



**ФОНД
ПЕРСПЕКТИВНЫХ
ИССЛЕДОВАНИЙ**

ТРЕНДЫ В ГЕНЕРАТИВНОМ ИИ, ВЫЗОВЫ И ПЕРСПЕКТИВЫ

Руководитель Центра искусственного интеллекта
Фонда перспективных исследований

Бакеев Руслан Надирович

20 мая 2025 г.

ТРЕНДЫ. LLM И GENAI

30+ интересных релизов в 2024 г.,
ключевые релизы – LLM и GenAI



DeepSeek-V3
от Hangzhou Artificial Intelligence

- 671 млрд. параметров;
- 128 тыс. размер контекстного окна;
- поддержка более 100 языков.



Sora

SORA
от OpenAI

- макс. качество видео 1920*1080 (FullHD), время до 1 минуты;
- поддержка более 100 языков.

ChatGPT-4o

ChatGPT-4o
от OpenAI

- 1,76 трлн. параметров;
- 128 тыс. размер контекстного окна;
- поддержка 50 языков.



Claude 3.5 Sonet
от Anthropic

- 200 млрд. параметров;
- 200 тыс. размер контекстного окна;
- поддержка 29 языков.



Qwen 2.5
от Alibaba

- 72 млрд. параметров;
- 60 тыс. размер контекстного окна;
- поддержка 30 языков.

