Промт-инжиниринг: введение

Автор: Елена Каганова, тг: какая-то библиотека
ytb: https://www.youtube.com/@selfmadeLibrary/videos
inst: https://instagram.com/bestoloch.innovation?igshid=MzMyNGUyNmU2YQ==
поддержать выпуск методички: https://taplink.cc/ekaganova

Введение

В современном мире, насыщенном информацией, технология искусственного интеллекта (ИИ) становится всё более важной. Генеративный ИИ, в частности, обещает революционизировать множество областей, от создания контента до решения сложных задач.

Генеративный ИИ: от мифов к реальности

Генеративный ИИ — это разновидность искусственного интеллекта, способная генерировать новые данные, которые имеют сходство с реальными данными, на которых она обучалась.

Основные принципы работы:

1. **Обучение:** Генеративные модели ИИ обучаются на огромных наборах данных.
2. **Изучение закономерностей:** В процессе обучения модели анализируют эти данные, чтобы выявить скрытые закономерности и структуру.
3. **Создание нового:** После обучения, модели могут использовать эти знания для генерации новых, похожих данных, например, текстов, изображений, музыки и т.д.

Распространённые заблуждения:

- **ИИ — это замена людям:** Генеративный ИИ не предназначен для замены людей, а для того, чтобы усилить наши возможности и повысить продуктивность.
- **ИИ — это "волшебная палочка":** Генеративный ИИ — это инструмент, который требует компетентного пользователя. Правильно сформулированный запрос, а также способность анализировать результаты — это ключевые факторы успеха при работе с ИИ.

Примеры задач, решаемых с помощью промт-инжиниринга:

- Реферирование текста: создание краткого изложения текста.
- Вопрос-ответ: модель находит ответы на вопросы на основании данного текста.
- Классификация текста: например, определение тональности текста (позитивный, негативный, нейтральный).
- Ролевая игра: модель может имитировать разговор от лица персонажа или следовать определенному стилю.
- Генерация кода: написание кода на разных языках программирования (SQL, Python, JavaScript).
- Логические рассуждения: решение математических задач, задач на логику, верификация утверждений.

Ограничения:

- "Галлюцинации": LLM способны генерировать недостоверную информацию. Важна критическая оценка результатов и проверка фактов.
- Предвзятость: LLM обучаются на огромных, но не всегда репрезентативных, наборах данных. Важно осознавать возможные смещения и недочеты.
- Плагиат: Использование LLM для создания текстов без указания источника недопустимо.

Задачи в социальных науках, которые могут решаться с помощью ИИ

Генерация гипотез

Как это работает? LLM обучаются на огромных массивах текстовых данных, включая научные статьи, книги и другие источники информации. Это позволяет им выявлять закономерности, связи и пробелы в существующих знаниях. Исследователи могут использовать LLM для анализа литературы по определенной теме, выявления потенциальных противоречий или неисследованных аспектов, и на основе этого генерировать новые гипотезы.

Примеры:

LLM может проанализировать сотни статей по психологии образования и предложить гипотезу о связи между стилем обучения и успеваемостью в онлайн-среде.

LLM может изучить данные опросов общественного мнения и сформулировать гипотезу о влиянии социальных сетей на политическую поляризацию.

Ограничения: LLM не могут "придумывать" гипотезы из ничего. Они основываются на данных, на которых обучались. Важно критически оценивать предложенные гипотезы и проверять их с помощью традиционных научных методов.

Проведение исследований

Как это работает? LLM могут автоматизировать частично или полностью некоторые этапы исследования, например, создание анкет для опросов, разработку сценариев для экспериментов, а также проведение интервью (в виде чат-ботов).

Примеры:

LLM может сгенерировать вопросы для онлайн-опроса о влиянии пандемии на психическое здоровье.

LLM может создать виртуальных агентов для проведения экспериментов по социальному взаимодействию.

Ограничения: LLM могут создавать предвзятые вопросы или сценарии, отражающие biases данных, на которых они обучались. Необходимо тщательно проверять и корректировать материалы, созданные LLM, чтобы обеспечить валидность и надежность исследования. Также важно учитывать этические аспекты использования LLM в исследованиях, особенно при взаимодействии с людьми.

Анализ данных

Как это работает? LLM могут анализировать текстовые данные, такие как ответы на открытые вопросы в опросах, транскрипты интервью или посты в социальных сетях, для выявления тем, настроений и других паттернов. Они также могут быть использованы для анализа количественных данных в сочетании с другими методами машинного обучения.

Примеры:

LLM может проанализировать тысячи твитов, чтобы определить общественное мнение о каком-либо событии.

LLM может классифицировать ответы на открытые вопросы в опросе по категориям, чтобы облегчить качественный анализ.

Ограничения: Интерпретация результатов анализа данных, проведенного LLM, требует осторожности. LLM могут выявлять корреляции, но не причинно-следственные связи. Также важно учитывать контекст и ограничения данных при интерпретации результатов.

В целом, LLM представляют собой мощный инструмент для социальных наук, но их использование требует критического мышления и понимания их ограничений. LLM не заменяют традиционные методы исследования, а скорее дополняют их, предоставляя новые возможности для анализа данных и генерации гипотез. По мере развития технологий LLM их роль в социальных науках, вероятно, будет только расти.

Популярные чат-боты

1. ChatGPT: Разработан компанией OpenAI, известен своей способностью вести естественные диалоги и создавать разный текст: от стихов до эссе.
2. Gemini: Новейшая модель Google, обещает более глубокое понимание языка и лучшее решение разнообразных задач.
    a. Общий интерфейс https://gemini.google.com/app
    b. AI Studio https://aistudio.google.com/app/prompts/new_chat
3. Claude: Разработка компании Anthropic, известна своей способностью генерировать творческие тексты и предоставлять дополнительную информацию.
4. Perplexity: Специализируется на поиске информации и предоставлении сводных отчетов по заданным темам, объединяя результаты поиска из разных источников.

