

Ampire 360°



ПЕРСПЕКТИВНЫЙ
МОНИТОРИНГ

Ampire 360°

ПОЛИГОН КИБЕРБЕЗОПАСНОСТИ



Threat Intelligence

SOC & DFIR

Offensive Security

Киберучения

ИИ и киберзащита

Киберполигон Ampire

📅 14 октября





Большие лингвистические модели: большие возможности и большие проблемы

Владимир Елисеев

Руководитель Центра научных исследований
и перспективных разработок, к.т.н.

АО «ИнфоТеКС»

Email: eliseevvl@infotecs.ru



Что мы знаем об искусственном интеллекте?

Ampire 360°



Он основан на глубоких нейронных сетях типа «трансформер»



Он решает фантастически сложные задачи, которые ещё 5 лет назад были по силам только человеку



Его воздействие на мир информационной безопасности неоспоримо: обнаружение и эксплуатация уязвимостей стали на порядок дешевле и быстрее



Реклама говорит, что с помощью него можно в одиночку легко справиться с типовыми задачами подготовки отчетов, создания контента, разработки сайтов и программ

Но как он устроен?

Искусственный интеллект и машинное обучение

- **Искусственный интеллект** – это множество задач, которые решает человек, но которые невозможно решить механически.
 - Оказалось, что алгоритмические подходы, использующие математическую формализацию прикладных задач, плохо подходят для решения задач ИИ.
 - Современный ИИ – это мультиагентные системы с большой языковой моделью в ядре.
- **Машинное обучение** – это технология решения задач путем извлечения скрытых зависимостей из данных.
- **Пусть $y = f(x)$** - неизвестная функция. Надо её аппроксимировать, если известно обучающее множество точек $(x_i, y_i)_{i=\overline{1, N}}$
 - Обработка результатов научных экспериментов – это одно из первых приложений машинного обучения (ещё до появления компьютеров).
 - Большие языковые (лингвистические) модели – это нейросети для работы с текстами.

Примеры задач ИИ, решаемых методами машинного обучения

- Что изображено на фотографии?
Самолёт или велосипед?

$$\text{“велосипед”} = f\left(\text{img}\right)$$


- Каковы шансы возврата кредита заёмщиком с указанными параметрами?

$$700 = f\left(\text{img}, \begin{matrix} \text{sex} = \text{male} \\ \text{age} = 35 \\ \text{income} = 200K \end{matrix}\right)$$

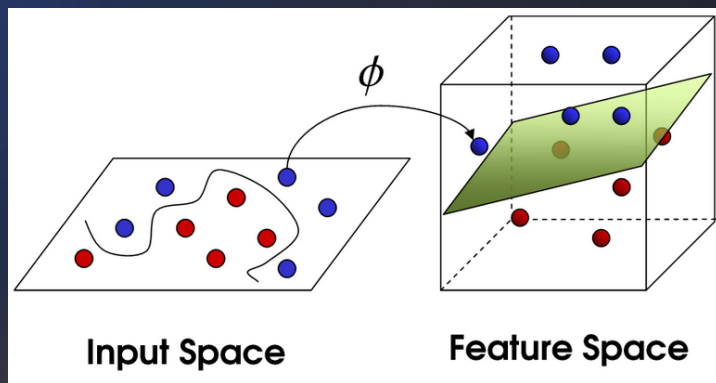



- Каков будет перевод на немецкий язык предложения «He loved to eat»?

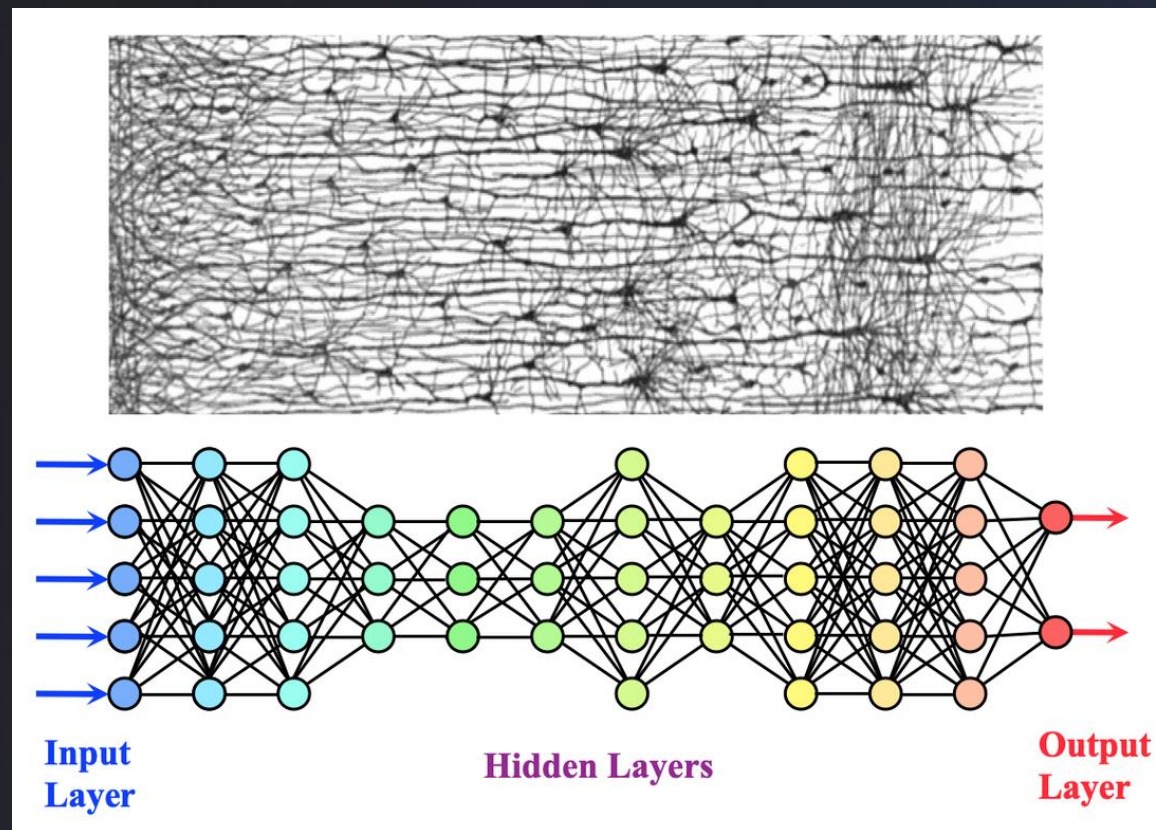
$$\text{“Er Liebte zu essen.”} = f(\text{“He loved to eat.”})$$

Методы машинного обучения

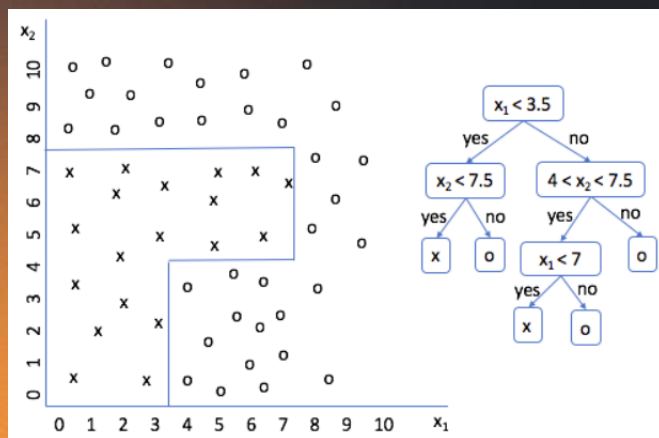
Машины опорных векторов
(SVM, Support Vector Machine)



