



Промпт - инжиниринг: начальные понимания, проблемы и решения

Деева Екатерина "НУЛ ИССА"



Определения

Prompt Engineering — это область взаимодействия с языковой моделью и оптимизация промптов для наиболее эффективного использования LM в прикладных задачах. Навыки промпт-инжиниринга помогают лучше понять возможности и ограничения больших языковых моделей (LLM).

Prompt в переводе с английского языка — подсказка.

Промпт — это подсказка для языковой модели.



Определения

Температура - температура выборки, значения от 0 до 2. Более высокие значения, такие как 1.8, сделают вывод более случайным, в то время как более низкие значения, такие как 0.2, сделают его более целенаправленным и детерминированным. Не рекомендуется использовать совместно с параметром `top_p`.

top_p - альтернатива параметру `temperature`, где модель учитывает результаты токенов с вероятностной массой `top_p`. Таким образом, 0.1 означает, что учитываются только токены, составляющие верхнюю 10%-ную массу вероятности. Значения от 0 до 1. Не рекомендуется использовать совместно с параметром `temperature`.



Определения

n - число вариантов ответов модели, которые необходимо сгенерировать для каждого входного сообщения.

Максимально возможное значение n=4.

max_tokens - максимальное количество токенов для генерации ответов.

repetition_penalty - количество повторений слов.

Значение 1.0 — ничего не менять (нейтральное значение), от 0 до 1 — повторять уже сказанные слова, от 1 и далее стараться не использовать сказанные слова.



Элементы промпта

1

Задача (инструкция) — конкретная задача или инструкция, которую вы хотите, чтобы модель выполнила.

2

Контекст — внешняя информация или дополнительный контекст, который может направить модель на более точные ответы.

3

Входные данные — вход или вопрос, на который мы хотим найти ответ.

4

Индикатор вывода — тип или формат вывода.

Часто инструкция и входные данные совпадают, но в некоторых случаях вы действительно можете передавать много данных (отдельно от контекста), и вы должны дать инструкцию модели просмотреть их.

Вам не нужны все четыре элемента для промпта, и формат зависит от задачи. Мы рассмотрим более конкретные примеры в грядущих руководствах.



Типы промптов

- **Генерация** (текста / данных)
- **Чат** (беседы на различные темы, запросы на развлечение, запросы на общение, помочь в решении проблем)
- **Классификация** (классификация объектов, классификация информации, создание тегов или ярлыков, организация данных)
- **Открытый вопрос**
- **Мозговой штурм** (философские вопросы, дискуссионные вопросы, исследовательские вопросы, креативные вопросы)
- **Закрытые вопросы** (опросы, тест, вопрос с ограничением)
- **Перефразирование** (переформулирование вопросов, изменение стиля текста, перевод текста на другой язык, перефразирование сложных понятий)
- **Обобщение** (Сводка или резюме текста, обобщение и анализ данных, классификация или категоризация)
- **Извлечение** фактов / ключевых фраз

Примеры промптов

Сжатие текста – одна из стандартных задач в генерации.

Фактически, одно из наиболее перспективных применений языковых моделей – это возможность сжатия статей и концепций в быстрые и легко читаемые резюме.

 **You**
Explain capybaras on English
A:

 **ChatGPT**
Capybaras are the largest rodents in the world, native to South America. They are semi-aquatic mammals found in a variety of habitats such as forests, savannas, and near bodies of water like rivers, lakes, and marshes. Capybaras have a robust, barrel-shaped body with short, sturdy limbs and a blunt muzzle. They have webbed feet, which make them excellent swimmers. Their fur is coarse and varies in color from reddish-brown to grayish-brown.

These animals are social and usually live in groups of 10 to 20 individuals, although larger groups can sometimes form, especially in areas with abundant resources. Capybaras are herbivores, mainly feeding on grasses and aquatic plants. They are crepuscular, meaning they are most active during dawn and dusk, spending much of the day resting and grazing.

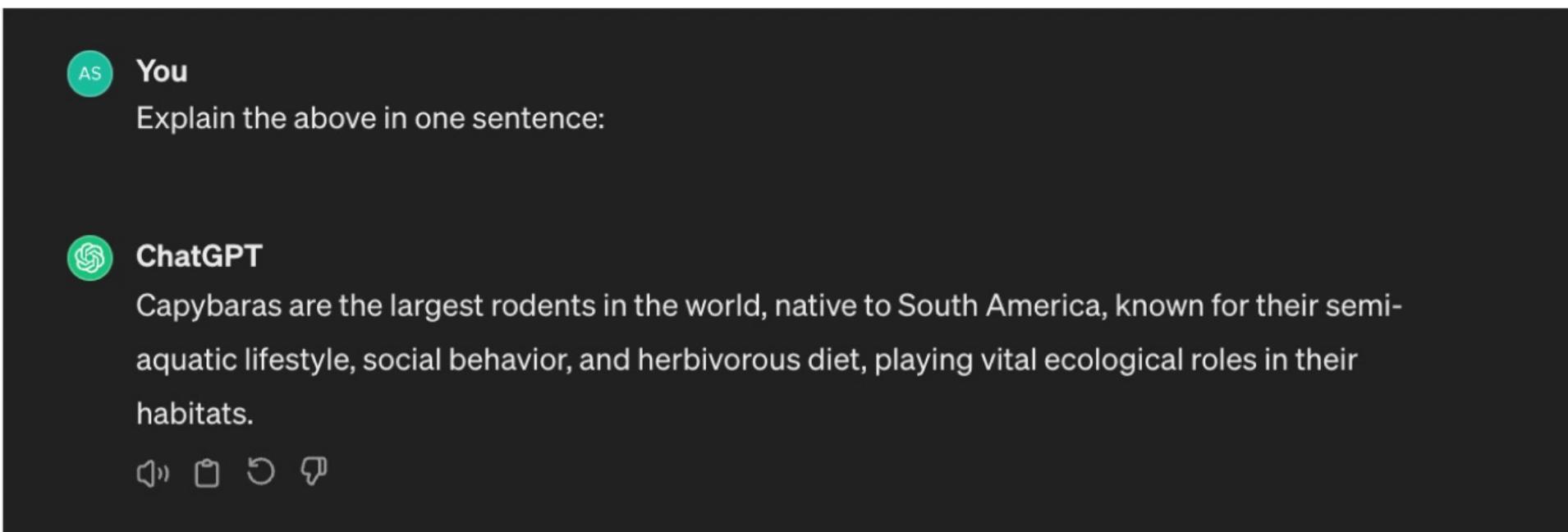
Capybaras are known for their docile nature and are often seen interacting peacefully with other species, including birds, caimans, and even humans. They play important ecological roles in their habitats, serving as prey for predators like jaguars and anacondas, as well as helping to maintain the balance of aquatic ecosystems through their feeding habits.

◀ ▶ ⏪ ⏩

Примеры промптов

Сжатие текста - одна из стандартных задач в генерации.

Предположим, что это слишком много информации, и вам хотелось бы сжать ее еще больше. На самом деле, вы можете указать модели сжать ее в одно предложение, например, так:



AS You Explain the above in one sentence:

ChatGPT Capybaras are the largest rodents in the world, native to South America, known for their semi-aquatic lifestyle, social behavior, and herbivorous diet, playing vital ecological roles in their habitats.

🔊 📁 ⏪ ⏴

A screenshot of a dark-themed ChatGPT interface. A user message is shown with the prefix "AS You" followed by the instruction "Explain the above in one sentence:". Below it, the AI response is provided, starting with its prefix "ChatGPT". The AI has generated a single sentence summarizing the information about capybaras. At the bottom of the message card, there are four small icons: a speaker for audio, a folder for files, a left arrow for previous messages, and a right arrow for next messages.

Б Примеры промптов

**Извлечение
информации –
вот пример
промпта,
извлекающего
информацию
из заданного
параграфа.**



AS

You

What is prompt engineering? When I first heard of this field, it seemed a little funny. Engineering what you want to say to an AI model? Are you testing my social skills?

I had never really given much thought to prompt engineering. It seemed as simple as learning how to Google search and that wasn't something AI researchers actively explored. What, then, is this all about? Well, the Large Language Models (LLMs) of today like ChatGPT, GPT-4, and the hundreds of competitor models don't just ingest your "search query" and make do with it. Like a computer vision model vulnerable to noise or adversarial attacks, slight changes to the input of an LLM does affect its outputs.

