogueira T.A., Souza L.S.** Machine learning models for predicting subsidence in coal mining areas. This study delves into the development of machine learning models to predict subsidence, offering insights into ground movement in coal mining areas; 2021.
- [25] **Tan W., Zhang W., Liu S., Zheng L., Wang X., An B.** True knowledge comes from practice: aligning LLMs with embodied environments via reinforcement learning. Available at: <https://arxiv.org/pdf/2401.14151v1> (accessed March 3, 2025).
- [26] **Tan B., Zhu Y., Liu L., Wang H., Zhuang Y., Chen J., Xing E., Hu Z.** Redco: a lightweight tool to automate distributed training of LLMs on any GPU/TPU. Available at: <https://arxiv.org/pdf/2310.16355> (accessed March 17, 2025).
- [27] **Tunstall L., Von Werra L., Wolf T.** Natural language processing with transformers: building language applications with hugging face revised edition. O'Reilly Media, Inc.; 2022: 406.
- [28] **Bhattacharya A.** Applied machine learning explainability techniques: make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more. Birmingham: Packt Publishing; 2022: 305.
- [29] **Lopez A., Nocita M., Guastella D., Palombo A.** A review on the use of remote sensing technologies for the estimation of green house gases emissions and removals in forests. Journal of Environmental Management. 2020: 277.
- [30] **Li J., Cheng X., Zhao W.X., Nie J.Y., Wen J.** A large-scale hallucination evaluation benchmark for large language models. Available at: <https://arxiv.org/pdf/2305.11747v3> (accessed March 07, 2025).
- [31] Practical steps to reduce hallucination and improve performance of systems built with large language models. 2024. Available at: <https://newslettervictordibia.com> (accessed January 26, 2024).
- [32] **Jobin A., Ienca M., Vayena E.** The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019; 1(9):389–399.
- [33] **Patel A., Arasanipalai A.** Applied natural language processing in the enterprise: teaching machines to read, write, and understand. O'Reilly Media; 2021: 336.
- [34] **Mille R.E.** Prompt engineering bible: join and master the AI revolution. Independently Published; 2023: 249.
- [35] **Caelen O., Blete M.-A.** Developing apps with GPT-4 and ChatGPT: build intelligent chatbots, content generators, and more. O'Reilly Media; 2023: 157.
- [36] **Нобак М.** Объекты. Стильное ООП. СПб.: Питер; 2023: 304.
- [37] **Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kuttler H., Lewis M., Yih W., Rocktaschel T., Riedel S., Kiela D.** Retrieval-augmented generation for knowledge-intensive NLP tasks. Available at: <https://arxiv.org/abs/2005.11401> (accessed January 26, 2025).
- [38] **Ashok K.L., Karthika R.D.** Deep learning approach for natural language processing, speech, and computer vision: techniques and use case. CRC Press; 2023: 246.
- [39] **Fregly C., Barth A., Eigenbrode S.** Generative AI on AWS: building context-aware multimodal reasoning applications. O'Reilly Media, Inc.; 2024: 309.
- [40] **Magajna T.** Natural language processing with Flair: a practical guide to understanding and solving NLP problems with Flair. Packt Publishing; 2022: 199.
- [41] **Amaratunga T.** Understanding large language models: learning their underlying concepts and technologies. Apress; 2023: 166.
-

MATHEMATICAL MODELLING

DOI:10.25743/ICT.2025.30.4.004

New approaches to solving applied problems of geoecology and nonlinear geomechanics based on large language model

Yu. I. SHOKIN, V. P. POTAPOV, S. E. POPOV*

Federal Research Center for Information and Computational Technologies, 650025, Kemerovo, Russia

*Corresponding author: Semyon E. Popov, e-mail: popov@ict.sbras.ru

Received April 18, 2025, revised June 26, 2025, accepted July 01, 2025.

Abstract

Purpose. The article addresses analyzing of the introduced new approaches for solving applied problems in geoecology and nonlinear geomechanics using large language models — a rapidly developing branch of artificial intelligence. The authors seek to demonstrate how LLMs can address limitations of traditional computational methods, especially when working with large, heterogeneous, and multimodal datasets.

Methodology. The authors review existing computational approaches (e.g., boundary integral equation and finite element methods) and highlight their constraints in handling unstructured textual and multimodal data. They propose the integration of LLMs, trained on extensive textual corpora, to extract, synthesize, and generate domain-specific knowledge. The article discusses several practical applications, such as modelling nonlinear soil behavior, predicting environmental impacts, optimizing mining operations, and implementing transfer learning. The integration of LLMs with geographic information systems and real-time monitoring systems is also explored.