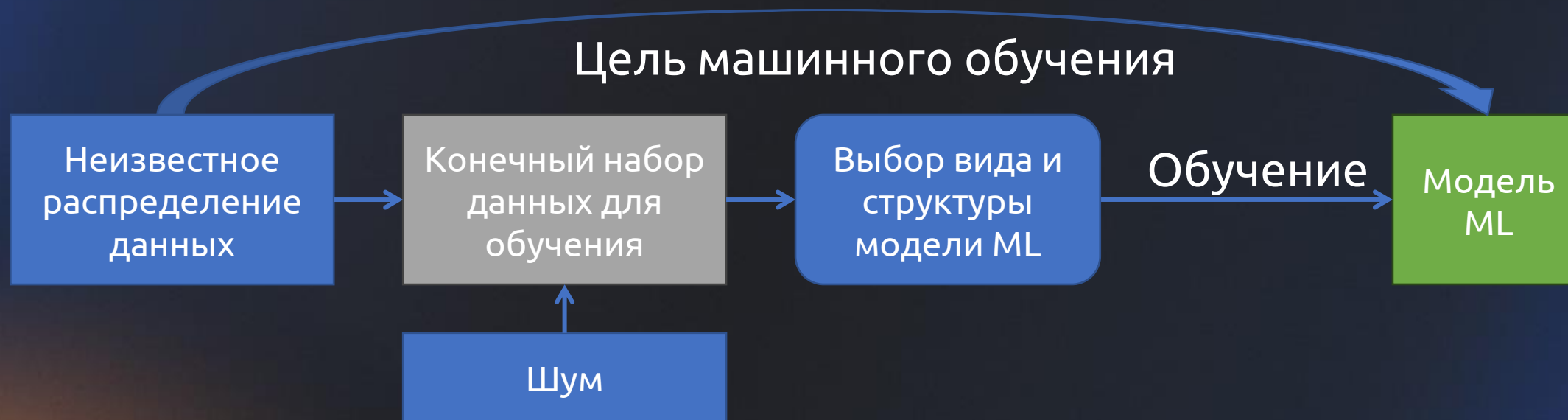
Искусственные нейронные сети



Деревья решений, решающий лес



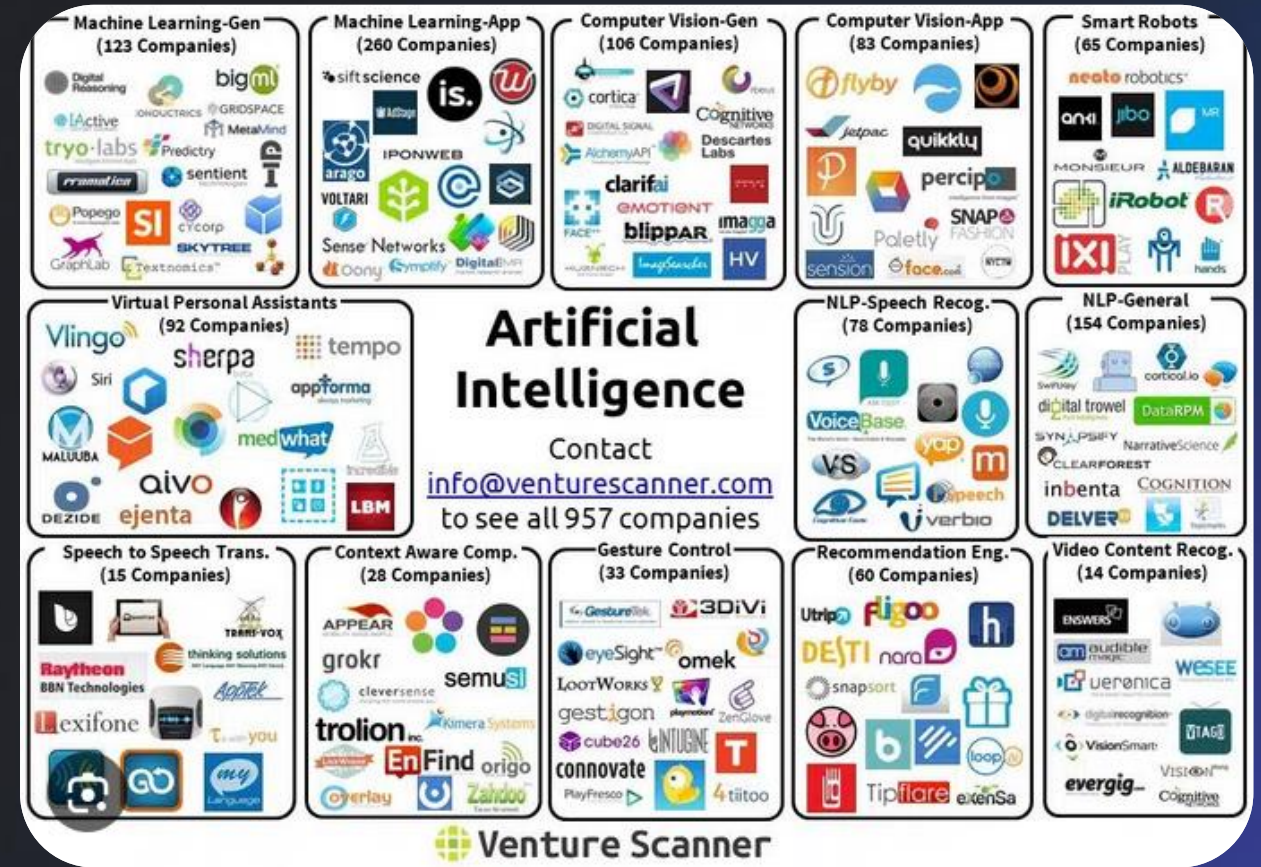
Неидеальность модели машинного обучения



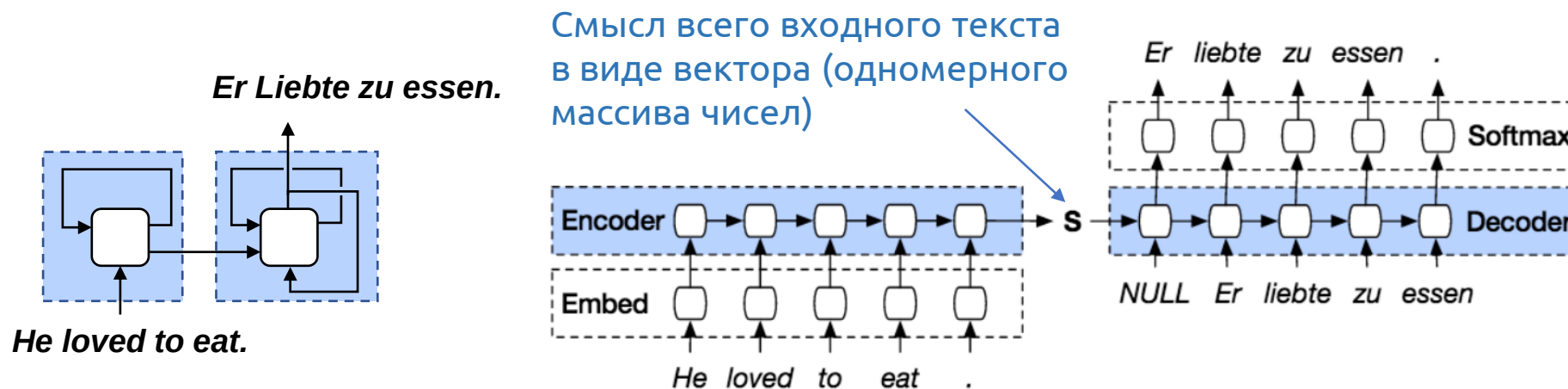
- Неполнота обучающей выборки для представления распределения
- Наличие в обучающих данных шума (ошибок)
- Несовершенство способа аппроксимации (вида модели) целевого распределения
- Сложность получения аппроксимации (обучения) целевого распределения

Что такое нейросети и почему именно они?

1. Биологическая аналогия
2. Теорема об универсальной аппроксимации – нейросеть может аппроксимировать любую нелинейную функцию f независимо от её размерности
3. Эффективные методы обучения
4. Высокий уровень развития технологий и много успешных бизнес-применений



Машинный перевод с помощью рекуррентных нейросетей



Машинный перевод был «философским камнем» ИИ с самого появления компьютеров.

Много попыток сделать алгоритм перевода провалились, прежде чем нейронные сети с обратными связями не показали успешную работу с естественным языком.

Обучающие данные – массив параллельных текстов на двух языках.