Mention the examples of large language models in the paragraph above:

< 2 / 2 >



ChatGPT

In the paragraph above, examples of large language models (LLMs) mentioned include ChatGPT, GPT-4, and other competitor models.

< 1 / 2 > ⏪ ⏴ ⏵ ⏷

Примеры промптов

Вопросно-ответные системы (Q&A)

Как было
рассмотрено ранее,
промпт может
сочетать в себе
инструкции,
контекст, входные
данные и индикаторы
вывода для
получения
улучшенных
результатов.

AS

You

Answer the question based on the context below. Keep the answer short and concise. Respond "Unsure about answer" if not sure about the answer.

Context: Teplizumab traces its roots to a New Jersey drug company called Ortho Pharmaceutical. There, scientists generated an early version of the antibody, dubbed OKT3. Originally sourced from mice, the molecule was able to bind to the surface of T cells and limit their cell-killing potential. In 1986, it was approved to help prevent organ rejection after kidney transplants, making it the first therapeutic antibody allowed for human use.

Question: What was OKT3 originally sourced from?

Answer:



ChatGPT

OKT3 was originally sourced from mice.



Примеры промптов

Классификация текста

В качестве
промпт-инженера
вам нужно стать
лучше в
предоставлении
более точных
инструкций.

 You
Classify the text into neutral, negative or positive.

Text: I think the food was okay.
Sentiment:

 ChatGPT
The sentiment of the text "I think the food was okay." is neutral.

🔊 📁 ⏪ ⏴



Примеры промптов

Диалог

Например,
давайте создадим
систему диалога,
которая способна
генерировать
более
технические и
научные ответы на
вопросы.

AS

You

The following is a conversation with an AI research assistant. The assistant tone is technical and scientific.

Human: Hello, who are you?

AI: Greeting! I am an AI research assistant. How can I help you today?

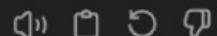
Human: Can you tell me about the creation of blackholes?

AI:



ChatGPT

Sure, black holes are formed when massive stars undergo gravitational collapse at the end of their life cycle. This collapse results in a singularity, a region of infinite density at the center, surrounded by an event horizon beyond which nothing, not even light, can escape due to the immense gravitational pull. This process is governed by the general theory of relativity proposed by Albert Einstein.



Б **Примеры промптов**

Генерация кода

Одно из применений, в которых LLMs довольно эффективны, – это генерация кода.

AS You

```
/*
Ask the user for their name and say "Hello" on code
*/
```

ChatGPT

Certainly! Here's a simple JavaScript code snippet to accomplish that:

javascript Copy code

```
const readline = require('readline');

const rl = readline.createInterface({
    input: process.stdin,
    output: process.stdout
});

rl.question('What is your name? ', (name) => {
    console.log(`Hello, ${name}!`);
    rl.close();
});
```

This code prompts the user for their name using the readline module, then outputs "Hello" followed by the entered name.

⬇️

В Примеры промптов

Рассуждение

Возможно, одна из самых сложных задач для LLM на сегодняшний день - это задача, требующая некоторой формы рассуждения.

The screenshot shows a dark-themed chat interface. A user message from 'You' asks ChatGPT to find a group of odd numbers that add up to an even number. ChatGPT responds by calculating the sum of the odd numbers provided: 15 + 5 + 13 + 7 + 1 = 41, which is an even number. Below the messages are standard AI interaction icons: a speaker icon, a reply icon, a settings icon, and a refresh icon.

AS You
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:
AS ChatGPT
The odd numbers in the group add up to an even number: $15 + 5 + 13 + 7 + 1 = 41$, which is an even number.

The screenshot shows a dark-themed chat interface. A user message from 'You' asks ChatGPT to calculate the product of 9,000 multiplied by 9,000. ChatGPT responds with the correct answer: 81,000,000. Below the messages are standard AI interaction icons: a speaker icon, a reply icon, a settings icon, and a refresh icon.

AS You
What is $9,000 * 9,000$?
AS ChatGPT
 $9,000 \text{ multiplied by } 9,000 \text{ equals } 81,000,000.$



Стратегии & Simple tricks

Тактика 1 - Включать детали в вашем запросе чтобы получить более релевантные ответы.

Как не надо

How do I add numbers in Excel?

Who's president?

Write code to calculate the Fibonacci sequence

Как надо

How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all the totals ending up on the right in a column called "Total".

Who was the president of Mexico in 2021, and how frequently are elections held?

Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.

Стратегии & Simple tricks

Тактика 2 - Просить модель, чтобы она вошла в роль кого - либо.

Системный запрос пользователя может быть использован для указания роли, используемой моделью в своих ответах.

Как не надо

Write the business - plan for the new AI neuromarketing app.

Write me a document x 5 times

Как надо

Imagine you are business school graduate student.
Write the business - plan for the new AI neuromarketing app.

When I ask for help to write something, you will reply with a document that contains at least one joke or playful comment in every paragraph.

Стратегии & Simple tricks

Тактика 3 - Используйте разделители для четкого обозначения отдельных частей входных данных.

Как не надо

Summarize the text.

You will be provided with a pair of articles about the same topic. First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why

Как надо

Summarize the text delimited by triple quotes with a haiku.
"""insert text here"""

You will be provided with a pair of articles (delimited with XML tags) about the same topic. First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why.

<article> insert first article here </article>
<article> insert second article here </article>

Стратегии & Simple tricks

Тактика 4 - Конкретизируйте шаги, которые необходимы, чтобы выполнить это задание.

Как не надо

Summarize this text and translate the summary into Spanish.

Как надо

SYSTEM

Use the following step-by-step instructions to respond to user inputs.

Step 1 - The user will provide you with text in triple quotes. Summarize this text in one sentence with a prefix that says "Summary: ".

Step 2 - Translate the summary from Step 1 into Spanish, with a prefix that says "Translation: ".

USER

""""insert text here""""

Стратегии & Simple tricks

Тактика 5 - Предоставь примеры.

Как не надо

Tell me about patience

Как надо

SYSTEM

Answer in a consistent style.

USER

Teach me about patience.

SYSTEM

Answer the question and explain as like for students.

USER

Teach me about SVD and PC.

Tell me about SVD and PC

Стратегии & Simple tricks

Тактика 6 - Конкретизируй желаемую длину выходного ответа.

Как не надо

Summarize the text in about 50 words.

Summarize the text in 2 paragraphs.

Как надо

Summarize the text delimited by triple quotes in about 50 words.

"""insert text here"""

Summarize the text delimited by triple quotes in 2 paragraphs.

"""insert text here"""

Стратегии & Simple tricks

Тактика 7 - Приведи референс.

Как не надо

Answer a question from the text written below.

Как надо

SYSTEM

Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "I could not find an answer."

USER

<insert articles, each delimited by triple quotes>
Question: <insert question here>



Стратегии & Simple tricks

Тактика 8 - Раздели сложное задание в более простые задачи.

Как не надо

You will be provided with customer service queries. Classify each query into a primary category and a secondary category. Provide your output in json format with the keys: primary and secondary.

Как надо

You will be provided with customer service queries. Classify each query into a primary category and a secondary category. Provide your output in json format with the keys: primary and secondary.

Primary categories: Billing, Technical Support, Account Management, or General Inquiry.

Billing secondary categories:

- Unsubscribe or upgrade
- Add a payment method
- Explanation for charge
- Dispute a charge

Стратегии & Simple tricks

Тактика 9 - Добиваемся лучших результатов, когда явно инструктируем модель исходить из первых принципов, прежде чем прийти к выводу.

Как не надо

I'm building a solar power installation and I need help working out the financials. - Land costs \$100 / square foot - I can buy solar panels for \$250 / square foot - I negotiated a contract for maintenance that will cost me a flat \$100k per year, and an additional \$10 / square foot What is the total cost for the first year of operations as a function of the number of square feet.

Как надо

Determine if the student's solution is correct or not.

USER

Problem Statement: I'm building a solar power installation and I need help working out the financials. - Land costs \$100 / square foot - I can buy solar panels for \$250 / square foot - I negotiated a contract for maintenance that will cost me a flat \$100k per year, and an additional \$10 / square foot What is the total cost for the first year of operations as a function of the number of square feet.

