РЕФЕРАТ

Марчук Иван ИУ6-31М

Проблемы появления сильного ИИ

В реферате исследуются концепции сильного и сверхсильного искусственного интеллекта (ИИ), способного адаптироваться к изменяющимся условиям и решать сложные задачи. Анализируются ключевые характеристики, необходимые для реализации сильного ИИ, включая мышление, память, планирование и обучение. Обсуждаются современные технологические достижения и ограничения, такие как сложность моделей, ресурсозатраты, регуляторные барьеры и уязвимости инфраструктуры. Особое внимание уделено альтернативным подходам, таким как модульная архитектура и использование слабых ИИ в качестве строительных блоков для сильного ИИ.

Введение

Сильный или общий ИИ (AGI) [1] – это ИИ, который может ориентироваться в меняющихся условиях, моделировать и прогнозировать развитие ситуации. Если ситуация выходит за стандартные алгоритмы, то он должен самостоятельно найти ее решение. Например, решить задачу «поступить в университет». Или изучить правила игры в шашки, и вместо шахмат начать играть в шашки.

Также, различные исследователи выделяют суперсильный ИИ (ASI) [1]. Это ИИ, который может не только решать сложные задачи, но и делать это практически моментально. Если слабых ИИ уже тысячи, под каждую задачу, сильных ИИ будут десятки (скорее всего будет разделение по направлениям), но вот суперсильный ИИ будет 1 на государство или даже на всю планету.

Какими качествами должен обладать сильный ИИ?

Мышление [2] - использование таких методов как дедукция, индукция, ассоциация и т.д., которые направлены на выделение фактов из информации, их представление (сохранение). Это позволит точнее решать задачи в условиях неопределённости.

Память - использование различных типов памяти (кратковременная, долговременная). Память может быть использована для решения задач опираясь на предыдущий опыт. Даже если Вы попробуете пообщаться с ChatGPT 4, то увидите, что алгоритм обладает небольшой краткосрочной памятью, и через 10-15 сообщений забывает, с чего все начиналось.

Планирование - тактическое и стратегическое. Да, уже есть исследования, которые утверждают, что ИИ может планировать свои действия и даже обманывать человека для достижения своих целей. Но сейчас это все равно только в стадии зарождения. И чем глубже будет планирование, особенно в условиях неопределенности, тем больше нужно мощностей. Ведь одно дело планировать игру в шахматы на 3-6 шагов в глубину, где все правила четкие, и совсем другое найти правила в ситуации неопределенности.

Обучение - имитация действий другого объекта и обучение через проведение экспериментов. Сейчас ИИ учится на больших массивах данных, но он сам не моделирует и не проводит экспериментов. Хотя, мы не до конца понимаем, как работает тот же Chat GPT внутри. И это одна из главных проблем. Но обучение требует формирования долгосрочной памяти и сложных взаимосвязей. А это, проблема для ИИ.

Сейчас сильного ИИ нет ни у кого. Мы лишь на стадии перехода от слабого к промежуточному. Да, ChatGPT от OpenAI, LaMDA от Google и другие большие языковые модели (LLM) умеют генерировать текст / иллюстрацию / видео через анализ запроса и обработку больших данных. Но они лишь транслируют то, чему обучили их создатели. Они ищут наиболее вероятные варианты сочетания слов, или слов и изображений, пытаются имитировать человеческую деятельность. А в их ответах много «брака» и «галлюцинаций». К реальному взаимодействию с миром они еще не готовы.

Ограничения на пути к сильному ИИ

Во-первых, появление сильного или суперсильного ИИ — это очень затратный и сложный процесс с точки зрения регуляторных ограничений [3]. Эпоха бесконтрольного развития публичных ИИ заканчивается. На него будет накладываться все больше и больше ограничений.

В риск-ориентированном подходе сильный и суперсильный будут на верхнем уровне риска. А значит и ограничения будут заградительные. Уже сейчас разработчики ИИ, в том числе ChatGPT, сталкиваются с судебными исками о нарушении авторских прав. И это до введения жестких правил.

Во-вторых, это сложная задача с технической точки зрения, причем сильный ИИ будет и очень уязвим.

Сейчас, в середине 2020-х, для создания и обучения сильного ИИ нужны гигантские вычислительные мощности и сложные ИИ-модели. Придется экспоненциально увеличивать количество нейронов и выстраивать связи между ними. Если человеческие нейроны могут быть в нескольких состояниях, а

активация может происходить «по-разному» (да простят меня биологи за такие упрощения), то машинный ИИ так не может. То есть, условно, машинные 80-100 млрд нейронов не равны 80-100 млрд у человека. Машине потребуется больше нейронов. Тот же GPT4 оценивают в 100 трлн параметров (условно нейронов), и он все равно уступает человеку.

Факторы, мешающие появлению сильного ИИ.

Первый фактор - рост сложности [4]**.** Рост сложности всегда приводит к проблемам надежности, увеличивается количество точек отказа. Такие модели сложно как создавать, так и поддерживать от деградации во времени, в процессе работы. ИИ-модели нужно постоянно «обслуживать», направляя дообучение в правильную сторону.