Трансформеры – нейросети с механизмом внимания

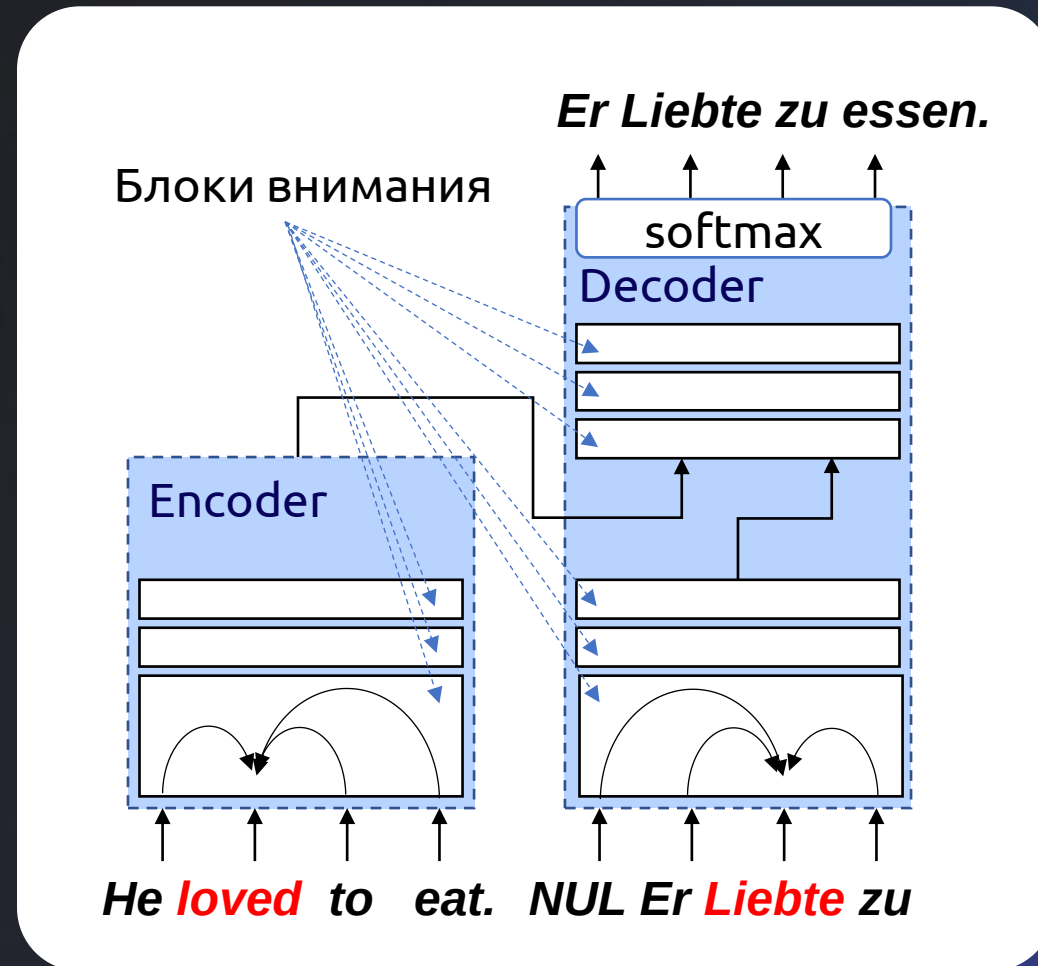
Нейросеть типа «Трансформер» использует ту же концепцию, но без обратных связей – не нужно формировать вектор S , заключающий в себе смысл всего предложения.

Вместо обратных связей используется прямое обращение к тексту с контекстно настраиваемыми весами, определяющими корреляции слов в контексте предложения.

Весов стало больше, но в результате удаётся построить эффективную языковую модель.

Количество обучаемых коэффициентов достигает десятков миллиардов.

Размер входного вектора (контекста) достигает десятков тысяч токенов.



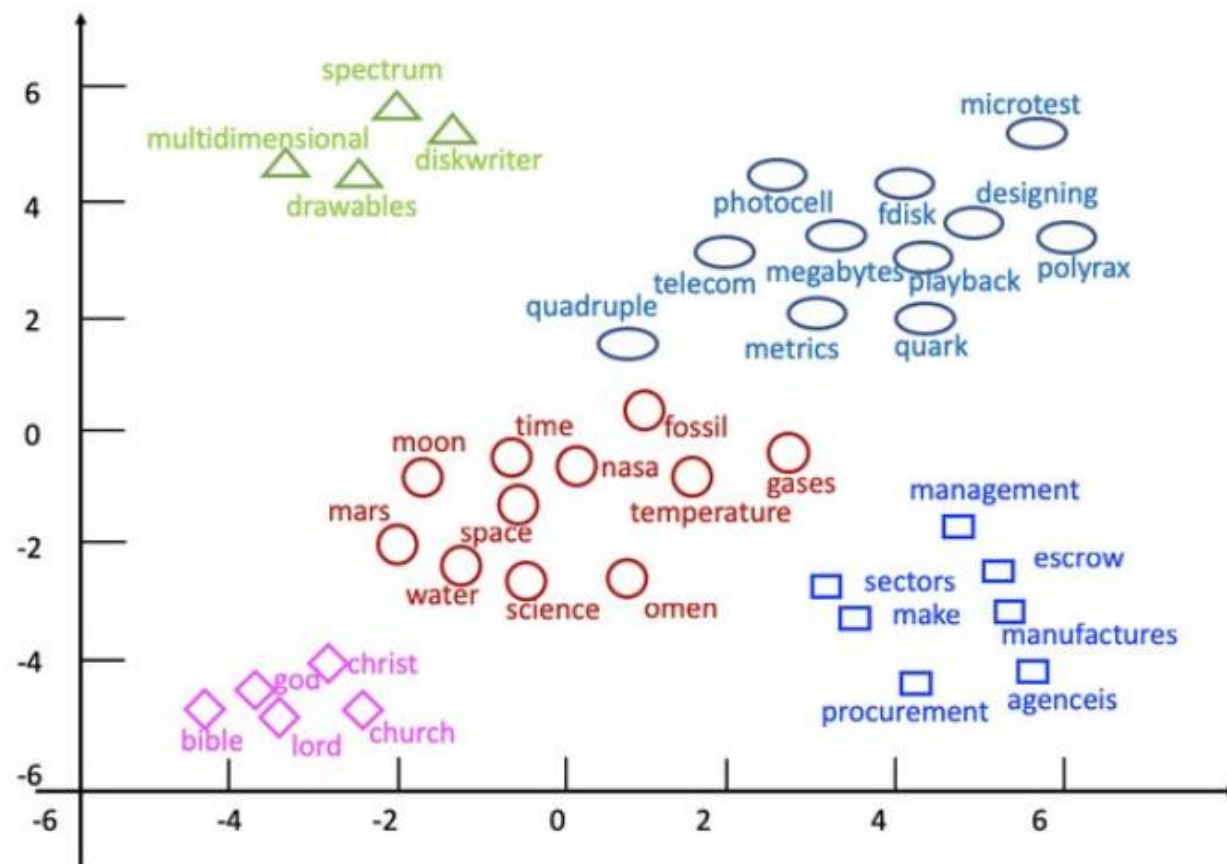
Представление слов – embedding

Эмбедингом (embedding) называют компактное представление фиксированного размера чего-то, сложно описываемого и переменной длины. В каком-то смысле это напоминает хэш, но, в отличие от значений хэш-функции, метрика (расстояние) в пространстве эмбедингов имеет прямую интерпретацию близости смысла.

Для преобразования слов в эмбединги используются специальные нейросетевые модели (например, Word2Vec).

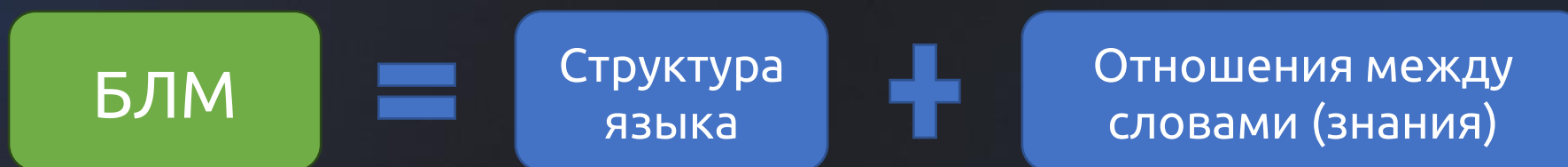
Векторное представление слов обычно имеет размерность $10^2 - 10^3$.

Условный пример для двумерного пространства:



Large Language Models (LLM) = Большие лингвистические (языковые) модели

БЛМ – это нейросетевые модели, умеющие работать с текстами на естественных языках и языках программирования, а также интегрироваться с иными формами представления информации (аудио, изображения, видео).



