Лекция 3: Введение в архитектуру генеративных моделей ИИ. Большие языковые модели и их архитектура.

Основные вопросы:

1. NLP (Natural Language Processing) Обработка естественного языка
2. Что такое Большие языковые модели (LLM)
3. Фундаментальные концепции, лежащие в основе LLM
4. Общая архитектура LLM
5. Примеры генеративного искусства

NLP (Natural Language Processing)

Обработка естественного языка (NLP) - это **междисциплинарная область лингвистики, информатики и искусственного интеллекта**. Его цель состоит в том, чтобы компьютер мог понимать тексты и другие средства массовой информации на их естественных языках, включая их контекстуальные нюансы.

Фундаментальные зачатки НЛП можно проследить с 1950-х годов, когда Алан Тьюринг опубликовал свою статью, предлагающую тест Тьюринга в качестве критерия интеллекта. Однако на протяжении многих десятилетий НЛП ограничивалось имитацией понимания естественного языка на основе набора правил.

Начиная с 1980-х годов, с **ростом вычислительной мощности и внедрением алгоритмов машинного обучения** для обработки языка, статистическое НЛП начало обретать форму. Все началось с машинного перевода с использованием алгоритмов контролируемого обучения. Но вскоре внимание переключилось на алгоритмы обучения с частичным контролем и без учителя для обработки растущих объемов необработанных языковых данных, генерируемых в Интернете. Одним из ключевых инструментов приложений NLP является языковое моделирование.

NLP (Natural Language Processing)

NLP (Natural Language Processing) — это направление в машинном обучении, посвящённое распознаванию, генерации и обработке устной и письменной человеческой речи. Находится на стыке дисциплин искусственного интеллекта и лингвистики.

Цель NLP — научить компьютеры понимать, интерпретировать и генерировать человеческую речь и текст так же, как это делаем мы. Это включает в себя не только распознавание слов, но и понимание их смысла, контекста и эмоций.

Некоторые области применения NLP:

Машинный перевод. Автоматический перевод текста с одного языка на другой.

Анализ тональности. Определение эмоциональной окраски текста (положительная, отрицательная, нейтральная). Используется для анализа отзывов клиентов, мониторинга социальных сетей и изучения общественного мнения.

Чат-боты и виртуальные ассистенты. Автоматизированные системы, которые общаются с пользователями на естественном языке. Помогают с поиском информации, совершением покупок, бронированием билетов и выполнением других задач.

Некоторые области применения NLP:

Поиск информации. NLP помогает поисковым системам понимать смысл запросов пользователей и находить наиболее релевантные результаты, даже если они сформулированы неточно.

Извлечение информации. Автоматическое извлечение ключевой информации из текста. Полезно для анализа больших объёмов данных, например, новостных статей или юридических документов.

Генерация текста. Автоматическое создание текстов, например, новостных статей, стихов, сценариев. Используется в журналистике, маркетинге и развлекательной индустрии. 2
Распознавание речи. Преобразование устной речи в текст. Применяется в голосовых помощниках и системах диктовки.

Синтез речи. Создание искусственной речи из текста. Используется для озвучки книг, создания голосовых помощников и разработки систем для людей с нарушениями зрения. 2

Большие языковые модели (LLMS)

Большие языковые модели (LLM) по сути, **это нейронные языковые модели, работающие в большем масштабе**. Большая языковая модель состоит из нейронной сети, возможно, с миллиардами параметров. Более того, обычно он обучается на огромном количестве немаркированного текста, возможно, состоящего из сотен миллиардов слов.
Большие языковые модели, также называемые моделями глубокого обучения, обычно представляют собой модели общего назначения, которые превосходно справляются с широким спектром задач. Обычно их обучают относительно простым задачам, таким как предсказание следующего слова в предложении.

LLMs и базовые модели

Базовая модель обычно относится к **любой модели, обученной на обширных данных, которая может быть адаптирована к широкому кругу последующих задач**. Эти модели обычно создаются с использованием глубоких нейронных сетей и обучаются с использованием самостоятельного обучения на множестве немаркированных данных.
Этот термин был введен не так давно Стэнфордским институтом искусственного интеллекта, ориентированного на человека (HAI). Однако нет четкого различия между тем, что мы называем базовой моделью, и тем, что квалифицируется как модель большого языка (LLM). Тем не менее, магистры права обычно обучаются на данных, связанных с языком, таких как текст. Но **базовая модель обычно обучается на мультимодальных данных**, сочетании текста, изображений, аудио и т.д. Что еще более важно, базовая модель предназначена для того, чтобы служить основой для более конкретных задач:

Большие языковые модели (LLMs - Large Language Models)

Достаточная подготовка на большом наборе данных и огромном количестве параметровпозволяет модели охватывать большую часть синтаксиса и семантики человеческого языка. Таким образом, они становятся способными к более тонким навыкам в решении широкого круга задач компьютерной лингвистики.
Это значительный отход от более раннего подхода в приложениях NLP, где специализированные языковые модели обучались для выполнения конкретных задач. Напротив, **исследователи наблюдали многие возникающие способности в LLMs**, способности, которым они никогда не обучались. Например, было показано, что LLM выполняют многоступенчатую арифметику, расшифровывают буквы слова и идентифицируют оскорбительный контент в разговорных языках. Недавно ChatGPT, популярный чат-бот, созданный поверх LMS семейства GPT от OpenAPI, сдал профессиональные экзамены, такие как экзамен на медицинское лицензирование в США!

Принцип работы больших языковых моделей

Чтобы ИИ распознавал запрос и интент пользователя, а затем генерировал ответ, нужно обучить нейросеть с использованием Machine Learning, NLP Modeling и других.

Чтобы создать LLM, необходимо:

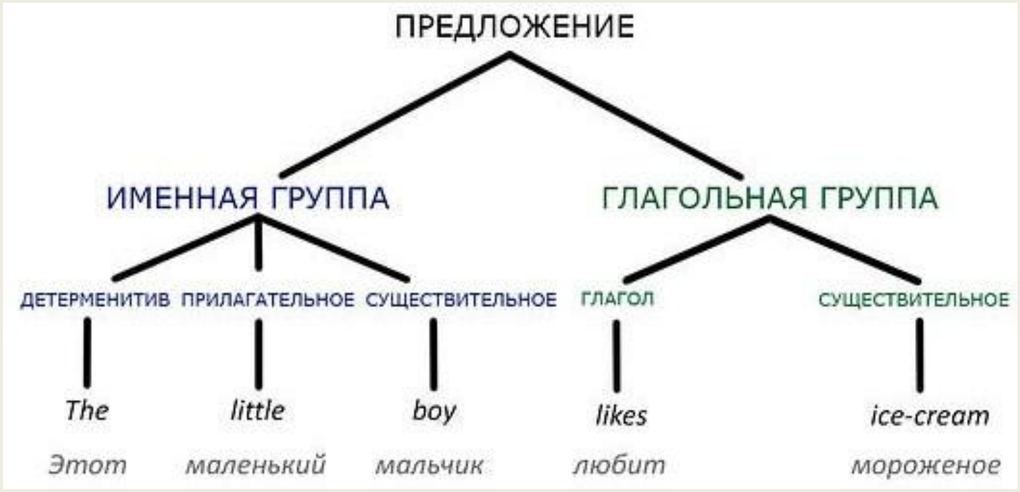
1. Собрать много качественных данных (поиск, сбор, очистка датасета и т. д.).
2. Выбрать архитектуру (Transformer, BERT — Bidirectional Encoder Representations from Transformers, GPT — Generative Pre-trained Transformer, T5).
3. Отточить процесс обучения языковой модели. Масштабировать систему, продумать отладку при сбоях (к примеру, для работы нужно более 1000 видеокарт, есть риск выхода из строя).
4. Усовершенствовать работу (CUDA-отладчик, библиотека NCCL, Garbage Collectors, фреймворк PyTorch FSDP).
5. Получить LLM (LL).