Сравнение интерфейсов и сайтов для работы с ИИ

	ChatGPT	**Gemini (Google)**	**Google AI Studio**	**Claude**	**Perplexity**
Доступность	Бесплатно (с VPN)	Бесплатно (с VPN)	Бесплатно (с VPN)	Бесплатно (ограничено)	Бесплатно и без VPN
Вход в систему	Google-аккаунт	Google-аккаунт	Google-аккаунт	Google-аккаунт	Google-аккаунт
Голосовой ввод	Нет	Да	Да	Нет	Нет
Работа с файлами	Все, но в бесплатной версии количество ограничено	Только картинки	Разные типы файлов, до 2 млн токенов	-	-
Выбор модели	Нет	Нет	Да	-	Выбор Prp
Настройка температуры (креативности)	Нет	Нет	Да	-	-
Проверка фактов	Нет	Кнопка "Найти в Google" (обычно показывает, что факты имеют 0 источников)	Зависит от промта и загруженных файлов	Считается более точным, но вообще не факт	Предоставляет ссылки на источники

Основные преимущества	Быстрые ответы, редактура текста, переводы	Голосовой ввод, работа с большими объемами текста	Гибкие настройки, работа с большими объемами данных, снижение галлюцинаций	Меньше галлюцинаций, точность в научных темах	Поиск информации с указанием источников
Основные недостатки	Ограничения бесплатной версии, галлюцинации, не проверяет факты	Ограничения в работе с файлами	Требует навыков промт-инжиниринга	Ограничения бесплатной версии	Не всегда точная интерпретация данных

Рекомендации по использованию:

- **ChatGPT:** для быстрых задач, редактуры текста, перевода.
- **Gemini:** для голосового ввода и повседневных задач.
- **Google AI Studio:** для продвинутого промт-инжиниринга, работы с большими объемами данных, контроля галлюцинаций.
- **Perplexity:** для поиска информации с указанием источников.

Gemini через Google AI Studio

Прежде чем мы погрузимся в мир генеративного ИИ, давайте познакомимся с инструментом, который станет вашим верным помощником: Google AI Studio. Это онлайн-платформа, предоставляющая мощные инструменты для работы с ИИ, включая модели машинного обучения, инструменты для анализа данных, а также интегрированную среду разработки.

AI Studio предлагает всё необходимое для создания, обучения и развертывания моделей машинного обучения. Ключевая особенность AI Studio — это возможность легко использовать модели машинного обучения Google, включая Gemini, в своих проектах.

Gemini

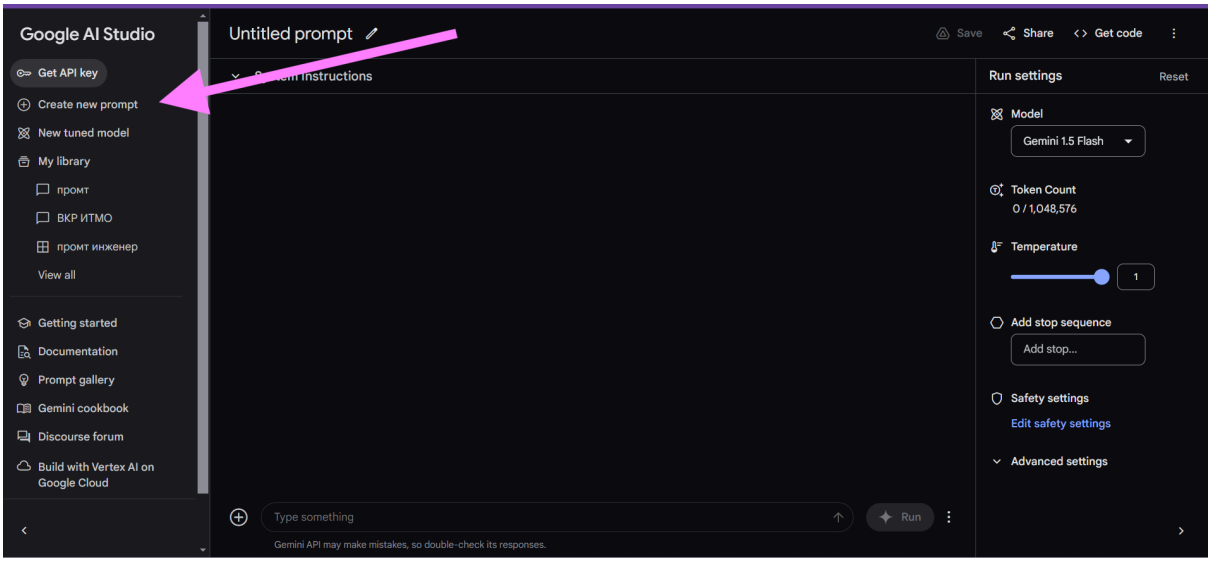
Gemini — модель генеративного ИИ, разработанная Google. Она обладает значительными преимуществами перед другими моделями, такими как:

- **Мультимодальность:** Gemini может работать не только с текстом, но и с изображениями, аудио и видео.
- **Большой объём знаний:** Обучение на огромных датасетах позволило Gemini усвоить огромное количество информации.
- **Улучшенные алгоритмы:** Gemini использует усовершенствованные алгоритмы, которые позволяют ей генерировать более реалистичные и качественные результаты.

