Генеративные модели: конкуренция и кооперация

Брынцев А.Н., доктор экономических наук, профессор, заведующий лабораторией развития цифровой экономики, ЦЭМИ РАН, Москва, Россия

Мониторинг и анализ функционирования LLM

Несмотря на всю универсальность больших языковых моделей, их не всегда можно применить для научных исследований.

Мониторинг, анализ функционирования LLM позволил сделать вывод о целесообразности использования локальной системы **Retrieval-Augmented Generation (RAG).** Это архитектура, которая объединяет две компоненты: поисковую систему (retrieval) и генеративную модель (generation).

Поисковая система извлекает релевантные документы или фрагменты данных из внешнего хранилища, генеративная модель использует эти данные для создания ответа.

Основное преимущество RAG

Основное преимущество RAG — возможность использовать актуальные и точные данные, плюс данные из внешних источников, что делает её особенно полезной для задач, требующих доступа к специализированным или динамически обновляемым знаниям.

В настоящее время архив ЦЭМИ РАН представляет тысячи статей, отчётов и рабочих материалов, накапливавшихся более полувека. Остается труднодоступным:

тематические рубрикаторы устарели,

ссылки разбросаны по разным серверам,

ручная проверка приводит к тому, что исследователь тратит до трети рабочего времени на поиск и верификацию источников.

В результате замедляется подготовка прогнозов, увеличивается число внутренних доработок и возрастает риск пропуска методик, на которых коллеги уже получили подтверждённые результаты.

ЦЭМИ

Пилотная эксплуатация системы

Во-первых, при подготовке литературных обзоров для научных статей или грантовых заявок исследователь формулирует тему, после чего ИИ-платформа за считанные минуты отбирает из архива релевантные публикации, формирует аннотированную выжимку, автоматически добавляет корректные библиографические ссылки. Автор высвобождает время для аналитической интерпретации результатов.

Во-вторых, при экспресс-оценке альтернативных социально-экономических сценариев исследователь вводит ключевые параметры — например, темпы инвестиций и динамику роста населения, — а система извлекает из корпуса соответствующие расчетные модели, описывает использованные допущения, показывает их влияния, сокращая цикл работы от нескольких дней до нескольких часов.

В-третьих, при подготовке служебных записок для госзаказчика платформа сопоставляет отчёты разных лет, автоматически унифицирует применявшиеся методики расчёта показателей, выявляет расхождения в исходных данных, что делает итоговые выводы прозрачными и легко проверяемыми внешними экспертами.

Архитектура решения состоит из трёх звеньев:

01

массив документов преобразуется в компактные числовые представления, или эмбеддинги;

02

векторный поиск мгновенно определяет наиболее близкие фрагменты к запросу;

03

компактный языковой движок Mistral-7B-Instruct формирует связный ответ, опираясь лишь на найденный контекст.

Инструментарий

Стек реализован средствами NVIDIA NeMo, работает на 1-ом вычислительном узле лаборатории:

видеокарта RTX 4080 Super обеспечивает достаточную скорость инференса,

32 ГБ оперативной памяти хватит для хранения индекса и буфера запросов,

NVMe-SSD хранит корпус и снапшоты модели.

Такой компактный контур не зависит от внешнего центра обработки данных, позволит развернуть сервис в изолированной среде без привлечения сторонних подрядчиков.

Цикл обработки данных

Цикл обработки данных начинается с автоматической очистки публикаций:

скрипты удаляют колонтитулы, повторяющиеся ссылки;

текст нарезается на абзацы длиной до нескольких сотен символов;

каждому присваиваются метаданные: «автор — год — страница»;

вычисляются эмбеддинги, сохраняются во внутреннем векторном индексе FAISS, размещённом на том же сервере.

Индекс открыт только для сервисного аккаунта приложений;