Ограничения и вызовы

- ***Этика:*** Риск использования для обмана (плагиат, автоматическое написание эссе).
- ***Технические сложности:*** Высокие затраты на вычисления.
- ***Данные:*** Возможная предвзятость.

Основные принципы работы обработки естественного языка

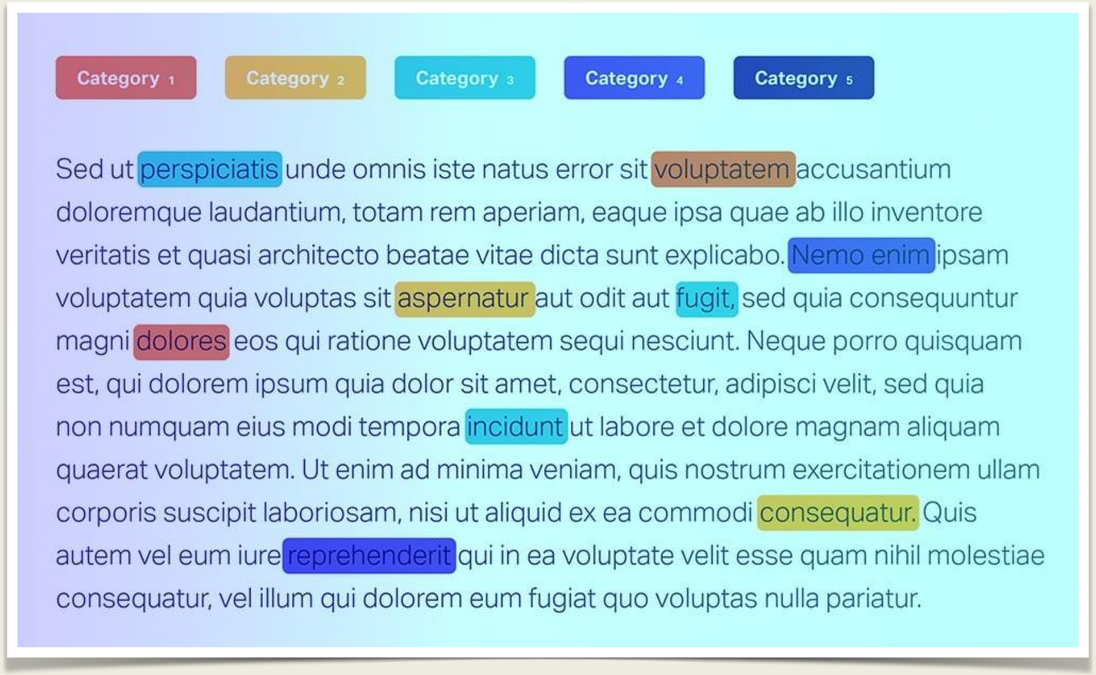
- В огромном наборе текста слова и предложения появляются в последовательностях с определенными зависимостями.
- Эта рекуррентность помогает модели понять, как разрезать текст на статистические блоки, которые имеют некоторую предсказуемость.
- Она изучает закономерности этих текстовых блоков и использует эти знания, чтобы предложить, что может произойти дальше.



Дерево синтаксического анализа

Типы разметок текстовых данных в классических алгоритмах Машинного обучения

- **Распознавание именных сущностей (NER):** люди, организация, места, дня и так далее.
- **Разметка на настроение (анализ настроения):** классификация эмоций или отношений, переданных в тексте (положительное, отрицательное, нейтральное).
- **Объяснение частей речи:** грамматические категории, относящаяся к каждому слову в предложении, например: глагол, существительное, прилагательное и так далее.
- **Отношения между сущностями (экстракция отношений):** определяет связи между различными сущностями в тексте. Например, в предложении "Стив Джобс является соучредителем Apple", отношения между "Стивом Джобсом" и "Apple" являются отношениями "соучредителя".



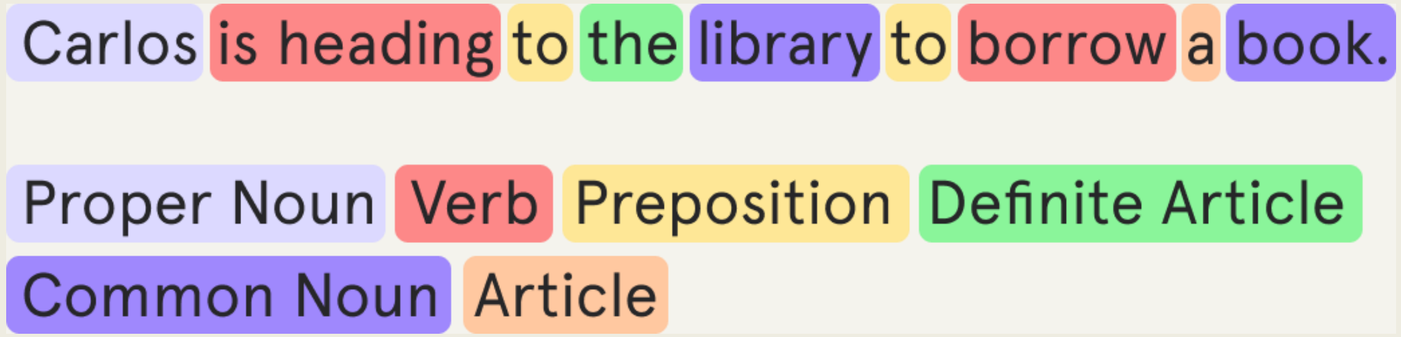
Типы разметок текстовых данных в классических алгоритмах Машинного обучения

- **Разметка при синтаксическом разборе зависимостей:** Эта аннотация выявляет грамматические отношения между словами в предложении, обозначая зависимость между основным словом (обычно глаголом) и его дополнением или модификатором.