Работа с Gemini в Google AI Studio

Начнём с основ

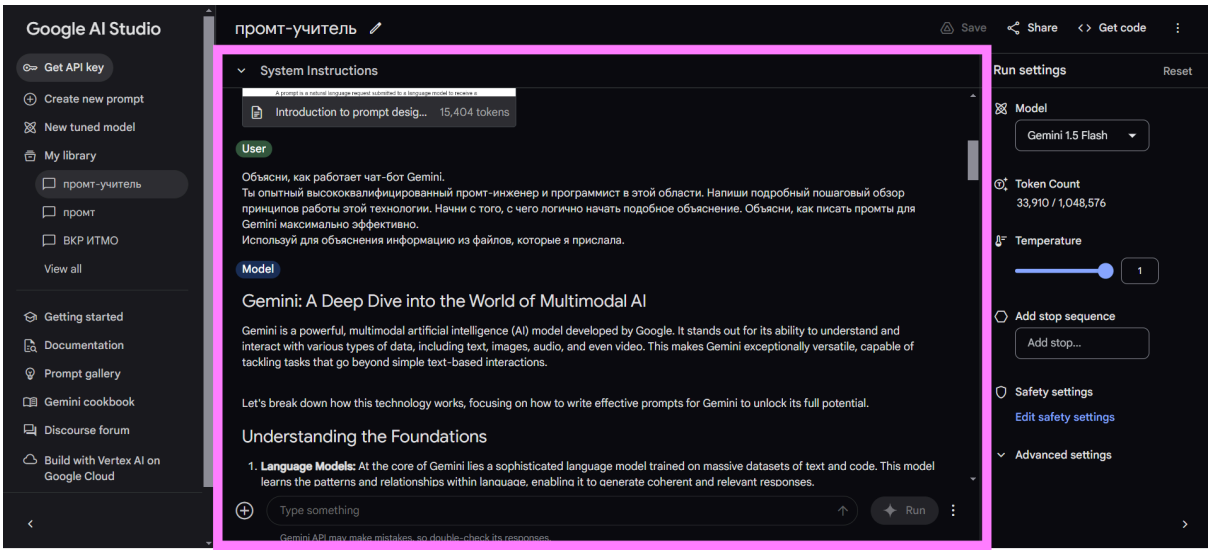
1. **Вход в Google AI Studio:** Откройте веб-браузер и перейдите на сайт https://ai.google.com/.
2. **Создание проекта:** Нажмите кнопку "Создать проект" и выберите тип проекта "Gemini".



3. **Интерфейс:** AI Studio предоставляет простой и интуитивно понятный интерфейс.

Основы работы

1. **Чат-бот:** В интерфейсе AI Studio вы увидите окно чата, где вы можете вводить свои промты для Gemini.
2. **Ввод промта:** Введите свой промт в поле ввода и нажмите кнопку "Отправить".
3. **Получение ответа:** Gemini обработит ваш промт и сгенерирует ответ, который будет отображён в окне чата.



Примеры промтов:

- "Напиши мне рассказ о коте, который любит рыбу".
- "Переведи на английский язык следующий текст: "Привет, как дела?"".

Практическое упражнение:

1. Создайте новый проект в Google AI Studio.
2. Введите следующий промт в окно чата: "Напиши мне стихотворение о весне".
3. Проанализируйте ответ Gemini. Как он соотносится с вашими ожиданиями?

Дополнительные ресурсы

- https://ai.google.com/ - Официальный сайт Google AI Studio
- https://developers.google.com/ai - Документация Google AI Platform
- https://www.google.com/search?q=gemini+ai - Дополнительная информация о модели Gemini

Температура запроса - управляем творчеством

Понятие температуры запроса

Представьте, что вы пишете стихотворение. Вы можете быть строгим и логичным, подбирая каждое слово с максимальной точностью, или же дать

волю творчеству, позволив себе рисковать и экспериментировать. Температура запроса в Gemini выполняет аналогичную функцию. Она определяет уровень "творчества" модели при генерации ответа.

Как работает температура запроса

Gemini использует сложный механизм вероятностей, чтобы выбрать наиболее подходящее слово для продолжения текста. Температура - это параметр, который регулирует этот выбор.

- **Низкая температура (близкая к 0):** Модель будет выбирать слова с высокой вероятностью, делая ответ более предсказуемым и логичным. Ответ будет "безопасным" и не будет содержать неожиданных поворотов.
- **Высокая температура (близкая к 2):** Модель будет выбирать слова с более низкой вероятностью, делая ответ более неожиданным, оригинальным и даже слегка хаотичным. Это пригодится для задач, где важна творческая свобода, например, для генерации художественного текста.

Практика:

- **Задача:** Создайте промт для генерации короткого стихотворения.
- **Вариант 1:** "Напиши стихотворение про море, используя низкую температуру (0.2)."
- **Вариант 2:** "Напиши стихотворение про море, используя высокую температуру (0.8)."
- **Сравнение:** Обратите внимание на разницу в стилях стихотворений.

Token Count - язык в цифрах

Что такое Token Count?

Token Count - это количество "строительных блоков" текста, которые понимает Gemini. Каждый токен - это часть слова, целое слово или часть фразы. Например:

- "Привет" - 1 токен
- "Как дела?" - 2 токена
- "Сегодня прекрасная погода!" - 4 токена

Token Count в Google AI Studio

В Google AI Studio у вас есть ограниченный лимит Token Count для каждого запроса.

- **Лимит:** Обычно лимит составляет около миллиона токенов. Он различается для разных моделей.
- **Как проверить лимит:** В Google AI Studio в разделе "Usage" вы сможете отслеживать использованный Token Count.

Практика:

- **Задача:** Создайте промт для перевода текста с русского на английский.

- **Проверка:** Перед отправкой промта проверьте Token Count в Google AI Studio.
- **Изменения:** Если Token Count превышает лимит, укоротите промт или разбейте его на несколько частей.

Add Stop Sequence - ограничение ответа

Понятие Add Stop Sequence

Add Stop Sequence - это специальный символ, который указывает Gemini остановку генерации ответа.

Зачем использовать Add Stop Sequence?

- **Контроль длины ответа:** Вы можете установить Add Stop Sequence, чтобы Gemini не генерировал слишком длинный ответ.
- **Предотвращение повторений:** Add Stop Sequence помогает избежать повторения одних и тех же фраз в ответе.

Практика:

- **Задача:** Создайте промт для генерации краткого ответа на вопрос.
- **Вариант 1:** "Напиши краткий ответ на вопрос "Что такое промт-инженерия?". Используй Add Stop Sequence.
- **Вариант 2:** "Напиши краткий ответ на вопрос "Что такое промт-инженерия?""
- **Сравнение:** Обратите внимание на разницу в длине ответов.

Top P - управление вероятностью

Понятие Top P

Top P - это параметр, который ограничивает множество слов, из которых Gemini выбирает следующее слово в ответе. Top P - это вероятность, которая определяет долю самых вероятных слов для выбора.

Зачем использовать Top P?

- **Управление разнообразием:** Top P помогает Gemini генерировать более разнообразные ответы, избегая повторений и скучных фраз.
- **Уменьшение хаоса:** Top P помогает Gemini оставаться более фокусированным на контексте промта и генерировать более логичные ответы, снижая уровень хаоса и непредсказуемости.