Для раскрытия этой проблемы приведу пример из автоспорта. Например, гонки Формулы 1. Так, если взять пример в вакууме, отставание в 1 секунду можно устранить, если вложить 1 млн и 1 год. Но вот чтобы отыграть решающие 0,2 секунды может потребоваться уже 10 млн и 2 года работы. А фундаментальные ограничения конструкции машины могут заставить вообще пересмотреть всю концепцию гоночной машины. И чем сложнее машина, тем труднее её обслуживать. Если взять современные гиперкары той же формулы 1, то после каждого выезда требуются целые команды техников для приведения гиперкара в исходное состояние.

Если вернуться к вопросу деградации ИИ, то помимо технологий тут будет и влияние людей. Любой ИИ, особенно на раннем этапе, будет обучаться на основе обратной связи от людей (их удовлетворённость, начальные запросы и задачи). Примером тут может служить тот же ChatGPT4. Так, например, ChatGPT использует запросы пользователей для дообучения своей модели. И в конце 2023 года стали появляться статьи, что ИИ-модель стала «более ленивой». Её ответы становятся короче. Чат-бот либо отказывается отвечать на вопросы, либо прерывает разговор, либо отвечает просто выдержками из поисковиков и других

сайтов. Причем к середине 2024 года это уже стало нормой, когда модель просто приводит выдержки из Википедии.

А одна из возможных причин в том, что сами пользователи стали задавать все более простые и примитивные запросы. Ведь LLM (большие языковые модели) не придумывает ничего нового, эти модели пытаются понять, что вы хотите от них услышать и подстраиваются под запрос. Она ищет максимальную эффективность связки трудозатраты-результат, таким образом решая задачу максимизациии функции.

Второй фактор - количество данных. Да, мы можем увеличить текущие модели на порядки. Но тому же прототипу ChatGPT5 уже в 2024 году не хватает данных для обучения [5]. Сильному ИИ для первоначального обучения потребуется огромное количество качественных данных, а для наращивания мощности ещё больше.

Третий фактор - ИИ-модель будет привязана к своей «базе» [6]**.** Ей потребуются огромные и сложные дата-центры для работы, с мощными источниками энергии и качественным охлаждением. Так, по некоторым оценкам на 5 - 50 запросов для ChatGPT 4 уходит до 0,5л воды на охлаждение. Для более мощных моделей этот показатель будет также расти, хотя возможно это нивелируется технологическим развитием ИИ-ускорителей.

И какими бы ни были пропускными каналы интернета, все равно основные вычисления будут сосредоточены в небольших дата центрах, не будет распределенных сетей обработки данных.

Во-первых, распределенные вычисления все равно теряют в производительности и эффективности. Кроме того, распределенная сеть не может гарантировать работу вычислительных мощностей постоянно.

Во-вторых, это уязвимость перед атаками на каналы связи и ту же распределенную инфраструктуру. Представьте, что вдруг 10% нейронов вашего мозга просто перестали работать (блокировка каналов связи или просто отключились), а остальные тупят или работают вполсилы. В итоге снова имеем риск получить сильный ИИ, который забывает кто он, где он, или еще что-то.

А уж если все придет к тому, что сильному ИИ потребуется тело для взаимодействия с миром, то реализовать это будет еще сложнее. Тогда ИИ-модель будет ограниченной, иначе как все это обеспечивать энергией и охлаждать? Откуда брать мощности для обработки данных? То есть это будет ограниченная ИИ с постоянным подключением к основному центру по беспроводной связи. А это снова уязвимость. Современные каналы связи дают выше скорость, но это сказывается на снижении дальности действия и проникающей способности. Кроме этого, такие каналы и проще подавить средствами радиоэлектронной борьбы. То есть мы получаем рост нагрузки на инфраструктуру связи и рост рисков.

Тут можно, конечно, возразить. Например тем, что можно взять предобученную модель и сделать ее локальной. Да, в таком виде все это может работать на одном сервере. Но такой ИИ будет очень ограничен, это будет «промежуточный» ИИ, и он будет «тупить» в условиях неопределенности и ему все равно нужна будет энергия. То есть это не про создание человекоподобных суперсуществ. Это будет большое количество сильных ИИ, но дорогих и с ограниченными возможностями, что не очень интересно рынку.

Все эти факторы приведут к геометрическому росту сложности и затрат на создание, развитие и поддержание сильного ИИ. Затраты на исследования и создания прототипов могут преодолеть 0,5 – 1 трлн долларов США.

При этом, слабые модели с узкой специализацией останутся более «свободными» и простыми для создания, а главное востребованными.

Все это приводит к вопросам об экономической целесообразности инвестиций в это направление. Тем более с учетом двух ключевых трендов в развитии ИИ:

- создание дешевых и простых локальных моделей для решения специализированных задач;
- создание ИИ-оркестраторов, которые будут декомпозировать запрос на несколько локальных задач и затем перераспределять это между разными локальными моделями.

В итоге мы имеем более простое и дешевое решение рабочих задач, нежели создание сильного ИИ.