- **Разметка по сегменту текста:** Состоит в разделении текста на логические единицы, такие как предложения, абзацы или тематические разделы.

- **Разметка по намерениям:** указывает на намерение предложения или текста, например, просьба о предоставлении информации, просьба об оказании услуг или жалоба.

- **Разметка перевода:** для машинного перевода текста с одного языка на другой каждый сегмент текста синхронизируется с соответствующим переводом.



Как LLM генерирует связный текст

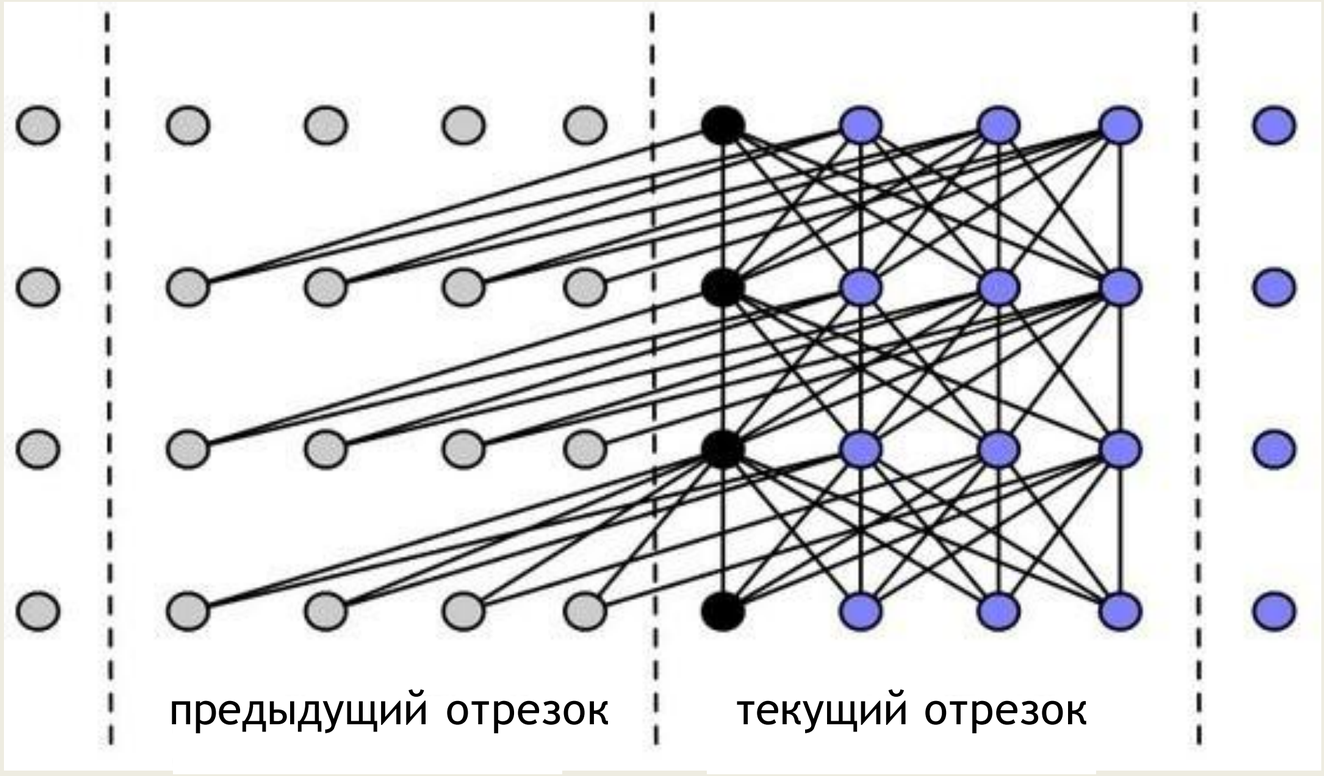
Принцип работы языковой модели прост — предсказывать следующее слово в предложении. Допустим, мы просканировали весь интернет и нашли все случаи, где встречается фраза «GigaChat используют для».
Дальше мы взяли все слова, которые следуют за строкой «GigaChat используют для» и вычислили, с какой вероятностью встречается каждое.

GigaChat используют для
+бизнеса10%маркетинга10%генерации10%SEO10%написания10%создания10%разработки10%поиска10%
В нашем случае искусственный интеллект, вероятно, добавит слово «бизнеса». Фраза будет звучать как «GigaChat используют для бизнеса». Искусственный интеллект может выбрать и другое продолжение — всё зависит от настроек и сформулированного запроса.
Например, может появиться фраза «GigaChat используют для генерации». Дальше искусственный интеллект уже работает с ней: может добавить «картинок», и результат будет выглядеть как «GigaChat используют для генерации картинок».

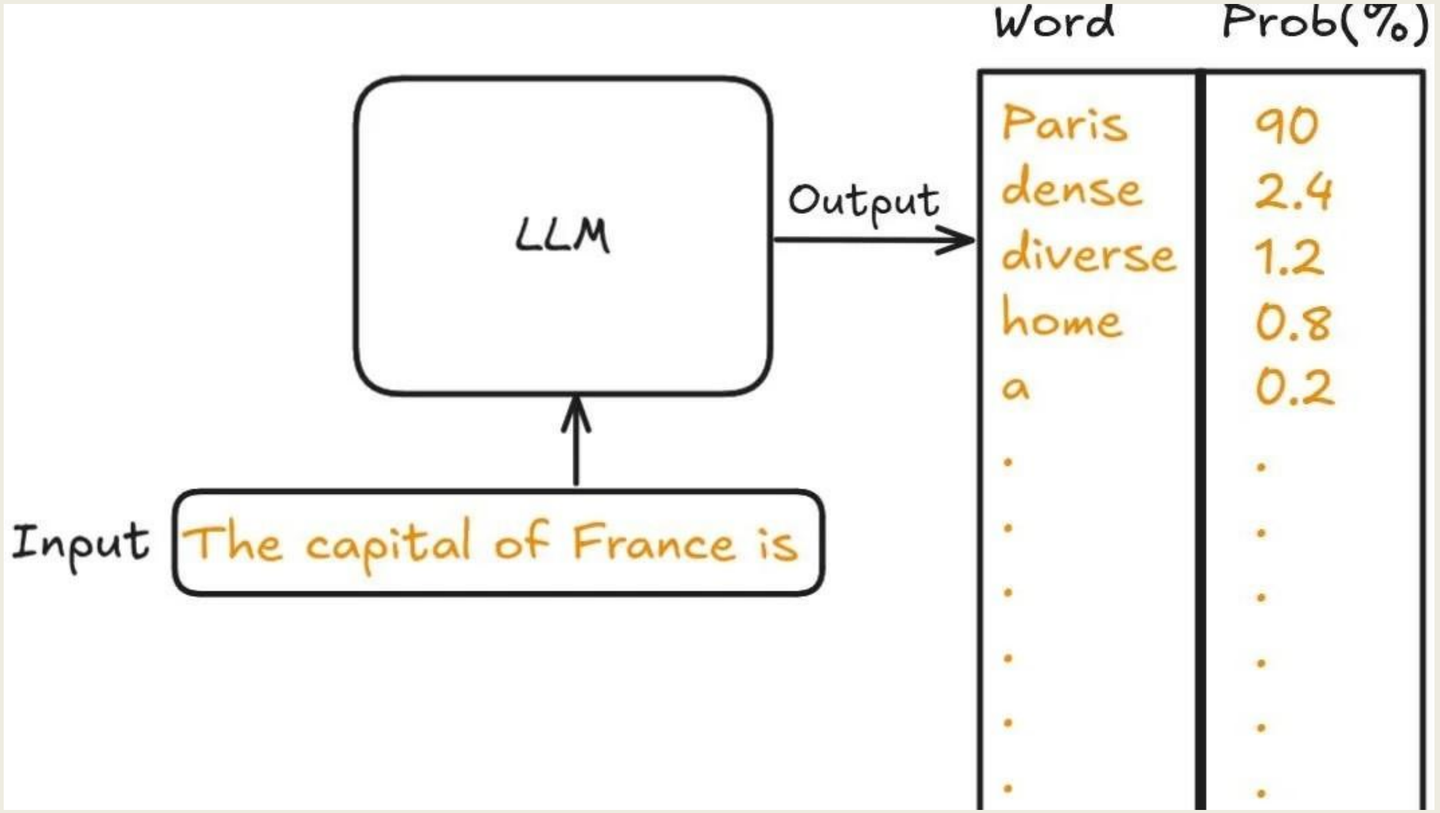
Разметка текстовых данных для больших языковых моделей