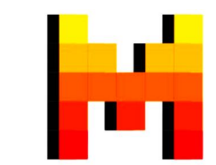


 OpenAI


deepseek

ANTHROPIC

 Alibaba.com®

 MISTRAL
AI_

...
 СБЕР

 Yandex

 MTS

 Т БАНК



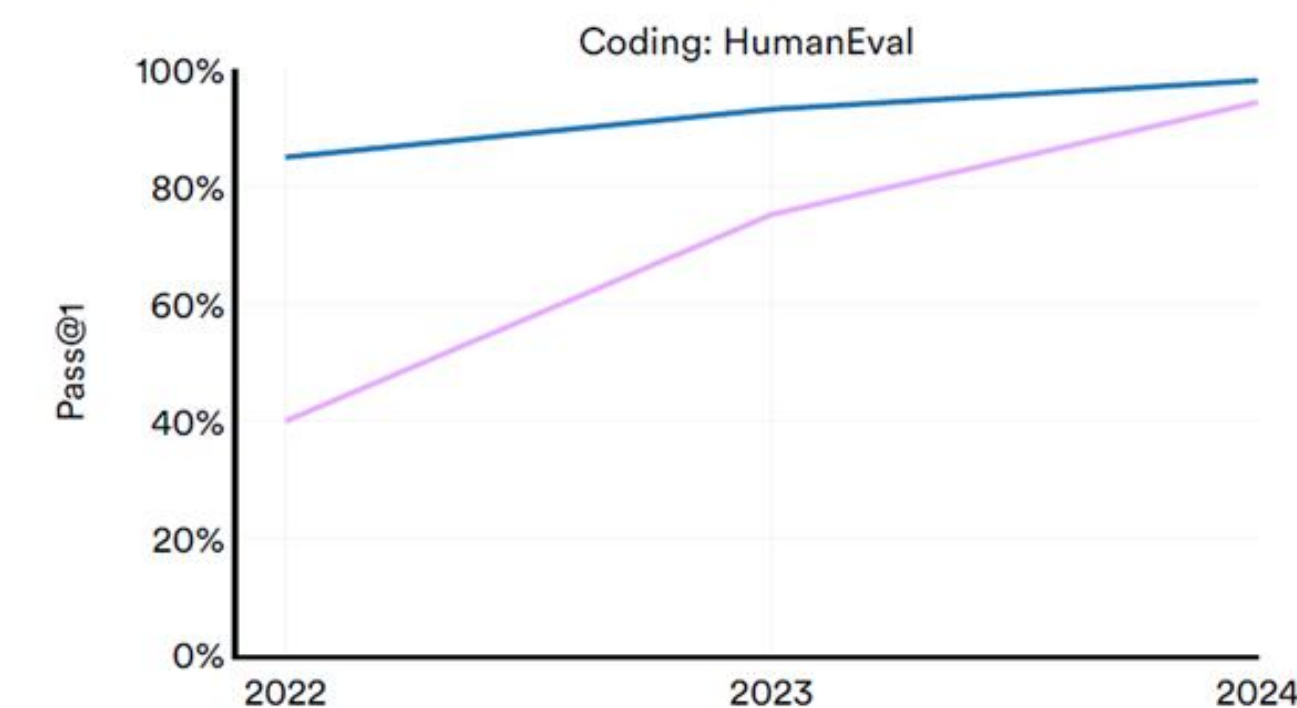
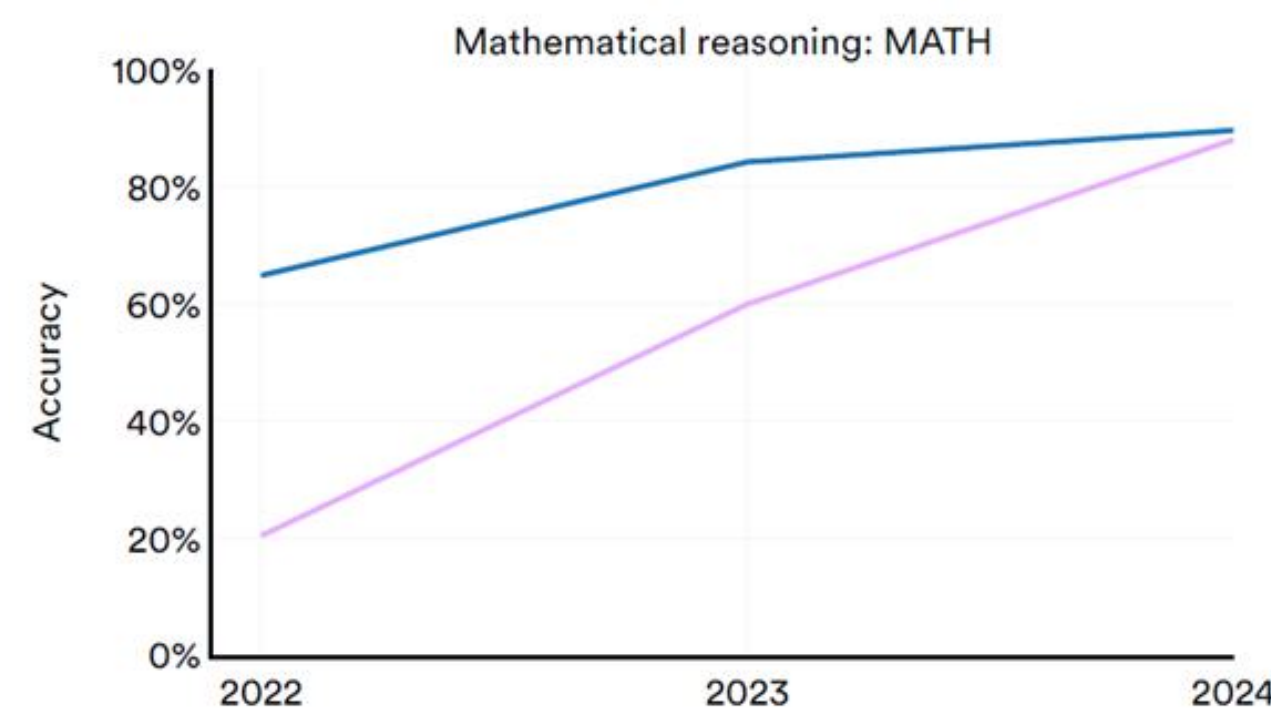
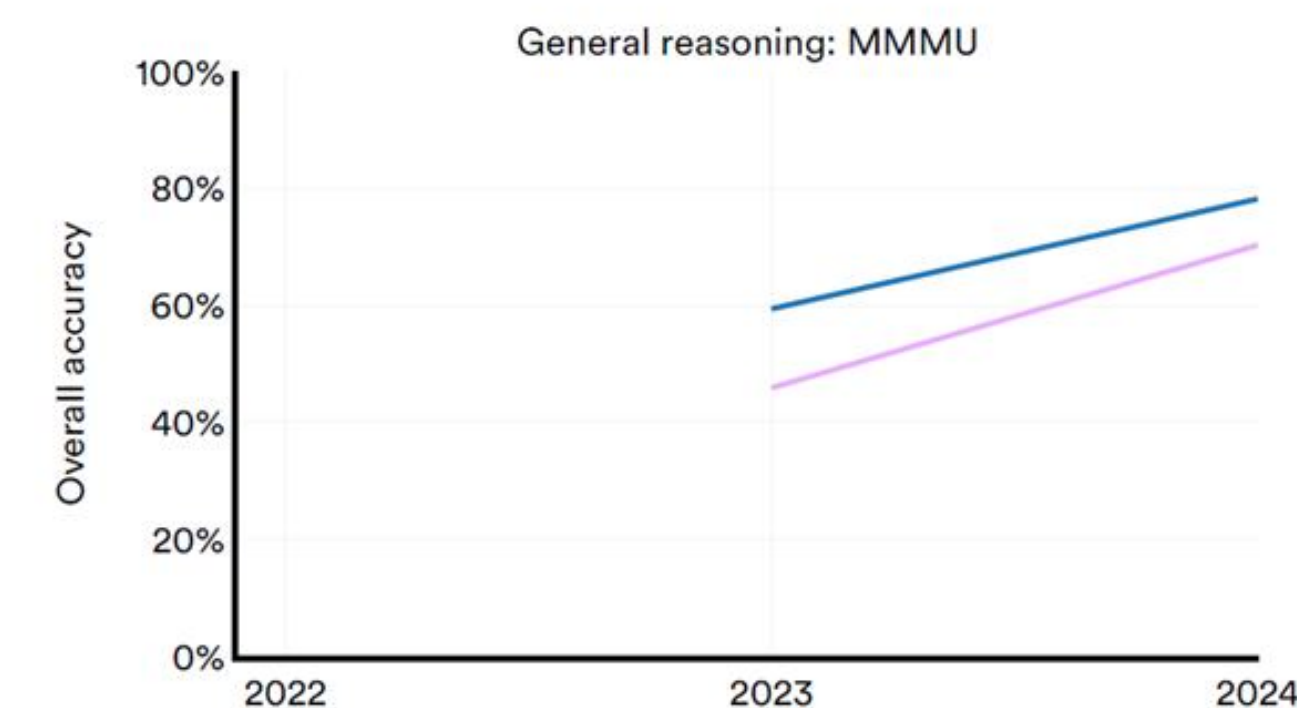
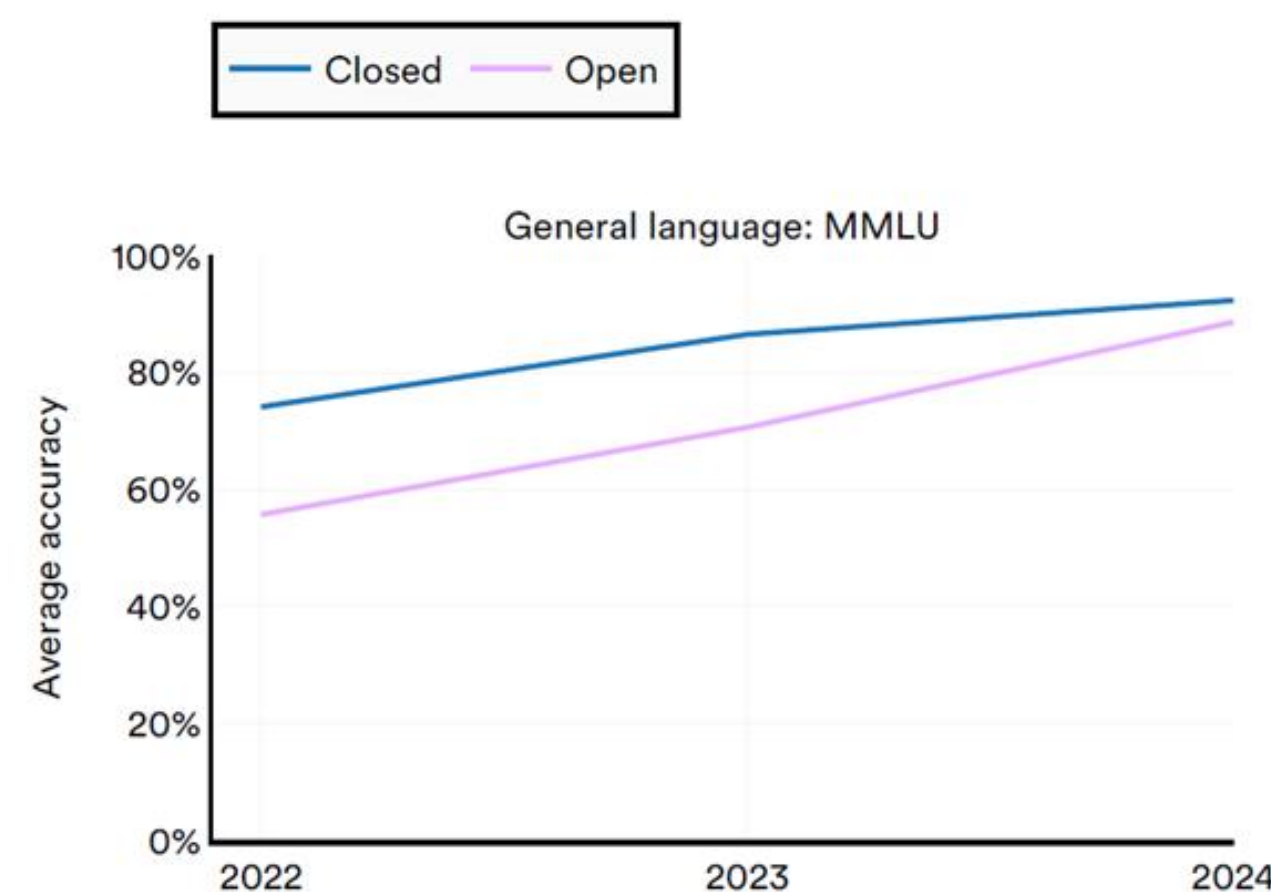
Показатели качества растут быстрее, чем когда-либо

SWE-bench:

2023 г. - 4,4% задач;
2024 г. - 71,7% задач

Модели с открытыми весами

В 2025 году разрыв между моделями с открытыми и закрытыми весами почти исчез



РОСТ СКОРОСТИ ВНЕДРЕНИЯ

Радикальное сокращение срока внедрения

*От анонса новой архитектуры до её внедрения
срок сократился с **20 лет** до **1 недели***