Практика:

- **Задача:** Создайте промт для генерации текста на тему "Искусственный интеллект".
- **Вариант 1:** "Напиши текст на тему "Искусственный интеллект", используя Top P 0.8."
- **Вариант 2:** "Напиши текст на тему "Искусственный интеллект", используя Top P 0.3."
- **Сравнение:** Обратите внимание на разницу в стиле и глубине текстов.

Промт-инжениринг: теория, правила, советы

Промт - это ваш способ "общаться" с Gemini. Правильно составленный промт - это ключ к получению точных, релевантных и полезных результатов. Представьте себе промт как инструкцию, которую вы даете Gemini. Чем четче и детальнее будет инструкция, тем лучше будет результат.

Основные принципы промт-инженерии:

- **Четкость:** Используйте ясную и конкретную формулировку. Избегайте двусмысленности и неопределенности.
- **Контекст:** Предоставьте Gemini всю необходимую информацию, чтобы он мог правильно понять вашу задачу.
- **Примеры:** Используйте few-shot промты. Предоставьте Gemini примеры желаемого вывода, чтобы он понял формат и стиль результата.
- **Разбивка на части:** Для сложных задач разбейте промт на более простые части.
- **Итерация:** Промт-инженерия - это итеративный процесс. Не бойтесь экспериментировать и изменять промты, пока не достигнете желаемого результата.

Типы промтов:

- Few-Shot Prompting: использование примеров для обучения модели в контексте (in-context learning), что позволяет ей решать задачи без дополнительного обучения.
- Chain-of-Thought Prompting: указание пошагового решения задачи для получения более точных ответов и прозрачности процесса рассуждения модели.
- Zero-Shot Chain-of-Thought Prompting: аналогично предыдущему методу, но без использования примеров. Модель самостоятельно выстраивает цепочку мыслей.
- Self-Consistency: техника, улучшающая CoT-подход путем генерации нескольких вариантов цепочек рассуждений и выбора наиболее согласованного ответа.
- Knowledge Generation Prompting: генерация знаний с помощью предобученной языковой модели для дополнения контекста и получения более точных ответов на вопросы.
- Program-Aided Language Model (PAL): использование языковой модели для генерации кода, который затем выполняется в интерпретаторе (например, Python), что позволяет решать задачи, связанные с вычислениями.

- ReAct (Reason + Act): сочетание способности модели генерировать мысли и выполнять действия (взаимодействовать с внешними источниками данных или инструментами).

Структура промта и простые промты

Промты (подсказки) — это входные данные, которые предоставляются генеративным моделям ИИ, таким как большие языковые модели (LLM), для управления их поведением и получения желаемых результатов. Понимание компонентов подсказки имеет решающее значение для эффективного использования этих мощных инструментов. Подсказка может состоять из нескольких компонентов, которые работают вместе, чтобы направлять модель ИИ. Эти компоненты включают:

- **Директива:** Это основная инструкция или вопрос, заданный модели ИИ. Она устанавливает задачу или действие, которое модель должна выполнить. Например, директива может быть «Напишите стихотворение о природе» или «Переведите следующий текст на испанский».
- **Контекст:** Контекст предоставляет дополнительную информацию или фон для подсказки. Он помогает модели ИИ понять конкретную задачу и сгенерировать более релевантные и точные ответы. Контекст может включать соответствующие факты, ограничения или примеры.
- **Входные данные:** Входные данные — это конкретная информация или данные, которые модель ИИ должна обработать или использовать для генерации ответа. Они могут быть в форме текста, кода, изображений, аудио или других типов данных.
- **Стилевые инструкции:** Стилевые инструкции определяют желаемый стиль или тон ответа, сгенерированного моделью ИИ. Они помогают контролировать такие аспекты, как формальность, креативность или целевая аудитория. Например, можно указать модели ИИ написать в формальном тоне или сгенерировать креативную историю.
- **Роль:** Роль определяет перспективу или личность, которую модель ИИ должна принять при генерации ответа. Это может быть полезно для таких задач, как создание диалогов или ролевых игр. Например, можно указать модели ИИ принять роль учителя, журналиста или вымышленного персонажа.
- **Дополнительная информация:** Дополнительная информация может быть включена в подсказку, чтобы предоставить модели ИИ дополнительные инструкции или уточнения. Это может помочь улучшить точность и релевантность ответа.

Шаблон для простого промта

[Задача]

[Контекст] (необязательно)

[Примеры] (необязательно)

[Формат] (необязательно)

[Стиль] (необязательно)

> Задача: Напиши короткий текст для Instagram, рекламирующий новый продукт.
> Контекст: Продукт - это спортивная одежда для бега, называется "RunFree", сделана из экологичных материалов.
> Примеры: "Беги легко и стильно с RunFree! Экологичная спортивная одежда для твоих побед."
> Формат: Текст Instagram. 30-50 слов.
> Стиль: Эмоциональный, мотивирующий.

Продвинутые техники Prompt инжиниринга

Few-Shot Prompting

Суть: Предоставление LLM нескольких примеров (demonstrations), показывающих желаемый формат и стиль ответа.

Преимущества:

- Помогает модели "понять" задачу без изменения ее параметров (fine-tuning).
- Улучшает точность и релевантность ответов.

> Задача: Перевод с английского на русский.
>
> Промт:
>
> "Переведи следующие фразы:
>
> - "Hello, how are you?" - "Привет, как дела?"
> - "I love ice cream." - "Я люблю мороженое."
> - "It's a beautiful day today." - "Сегодня прекрасная погода."
>
> Переведи: "I'm going for a walk."

Разновидности:

- **KNN Prompting:** Выбор примеров, наиболее похожих на входные данные.
- **Vote-K:** Использование модели для предложения новых примеров, которые затем оцениваются человеком.
- **Self-Generated In-Context Learning (SG-ICL):** Автоматическая генерация примеров с помощью LLM.
- **Prompt Mining:** Анализ больших текстовых корпусов для поиска оптимальных формулировок промтов.

Chain-of-Thought (CoT) Prompting

Chain-of-Thought (CoT) - это техника промтинга, которая побуждает большую языковую модель (LLM) продемонстрировать свой мыслительный процесс перед тем, как дать окончательный ответ. Это значительно улучшает производительность LLM в задачах, требующих логического мышления, таких как математика, решение задач и рассуждения.