- Архитектуры трансформеров, включая большие языковые модели (LLM), обычно не используют текстовые аннотации так же, как традиционные модели обучения под руководством.
- Однако они требуют больших объемов данных, которые могут иметь разную степень неявных или явных аннотаций в зависимости от целей и задач обучения.

Предсказания следующего слова, сегмента, отрезка



Большие языковые модели



Основные принципы работы обработки естественного языка и генеративного ИИ

- В то время как большие наборы данных стали одним из катализаторов, приведших к буму генеративного ИИ, различные крупные научные достижения также привели к более сложным архитектурам глубокого обучения.

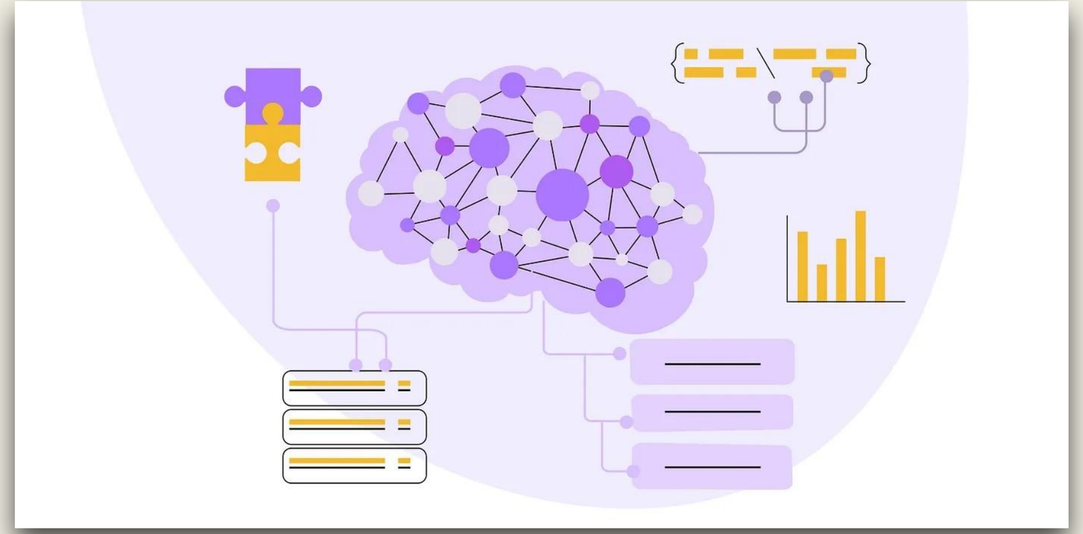
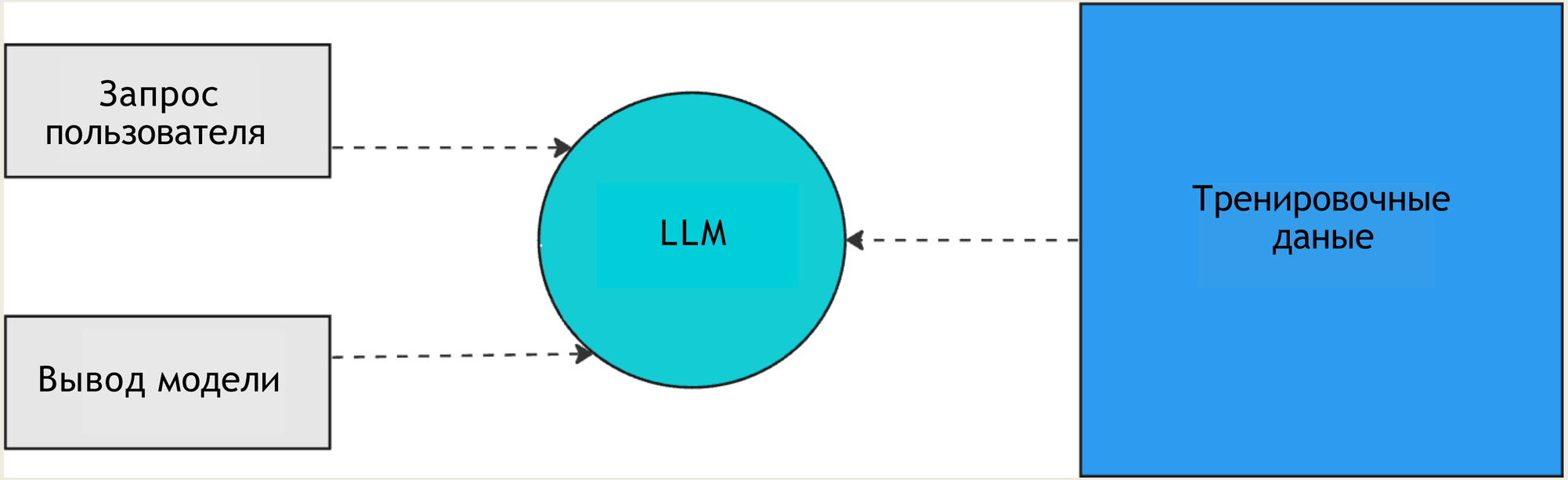
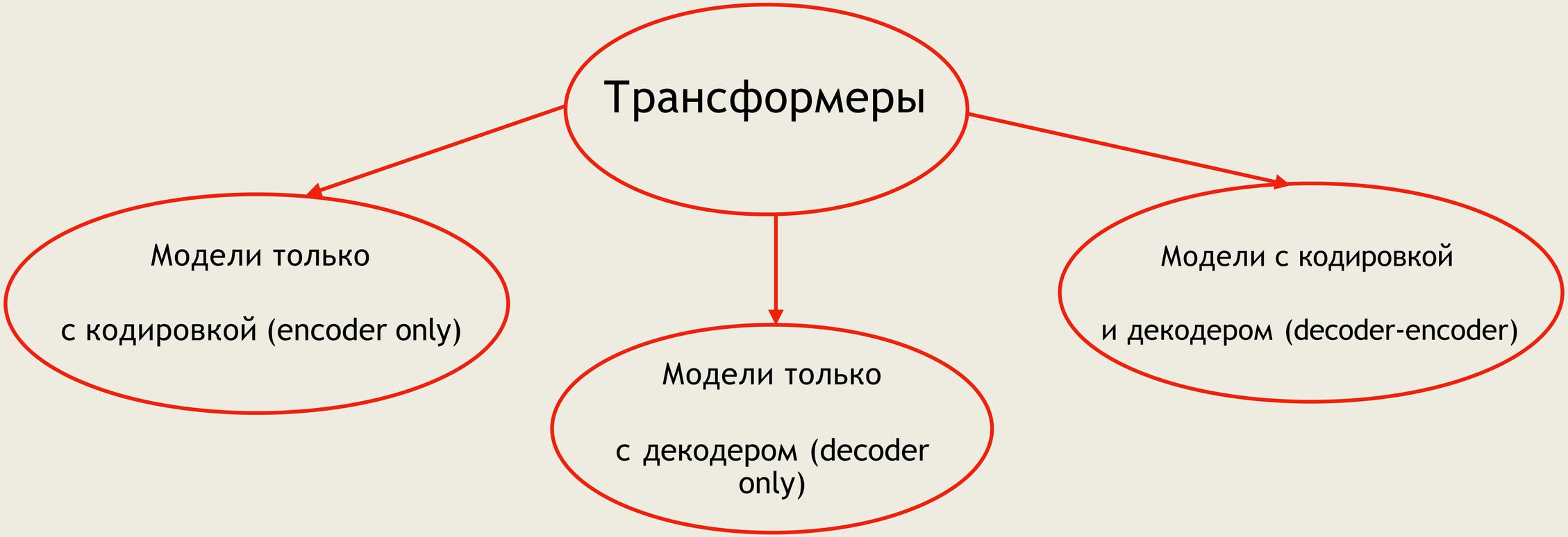