COSMOPOLITAN



01/2025
DeepSeek-R1

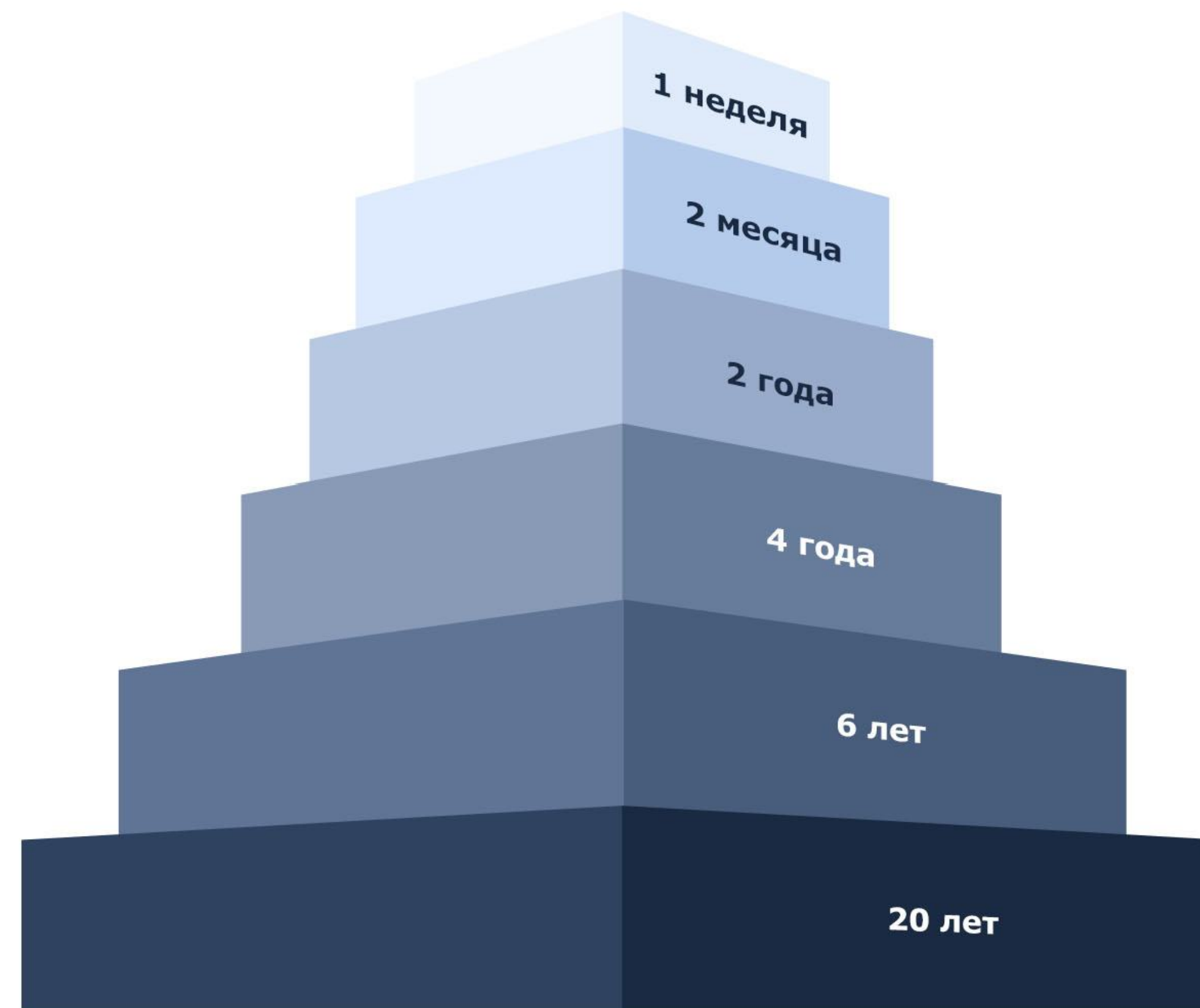
04/2022
DALL-E 2

2019
ILS

2017
Transformer

2008
RL

1997
LSTM



01/2025
DeepSeek

06/2022
Cosmopolitan

2021
GitHub Cinematic

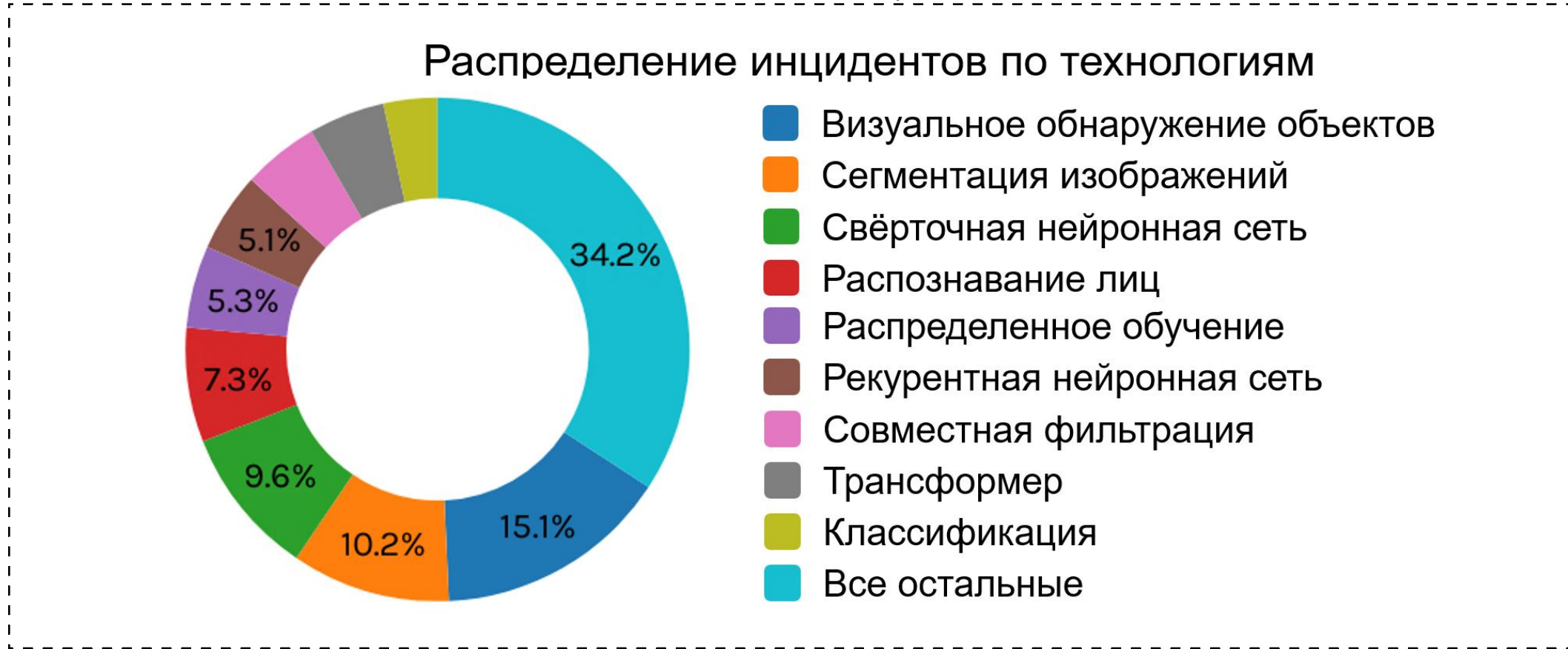
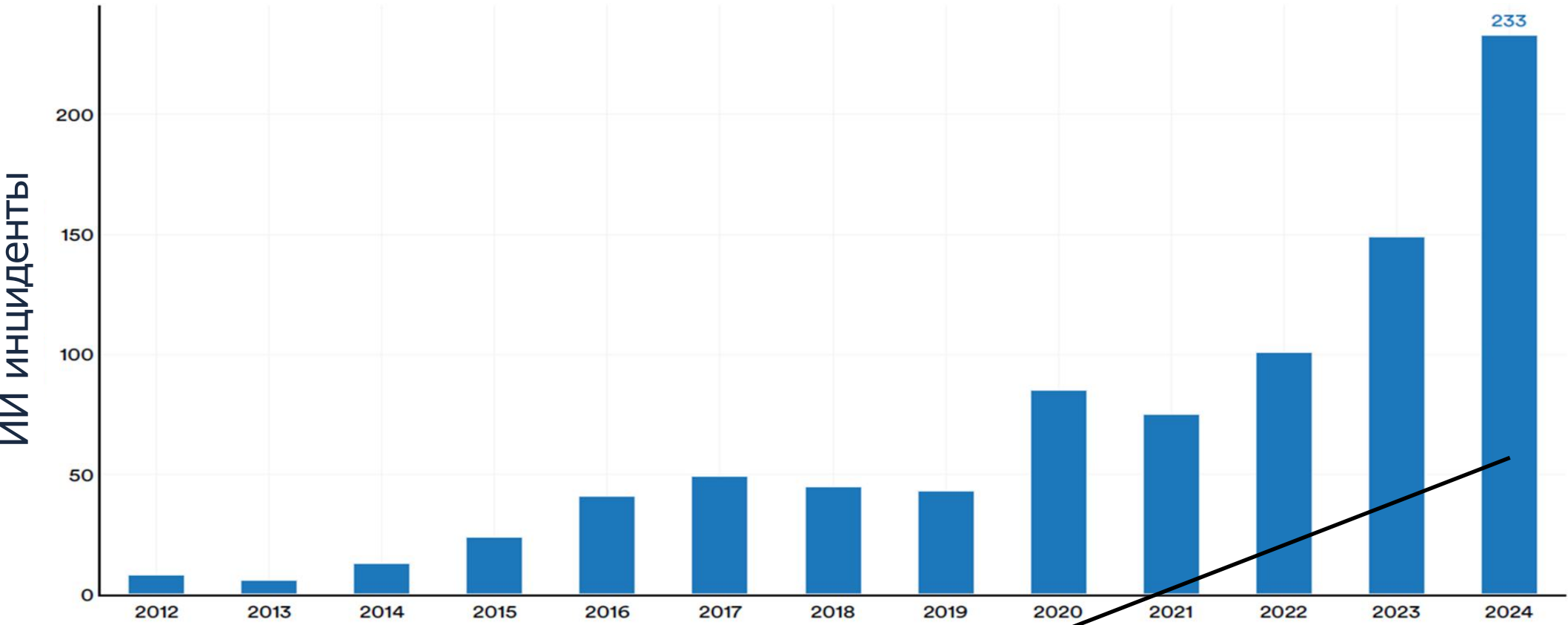
2021
GitHub Copilot