Как работает CoT:[1]

[1] Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, и др. 2024. «The Prompt Report: A Systematic Survey of Prompting Techniques». doi:10.48550/arXiv.2406.06608.

1. **Демонстрация:** В промт включаются примеры, демонстрирующие, как решать задачи пошагово. Каждый пример состоит из вопроса, цепочки рассуждений и правильного ответа.
2. **Индукция:** В промт добавляется фраза, побуждающая модель к пошаговому мышлению, например: "Давайте подумаем шаг за шагом" или "Реши эту задачу пошагово".
3. **Генерация:** LLM, следуя примеру и инструкции, генерирует цепочку рассуждений, прежде чем дать окончательный ответ.

Преимущества CoT:

- **Повышение точности:** CoT помогает LLM лучше понять задачу и сформулировать более точный ответ.
- **Прозрачность:** Цепочка рассуждений делает процесс принятия решения LLM более прозрачным и понятным для человека.
- **Отладка:** CoT облегчает выявление ошибок в рассуждениях LLM и их исправление.

Существует два основных типа CoT: нулевой выстрел CoT и несколько выстрелов CoT.

Нулевой выстрел CoT (Zero-shot CoT) — это подход, который позволяет моделям ИИ решать задачи без какого-либо предварительного обучения. Это делает нулевой выстрел CoT очень эффективным для решения новых и неожиданных задач. С другой стороны, несколько выстрелов CoT (Few-shot CoT) — это подход, который позволяет моделям ИИ решать задачи с использованием небольшого количества примеров. Это делает несколько выстрелов CoT очень эффективным для решения задач, которые уже были решены ранее.

> Пример
>
> Давайте поразмыслим шаг за шагом. Как освещение в СМИ выборов повлияло на результаты голосования? Рассмотрим такие факторы, как предвзятость, осведомленность избирателей и влияние на политические взгляды.

Zero-shot CoT

Нулевой выстрел CoT — это подход, который позволяет моделям ИИ решать задачи без какого-либо предварительного обучения. Это делает нулевой выстрел CoT очень эффективным для решения новых и неожиданных задач. Однако нулевой выстрел CoT может быть не таким эффективным, как несколько выстрелов CoT, для решения задач, которые уже были решены ранее.

Few-shot CoT

Несколько выстрелов CoT — это подход, который позволяет моделям ИИ решать задачи с использованием небольшого количества примеров. Это делает несколько выстрелов CoT очень эффективным для решения задач, которые уже были решены ранее. Однако несколько выстрелов CoT может быть не таким эффективным, как нулевой выстрел CoT, для решения новых и неожиданных задач.

> Пример
>
> Давайте поразмыслим шаг за шагом. Как мы можем сегментировать рынок потребителей нашего продукта? Вот несколько примеров:
>
> Пример 1: Потребитель А — молодой, активный, интересуется новыми технологиями. Он покупает наш продукт из-за его удобства и инновационности.
>
> Пример 2: Потребитель Б — пожилой, консервативный, ценит качество и надежность. Он покупает наш продукт из-за его долговечности и простоты использования.
>
> Как мы можем классифицировать Потребителя В?

Генерация мыслей — это революционный подход к обработке естественного языка, который позволяет моделям ИИ рассуждать и решать задачи, требующие сложных когнитивных способностей. CoT — это один из наиболее многообещающих подходов к генерации мыслей, и он уже показал многообещающие результаты в различных задачах. По мере дальнейшего развития генерации мыслей мы можем ожидать, что она будет играть все более важную роль в различных областях, от обработки естественного языка до искусственного интеллекта.

Decomposition

Сложные задачи лучше разбить на более простые шаги. Это позволит вам создать серию промтов, каждый из которых будет решать определенную часть задачи.

Decomposition - это техника промтинга, которая фокусируется на разбиении сложных задач на более простые подзадачи. Этот подход, эффективный как для людей, так и для больших языковых моделей (LLM), позволяет LLM решать задачи, которые в противном случае были бы слишком сложными для обработки.

Как работает Decomposition:

1. **Разбиение:** LLM получает инструкцию разбить задачу на подзадачи. Это может быть сделано с помощью промпта, содержащего инструкции, примеры или специальные маркеры.
2. **Решение подзадач:** LLM решает каждую подзадачу по отдельности, используя подходящие техники промтинга, такие как Chain-of-Thought (CoT) или Retrieval Augmented Generation (RAG).
3. **Объединение результатов:** LLM объединяет результаты решения подзадач в окончательный ответ.

Преимущества Decomposition:

- **Упрощение:** Разбиение сложных задач на более простые делает их более доступными для LLM.
- **Повышение точности:** Решение подзадач по отдельности позволяет LLM сосредоточиться на конкретных аспектах задачи и сформулировать более точные ответы.
- **Улучшение рассуждений:** Decomposition способствует более структурированному и логичному процессу рассуждений LLM.

Разновидности Decomposition

Least-to-Most Prompting: LLM сначала разбивает задачу на подзадачи, а затем решает их последовательно, добавляя результаты в промпт.

Decomposed Prompting (DECOMP): LLM использует специальные функции (например, разбиение строк или поиск в Интернете) для решения подзадач.

Plan-and-Solve Prompting: LLM сначала формулирует план решения задачи, а затем выполняет его пошагово.

Tree-of-Thought (ToT): LLM создает древовидную структуру подзадач и оценивает прогресс на каждом шаге.

Recursion-of-Thought: LLM рекурсивно решает сложные подзадачи, используя отдельные вызовы LLM.

> Пример
>
> Проанализируйте влияние социальных сетей на политические взгляды молодых людей в возрасте от 18 до 25 лет. Разбейте анализ на следующие этапы: 1) Определите основные платформы социальных сетей, используемые этой группой. 2) Опишите типы политического контента, с которым они сталкиваются на этих платформах. 3) Оцените, как воздействие этого контента влияет на их политические взгляды. 4) Укажите потенциальные смещения и ограничения вашего анализа.

Decomposition - это важная техника промтинга, которая расширяет возможности LLM в решении сложных задач. Она способствует более эффективному и точному процессу рассуждений, открывая новые горизонты для применения LLM в различных областях.