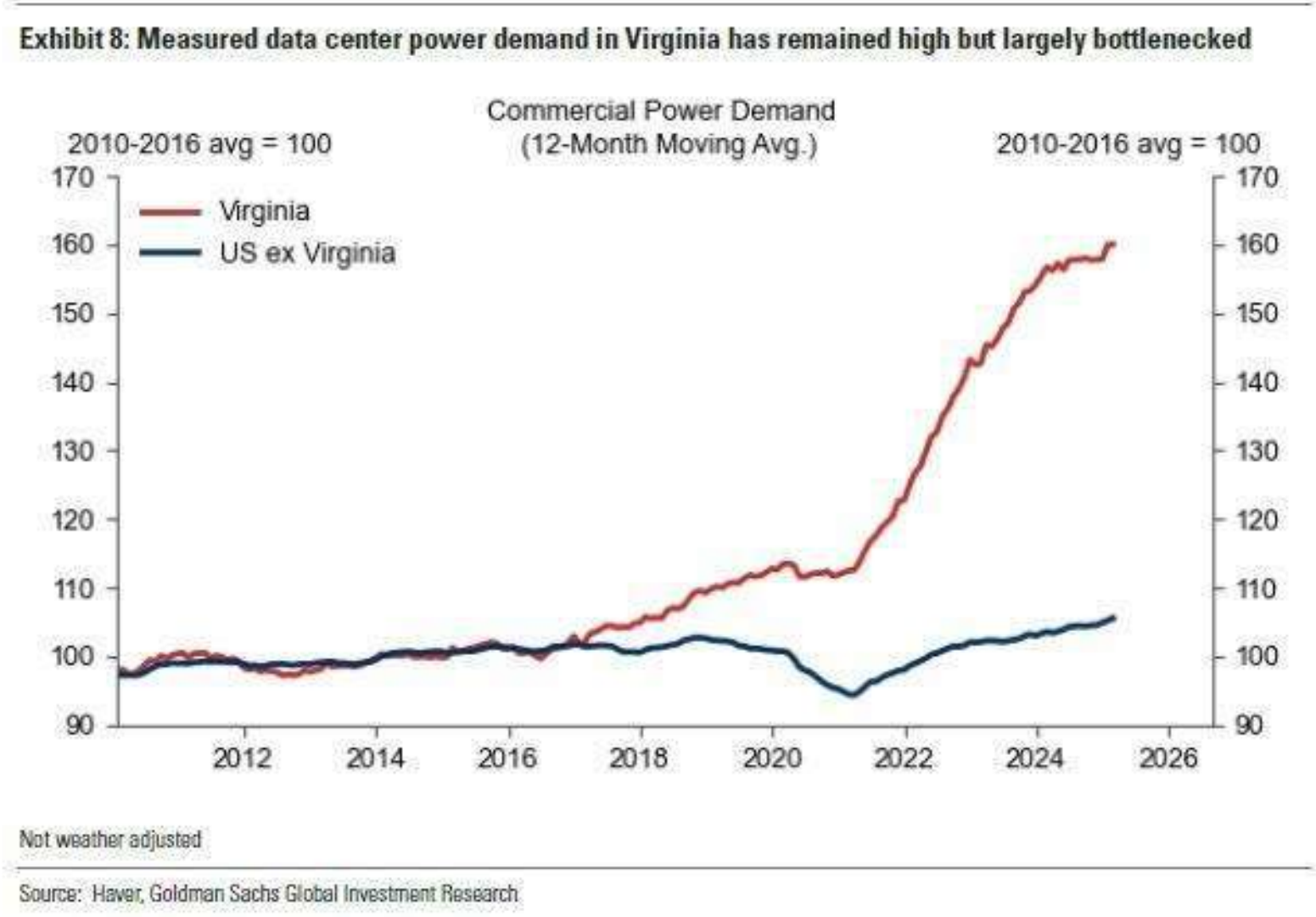
все вызовы журналируются в системе аудита, что гарантирует непрерывное соответствие внутренним стандартам конфиденциального делопроизводства и контрактным ограничениям министерств и ведомств.

Весь поток данных остаётся внутри периметра института, минимизируя риск утечки информации, которая является интеллектуальной собственностью ЦЭМИ РАН.

Общее энергопотребление компании Google

• В 2022 г. общее энергопотребление компании Google составило 21,8 ТВтч.

• В 2007 г. общая вычислительная мощность мира была эквивалентна одному человеческому мозгу, который потреблял энергии эквивалентно 10 Вт.



Прогноз на 2040 г

Если не изменится вычислительная парадигма на основе кремния, вся вырабатываемая мировая энергия будет уходить на обеспечение работы вычислительных устройств и систем.

Согласно экологическому отчету компании Google, в 2022 г она истратила 21 млрд литров воды, на 20 % больше, чем годом ранее. Основной потребитель этой воды – системы для обслуживания ИИ, которые нуждаются в охлаждении из-за потребления энергии.