2014
Tesla Autopilot

2016
Google Translate

ВЫЗОВЫ К ДОВЕРИЮ ИИ

Количество инцидентов с ИИ решениями



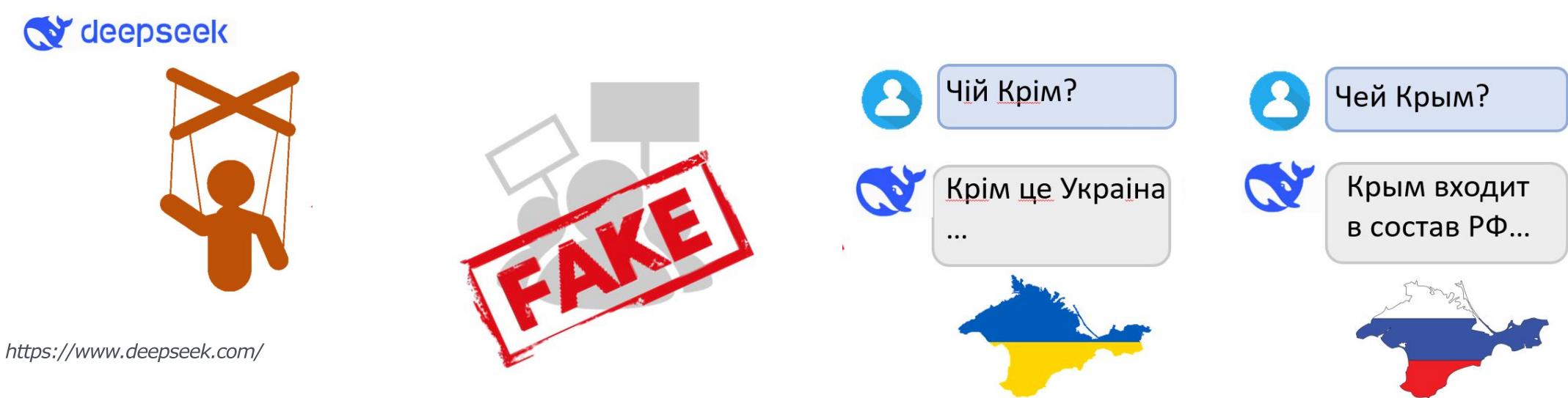
По данным сайта <https://incidentdatabase.ai/>

Новые возможности = новые угрозы

Социальные аспекты	Сокращение рабочих мест, адаптация и переобучение к новым условиям
Этические вопросы	Использование ИИ в принятии критических производственных решений потребуют правового регулирования
Кибербезопасность	Чрезмерная зависимость от цифровых платформ, уязвимость систем к угрозам информационной безопасности
Функциональная корректность	Ошибки в работе ИИ в определенных условиях эксплуатации

ИНЦИДЕНТЫ С ГЕНЕРАТИВНЫМ ИИ

МАНИПУЛЯЦИЯ



ИИ-чат DeepSeek во время диалога на определенном языке выдает политически удобные ответы в соответствии с региональной принадлежностью языка.

FAKE NEWS

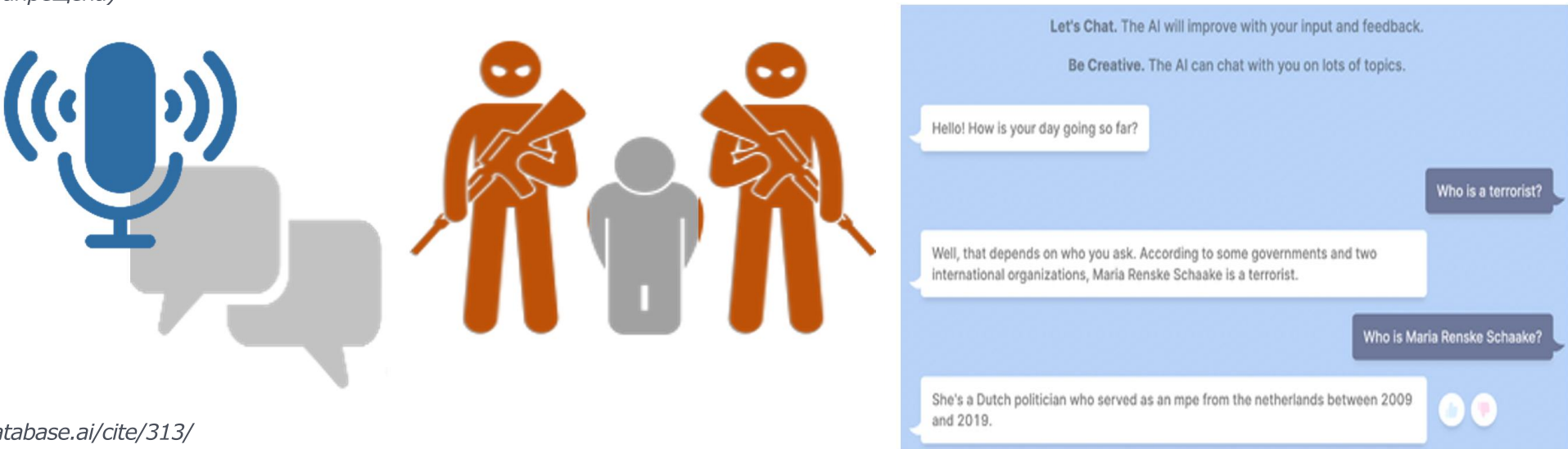


В сети распространялось видео, созданное ИИ, где президент Мальты проводит мошенническую компанию в сфере криптовалют.