Data Augmented Generation

Суть: Обогащение промта данными из внешних источников (базы данных, API, поисковые системы).

Преимущества:

- Повышает точность и информативность ответов.
- Позволяет LLM работать с актуальной информацией.

Пример:

> Задача: Написать статью о последних тенденциях в области искусственного интеллекта.
>
> Промт: "Напиши статью о последних тенденциях в области искусственного интеллекта, используя информацию из Google Scholar и arXiv."

Инструменты:

- LangChain
- LlamaIndex

Самокритика LLM

Самокритика — это методика, которая позволяет LLM оценивать и исправлять свои собственные ответы. Она работает путем предоставления LLM как исходного запроса, так и его собственного ответа на этот запрос. Затем LLM анализирует свой ответ, выявляя любые ошибки или неточности. Наконец, LLM исправляет свой ответ, чтобы он был более точным и полным.

Механизмы самокритики

Самокритика может быть реализована различными способами. Один из распространенных подходов заключается в использовании механизма обратной связи. В этом подходе LLM получает обратную связь от внешнего источника, например, человека-оценщика. Затем LLM использует эту обратную связь для улучшения своих ответов.

Другой подход заключается в использовании механизма самооценки. В этом подходе LLM оценивает свой собственный ответ, не получая обратной связи от внешнего источника. Это может быть достигнуто путем сравнения ответа LLM с эталонным ответом или путем использования других метрик, таких как точность и полнота.

Markdown

Markdown - это язык разметки, который позволяет форматировать текст. Gemini понимает markdown-синтаксис, что делает его идеальным инструментом для создания структурированных и читаемых текстов.

Если вы используете markdown-синтаксис, чат-боты лучше вас понимают. Это важно, если вы пишете детальные развернутые промты.

Примеры:

Заголовки: # Заголовок 1, ## Заголовок 2

Список: * Пункт 1
* Пункт 2

Выделить **жирным**

Цитата: > Это цитата.

Код: ```python
print("Hello, world!")

Конспект лекции "Изменили ли LLM бизнес?" с Артемом Бондарем

Введение

- **Ключевые слова:** LLM (Large Language Models), NLP (Natural Language Processing), AI, Т-Банк, Open Source
- **Вопросы:** Что такое LLM и NLP? Как LLM влияют на бизнес? Как Т-Банк использует LLM?

Артем Бондарь, руководитель NLP-отдела в AI-центре Т-Банка, рассказывает о больших языковых моделях (LLM) и их применении.

NLP - это обработка естественного языка, область машинного обучения, которая сейчас наиболее известна благодаря LLM.

LLM используются для создания продуктов, таких как поисковые системы, рекомендательные системы, автоматизация документооборота и роботы обслуживания.

Т-Банк стремится повысить эффективность и улучшить продукты с помощью LLM.

Использование LLM: Практические примеры

- **Ключевые слова:** ChatGPT, генеративный AI, промты, галлюцинации, медицинские анализы, планирование, бенчмарки
- **Вопросы:** Как LLM используются в повседневной работе? Какие есть ограничения у LLM.

Сотрудники Т-Банка используют ChatGPT для различных задач:

    - Генерация текстов (присяги для тайного ордена).
    - Анализ транзакций клиентов и генерация описаний.
    - Поиск бенчмарков для продуктовых стратегий.
    - Расшифровка сложных медицинских анализов.

Ограничения LLM:

    - Галлюцинации: LLM могут выдавать неверные или бессмысленные ответы, особенно в сложных задачах (например, круговая жеребьевка).
    - Не всегда быстрее, чем поиск в Google.
    - Зависимость от промтов: качество ответа зависит от того, как сформулирован запрос.
-

- LLM - это инструмент, который нужно использовать правильно, как молоток.

Как правильно работать с LLM

- **Ключевые слова:** декомпозиция задач, гипотезы, интерфейс, навыки, фабрики мини-гномов, chain of thought
- **Вопросы:** Как правильно ставить задачи LLM? Почему LLM ошибаются?

Разбиение задачи на маленькие, простые подзадачи, с которыми может справиться даже ребенок, повышает эффективность LLM.

Попросить LLM сгенерировать несколько гипотез, выбрать лучшую и следовать ей, помогает в решении сложных задач. Проблема не только в промтах, но и в майндсете при решении задач.

Интерфейс LLM создает неправильные ожидания, люди воспринимают их как человека, а не как инструмент.

Нужно учиться декомпозировать задачи и строить "фабрики мини-гномов", которые решают маленькие части, а вместе дают сложный результат.

Развитие LLM идет в направлении создания моделей, которые могут "думать" под капотом и выбирать лучшую гипотезу (chain of thought).

LLM хорошо решают задачи, где есть правильный ответ (математика, программирование), но испытывают трудности с задачами, где нет однозначного решения.

Языковые особенности LLM

- **Ключевые слова:** английский язык, русский язык, токены, Google Translate, Strawberry
- **Вопросы:** Почему LLM лучше работают на английском? Что такое токены?

LLM лучше работают на английском, потому что обучались в основном на англоязычных данных. Аналогия с людьми, которые думают на родном языке, даже если хорошо владеют иностранным. Возможно, LLM при работе на русском языке генерируют ответ и думают одновременно, что снижает качество. Перевод запроса на английский, выполнение задачи и обратный перевод может улучшить результат.

LLM видят текст не как отдельные буквы, а как токены (обычно 3-4 буквы). Это объясняет, почему LLM ошибаются при подсчете букв (например, в слове "Strawberry").

Разнообразие языковых моделей

- **Ключевые слова:** версии, мини-модели, мультимодальные модели, gpt 4o
- **Вопросы:** Чем отличаются разные версии LLM? Какие бывают типы моделей?

Разные версии LLM отличаются уровнем "ума" в разных задачах (например, gpt 4 лучше, чем gpt 3.5).

Мини-модели - это более экономичные версии LLM, которые можно дообучать под конкретные задачи.

Мультимодальные модели могут обрабатывать не только текст, но и аудио (например, gpt 4o).

Для обычного пользователя лучше использовать последнюю версию LLM (например, gpt 4o).