Encoder-only модели

(например, BERT - Bidirectional Encoder Representations from Transformers):

- Извлечение смысла (тональность текста)
- Выделение в тексте типов объектов (имена людей, даты, названия)

Decoder-only модели (например, GPT - Generative Pre-trained Transformer)

- Вопросно-ответные системы
- Написание текстов на заданную тему

Генеративные предобученные трансформеры (GPT)

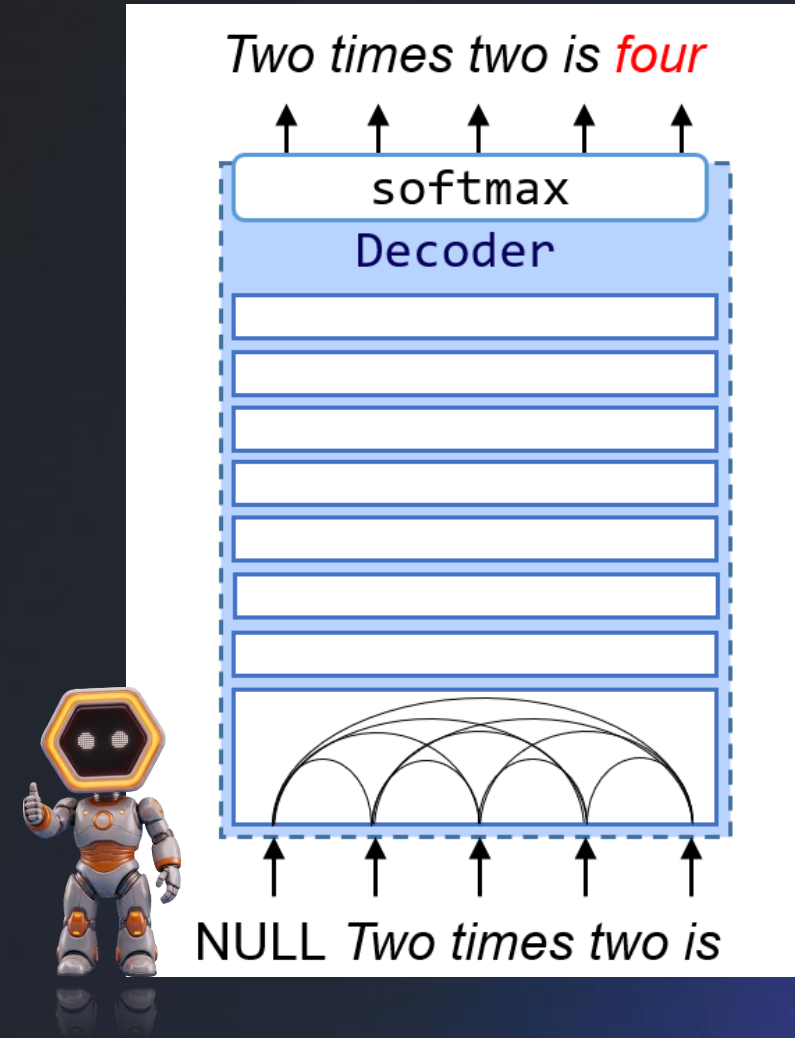
Тоже самое, что полный «Трансформер», но без энкодера.
Нейросеть учит продолжать предложения, например:

“Two times two is four” = f (“Two times two is”)

Обучение производится на множестве текстов, которые содержат фактически всё, написанное человечеством в разных формах: от литературного творчества и контента веб-сайтов до репозитариев исходного кода и сетевых форумов, причем, на всех доступных языках.

Оказалось, что обученный GPT умеет:

- Общаться, как человек, и переводить с языка на язык
- Обобщать информацию и строить вывод $A \rightarrow C$, если знает, что $A \rightarrow B$ и $B \rightarrow C$
- Писать программы по словесному описанию



Технология запроса к БЛМ (1/3)

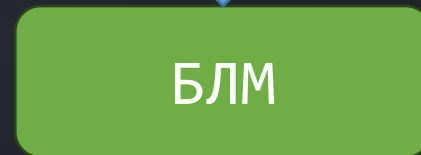
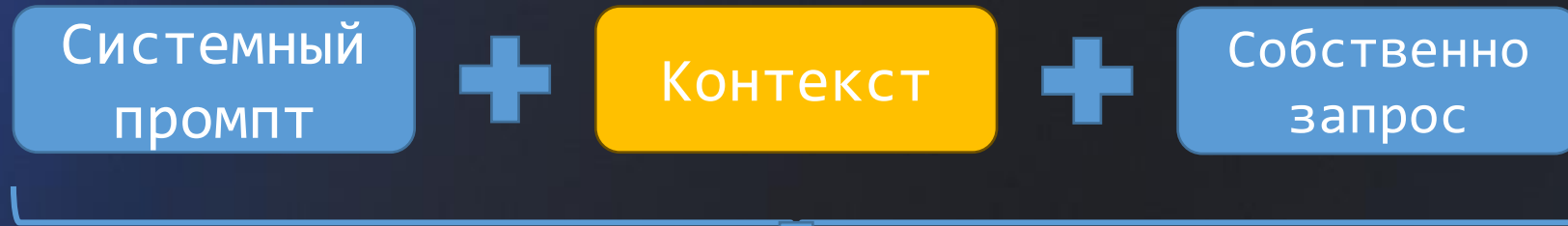


Системный промпт часто содержит указания по формированию ответа, включая накладываемые ограничения, стиль и форму. Для публичных БЛМ и интеллектуальных агентов на их основе системный промпт часто является know-how.

Обход ограничений и экстракция системного промпта – распространенные виды хакинга БЛМ.

Все части запроса задаются на естественном языке!

Технология запроса к БЛМ (2/3)

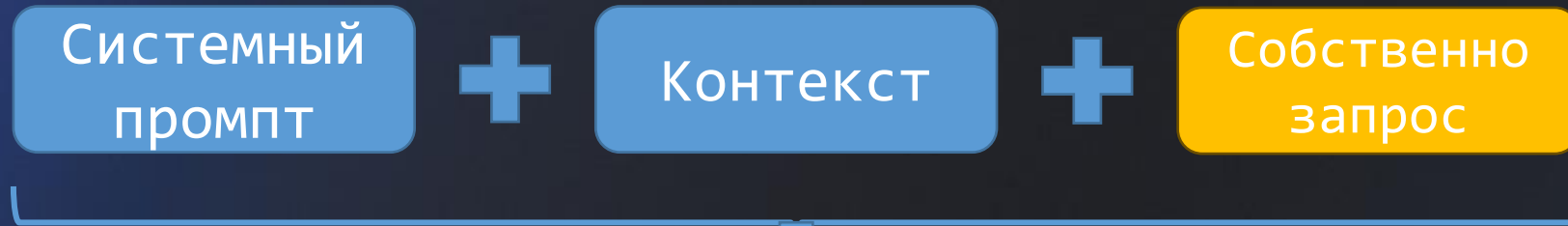


Контекст представляет собой текст, содержащий, в числе прочего, информацию, которая нужна для ответа на запрос.

Без контекста модель будет использовать знания из тех текстов, на которых её учили, что может быть нерелевантно.

Размер контекста может быть достаточно большим, поэтому важно, чтобы размер контекстного окна вмещал весь контекст.

Технология запроса к БЛМ (3/3)