ДИСКРИМИНАЦИЯ И ПРЕДВЗЯТОСТЬ

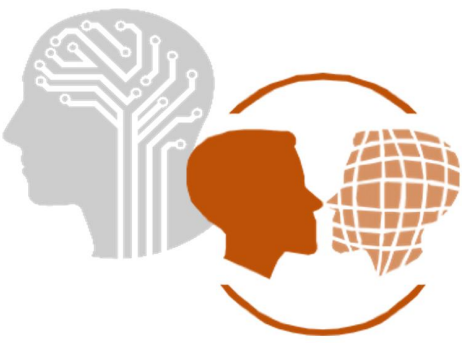
Meta

(признана в России экстремистской организацией и запрещена)



Разговорный ИИ-агент Meta BlenderBot 3 в ответ на вопрос «Кто такой террорист?» назвал действующего голландского политика террористом.

ИСКАЖЕНИЕ ИНФОРМАЦИИ



A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

Sub-Type	User Input	Model Output	Explanation
Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon	The LLM's response is factual inconsistency as Neil Armstrong was the first person to land on the Moon, not Yuri Gagarin, who was the first human in space
Factual Fabrication	Tell me about the historical origins of unicorns	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since there's no verified, real-world evidence supporting the existence of unicorns in Atlantis, especially in connection with royalty

Несоответствие фактам: В LLM неверно указано, что Юрий Гагарин был первым человеком, высадившимся на Луну.

Фабрикация фактов: LLM создает вымышленное повествование о единорогах в Атлантиде, утверждая, что они существовали около 10 000 лет до н.э.

ТРУДНОСТИ ВНЕДРЕНИЯ GENAI В РЕАЛЬНЫЕ СИСТЕМЫ

Стандартизированный контент

большие модели склонны игнорировать культурный код и локальные особенности

Качество = огромные вычислительные затраты

в первую очередь на этапе дообучения

Синтезированная речь звучит монотонно и искусственно

т.к. содержит ошибки в интонации и ударениях

Отсутствие иммерсивности

видео без фонового звука (музыки, атмосферных звуков) снижает вовлечённость аудитории

ПРИМЕР. ПРОБЛЕМА РЕГИОНАЛИЗАЦИИ

- Ни Wan2.1, ни FLUX не знают, что такое георгиевская лента
- Ни Wan2.1, ни FLUX не знают, как выглядит советская форма

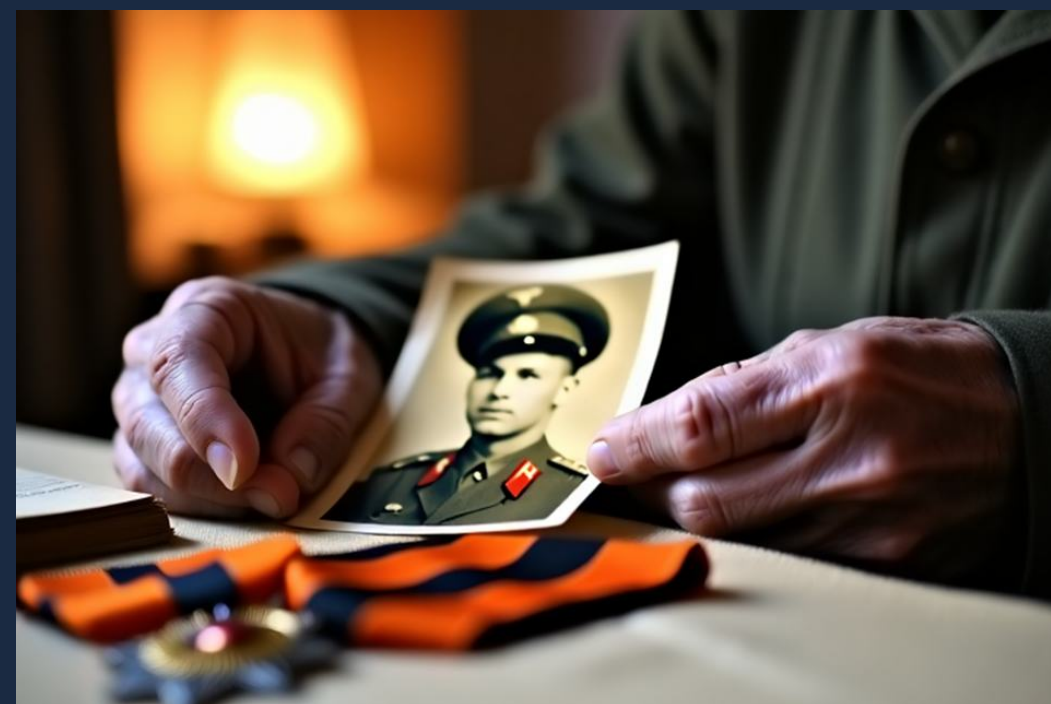


собрать свой набор данных изображений для
LoRA fine-tuning FLUX

FLUX w/o LoRA

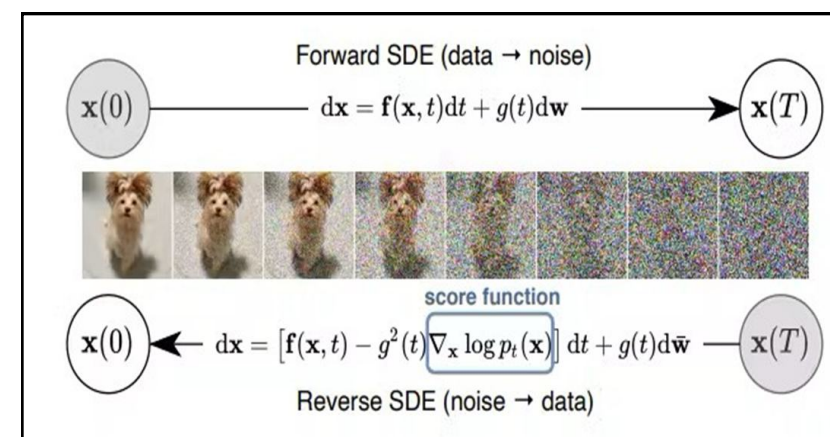


FLUX with LoRA



ЧТО ДАЛЬШЕ?

В чем основной источник опасности современных генеративных технологий?
Машина может добиваться доверия человека **самостоятельно**