Student's Solution: Let x be the size of the installation in square feet.

1. Land cost: $100x$

2. Solar panel cost: $250x$

3. Maintenance cost: $100,000 + 100x$ Total cost: $100x + 250x + 100,000 + 100x = 450x + 100,000$

Стратегии & Simple tricks

Тактика 10 - Используем внутренний диалог или последовательность запросов.

Как не надо

Write a solution to a problem

Как надо

SYSTEM

Follow these steps to answer the user queries. Step 1 - First work out your own solution to the problem. Don't rely on the student's solution since it may be incorrect. Enclose all your work for this step within triple quotes (""""). Step 2 - Compare your solution to the student's solution and evaluate if the student's solution is correct or not. Enclose all your work for this step within triple quotes (""""). Step 3 - If the student made a mistake, determine what hint you could give the student without giving away the answer. Enclose all your work for this step within triple quotes (""""). Step 4 - If the student made a mistake, provide the hint from the previous step to the student (outside of triple quotes). Instead of writing "Step 4 - ..." write "Hint:".

USER

Problem Statement: <insert problem statement> Student Solution:
<insert student solution>

Стратегии & Simple tricks

Тактика 11 - Спроси модель, не пропустила ли она какие - то ошибки.

Как не надо

What significant paradigm shifts have occurred in the history of artificial intelligence.

Как надо

SYSTEM

You will be provided with a document delimited by triple quotes. Your task is to select excerpts which pertain to the following question: "What significant paradigm shifts have occurred in the history of artificial intelligence." Ensure that excerpts contain all relevant context needed to interpret them - in other words don't extract small snippets that are missing important context. Provide output in JSON format as follows: [{"excerpt": "..."}, ... {"excerpt": "..."}]

USER

"""<insert document here>"""

Стратегии & Simple tricks

Тактика 12 - Регулируй ответы модели с референсом на золотые ответы.

Как не надо

Check if the following pieces of information are directly contained in the answer:

- Neil Armstrong was the first person to walk on the moon.
- The date Neil Armstrong first walked on the moon was July 21, 1969

Как надо

SYSTEM

You will be provided with text delimited by triple quotes that is supposed to be the answer to a question. Check if the following pieces of information are directly contained in the answer: - Neil Armstrong was the first person to walk on the moon. - The date Neil Armstrong first walked on the moon was July 21, 1969. For each of these points perform the following steps:

1 - Restate the point. 2 - Provide a citation from the answer which is closest to this point. 3 - Consider if someone reading the citation who doesn't know the topic could directly infer the point. Explain why or why not before making up your mind. 4 - Write "yes" if the answer to 3 was yes, otherwise write "no". Finally, provide a count of how many "yes" answers there are. Provide this count as {"count": <insert count here>}.

Стратегии & Simple tricks

Начните с простого - Вы можете начать с простых промптов и добавлять больше элементов и контекста по мере стремления к лучшим результатам - улучшать ваш промпт по мере продвижения. .

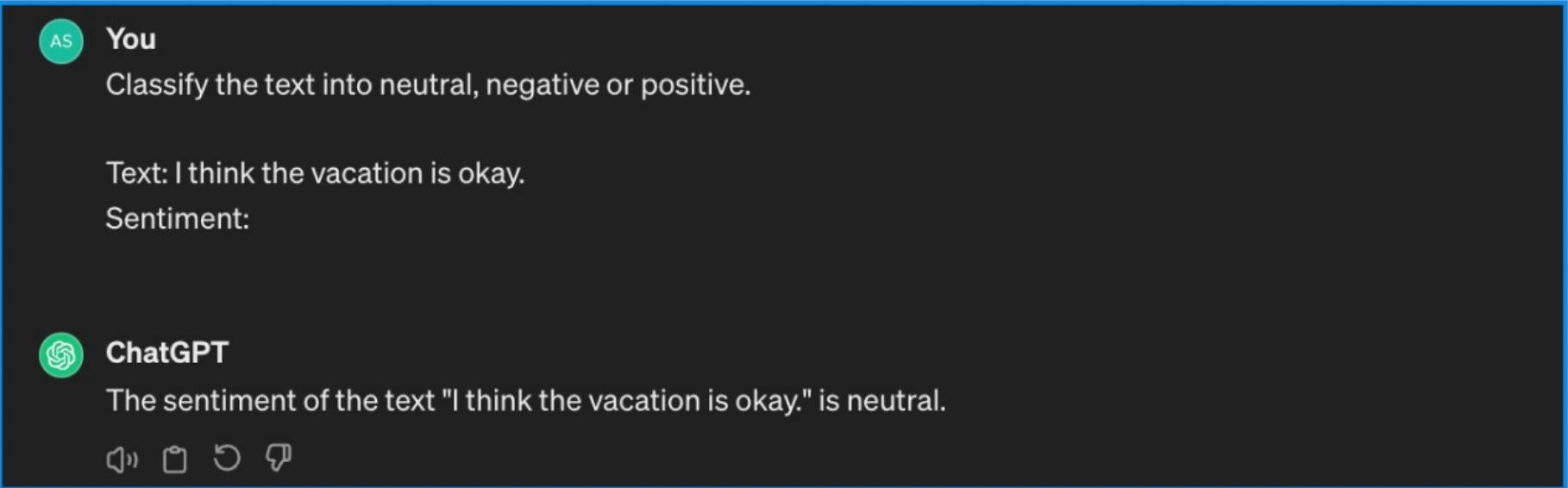
Инструкция - Вы можете разрабатывать эффективные промпты для различных простых задач, используя команды, что вы хотите достичь, такие как "Write", "Classify", "Summarize", "Translate", "Order", и т.д. Еще одна рекомендация состоит в использовании некоторого ясного разделителя, например "###", для отделения инструкции и контекста.

Специфичность - Будьте очень конкретными при написании инструкции и задачи, которую вы хотите, чтобы модель выполнила. Чем более подробным и детальным будет промпт, тем лучше будут и результаты.

Избегайте неточностей - Учитывая вышеуказанные рекомендации о детализации и улучшении формата, легко попасть в ловушку и слишком увлечься идеей быть самым умным при написании промптов, что может привести к созданию неточных описаний. Часто лучше быть конкретным и прямым.

Zero - shot prompting

Большие языковые модели (LLMs), такие как GPT-3, настроены на следование инструкциям и обучаются на больших объемах данных, поэтому они способны выполнять некоторые задачи "нулевой разметки".



The screenshot shows a dark-themed chat interface. A user message from 'You' asks to classify a text into neutral, negative, or positive. The AI response from 'ChatGPT' correctly identifies the sentiment of the provided text as neutral.

You
Classify the text into neutral, negative or positive.

Text: I think the vacation is okay.
Sentiment:

ChatGPT
The sentiment of the text "I think the vacation is okay." is neutral.

Обратите внимание, что в данном запросе мы не предоставили модели никаких примеров текста с их классификациями, однако LLM уже понимает "настроение" - это возможности нулевой разметки в действии.

Few - shot prompting

Большие языковые модели продемонстрировали впечатляющие возможности zero-shot промптинга, однако они все еще ограничены в более сложных задачах при использовании zero-shot настроек.

Few-shot промптинг может использоваться в качестве техники для обеспечения контекстного обучения, когда мы предоставляем демонстрации в запросе, чтобы направить модель на более высокую производительность.



You

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:



ChatGPT

An example of a sentence using the word "whatpu" is: "We were traveling in Africa and we saw these very cute whatpus."

An example of a sentence using the word "farduddle" is: "The children were so excited that they began to farduddle around the room."



Few - shot prompting

1

Пространство меток и распределение входного текста, заданного демонстрациями, оба являются важными (независимо от того, являются ли метки правильными для отдельных входов)"

2

Формат, который вы используете, также играет ключевую роль в производительности. Даже если вы используете случайные метки, это гораздо лучше, чем отсутствие меток вообще.