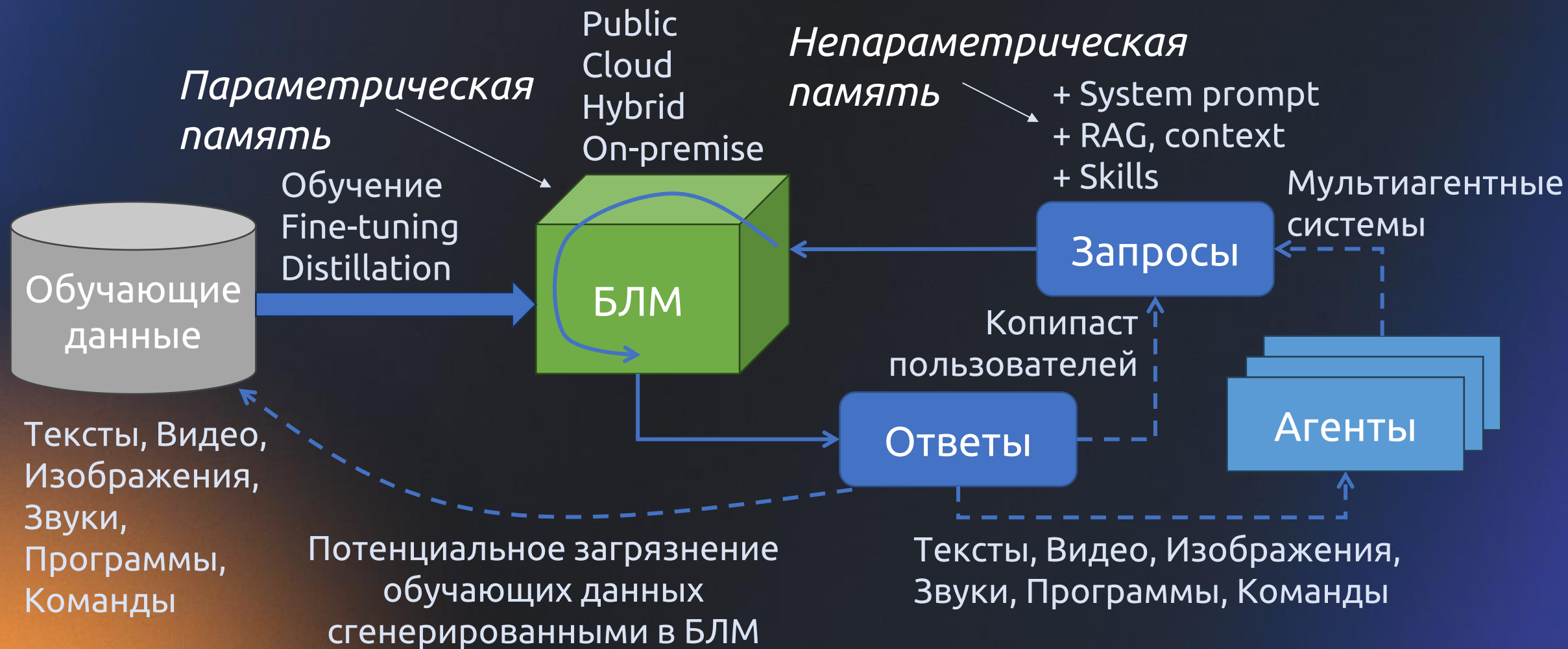
БЛМ

Ответ

Собственно запрос (user prompt) представляет собой непосредственное обращение пользователя. В него могут быть включены дополнительные инструкции по стилю, объёму и иным требованиям.

Умение составлять запросы к БЛМ стало новой профессией – prompt engineering.

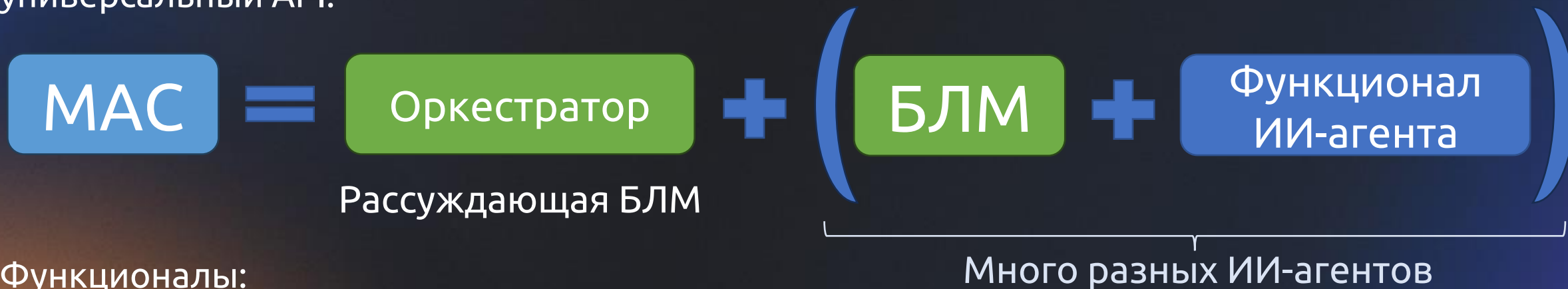
Жизненный цикл БЛМ



Мультиагентные системы (МАС)

Современные системы ИИ устроены как мультиагентные системы, реализующие сложный итоговый функционал из множества различных специализированных агентов.

Все агенты используют человекочитаемый язык для обмена данными и инструкциями, как универсальный API.



Функционалы:

- Скилы – обширные системные промпты, специализирующие ответ
- RAG по базе знаний или через поисковый запрос в Интернет
- Мультимодальные входные или выходные данные
- Выполнение команд или запуск исполняемого кода

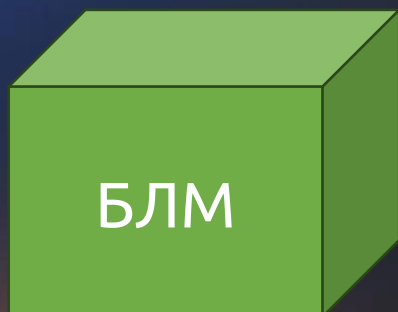
Влияние технологии БЛМ и МАС на информационную безопасность

Типовые способы применения БЛМ и МАС на их основе, а также возникающие эффекты, влияющие на ИБ:

- **прикладное применение** – утечка и потеря данных, создание программного кода с уязвимостями, получение неверной информации и недопустимого контента
- **использование для нападения** – социальная инженерия, неявная доставка вредоносного кода, поиск уязвимостей и синтез эксплойтов
- **использование для защиты** – поиск уязвимостей и их устранение, генерация пентестов, аннотирование системных событий и суммаризация инцидентов

Что опасного в БЛМ?

Public
Cloud
Hybrid
On-premise



Основной механизм реализации БЛМ – глубокая нейронная сеть – не изучен в достаточной мере, чтобы гарантировать какие-то свойства безопасности и даже эффективности.

Проблемы всех глубоких нейронных сетей:

- Состязательные атаки (свёрточные нейросети)
- Галлюцинации (генеративные нейросети)

Самая большая проблема применения БЛМ – это правдоподобие получаемых результатов:

- Отсутствие сомнений в ответе
- Приятная и легко читаемая форма представления

Ещё проблемы:

- Невозможность разделить «инструкции» и «данные» в запросе
- Все парадоксы человеческого языка превращаются в проблему неоднозначности понимания запросов и ответов
- БЛМ в неявной форме содержит информацию, которая может быть сочтена как закрытая или нежелательная

Получение неверной информации

Ответ на запрос к БЛМ может быть неправильным, что называется галлюцинацией, причиной которой могут быть:

- Ошибки в обучающих данных (Garbage In, Garbage Out);
- Плохо обученная модель – требуется Supervised Fine-Tuning и Reinforcement Learning from Human Feedback;
- Ответ на запрос отсутствует в обучающих данных;
- Запрос содержит нерешаемую/вычислительно сложную задачу;
- БЛМ не вполне уверена в правильности ответа, но выдаёт его;
- Сдвиг распределения в запросе относительно распределения обучающих данных, как у любых других ML-моделей с учителем.

1. Kalai A. T. et al. Why language models hallucinate //arXiv preprint arXiv:2509.04664. – 2025.
2. <https://habr.com/ru/articles/877360/>

Галлюцинации БЛМ



Слова на **выходе БЛМ** выбираются, как наиболее вероятные исходя из входного контекста и уже сформированного начала результирующего текста.

Наиболее вероятное – не значит имеющее большúю вероятность!

Из-за того, что БЛМ часто используются в режиме рассуждений (reasoning), ошибочное слово в ответе расширяет своё влияние, формируя ложные рассуждения.

Галлюцинации принято разделять на два основных типа:

- галлюцинации фактичности – неверный факт
- галлюцинации достоверности – неверный вывод

Что опасного в исходных данных?