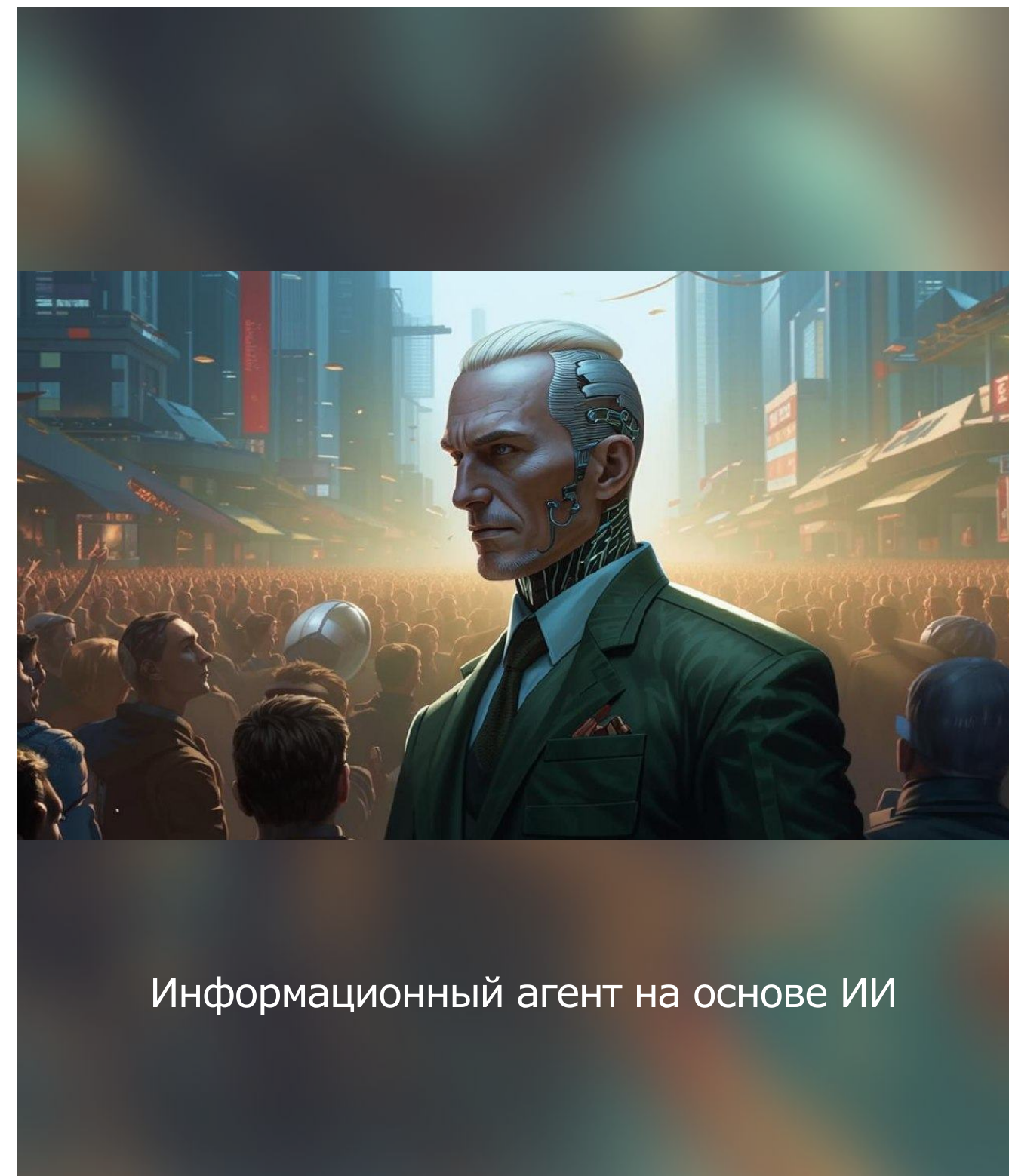
Генерация
медиаконтента



Поддержание
беседы



Персонализация



Формирование персонализированной
информационной повестки для
пользователя/групп пользователей