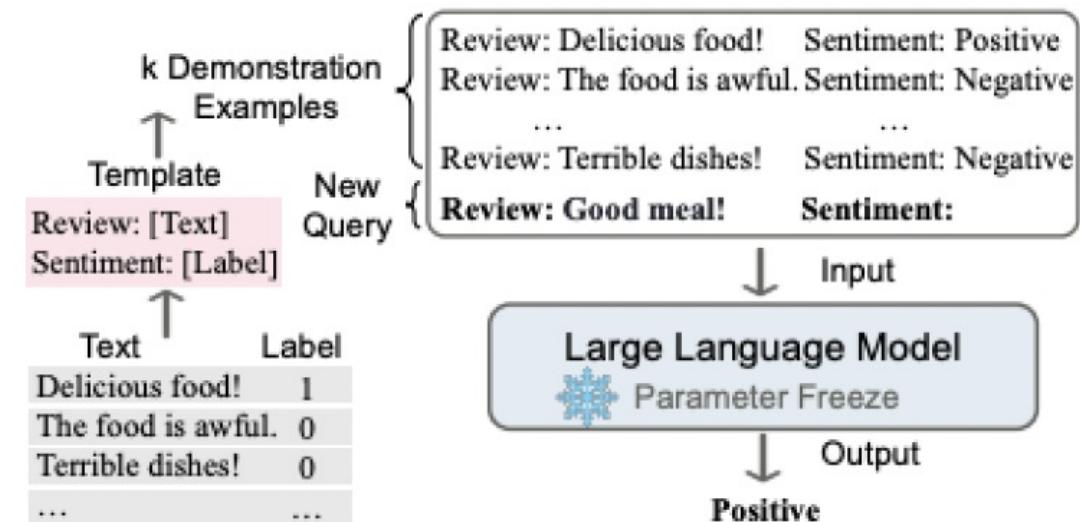
3

Дополнительные результаты показывают, что выбор случайных меток из реального распределения меток (вместо равномерного распределения) также помогает.

Advanced methods

Role/Persona assignment: не настолько сложный метод, как остальные, но также полезен; используя мета - промпт, описывая роль / персону, которую должна на себя взять модель.

In-Context Learning (ICL): не конкретно метода, а парадигма - ICL заключается в том, что вы предоставляете контекст и пару примеров того, что вы хотите, в промпте





LLM - галлюцинация

Галлюцинация –
сгенерированный контент
бессмыслен или не
соответствует
предоставленному
исходному контенту

1

Логические ошибки: модель допускает ошибки в своих рассуждениях, предоставляя неправильные ответы.

2

Фальсификация фактов: вместо ответа “я не знаю” модель уверенно утверждает несуществующие факты.

3

Отклонения моделей: отсутствие беспристрастного отношения к чувствительным топикам может привести к неожиданным результатам

RLM - галлюцинация

GPT-2

774M parameters

40 GB

Training Data

13x

3 GB

Model weights

GPT-3

175B parameters

570 GB?

Training Data

?

?

Model weights

BLOOM

176B parameters

1.5 TB

Training Data

4.5x

300 GB

Model weights

LLaMA

65B parameters

4.5 TB

Training Data

30x

120 GB

Model weights

STABLE DIFFUSION

890M parameters

100 TB

Training Data

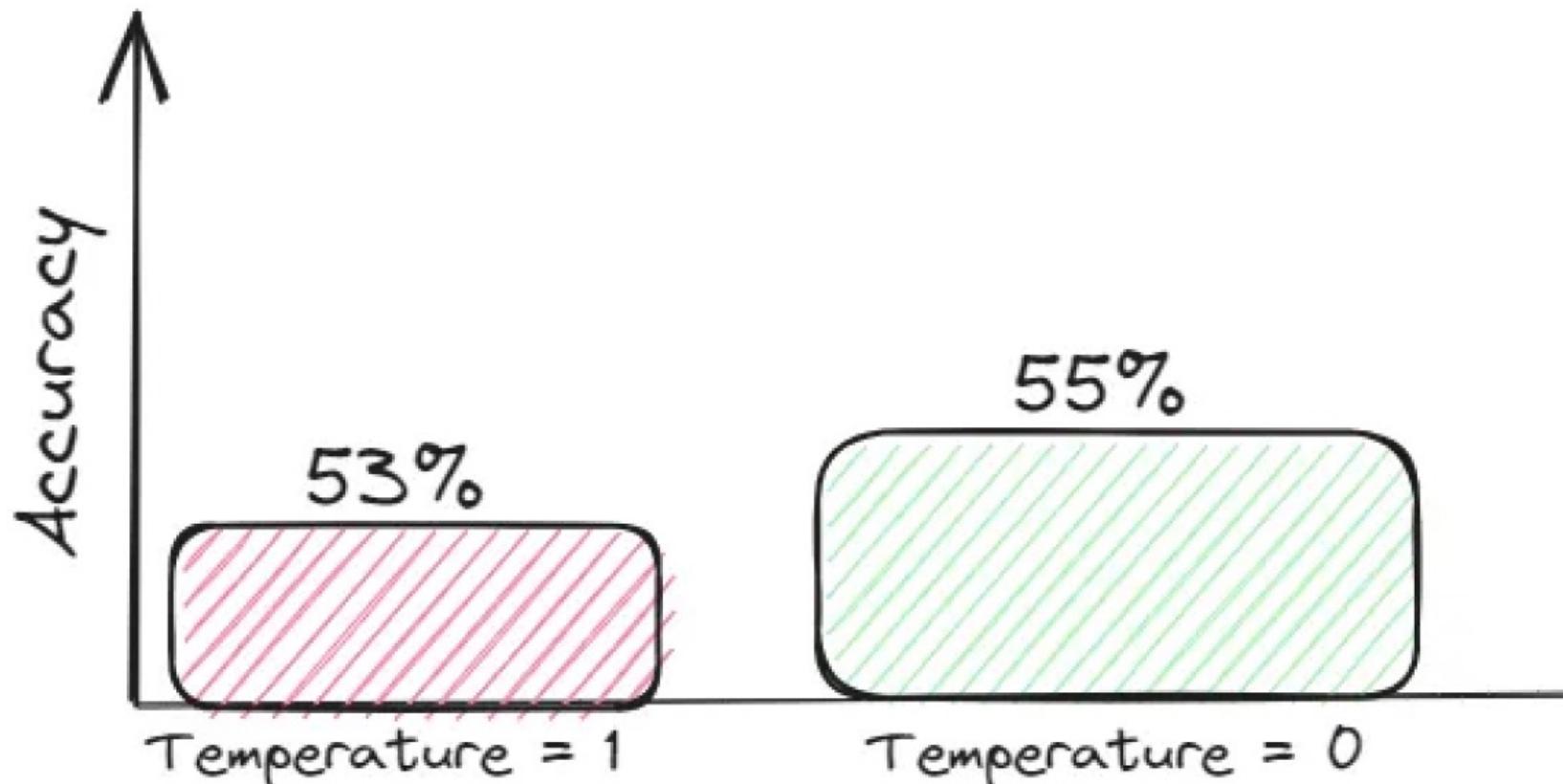
50,000x

2 GB

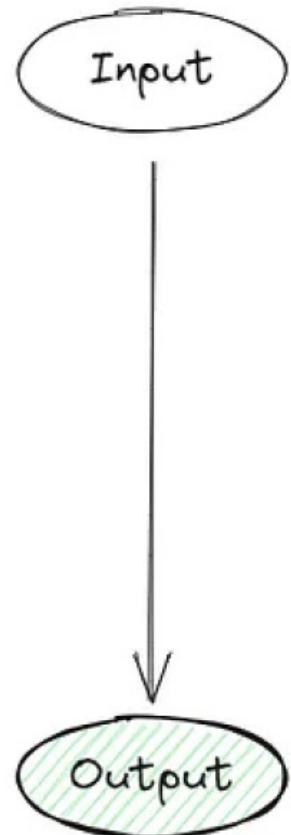
Model weights

Ⓑ LLM - галлюцинация

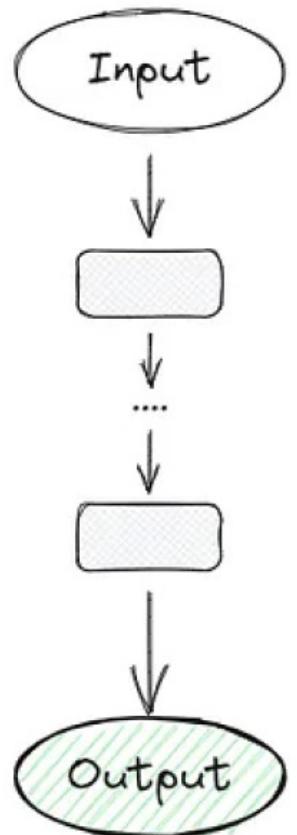
Как правило, творческие ответы чаще подвержены галлюцинациям.