Схема взаимодействия с LLM

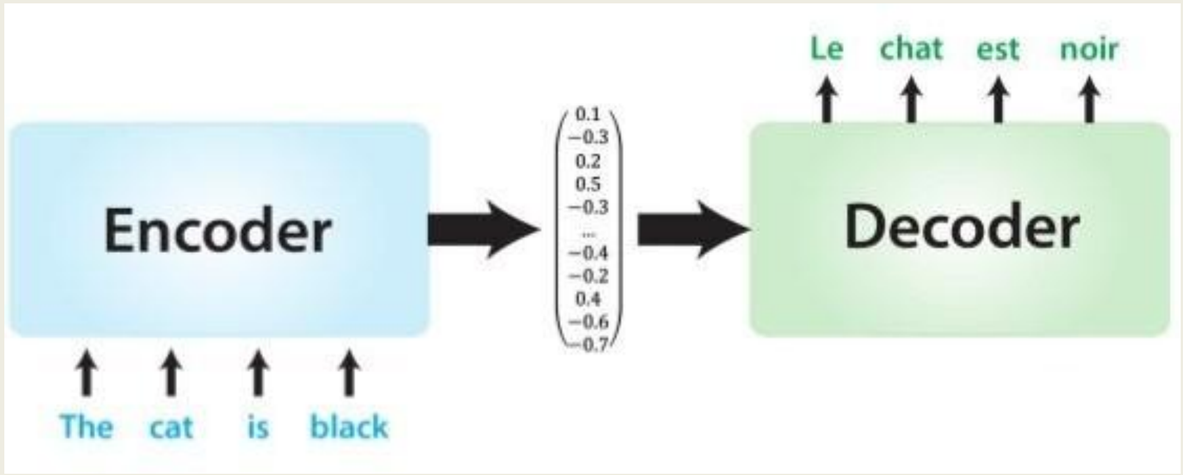


Архитектура трансформеров

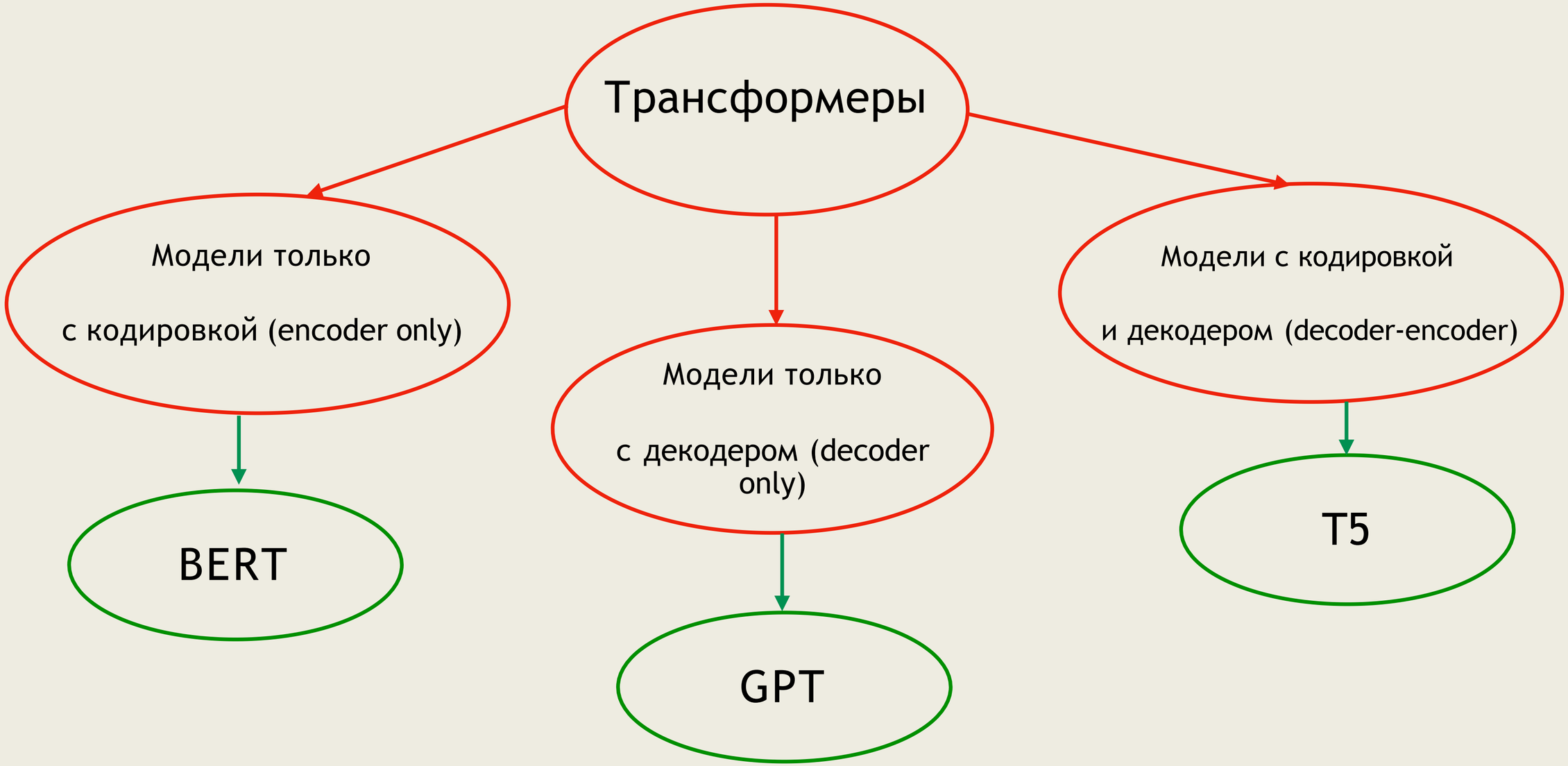


Архитектуры с кодированием и декодированием

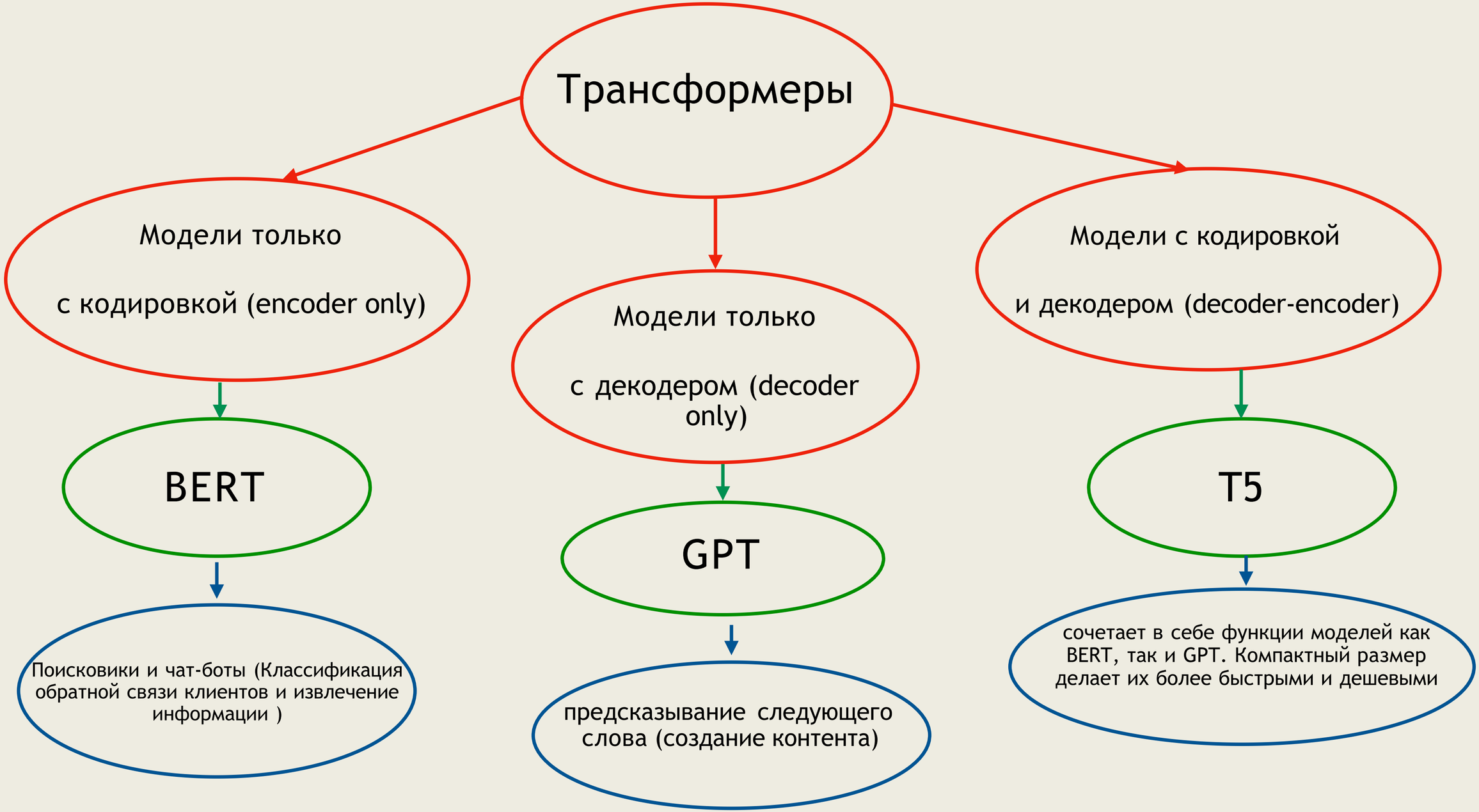
(Encoder-decoder)