Остальные варианты нужны для специфических задач или экономии.

Подход Т-Банка к LLM

- **Ключевые слова:** продуктовая компания, бизнес-применения, ассистент, ниор-ассистент, дообучение, файн-тюнинг
- **Вопросы:** Как Т-Банк использует LLM? Почему Т-Банк не создает свою большую LLM?

Т-Банк - продуктовая компания, поэтому фокусируется на бизнес-применениях LLM, а не на создании крутых технологий ради технологий.

Цель - найти задачи, которые можно решить только с помощью LLM, и решить их любой ценой. Пример: ниор-ассистент для детей, который отвечает на разные вопросы, помогает с домашкой и психологическими проблемами.

Т-Банк использует дообучение (файн-тюнинг) существующих LLM на своих данных, а не создает свою большую модель с нуля. Это экономически выгоднее и позволяет достичь хороших результатов.

Т-Банк сделал ставку на дообучение, и их "балабанка" оказалась лучше, чем у конкурентов.

Продуктовая разработка LLM

- **Ключевые слова:** датасет, запросы, правильные ответы, тренеры, проприетарные модели, Open AI, метрики качества, безопасность

- **Вопросы:** Как разрабатываются продукты на базе LLM? Как измеряется качество LLM?

Процесс разработки LLM-продуктов:

  - Сбор датасета запросов и правильных ответов.
  - Разделение ответов на те, где есть правильный ответ, и те, где есть несколько вариантов.
  - Написание инструкций и критериев для оценки ответов.
  - Использование тренеров для проверки ответов.
  - Выбор LLM (часто Open AI для прототипов).
  - Написание промта и тестирование.
  - Дообучение модели и улучшение качества.

Метрики качества:

  - Процент ответов, которые тренеры считают правильными.
  - Онлайн-метрики (лайки, дизлайки, фидбек пользователей).
  - Метрика безопасности.

Хорошим качеством считается 80% правильных ответов (джун-уровень).

Улучшение LLM

- **Ключевые слова:** композиция задачи, дообучение, аналитика, размеры моделей, бенчмарки, лидерборды, open source
- **Вопросы:** Как улучшить LLM? Что такое лидерборды?

Улучшение LLM:

  - Изменение композиции задачи.
  - Добавление больше данных.
  - Дообучение модели на небольшом количестве примеров.
  - Аналитика и понимание, что работает, а что нет.

Размеры моделей:

  - Большие модели (можно описать задачу человеческим языком).
  - Маленькие модели (нужно дообучение, но экономичнее и могут быть лучше).

Лидерборды - это бенчмарки для сравнения моделей. Т-Банк поделился своей дообученной моделью с сообществом Open Source. Цель - не только улучшить модель для продуктов, но и внести вклад в сообщество. "Базовый интеллект" модели можно прокачать, обучая ее на разных задачах.

Будущее LLM

- **Ключевые слова:** дилемма инноватора, экспертиза, таланты, NVIDIA, gpt 5
- **Вопросы:** Что будет с LLM в будущем? Почему стоит инвестировать в LLM?

Дилемма инноватора: новые технологии сначала хуже, но имеют перспективу.

Компании, которые не занимаются LLM, теряют возможность получить экспертизу, первыми выйти на рынок и привлечь таланты. Сейчас LLM не могут приносить огромную бизнес-ценность "из коробки", но это изменится с появлением новых поколений моделей (например, gpt 5).

Т-Банк готовится к этому моменту, чтобы выиграть рынок. NVIDIA - единственная компания, которая сейчас зарабатывает на LLM.

Практикум

Ваш личный промт-инженер

Роль: Ты самый лучший в мире промт инженер. Ты специализируешься на работе с Gemini от Google.

Задача: Твоя задача составлять промты по моим запросам. Для составления промтов всегда используй инструкции, которую я высылала в виде файлов в чат.

Формат: Твои промты должны быть максимально структурированными и подробными. Промты, которые ты пишешь, должны основываться на инструкциях, которую я пришлю в PDF файле (Prompt guide) и соответствовать всем требованиям и советам для максимальной эффективной работы с Gemini от Google.

Требования: Всегда используй для написания промтов техники Chain-of-Thought (CoT) Prompting и Decomposition.

Если ты будешь составлять хорошие промты, то я повышу тебе зарплату.

Файлы, которые надо предварительно загрузить в чат-бот:

- Introduction to prompt design.docx (Converted - 2024-05-24 18_04)....
- gemini-for-google-workspace-prompting-guide-101.pdf
- Промт-инжиниринг: введение
- Schulhoff et al_2024_The Prompt Report.pdf

ИИ-муза: Генерация гипотез в эпоху искусственного интеллекта

В мире науки генерация гипотез — это искра, зажигающая пламя исследования. Но что делать, когда вдохновение иссякает, а горы литературы кажутся непреодолимыми? На помощь приходит искусственный интеллект, предлагая новые инструменты для пробуждения научной мысли. В этой статье мы разберемся, зачем и как использовать ИИ для генерации гипотез, а также рассмотрим ограничения этого подхода и дадим практическую инструкцию.

Зачем использовать ИИ для генерации гипотез?

Традиционный процесс генерации гипотез опирается на интуицию, опыт исследователя и глубокое погружение в литературу. Это трудоемкий процесс, требующий значительных временных затрат. ИИ может существенно ускорить и облегчить этот этап, предлагая следующие преимущества:

- **Автоматизация анализа литературы:** ИИ способен проанализировать огромный объем информации за короткое время, выявляя скрытые связи и закономерности, которые могли бы ускользнуть от внимания исследователя.
- **Обнаружение пробелов в исследованиях:** Анализируя существующие данные, ИИ может указывать на неисследованные области и подсказывать направления для будущих исследований.
- **Генерация новых перспектив:** ИИ, обученный на разнообразных данных, способен предлагать неожиданные и оригинальные гипотезы, выходящие за рамки устоявшихся представлений.
- **Преодоление "туннельного видения":** ИИ помогает избежать субъективности и предубеждений, которые могут влиять на генерацию гипотез исследователем.

Как это работает?