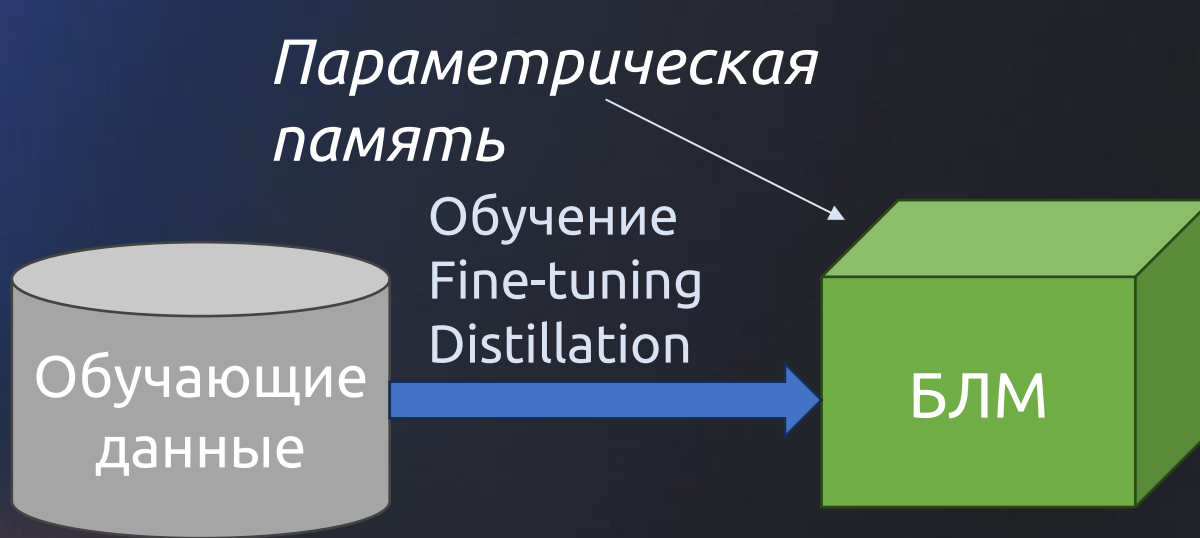


Во всех видах машинного обучения исходные данные определяют итоговый результат, выдаваемый обученной моделью. Обычно данные проверяют, исключают выбросы, недостоверные и заведомо неполные записи.

В случае с БЛМ проблема в объёме данных для проверки:

- Типичный размер обучающих данных для современной модели – терабайты токенов
- Непонятно, что считать выбросом или недостоверными данными?
- Возможны атаки типа «отравление данных»
- Количество информации, порождаемой БЛМ и публикуемой, как произведенной человеком, становится всё больше и она будет включаться в обучающие выборки новых БЛМ

Что опасного в обучении?

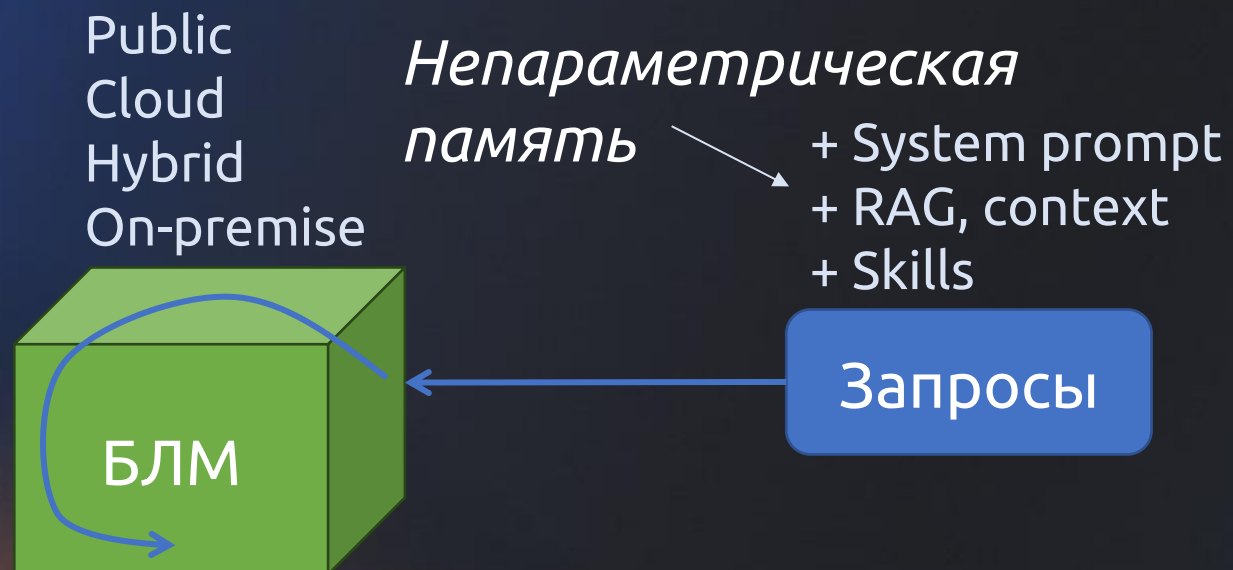


Обучающий алгоритм и программное обеспечение, его реализующее, являются базовой операцией, формирующей доверие к модели.

Обычно программное обеспечение может быть верифицировано на отсутствие недокументированных возможностей, а также могут быть приняты меры для исключения влияния на ПО в процессе его работы.

- Программные библиотеки, используемые для обучения БЛМ, достаточно сложные даже на уровне исходных текстов – они состоят из модулей на Python, C, C++ и других языках
- Программы для обучения БЛМ используют проприетарные средства программирования GPGPU, недоступные в исходных текстах
- БЛМ обучаются с помощью вычислительных кластеров с высокотехнологичными аппаратными ускорителями GPGPU, выпускаемыми только за рубежом

Что опасного в запросах?

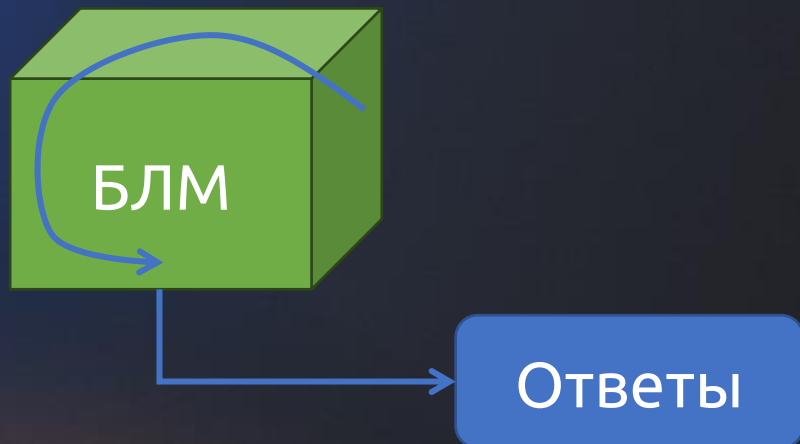


Запрос к БЛМ представляет собой текст на естественном языке, ответ на который пользователь хочет увидеть в наиболее естественной форме: тексте, картинке или структурированном документе.

- Запрос в самой непосредственной форме содержит все данные, которые пользователь хочет сообщить БЛМ для получения максимально точного ответа – **утечка!**
- Запрос в каком-то смысле наводит БЛМ на ожидаемый пользователем ответ – **верить ли?**
- Запрос адресован к БЛМ и может иметь целью получение не предусмотренного владельцем БЛМ ответа – **получение запрещенной или нежелательной информации**

Что опасного в ответах?

Public, Cloud, Hybrid, On-premise



Ответ БЛМ представляет собой реакцию на запрос, которую следует рассматривать, как наиболее вероятная с точки зрения массива обучающих данных последовательность слов с учетом контекста – текста, предшествующего запросу.

Контекст (непараметрическая память) может формироваться из системного промпта, скила, фрагментов уточняющих текстов (RAG), а также текста предшествующего диалога.

- Ответ не несёт семантической нагрузки, хотя и выглядит вполне разумно
- Ответ в дальнейшем может попасть как в параметрическую, так и непараметрическую память, что размывает уверенность в достоверности исходной информации для БЛМ
- Ответы БЛМ могут использоваться для проведения атак методом социального инжиниринга, а также широко применяются в т.н. «когнитивных войнах»
- Ответ часто «копипастится» человеком, получая статус ответственного заявления без должной проверки достоверности

Что опасного в агентах?



Агенты являются способом взаимодействия БЛМ с окружающим миром и, в отличие от собственно БЛМ, получают реальные полномочия, будь то поиск информации в Интернете, запуск программ и выполнение управляющих действий в информационных системах.