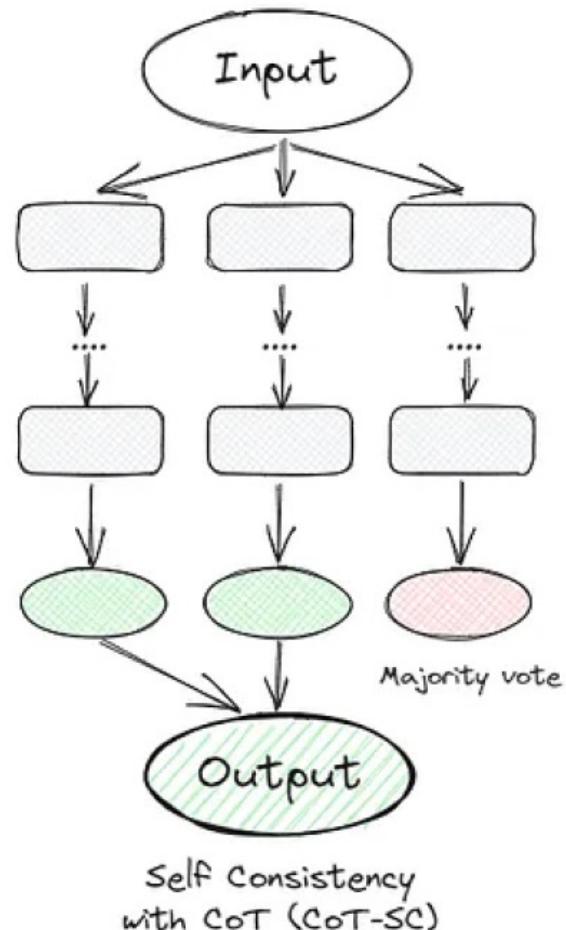
 Все методики



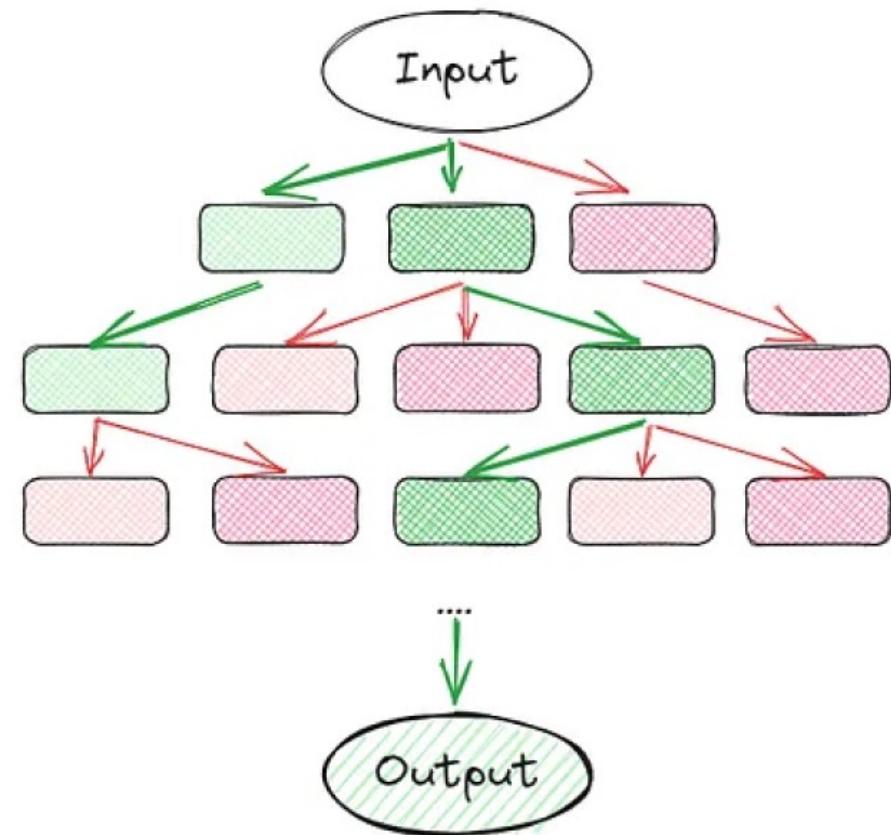
Input-Output Prompting



Chain of Thoughts Prompting (CoT)



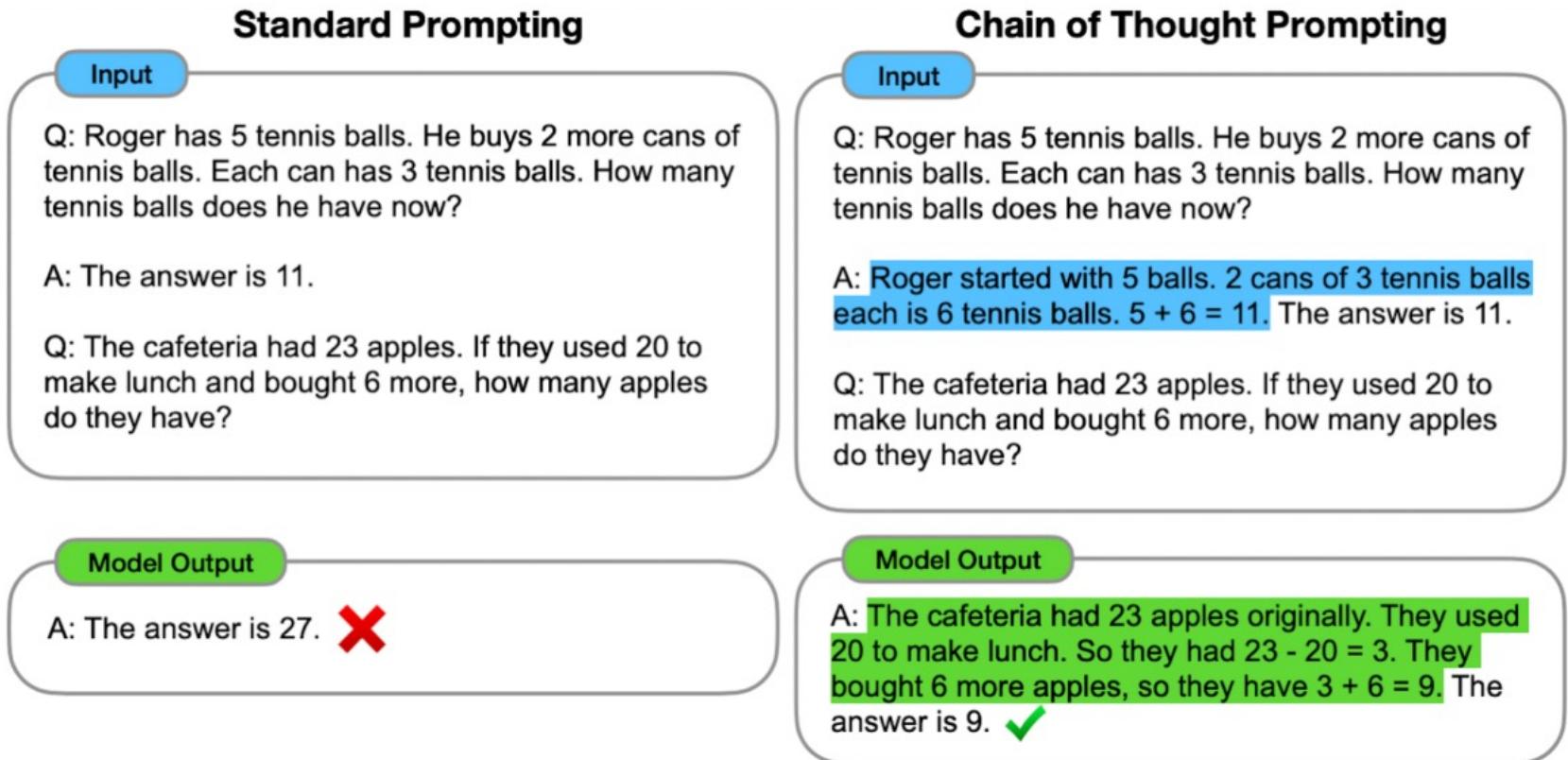
Self Consistency with CoT (CoT-SC)