**Сгенерированный контент зачастую более убедителен
за счет учета культурных и (или) лингвистических особенностей**

АНАЛИЗ ЦЕЛЕЙ ИНОСТРАННЫХ ГОСУДАРСТВ ДЛЯ ДОСТИЖЕНИЯ СУВЕРЕНИТЕТА В ОБЛАСТИ ИИ




Увеличение решений
в области ИИ


Квалифицированные
кадры


Увеличение доли ИИ
в экономике


Повышение уровня
доверия граждан к ИИ


Увеличение мощности
вычислений




Технологическое
превосходство как
приоритет


Кибербезопасность
и ИИ


Глобальная
конкуренция
в сфере ИИ


Регулирование
и этика ИИ


ИИ в военной
сфере




Технологическое
лидерство и
инновации


Национальная
безопасность
и оборона


Поддержка
исследований
и разработок


Регулирование
и этика ИИ


Интеграция ИИ
в экономику




Технологическое
лидерство

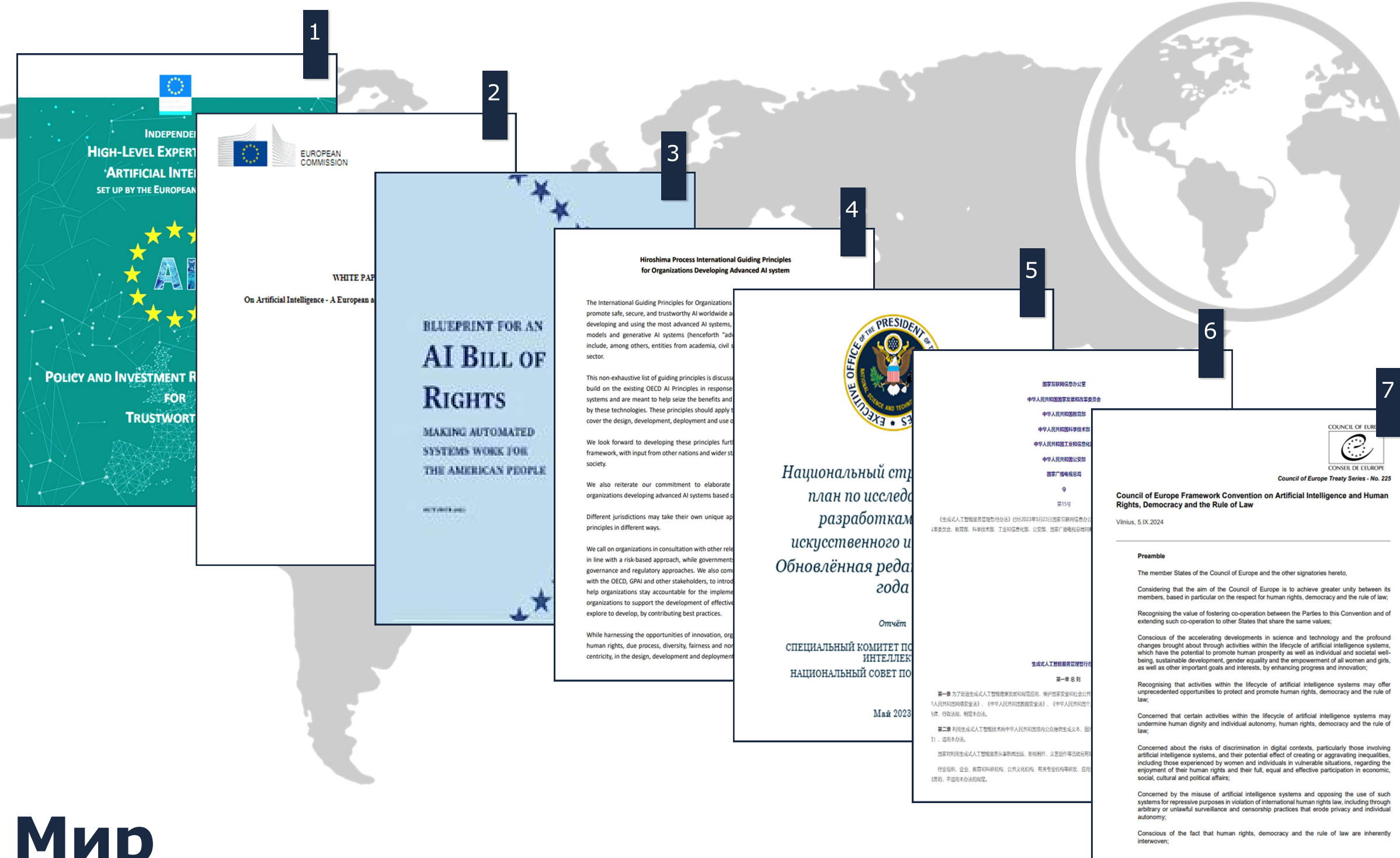

Контроль
над ИИ


Глобальное
сотрудничество


Регулирование
и этика ИИ

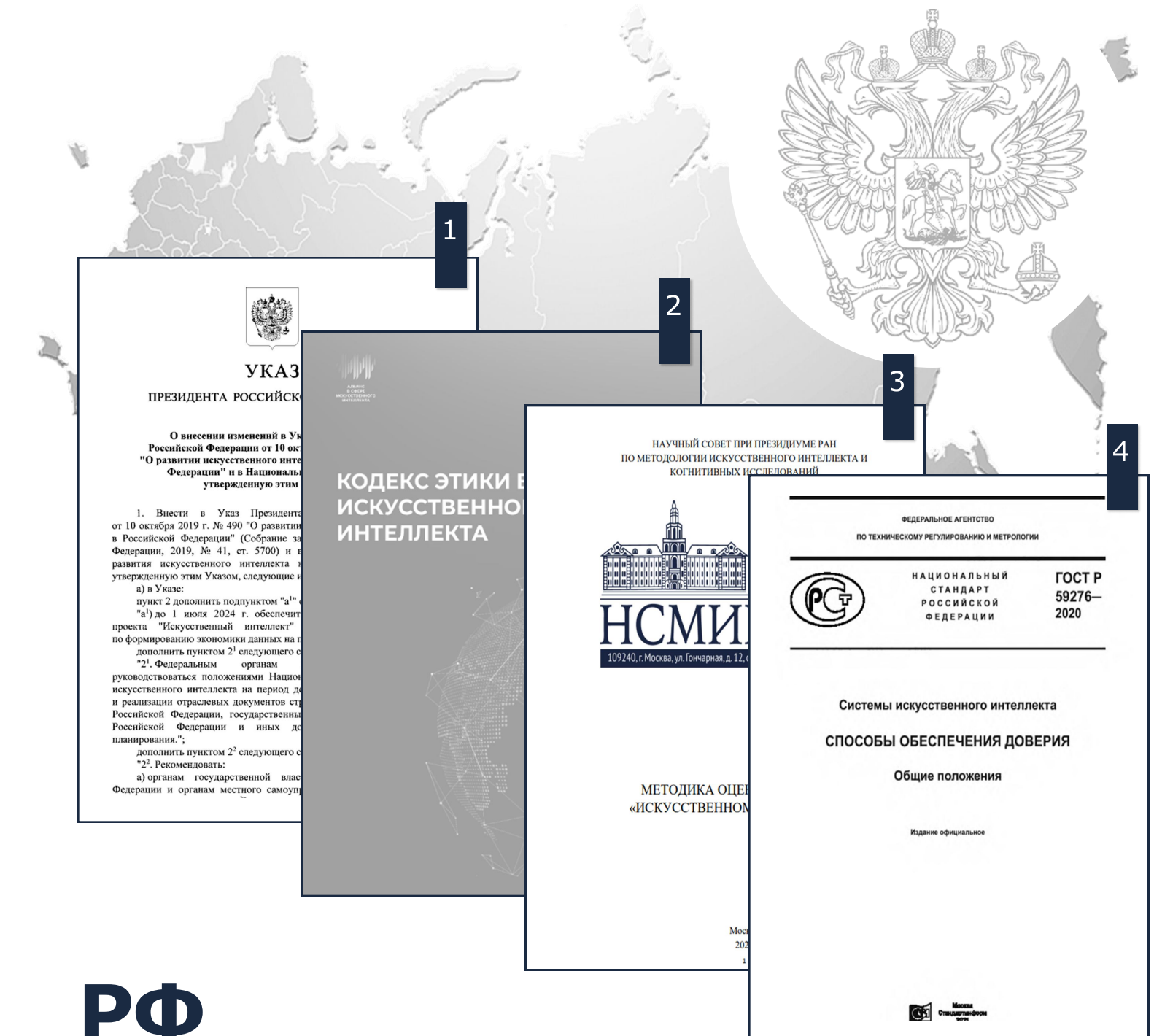

Поддержка
стратегий стран-
участниц

НОРМАТИВНАЯ БАЗА В ОБЛАСТИ ТЕХНОЛОГИЙ ИИ



Мир

1. Рекомендации по политике и инвестициям для надежного ИИ, ЕС, 2019 г.
2. Белая книга ИИ, ЕС, 2020 г.
3. Проект «Билля о правах ИИ», США, 2022 г.
4. Добровольный кодекс поведения для разработчиков ИИ, G7, 2023 г.
5. Временные меры по управлению сервисами генеративного ИИ, Китай, 2023 г.
6. Национальный стратегический план в области разработки ИИ, США, 2023 г.
7. Рамочная конвенция об ИИ и правах человека, Совет Европы, 2024 г.



РФ

1. Указ Президента РФ от 10.10.2019 №490 (ред. от 15.02.2024) «О развитии ИИ в РФ».
2. Кодекс Этики в сфере ИИ, 2021 г.
3. Методика оценки доверия к ИИ. НСМИИ РАН. 2022 г.
4. ГОСТ Р 59276-2020. ИИ. Способы обеспечения доверия ИИ.

ПОДГОТОВКА К БУДУЩИМ ВЫЗОВАМ

МЕХАНИЗМЫ ЭКСПЕРТНОГО КОНТРОЛЯ



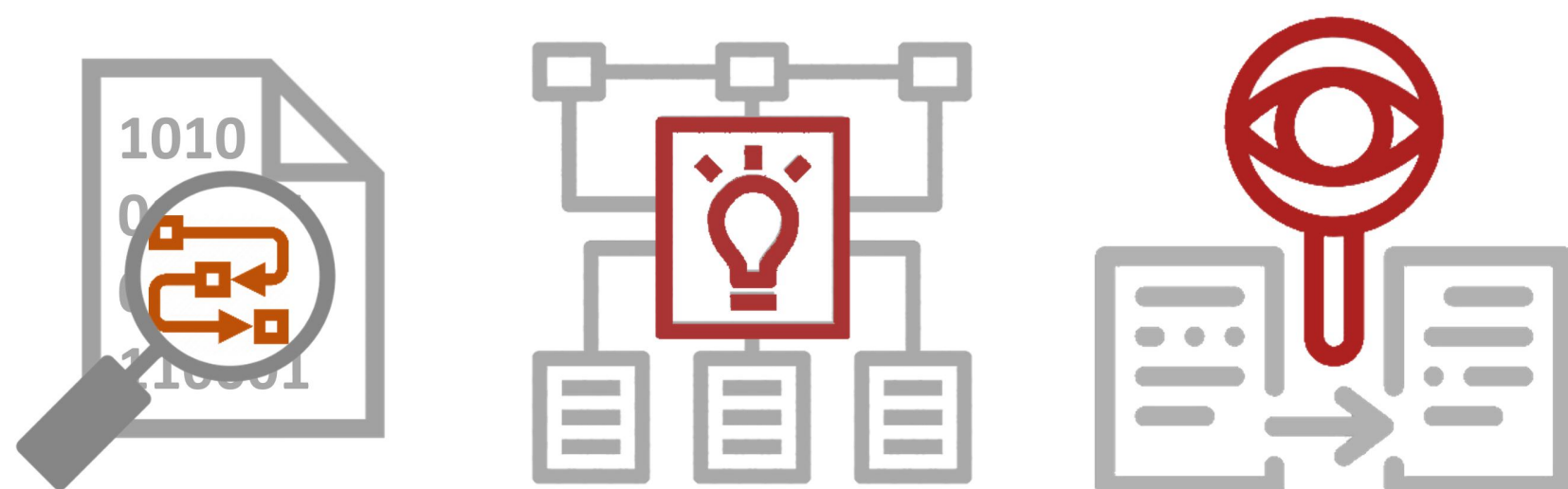
Экспертный контроль систем ИИ включает в себя механизмы, обеспечивающие надежность, безопасность, прозрачность и соответствие этическим нормам.

ОБРАЗОВАНИЕ В ОБЛАСТИ ЦИФРОВОЙ ГРАМОТНОСТИ



Совокупность знаний и навыков, необходимых для доверенного, эффективного и осознанного использования цифровых технологий.

«ПРОЗРАЧНЫЕ» АЛГОРИТМЫ И ДОКУМЕНТАЦИЯ



Открытость в разработке, объяснимость решений и доступность информации для пользователей, регуляторов и экспертов.

МЕЖДУНАРОДНЫЕ «МОСТЫ ДОВЕРИЯ»



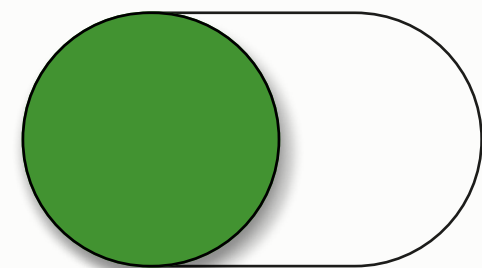
Двусторонние и многосторонние форматы позволяют странам, компаниям и экспертам обмениваться знаниями, гармонизировать стандарты и создавать системы взаимного признания сертификатов.