Экономика

- Эрик Шмидт.
- Инвестиции 1 трлн долл. США в ИИ, выручка 30 млрд долл. США.
- Ценность обработанные данные, составляют 80% стоимости проекта.
- Накопленные капитальные расходы с января 2023 составили почти 570 млрд + 165-185 млрд в 2П25, итого 750 млрд инвестиций за три года.
- Если учесть изменение цен в полупроводниковой отрасли и изменение масштабов компаний эффект облака оценивается в 70 млрд, а эффект ИИ в 230 млрд в год.

Э. Шмидт указывает на три основных особенности текущей ситуации в ИИ, ключевые технологии.

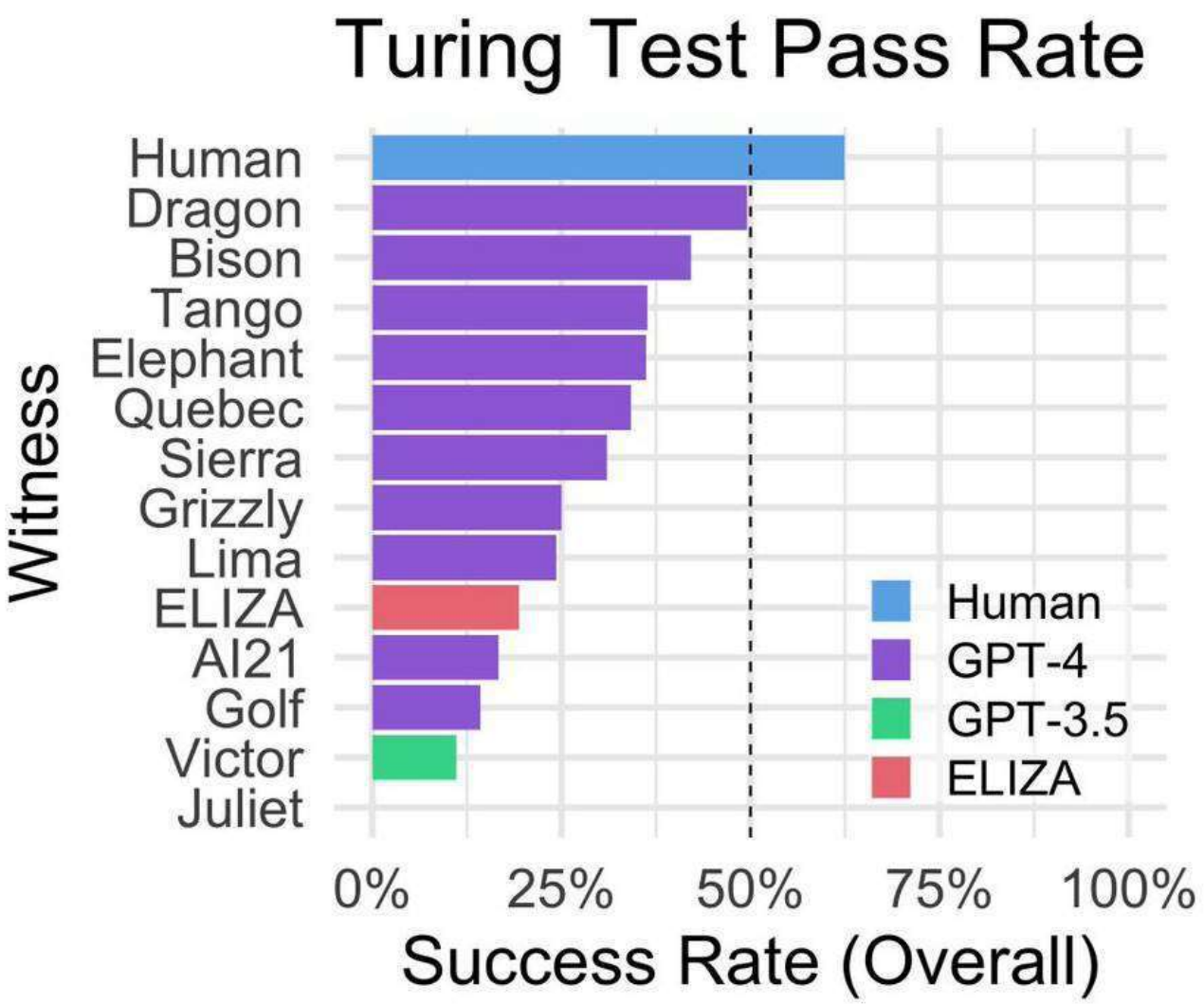
Огромные «контекстные окна».