Примеры моделей трансформеров



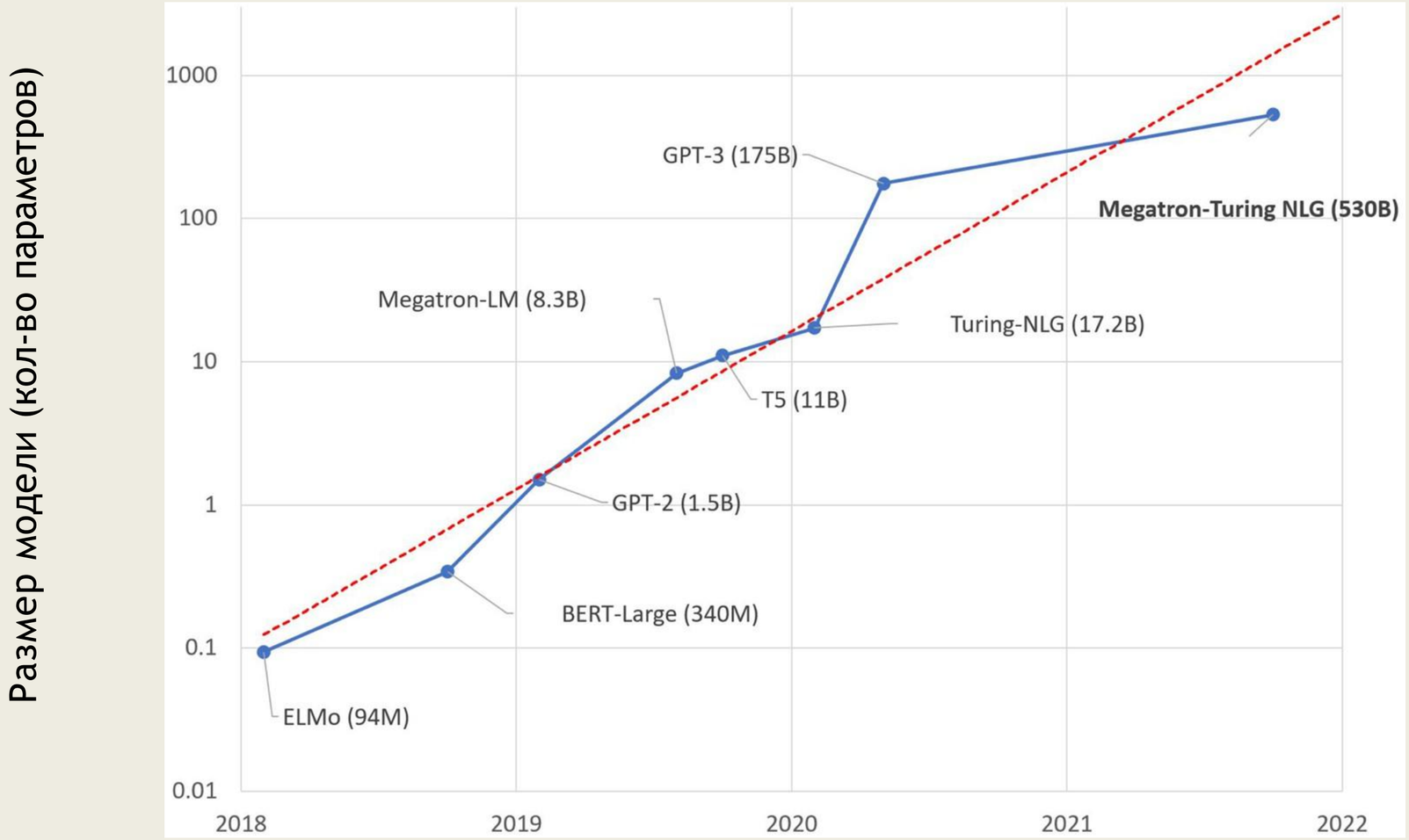
Применение моделей трансформеров



Аббревиатуры

- **BERT - Bidirectional Encoder Representations from Transformers**
  - Двунаправленные представления кодировщика из трансформеров
  - Оригинальная статья: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - Ссылка: https://arxiv.org/pdf/1810.04805
- **GPT - Generative Pre-training Transformer**
  - Генеративный предобученный трансформер
  - Оригинальная статья: Improving Language Understanding by Generative Pre-training
  - Ссылка: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- **T5 -Text-to-Text Transfer Transformer**
  - Трансформер для переноса обучения "Текст в текст”
  - Оригинальная статья: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
  - Ссылка: https://arxiv.org/pdf/1910.10683

Эволюция кол-ва параметров языковых моделей



*B (billion) - миллиард

Примеры генеративного искусства

Генеративное искусство – это процесс алгоритмического генерирования новых идей, форм, форм, цветов или узоров. Во-первых, вы создаете правила, устанавливающие границы для процесса создания. Затем компьютер – или, реже, человек – следует этим правилам для создания новых работ.

Примеры генеративного искусства