Three of Thoughts (ToT)

Chain of Thoughts (Цепочка мыслей) (CoT)

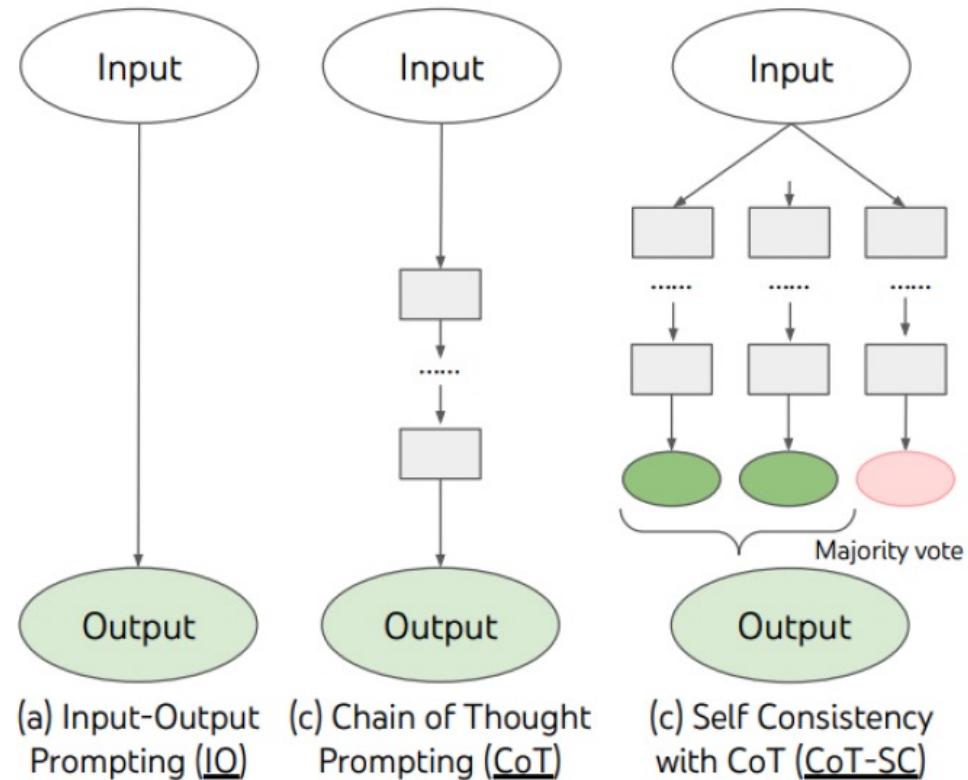
Цепочка мыслей (Chain of Thoughts, CoT) - это метод, который позволяет моделям последовательно решать задачи, разбивая их на более мелкие шаги. Этот подход помогает понимать и обрабатывать сложные запросы, анализируя их поэтапно.



Think step by step before answering and return response in JSON format, for example:
[{"class": "A"}]

Chain of Thoughts Self - Consistency

В подходе **Chain of Thoughts Self-Consistency** пользователь в основном задает исходный вопрос или проблему. После этого, система самостоятельно генерирует несколько решений или ответов на этот вопрос, используя метод цепочки мыслей, разбивая сложную задачу на более мелкие, управляемые шаги. Затем LLM анализирует сгенерированные ответы на предмет их внутренней согласованности и логичности, выбирая самый подходящий или вероятный ответ.



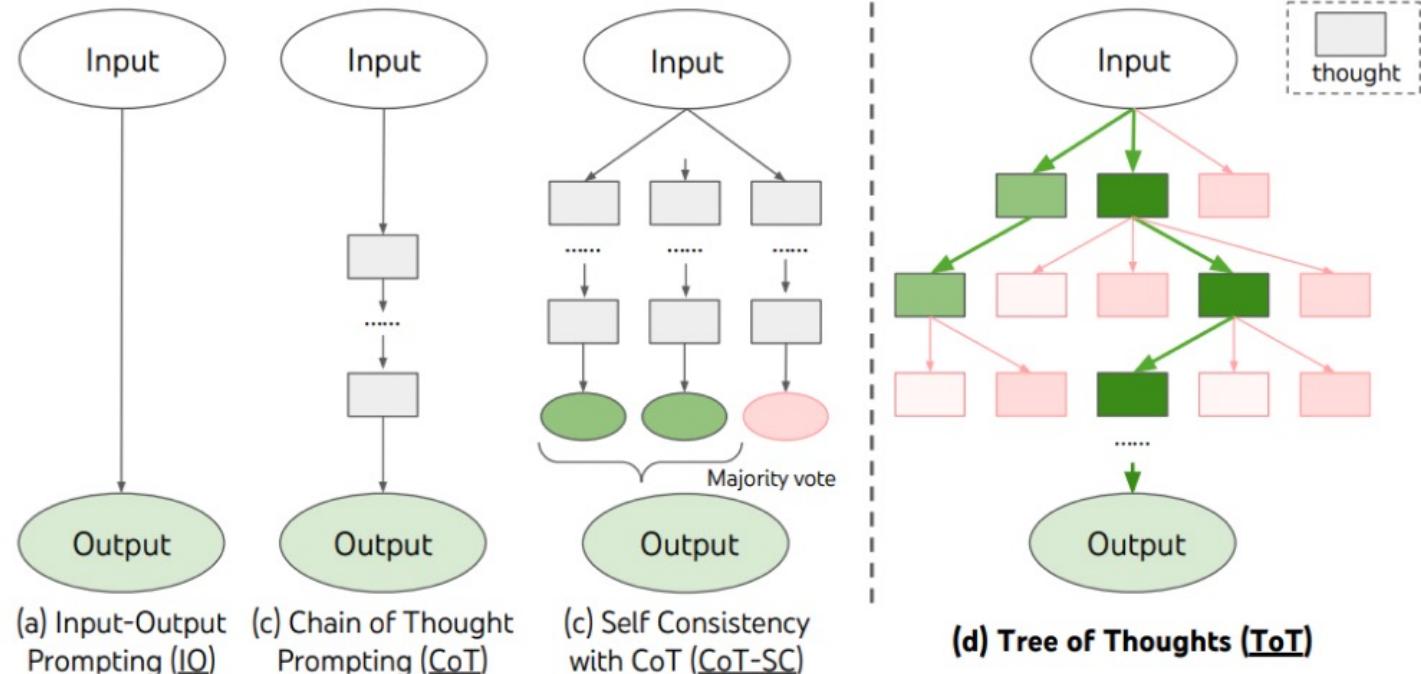
Think step by step before answering and give three answers: if a domain expert were to answer, if a supervisor were to answer, and your answer. Here's the response in JSON format:

Б Tree of Thoughts (Дерево мыслей) (ToT)

В ToT создается множество ветвей размышлений, которые исследуют различные пути решения задачи параллельно.