Findings. LLMs can significantly expand the range of solvable problems in geoecology and geomechanics. They enhance predictive accuracy, support knowledge transfer, and facilitate the creation of educational resources. However, challenges include high computational demands, data quality and availability, model interpretability, and ethical risks such as misinformation and privacy concerns.

Originality/value. The article is among the first to systematically evaluate the potential and challenges of LLMs in geoecological and geomechanical applications. It emphasizes the importance of prompt engineering and domain adaptation for effective LLM use, and highlights the transformative potential of these models for research, industry, and education in the field.

Keywords: geoecology, nonlinear geomechanics, large language models, artificial intelligence methods, neuroassistant, neuroadvisor, neuroconsultant, transfer learning.

Citation: Shokin Yu.I., Potapov V.P., Popov S.E. New approaches to solving applied problems of geoecology and nonlinear geomechanics based on large language model. Computational Technologies. 2025; 30(4):26–40. DOI:10.25743/ICT.2025.30.4.004. (In Russ.)

Acknowledgements. This research is supported by Russian Science Foundation (grant No. 23-17-00148, <https://rscf.ru/project/23-17-00148/>).

References

1. Philippe S., Souchet A.D., Lameris P., Petridis P., Caporal J., Coldeboeuf G., Duzan H. Image processing and pattern recognition: fundamentals and techniques. N.Y.: IEEE Press; 2010: 537.

2. **Oparin V.N., Kiryaeva T.A., Potapov V.P., Popov S.E.** Osobennosti primeneniya metodov iskusstvennogo intellekta v zadachakh nelineynoy geomekhaniki [Features of the application of artificial intelligence methods in problems of nonlinear geomechanics]. *Bezopasnost' i Monitoring Prirodnikh i Tekhnogennykh Sistem: Materialy i Doklady, Krasnoyarsk, 16–20 Oktyabrya 2023 Goda* [Safety and Monitoring of Natural and Man-Made Systems: Materials and Reports, Krasnoyarsk, October 16–20, 2023]. Krasnoyarsk: FITs IVT; 2023: 65–67. (In Russ.)
3. **Ugail H.** Deep learning in visual computing: explanations and examples. CRC Press; 2023: 140.
4. **Solanki A., Nayyar A., Naved M.** Generative adversarial networks for image-to-image translation. Academic Press; 2021: 427.
5. **Ghayoumi M.** Generative adversarial networks in practice. Boca Raton London: CRC Press; 2024: 671.
6. **Joe R., Matt H.** Fundamentals of data engineering: plan and build robust data systems. O'Reilly Media, Inc.; 2022: 446.
7. **Sinan O.** Quick start guide to large language models: strategies and best practices for using ChatGPT and other LLMs. Addison-Wesley: Pearson Education; 2023: 132.
8. **Vasilev I.** Python deep learning: understand how deep neural networks work and apply them to real-world tasks. 3rd ed. Birmingham: Packt Publishing; 2023: 361.
9. **Kulkarni A., Shivananda A., Kulkarni A., Gudivada D.** Applied generative AI for beginners: practical knowledge on diffusion models, ChatGPT, and other LLMs. Apress Berkeley; 2023: 212.
10. **Kubik S., Shubham S.** GPT-3: building innovative NLP products using large language models. O'Reilly Media; 2022: 151.
11. **He W., Liu P., Qian Q.** Machine learning contests: a guidebook. Springer: Posts & Telecom Press; 2023: 398.
12. **Kaplan J.** Generative artificial intelligence: what everyone needs to know. Oxford University Press; 2024: 241.
13. **IRMA.** Natural language processing: concepts, methodologies, tools, and applications. IGI Global; 2020: 1704.
14. **Baghbani A., Choudhury T., Costa S., Reiner J.** Application of artificial intelligence in geotechnical engineering: a state-of-the-art review. *Earth-Science Reviews*. 2022; (228):10.
15. **Jones B.E., Grantz K.H., Lindgren J.F.** Using natural language processing to extract environmental footprint data from text sources. *Environmental Science & Technology*. 2019; 53(13):7397–7406.
16. **Alto V.** Modern generative AI with ChatGPT and OpenAI models: leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4. Birmingham: Packt Publishing; 2023: 285.
17. **Bsharat S.M., Myrzakhan A., Shen Z.** Llama2 principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. Available at: <https://arxiv.org/abs/2312.16171> (accessed March 19, 2025).
18. **Jiang A.Q., Sablayrolles A., Mensch A., Bamford C., Albert Q., Chaplot D.S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L.R., Lachaux M.-A., Stock P., Le Scao T., Lavril T., Wang T., Lacroix T., El Sayed W.** Mistral 7B. Available at: <https://arxiv.org/pdf/2310.06825> (accessed March 19, 2025).
19. **Wang S., Sun Y., Zhao L., Li L.** Application of transfer learning in intelligent optimization of coal mining technology. 2020 IEEE 17th International Conference on Networking, Sensing and Control (ICNSC'2020). 2020: 1–6.
20. **Rothman D.** Transformers for natural language processing: build innovative deep neural network architectures for NLP with Python, PyTorch. BERT, RoBERTa, T5, GPT-2, architecture of GPT-3, and much more. Packt Publishing. 2021: 385.
21. **Johnson C., Zhang Y., Cai H.** A transfer learning framework for geotechnical engineering applications. Available at: <https://arxiv.org/pdf/2104.05704> (accessed March 17, 2025).
22. **Li Z., Guo W., Yuan Y., Wang H.** A spatiotemporal ensemble model for predicting urban land use changes. *Computers, Environment and Urban Systems*. 2022; (91):101688.
23. **Gupta A., Verma A.K.** Machine learning techniques for predicting ground vibration due to blasting in mines. *Journal of Rock Mechanics and Geotechnical Engineering*. 2022; 14(3):616–631.
24. **Santos B.F., Nogueira T.A., Souza L.S.** Machine learning models for predicting subsidence in coal mining areas. This study delves into the development of machine learning models to predict subsidence, offering insights into ground movement in coal mining areas; 2021.