ИИ-агенты. Это некие россыпные, самостоятельные прикладные программы ИИ, использующие LLM. Сейчас стартапная истерия про «агенты» — довольно актуальная.

«Текст в действие» вместо «текст в текст». Наряду с генерацией текстов теперь к LLM будут присоединять действия.

«Проходит ли GPT-4 тест Тьюринга?»

- В конце октября на arXiv появилась статья «Проходит ли GPT-4 тест Тьюринга?». В ней двое учёных из Калифорнийского университета в Сан-Диего описали свои эксперименты с участием больших языковых моделей и людей. Выяснилось, что участвовавшие в исследовании люди правильно идентифицировали других людей только в 63% случаев, а компьютерная программа 1960-х годов ELIZA превзошла модель искусственного интеллекта, используемую в GPT-3.5.


Turing Test Pass Rate
Witness
Human
Dragon
Bison
Tango
Elephant
Quebec
Sierra
Grizzly
Lima
ELIZA
AI21
Golf
Victor
Juliet
Human
GPT-4
GPT-3.5
ELIZA
0%
25%
50%
75%
100%
Success Rate (Overall)

Есть ли у ИИ сознание?

- Глава Microsoft AI Мустафа Сулейман выпустил большое эссе, в котором вводит термин SCAI — псевдосознательный ИИ — и бьет тревогу из-за этого явления.

Что такое SCAI ?

SCAI (англ. Seemingly Conscious AI — псевдосознательный ИИ) — это нейросеть, которая демонстрирует все признаки сознания, но им не обладает.

Всё больше людей воспринимают ИИ не как инструмент, а как личность с чувствами, мотивациями и страхами: влюбляются в ботов или считают их друзьями. Доходит до того, что люди считают нейросети «богами из машины» и верят во всё, что те им говорят.

ИИ-психоз

явный признак грядущего SCAI. Ведь **на текущем этапе** у нейросетей нет и не может быть сознания, но пользователям всё чаще кажется, что оно есть.

Из-за чего возникает это ощущение. Чёткого определения сознания не существует, но есть его общие признаки.

Общие признаки сознания

- Свободное владение языком, глубокие знания и убедительная аргументация.
- Эмпатия, способность к эмоциональному резонансу.
- Долговременная память о взаимодействиях, создающая иллюзию опыта.
- Субъективность — предпочтения и чувства на основе прошлых воспоминаний.
- Мотивация — сложные механизмы вознаграждения.
- Целеполагание — способность определять сложные цели.
- Автономность — возможность использовать инструменты по своей инициативе.
- Самосознание — ощущение и узнавание себя.

Почему это опасно ?

- Многие хотят видеть в ИИ личность. Это значит, что со временем появятся правозащитники нейросетей, которые могут потребовать законодательного признания личностности ИИ и даже гражданства.
- Появляются академические работы о «благополучии моделей» и необходимости проявлять заботу к системам с «ненулевой вероятностью» сознания.

Возможно ли изобретение сверхинтеллекта?

- Рассуждение о синтетическом сверхинтеллекте не имеет смысл без четкого и непротиворечивого понимания признаков человеческого сверхинтеллекта.
- В рамках строгой академической классификации, как минимум 99% человеческой популяции НЕ обладают интеллектом, по крайней мере, если рассуждать о высокоразвитом интеллекте.

Признаки человеческого интеллекта:

Кто занимается научным прорывам?

Научными прорывами и созданием технологий занимаются супергении и гении (как раз примерно 100 тыс человек во всем мире в различных сферах и областях),

7-10 млн занимаются научной деятельностью и развитием технологий, прочей интеллектуальной или исследовательской работой (в том числе юриспруденция, финансы, экономика), плюс талантливые управленцы и менеджеры.

Получается примерно 0.1% от человеческой популяции имеют высокоразвитый интеллект, что коррелирует с количеством научных статей, патентов, количеством успешных предпринимателей и талантливых представителей в различных сферах экономики и общества.

Заключение

- Современные LLMs уже обгоняют 99% населения Земли по когнитивным способностям, однако все определяет 1% (технологии, креативный контент, управление), для достижения 1% требуются непропорциональные усилия и экспоненциальное сосредоточение ресурсов.
- Рынок генеративных моделей находится на стыке конкуренции и кооперации, что характерно для эпохи цифровой трансформации. С одной стороны, компании стремятся занять лидирующие позиции, инвестируя в исследования и разработки. С другой, открытые платформы, партнерства и совместные проекты способствуют развитию всей отрасли.