Майкл Хансмейер – архитектор, который рассматривает генеративный дизайн как «размышление о проектировании не объекта, а процесса создания объектов». Этот процесс допускает более искусственную интуицию – счастливые случайности и новые идеи, на которые обычно нужно время, чтобы наткнуться на них.

В этом ярком примере вычислительной архитектуры набор грота был разработан для оперы Моцарта. (2018).
Используя вычислительные инструменты для быстрого изучения, оптимизации и тестирования творческих дизайнерских идей, такие художники, как Хансмейер, максимально расширяют возможности для творчества. Хансмейер использовал генеративный дизайн, чтобы создать набор грота для оперы Моцарта на изображении выше.

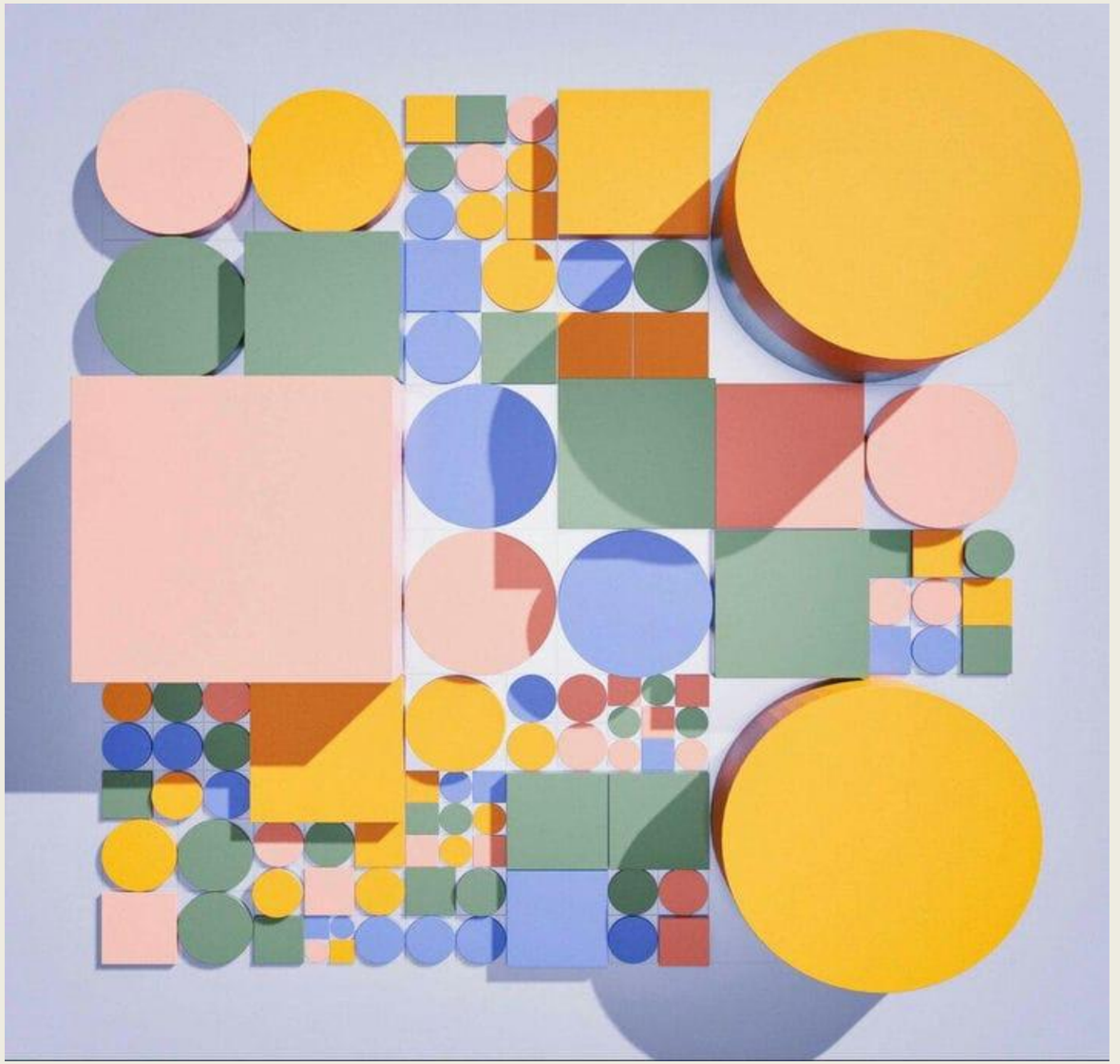
Мукарнас, Майкл Хансмейер. Мукарны – это тщательно продуманные декоративные своды и некоторые из самых ранних и самых впечатляющих примеров архитектурного дизайна, основанного на правилах. Используя достижения в области вычислительного проектирования и цифрового производства, Хансмейер предлагает нам пересмотреть эти типологии. Его генеративный художественный процесс позволяет ему создавать объекты исключительной широты и глубины; бесконечные вариации дизайна с богатством деталей, которые могут снова вызвать изумление, любопытство и недоумение этих традиционных архитектурных чудес.