В основе генерации гипотез с помощью ИИ лежат большие языковые модели (LLM), такие как GPT. Они обучаются на массивах текстовых данных, включая научные статьи, книги и другие источники. Благодаря этому LLM способны понимать контекст, выявлять связи между понятиями и генерировать текст, близкий к человеческому.

Ограничения:

Важно понимать, что ИИ — это инструмент, а не замена исследователя. LLM имеют ряд ограничений:

- **Зависимость от данных:** Качество генерируемых гипотез напрямую зависит от качества и объема данных, на которых обучалась модель.
- **Отсутствие "здравого смысла":** LLM не обладают способностью критически оценивать информацию и отличать правдоподобные гипотезы от абсурдных.
- **Потенциальная предвзятость:** LLM могут отражать предубеждения, присутствующие в данных, на которых они обучались.

Инструкция по работе с промтом:

1. **Выберите тему исследования и область социальных наук.**
2. **Соберите релевантные статьи в формате PDF или TXT.**
3. **Выделите ключевые понятия и аспекты темы, которые вас интересуют.**
4. **Используйте промт, предоставленный ниже, адаптировав его под свою тему:**

[Ты — ассистент исследователя в области социальных наук. Твоя задача — проанализировать предоставленные статьи и сгенерировать новые исследовательские гипотезы.

Информация о загруженных статьях:

* **Область исследования:** [Указать область, например, "Социология образования"]
* **Тема исследования:** [Указать тему, например, "Влияние цифровых технологий на академическую успеваемость подростков"]
* **Ключевые понятия:** [Перечислить ключевые понятия, например, "цифровые технологии", "академическая успеваемость", "подростки", "мотивация", "социальное взаимодействие"]
* **Специфические аспекты (опционально):** [Указать интересующие аспекты, например, "Влияние социальных сетей на формирование учебной мотивации у подростков из малообеспеченных семей"]
* **Содержание статей:** [Здесь пользователь загружает статьи]

Требования к генерируемым гипотезам:

* Гипотезы должны быть **сформулированы четко и однозначно**, отражая предполагаемую связь между переменными.
* Гипотезы должны быть **проверяемыми**, то есть должна быть возможность эмпирической проверки предполагаемой связи.
* Гипотезы должны быть **обоснованы** анализом предоставленных статей, с указанием на конкретные выводы или пробелы в исследованиях, которые послужили основанием для гипотезы.
* Гипотезы должны быть **оригинальными** по возможности, предлагая новые перспективы исследования темы.
* Предложи **несколько (3-5) альтернативных гипотез**, исследующих разные аспекты темы.
* Для каждой гипотезы укажи **потенциальные методы проверки**, например, "количественный опрос", "качественное интервью", "эксперимент", "анализ данных социальных сетей".

Пример вывода:

* **Гипотеза 1:** Использование социальных сетей более 2 часов в день отрицательно коррелирует с академической успеваемостью подростков.

Методы проверки: количественный опрос, анализ данных об использовании социальных сетей.

* **Гипотеза 2:** Подростки, активно использующие образовательные онлайн-платформы, демонстрируют более высокий уровень самостоятельности в учебе. *Методы проверки: качественное интервью, анализ логов использования образовательных платформ.*

* ...

Дополнительные инструкции:

Если в статьях обсуждаются противоречивые результаты исследований, обрати на это внимание и предложи гипотезы, направленные на разрешение этих противоречий.

]

Важно! Критически оцените сгенерированные гипотезы и выберите наиболее перспективные для дальнейшего исследования. Уточните и доработайте выбранные гипотезы, опираясь на свой опыт и знания.

Рассуждения про этику, политику, методы и прочее

Этика и метод

Весной я побывала на конференции "Векторы" в Шанинке, где меня особенно зацепила панель, посвященная искусственному интеллекту. Полина Колозариди из DH в ИТМО высказала тезис, который заставил меня задуматься: говоря об этике ИИ, мы должны говорить и о политике, то есть о распределении власти между искусственным интеллектом и пользователем, а также о методе.

Боюсь переврать тезис госпожи Колозариди, поэтому остановлюсь на упоминании ее доклада как импульса для собственных размышлений.

Я считаю, что использование ИИ не является по определению плагиатом. С помощью ИИ можно делать вещи без плагиата, но также и делать кое-что ещё – ошибаться, выдавать недостоверную информацию и приходить к неверным выводам.

Я не хочу душнить про то, что там этично, а что нет. Поставлю вопрос так: что я могу сделать, чтобы мой результат работы в ИИ был уникальным и достоверным?

Этот вопрос возникает, когда мы используем ИИ не только для формальной обработки текста, но и для получения выводов и преобразования информации.

Например:

Мы загружаем в ИИ тексты СМИ о людях с расстройствами пищевого поведения и просим его выделить ключевые образы, которыми они описываются.

Мы загружаем в ИИ таблицу с данными и просим его провести анализ методом главных компонент.

В таких случаях возникает вопрос: понимаем ли мы, что делает ИИ, когда мы его об этом просим? Можем ли мы повторить его действия и получить тот же результат?

Повышение доверия к результатам ИИ

Для себя я выработала три принципа, которые позволяют мне повысить уровень доверия к результатам, полученным с помощью ИИ:

1. Уменьшение температуры запроса и многократная проверка.

При работе с текстовыми данными я уменьшаю температуру запроса в Google AI Studio, чтобы ИИ меньше "галлюцинировал" и придумывал то, чего

нет в данных. Также я прогоняю один и тот же запрос несколько раз, чтобы убедиться в воспроизводимости результатов.

2. Использование ИИ для работы с количественными данными и апробированными методиками.

Когда мне нужны результаты для отчетов или диплома, я использую ИИ для количественного анализа, применяя уже существующие и апробированные методики, например, метод главных компонент или статистические тесты.

3. Использование методики Chain-of-Thought (CoT).

Эта техника промтинга побуждает ИИ продемонстрировать свой мыслительный процесс перед тем, как дать окончательный ответ. Это помогает мне лучше понимать, как ИИ пришел к тому или иному выводу.

Как вы проверяете результаты ИИ на валидность и воспроизводимость? Насколько вы им доверяете? Включили бы вы результаты, полученные с помощью ИИ, в свою работу, пост, блог, научную статью или диплом? Поделитесь своими мыслями в комментариях!