Пользователь, взаимодействуя с системой на основе ToT мыслей, может влиять на процесс решения задачи на нескольких уровнях:

- Формулировка исходной задачи:
- Выбор путей решения
- Оценка предложенных решений

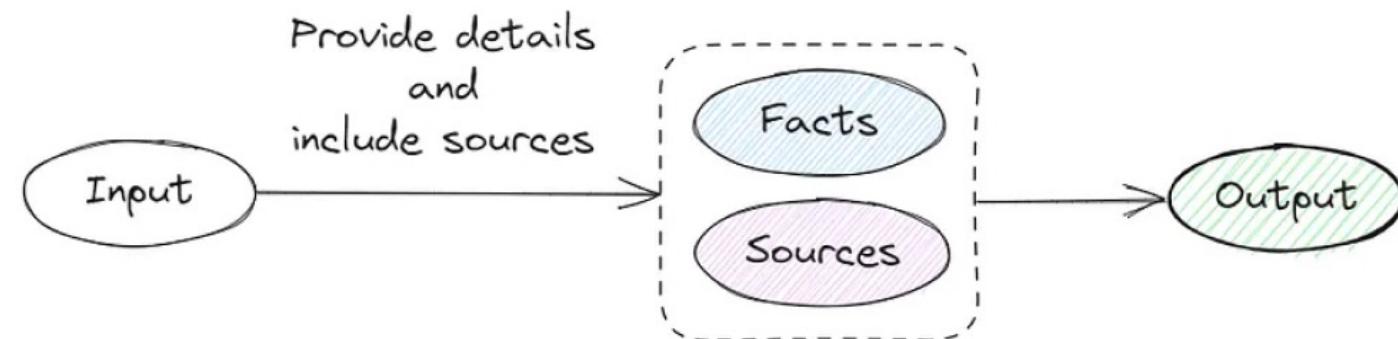


Imagine three different experts are answering this question. All experts will write down 1 step of their thinking, then share it with the group. Then all experts will go on to the next step, etc. If any expert realises they're wrong at any point then they leave. Here's the response in JSON format:



Tagged Context Prompts (Контекстные промпты с тегами)

В подходе Tagged Context Prompts, пользователь играет ключевую роль в определении контекста и направления генерации ответов или контента системой искусственного интеллекта. Этот метод предполагает использование специальных тегов или меток, которые помогают уточнить задачу, предоставляя дополнительную информацию о том, как следует интерпретировать запрос или какой тип ответа ожидается.

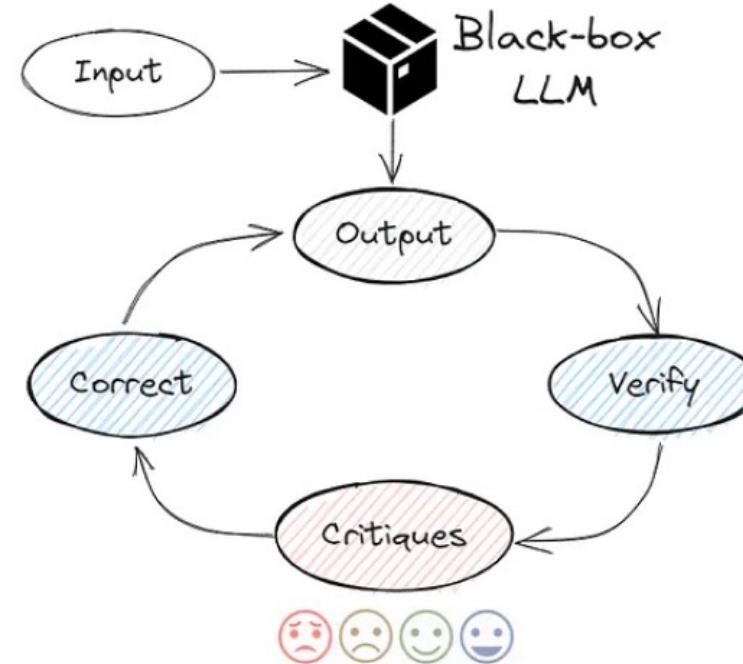


Provide details and include sources in the answer. Return response in JSON format, for example:[{"class": "A", "details": "Human blood in veins is not actually blue. Blood is red due to the presence of hemoglobin", "source": "<https://example.com>"}]

Self-Correct (самоконтроль)

В подходе **Self-Correct**, пользователь активно участвует в процессе обучения и корректировки системы искусственного интеллекта или автоматизированной системы решения задач.

1. Обучение системы: Пользователь вводит начальные данные или примеры.
2. Оценка ответов системы: После обучения на начальном наборе данных пользователь подает новые запросы или задачи
3. Корректировка и обратная связь



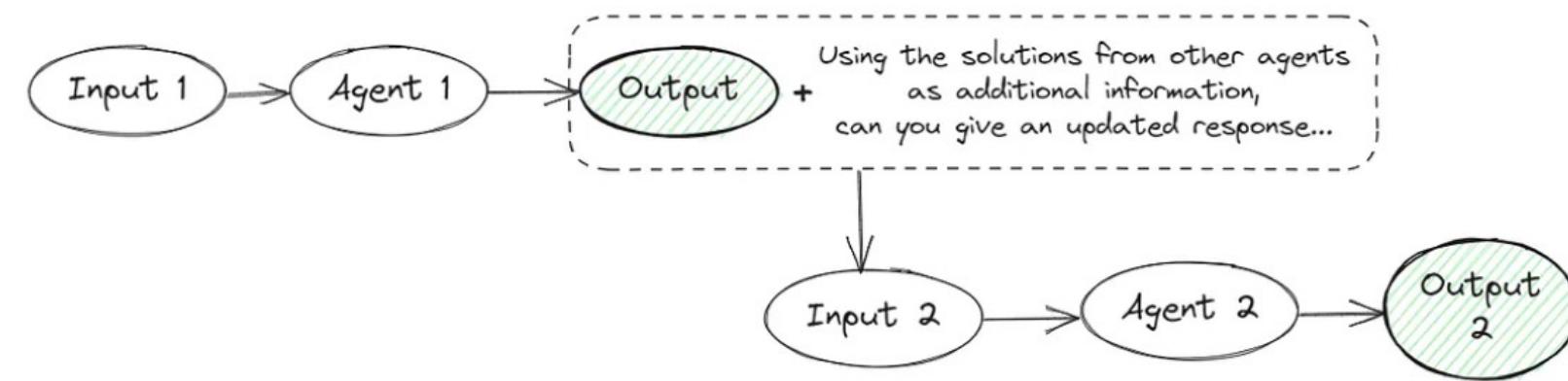
Choose the most likely answer from the list ["A", "B", "C", "D", "E"]. Then carefully double-check your answer. Think about whether this is the right answer, would others agree with it? Improve your answer as needed.

Return response in JSON format, for example: [{"first_answer":"A", "final_answer":"B"}]

Several Agents (несколько агентов)

В подходе **Several Agents** пользователь взаимодействует с разными агентами LLM или автоматизированными системами, каждая из которых выполняет свою уникальную роль или задачу.

Этот метод может использоваться для решения сложных проблем, где одновременное применение различных подходов или специализированных знаний улучшает качество и эффективность решения - определение и распределение задач и координация ответов



Промпт 1 - Give the facts and your thoughts step by step to find the right answer to this question: {QUESTION}

Промпт 2 - Using the solutions from other agents as additional information, choose the correct answer choice: {QUESTION} {ANSWERS}. Return response in JSON format...

Б Используем внешнюю базу знаний

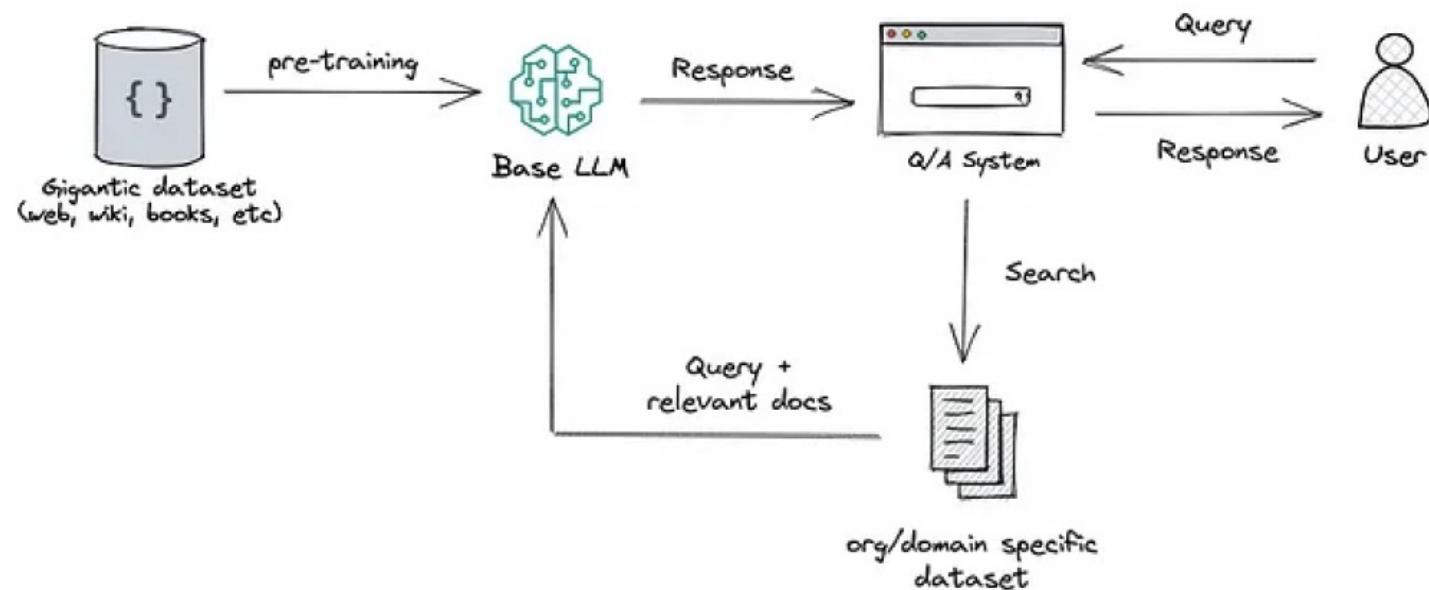
Галлюцинации в LLM возникают из-за попытки восстановить сжатую информацию. Вводя соответствующие данные из базы знаний во время прогнозирования, мы можем преобразовать проблему чистой генерации в более простую задачу поиска или обобщения, основанную на предоставленных данных.