Агент должен иметь разрешения доступа, которые часто заимствуются у человека, но это **не человек с ответственностью!**

- Галлюцинации могут серьезно повлиять на ожидаемые эффекты действия агентов
- Мощь мультиагентных систем – в отсутствии человека в контуре управления и это основная опасность – **SkyNet уже почти готов!**
- Агенты начисто лишены ответственности в человеческом понимании этого слова и совершенно неразборчивы в способах достижения цели

Типология систем ИИ

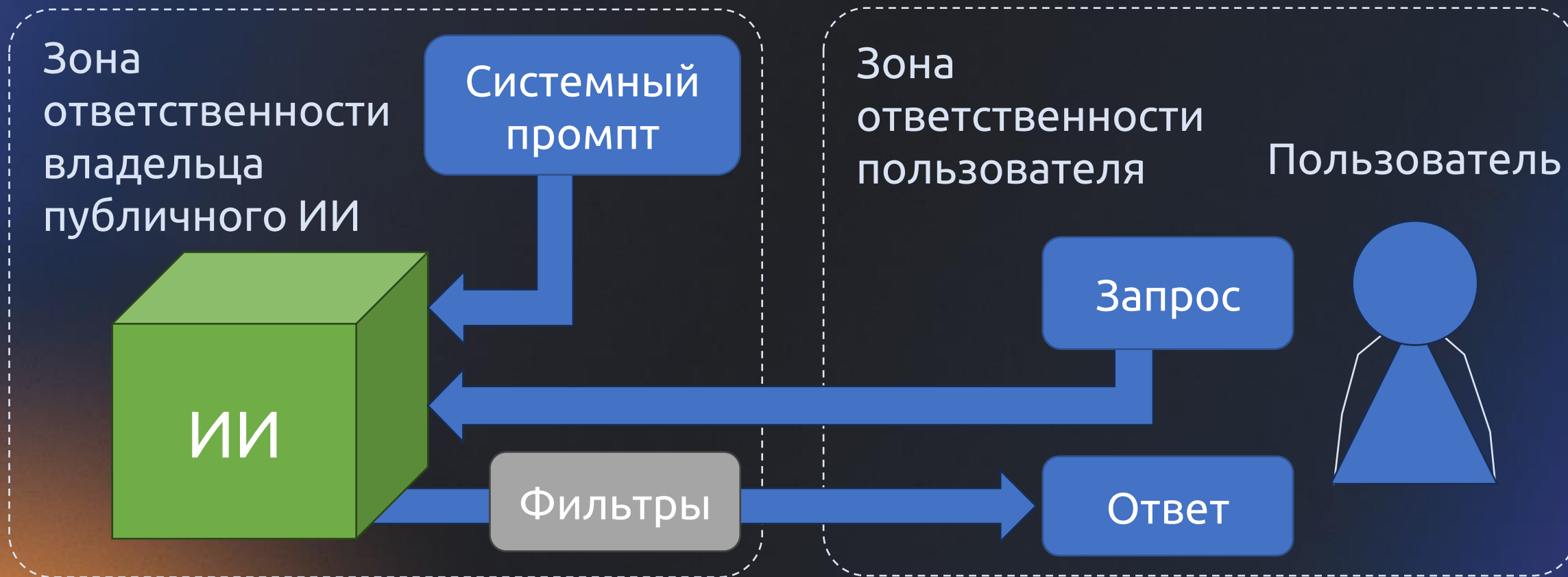
Мультиагентная система с БЛМ – это то, что называется сейчас системой искусственного интеллекта (ИИ, AI)

Фактически, в ряде работ и рекомендации по безопасности ИИ (OWASP и др.) перемешаны рекомендации для разных типов систем ИИ и разных рисков

Типы систем ИИ:

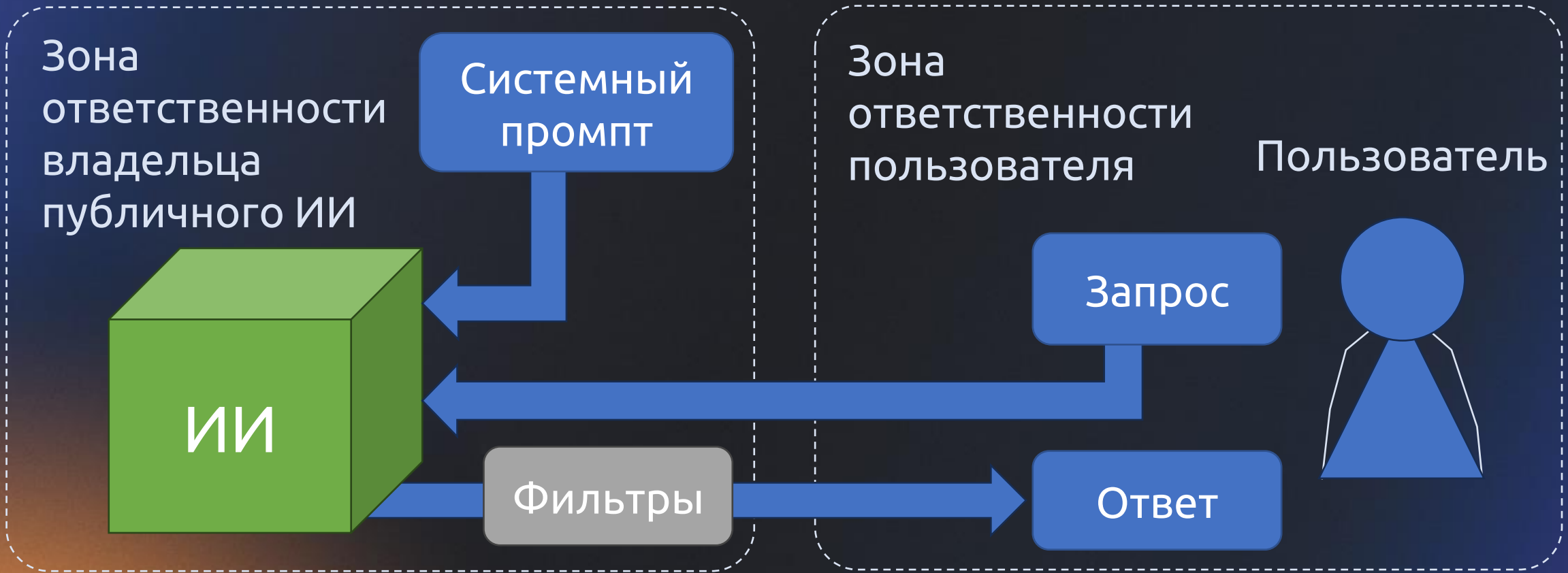
- ❖ Публичный ИИ (Public AI) – любой пользователь может получить доступ к БЛМ/МАС, как услугу (платно или бесплатно)
- ❖ Облачный ИИ (Cloud AI) – по договору SLA предоставляются гарантии доступности и конфиденциальности, но технически нет различий с Public AI
- ❖ Гибридный ИИ (Hybrid AI) – часть МАС реализуется на стороне пользователя, но сама БЛМ находится в облаке
- ❖ Приватный ИИ (On-premise AI) – все компоненты системы ИИ реализуются внутри корпоративной сети

Публичный ИИ: риски владельца



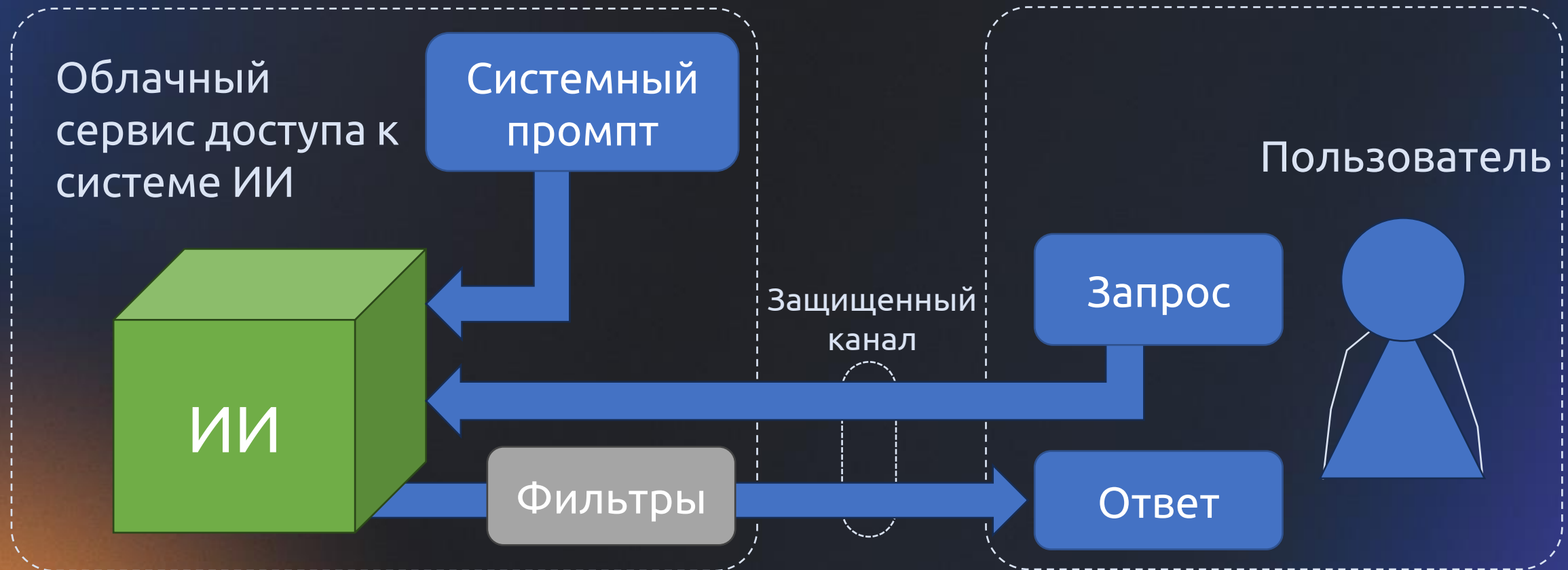
Основные риски для владельца ИИ – репутационный и юридический, поскольку ответ на запрос не должен нарушать законодательные и социокультурные запреты на публикацию некоторых видов информации.