Сильный ИИ на основе взаимодействия слабых

Однако слабые ИИ, успешно решающие отдельные задачи, такие как обработка текста, изображений или речи, могут стать строительными блоками для формирования более сложного и универсального интеллекта [7].

Идея модульной системы СИИ основывается на взаимодействии различных узкоспециализированных модулей. В такой архитектуре каждый модуль отвечает за выполнение своей задачи, а общий слой управления координирует их взаимодействие. Например, голосовой ассистент может сочетать модули для обработки речи, понимания текста и управления задачами, чтобы обеспечивать комплексную функциональность. Ключевая роль в этой системе отводится метауровню — дирижёру, который решает, какие модули активировать и как интерпретировать их результаты.

Однако, одной из главных проблем является ограниченная универсальность таких систем. Слабые ИИ оптимизированы для конкретных задач и плохо адаптируются к другим. Это приводит к необходимости создания сложного слоя координации, который способен учитывать контекст, разрешать конфликты между модулями и принимать согласованные решения. Кроме того, такие системы требуют значительных вычислительных ресурсов, что затрудняет их масштабирование и делает их менее эффективными по сравнению с более универсальными подходами.

Контекст и обобщение также остаются ключевыми проблемами. Современные слабые ИИ обладают узкой специализацией и не способны учитывать полный контекст или делать выводы за пределами своей области. Например, система, анализирующая изображения, не сможет эффективно работать с текстовыми данными. Это усложняет работу оркестратора и ограничивает возможности системы в реальных условиях, где задачи часто выходят за рамки заранее определённых сценариев.

Кроме того, такие системы имеют ограничения в креативности и адаптивности. СИИ, построенный на слабых ИИ, вряд ли сможет интуитивно решать новые задачи или находить нестандартные подходы. Его эффективность будет сильно зависеть от качества данных, на которых обучались модули. Скорее он будет похож на агрегатор слабых ИИ.

Будущее таких систем, вероятно, связано с развитием новых подходов к интеграции, более эффективным координационным алгоритмам и созданием модулей, способных к более глубокой адаптации, хотя опять же не такой полной как хотелось бы от СИИ.

Таким образом, хотя идея создания сильного ИИ на основе слабых ИИ выглядит привлекательной, её реализация связана с множеством технических и концептуальных ограничений. Это требует значительных усилий и инноваций в области искусственного интеллекта, нейронаук и компьютерных технологий. И в целом я убежден, что сильный ИИ — это не вопрос ближайшего будущего.

Вывод

Резюмируя вышесказанное, у сильного ИИ можно выделить несколько фундаментальных проблем.

1. Экспоненциальный рост сложности разработки и деградация сложных моделей.
2. Недостаток данных для обучения.
3. Стоимость создания и эксплуатации.
4. Привязанность к ЦОД и требовательность к вычислительным ресурсам.
5. Низкая эффективность текущих моделей по сравнению с человеческим мозгом.

Именно преодоление этих проблем определит дальнейший вектор развития всей технологии: либо все же сильный ИИ появится, либо мы уйдем в плоскость развития слабых ИИ и ИИ-оркестраторов, которые будут координировать работу десятков слабых моделей.

Однако на текущий момент сильный ИИ это не стремление ESG, экологии или коммерческому успеху. Его создание может быть только в рамках стратегических и национальных проектов, которые будет финансировать государство.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ:

1. Определения ИИ GOLOS, [Электронный ресурс]. – Режим доступа: https://golos.id/ru--tekhnologii/@aigents/opredeleniya-ii (дата обращения: 01.12.2024);
2. Каким бывает искусственный интеллект?, [Электронный ресурс]. – Режим доступа: https://ai.sber.ru/post/kakim_byvaet_iskusstvennyj_intellekt (дата обращения: 01.12.2024);
3. A critical review towards artificial general intelligence: Challenges, ethical considerations, and the path forward, [Электронный ресурс]. – Режим доступа: https://wjarr.com/content/critical-review-towards-artificial-general-intelligence-challenges-ethical-considerations (дата обращения: 01.12.2024);
4. В чем сила, Сильный ИИ?, [Электронный ресурс]. – Режим доступа: https://russiancouncil.ru/analytics-and-comments/analytics/v-chem-sila-silnyy-ii/ (дата обращения: 01.12.2024);
5. И целого интернета мало. Для создания больших языковых моделей нового поколения, включая GPT-5, попросту не хватает данных, [Электронный ресурс]. – Режим доступа: https://www.ixbt.com/news/2024/04/02/i-celogo-interneta-malo-dlja-sozdanija-bolshih-jazykovyh-modelej-novogo-pokolenija-vkljuchaja-gpt5-poprostu-ne-hvataet.html (дата обращения: 01.12.2024);
6. Exploring the Future of AGI: Insights, Challenges, and Perspectives, [Электронный ресурс]. – Режим доступа: https://aibrainpowered.com/2024/11/06/exploring-the-future-of-agi-insights-challenges-and-perspectives (дата обращения: 01.12.2024);
7. Сверхчеловеческий ИИ на основе открытых систем: Утопия или антиутопия?, [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/bothub/articles/835544/ (дата обращения: 01.12.2024).