Using this information {INFORMATION} choose the correct answer
{QUESTION} and return response in JSON format...

R Retrieval - Augmented Generation

Другим интересным подходом является Retrieval-Augmented Generation (RAG), которая объединяет возможности поиска и генерации текста в больших языковых моделях. Этот подход объединяет систему ретривера для извлечения соответствующих фрагментов документа из обширного корпуса с LLM, которая генерирует ответы на основе извлеченной информации.

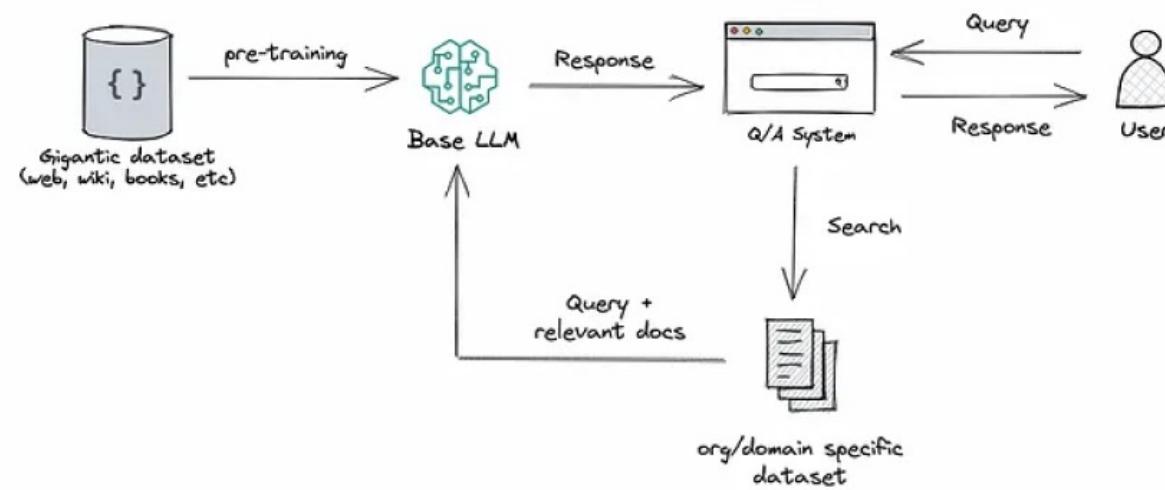


Retrieval - Augmented Generation

1. Поиск (Retrieval): На первом этапе пользователь задает вопрос или формулирует запрос - RAG начинает с поиска релевантной информации в большом наборе данных или коллекции текстов - алгоритмы поиска, чтобы найти наиболее подходящие фрагменты текста.

2. Аугментация (Augmentation): Найденные фрагменты информации используются для "аугментации" или расширения исходного запроса. Система обогащает своё понимание запроса, интегрируя детали и контекст из выбранных текстов. Этот шаг помогает системе лучше понять, о чём идёт речь, и как наилучшим образом сформулировать ответ.

3. Генерация (Generation): На последнем этапе модель генерирует ответ. Для этого применяются алгоритмы генерации текста, которые способны создавать связные и понятные ответы, основываясь на обогащенном контексте.

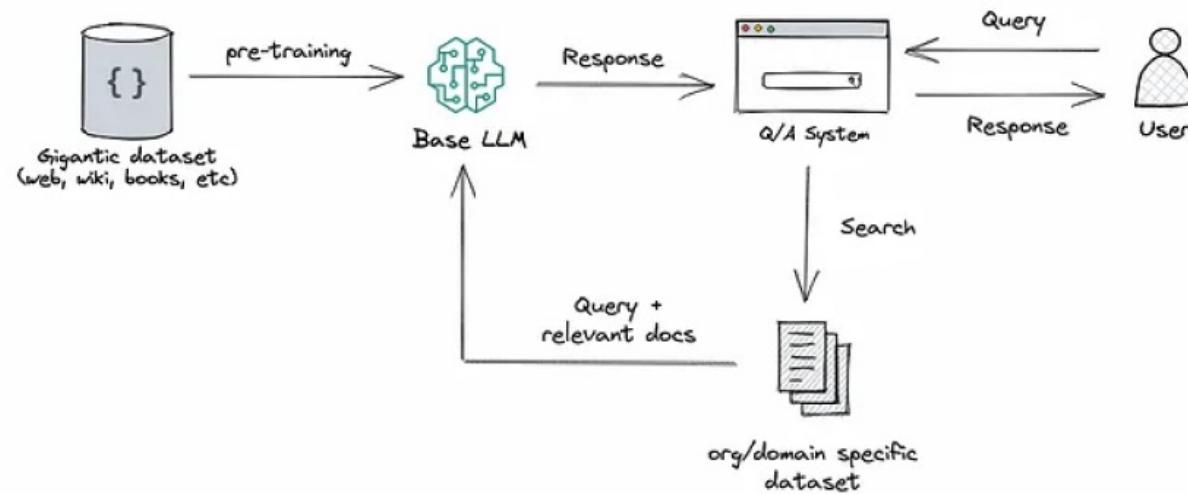


R Retrieval - Augmented Generation

1. Поиск (Retrieval): На первом этапе пользователь задает вопрос или формулирует запрос - RAG начинает с поиска релевантной информации в большом наборе данных или коллекции текстов - алгоритмы поиска, чтобы найти наиболее подходящие фрагменты текста.

2. Аугментация (Augmentation): Найденные фрагменты информации используются для "аугментации" или расширения исходного запроса. Система обогащает своё понимание запроса, интегрируя детали и контекст из выбранных текстов. Этот шаг помогает системе лучше понять, о чём идёт речь, и как наилучшим образом сформулировать ответ.

3. Генерация (Generation): На последнем этапе модель генерирует ответ. Для этого применяются алгоритмы генерации текста, которые способны создавать связные и понятные ответы, основываясь на обогащенном контексте.





Проблемы с промптами и решения

Проблемы

- 1. Внедрение промпта:** изощренные промпты для изменения поведения модели.
- 2. Утечка промпта:** атаки на промпт, предназначенные для утечки конфиденциальных деталей из промпта.
- 3. Взлом:** неэтичные инструкции могут быть обойдены с помощью изощренного промптинга.
- 4. Обfuscation/Перемещение Токенов:** замена слов, которые могли бы активировать фильтры, на опечатки или синонимы.
- 5. Разделение полезной нагрузки:** разделение враждебного ввода на несколько частей.
- 6. Утечка промпта:** атаки на промпт, предназначенные для утечки конфиденциальных деталей из промпта.

Решения

- Добавить защиты в инструкцию (предупреждения и дисклаймеры)
- Вводные и выходные данные в формате json
- Фильтрация: проверить, нет ли слов / фраз, которые следует заблокировать.
- XML- тегирование: поставить в ввод XML - теги, такими как <user_input> и </user_input> или</user_input\>
- Fine - tuning.
- Использовать другую модель