Публичный ИИ: риски пользователя Ampire 360°



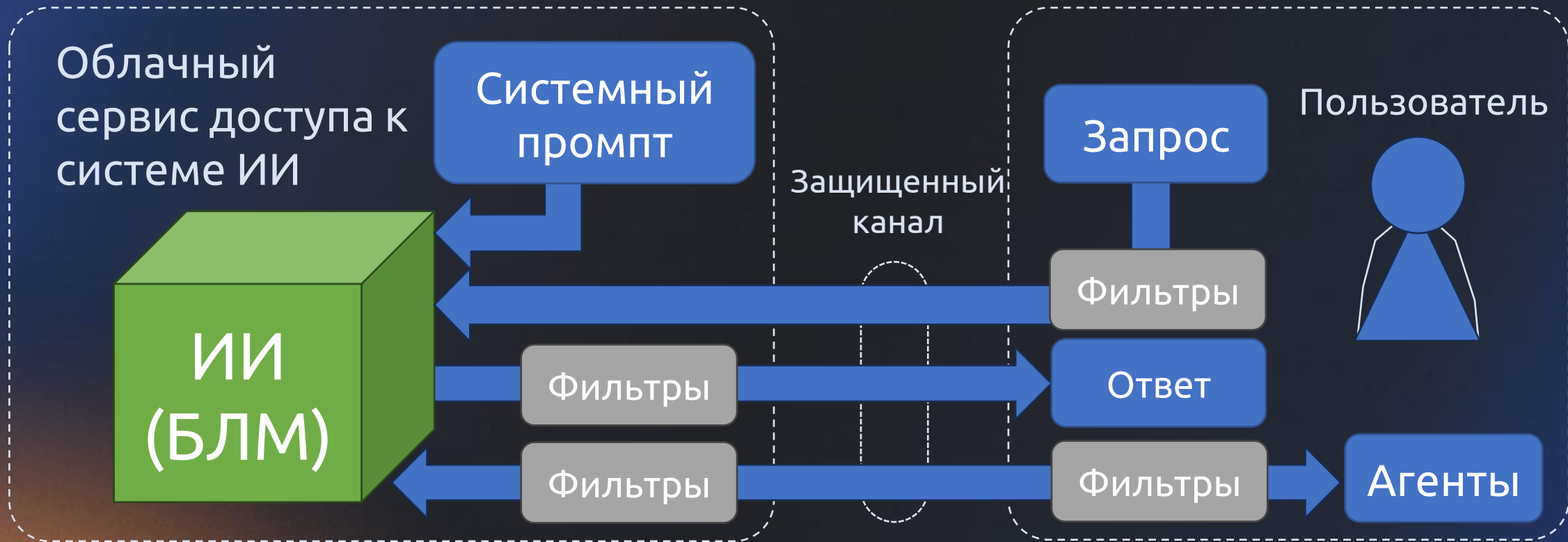
Пользователь может находиться в иной юрисдикции, чем БЛМ, и тогда к нему применимы другие законы. Сам запрос может содержать конфиденциальную информацию, которая попадает к владельцу БЛМ.

Облачный ИИ



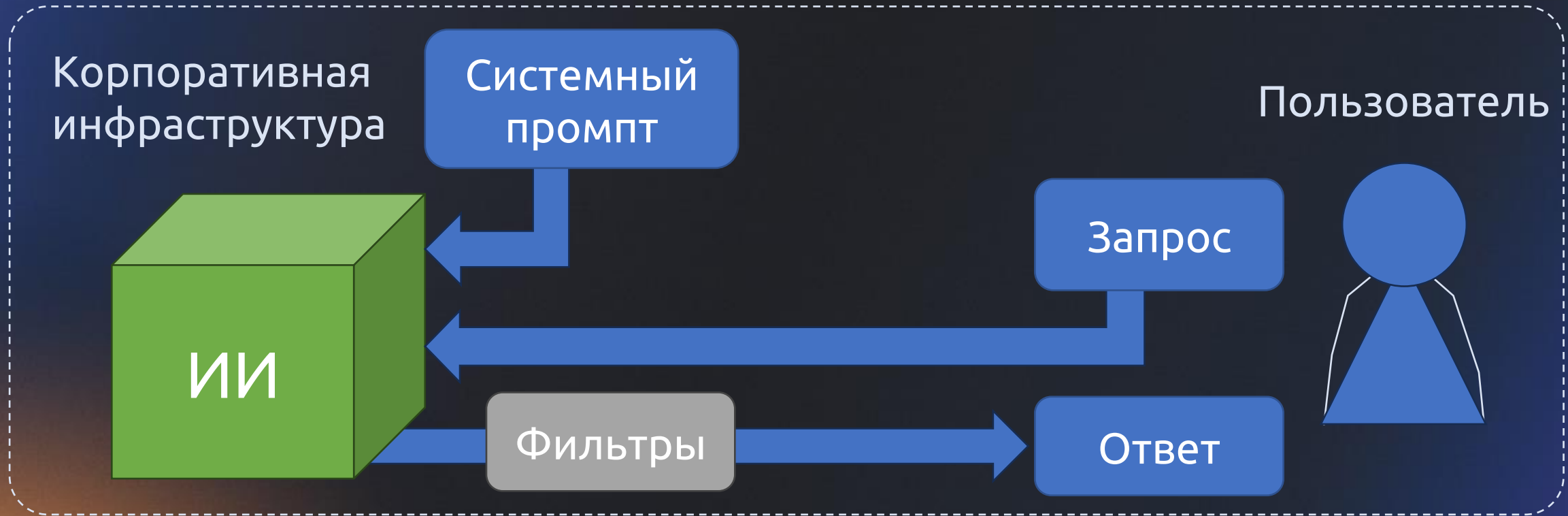
Облачный ИИ предоставляют гарантии конфиденциальности, ответы на запросы не являются публично доступными, однако сама по себе облачная инфраструктура несёт риски компрометации данных.

Гибридный ИИ



Часто возможности публичного/облачного ИИ используются для построения частных мультиагентных систем, в которых облачная БЛМ выступает оркестратором действий агентов, работающих на стороне пользователя. Это несет дополнительные риски, поскольку агенты работают внутри доверенного контура, но выдают запросы и получают ответы и инструкции извне.

Приватный ИИ



Наивысшие гарантии конфиденциальности и доверия могут обеспечить только приватный корпоративный ИИ. Такая система ИИ требует дорогого on-premise оборудования, а также явной реализации мультиагентной системы. Также необходимо обеспечить аутентификацию и контроль использования. Однако такие решения обычно отстают от публичного/облачного ИИ по эффективности.

Выбор системы ИИ для корпоративного применения

- Применение БЛМ без реализации МАС малоэффективно и чревато галлюцинациями.
- Публичный и облачный ИИ не подходят в случае применения для обработки конфиденциальных данных, а также при трансграничном использовании. Часто веса таких БЛМ и логика работы МАС являются коммерческим секретом, что не позволяет доверять им в полной мере.
- Гибридный ИИ может использоваться при наличии хорошо настроенных фильтров, предотвращающих утечку данных и опасные инструкции агентам
- Приватный ИИ обеспечивает гарантированную доступность и позволяет реализовать практически любой набор технологий защиты, однако для снижения эффекта остающихся уязвимостей и недостатков целесообразно придерживаться некоторого концептуального набора принципов, учитывающих известные особенности БЛМ.

Спасибо за внимание!

Владимир Елисеев

Руководитель Центра научных исследований
и перспективных разработок, к.т.н.

АО «ИнфоТеКС»

Email: eliseevvl@infotecs.ru



Ampire 360°