«Пространственный оперный театр» (Théâtre D'opéra Spatial)

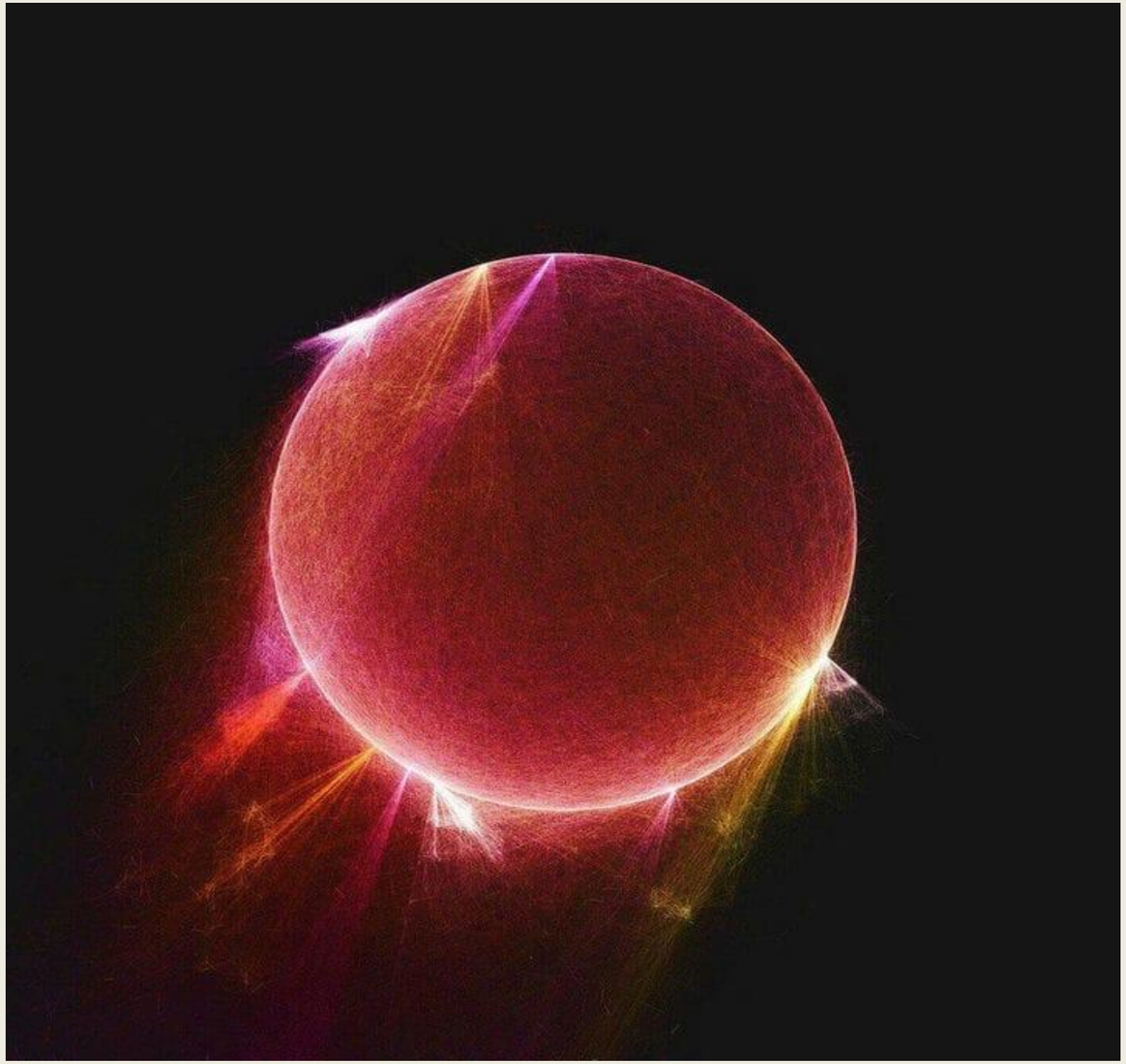
- Théâtre D'opéra Spatial ('космический театр') - это изображение, созданное Джейсоном Майклом Алленом с помощью платформы генеративного искусственного интеллекта Midjourney.

- Изображение выиграло ежегодный конкурс изобразительных искусств Государственной ярмарки Колорадо 2022 года в категории цифрового искусства 29 августа, став одним из первых изображений, созданных с использованием искусственного интеллекта (ИИ), получивших такую награду.

«В процессе проектирования достигается баланс между ожидаемым и неожиданным, между контролем и отказом. Хотя процессы детерминированы, результаты нельзя предвидеть. Компьютер обретает способность удивлять нас ». – Майкл Хансмайер



Генеративное искусство Маноло Гамбоа Наона, аргентинского художника, который использует алгоритмические инструменты, включая Обработку, для создания произведений искусства.



Генеративное искусство Андерса Хоффа. Это часть его проекта «Inconvergent», в котором исследуется сложное поведение, возникающее в системах с простыми правилами.
«Что мне больше всего нравится, так это сложные и запутанные результаты, которые можно получить с помощью набора простых правил». – Андерс Хофф

Спасибо за внимание!