25. **Tan W., Zhang W., Liu S., Zheng L., Wang X., An B.** True knowledge comes from practice: aligning LLMs with embodied environments via reinforcement learning. Available at: <https://arxiv.org/pdf/2401.14151v1> (accessed March 3, 2025).
26. **Tan B., Zhu Y., Liu L., Wang H., Zhuang Y., Chen J., Xing E., Hu Z.** Redco: a lightweight tool to automate distributed training of LLMs on any GPU/TPU. Available at: <https://arxiv.org/pdf/2310.16355> (accessed March 17, 2025).
27. **Tunstall L., Von Werra L., Wolf T.** Natural language processing with transformers: building language applications with hugging face revised edition. O'Reilly Media, Inc.; 2022: 406.
28. **Bhattacharya A.** Applied machine learning explainability techniques: make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more. Birmingham: Packt Publishing; 2022: 305.
29. **Lopez A., Nocita M., Guastella D., Palombo A.** A review on the use of remote sensing technologies for the estimation of green house gases emissions and removals in forests. Journal of Environmental Management. 2020: 277.
30. **Li J., Cheng X., Zhao W.X., Nie J.Y., Wen J.** A large-scale hallucination evaluation benchmark for large language models. Available at: <https://arxiv.org/pdf/2305.11747v3> (accessed March 07, 2025).
31. Practical steps to reduce hallucination and improve performance of systems built with large language models. 2024. Available at: <https://newslettervictordibia.com> (In Russ.) (accessed 26.01.2024)
32. **Jobin A., Ienca M., Vayena E.** The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019; 1(9):389–399.
33. **Patel A., Arasanipalai A.** Applied natural language processing in the enterprise: teaching machines to read, write, and understand. O'Reilly Media; 2021: 336.
34. **Mille R.E.** Prompt engineering bible: join and master the AI revolution. Independently Published; 2023: 249.
35. **Caelen O., Blete M.-A.** Developing apps with GPT-4 and ChatGPT: build intelligent chatbots, content generators, and more. O'Reilly Media; 2023: 157.
36. **Nobak M.** Objects. Stylish OOP. St. Petersburg: Peter; 2023: 304. (In Russ.)
37. **Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kuttler H., Lewis M., Yih W., Rocktaschel T., Riedel S., Kiela D.** Retrieval-augmented generation for knowledge-intensive NLP tasks. Available at: <https://arxiv.org/abs/2005.11401> (accessed January 26, 2025).
38. **Ashok K.L., Karthika R.D.** Deep learning approach for natural language processing, speech, and computer vision: techniques and use case. CRC Press; 2023: 246.
39. **Fregly C., Barth A., Eigenbrode S.** Generative AI on AWS: building context-aware multimodal reasoning applications. O'Reilly Media, Inc.; 2024: 309.
40. **Magajna T.** Natural language processing with Flair: a practical guide to understanding and solving NLP problems with Flair. Packt Publishing; 2022: 199.
41. **Amaratunga T.** Understanding large language models: learning their underlying concepts and technologies. Apress; 2023: 166.