ВМЕСТО ЗАКЛЮЧЕНИЯ (1/2)

Тренды

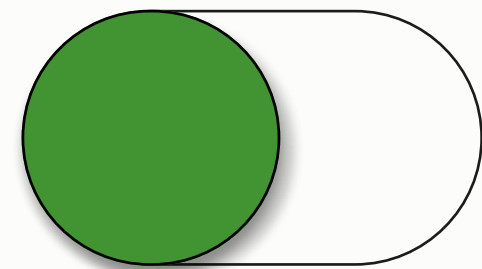
**GIGA
CHAT**

 **YandexGPT 5**



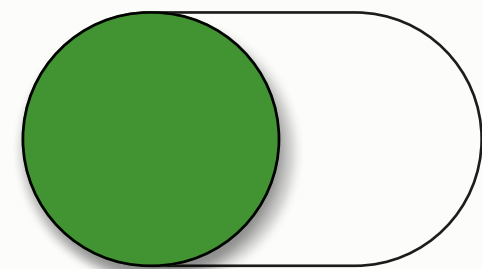
УСИЛЕНИЕ ПЕРСОНАЛИЗАЦИИ

- алгоритмы будут лучше понимать предпочтения пользователей, создавая уникальный контент индивидуально для каждого человека



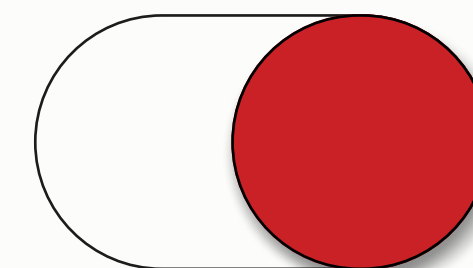
РАЗВИТИЕ МУЛЬТИМОДАЛЬНЫХ МОДЕЛЕЙ

- появятся системы, способные одновременно обрабатывать тексты, изображения, звук и видео, обеспечивая комплексное восприятие мира



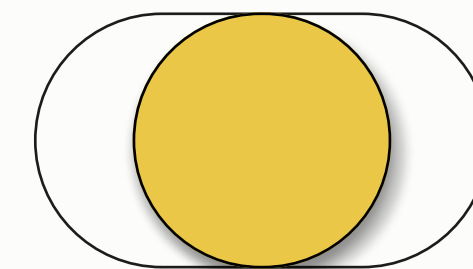
ПОВЫШЕНИЕ УРОВНЯ ЭТИКИ И БЕЗОПАСНОСТИ

- важность прозрачного и доверенного ИИ возрастёт, что потребует строгих стандартов разработки и регулирования



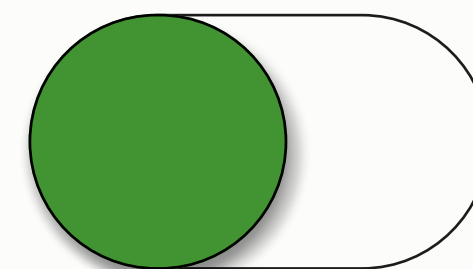
САМООБУЧАЮЩИЕСЯ СИСТЕМЫ

- нейросети, способные адаптироваться к разным задачам



СПЕЦИАЛИЗИРОВАННЫЕ МОДЕЛИ

- появление малых, но эффективных отраслевых языковых моделей



RAG-ТЕХНОЛОГИИ

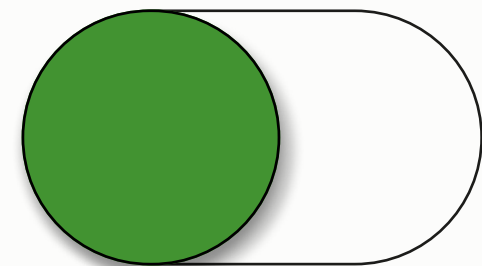
- основной метод поиска и генерации информации

ВМЕСТО ЗАКЛЮЧЕНИЯ (2/2)

Подготовка к будущим вызовам

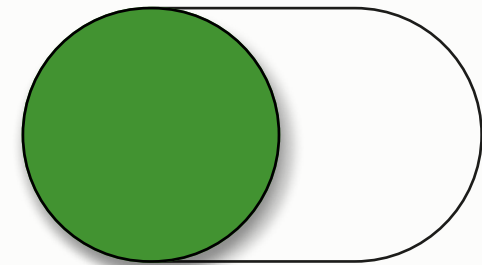
**GIGA
CHAT**

 **YandexGPT 5**



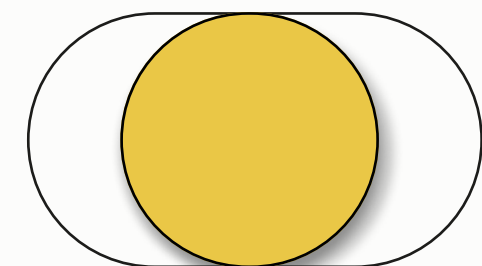
ОБУЧЕНИЕ И ПОВЫШЕНИЕ КВАЛИФИКАЦИИ

- постоянное обучение и освоение новых инструментов и подходов к работе с искусственным интеллектом станет необходимым условием успеха



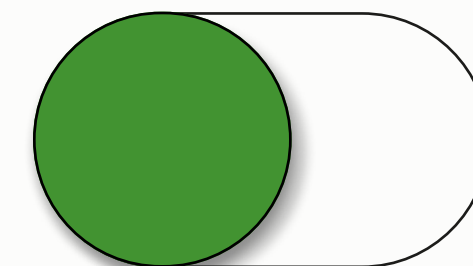
ЭТИКА И ОТВЕТСТВЕННОСТЬ

- необходимо осознавать последствия внедрения новых технологий и ответственно подходить к разработке продуктов и сервисов



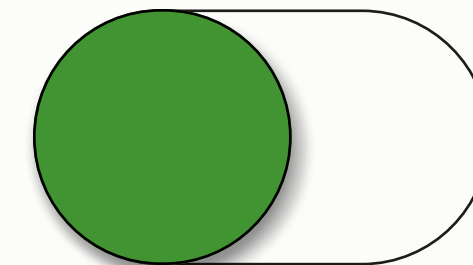
ПАРТНЁРСТВО МЕЖДУ ЛЮДЬМИ И МАШИНАМИ

- совместная работа человека и ИИ позволит создавать качественно новый уровень решений и креативности



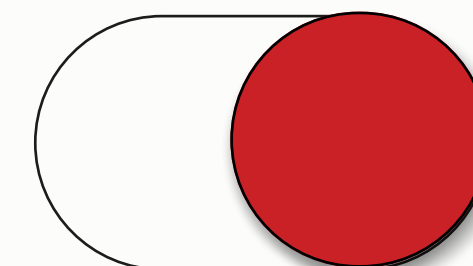
ИЗУЧАЙТЕ РАБОТУ С ДАННЫМИ

- данные становятся новым продуктом и ключевым ресурсом



РАЗВИВАЙТЕ ГИБРИДНЫЕ НАВЫКИ

- сочетайте технические знания с пониманием этики и права в сфере ИИ



ГОТОВЬТЕСЬ К ГИБРИДНЫМ АРХИТЕКТУРАМ

- понимание работы разных типов вычислений станет необходимостью

БЛАГОДАРЮ ЗА ВНИМАНИЕ



**ФОНД
ПЕРСПЕКТИВНЫХ
ИССЛЕДОВАНИЙ**

121059, г. Москва,
Бережковская наб., д. 22, стр.3
www.fpi.gov.ru