

IV ФОРУМ

13.05.2026

МОСКВА, КОНГРЕСС-ЦЕНТР МГТУ
ИМ. Н.Э. БАУМАНА

ТЕХНОЛОГИИ ДОВЕРЕННОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Практика построения доверенной среды ИИ

Масалович Андрей Игоревич

Алиса, как приготовить бараньи крылья?

Поиск **Алиса AI** [Картинки](#) [Видео](#) [Карты](#) [Товары](#) [Финансы](#) [Квартиры](#) [Переводчик](#) [Все](#)

🕒 Сегодня

алиса, как приготовить бараньи крылья

Бараньи крылья — это часть баранины, которая включает рёбра и мясо вокруг них. Для приготовления такого блюда можно использовать запекание или обжаривание. Вот два варианта рецептов.

Запекание в духовке

Ингредиенты:

- бараньи крылья (примерно 1 кг);
- соль — по вкусу;
- тимьян — 0,5 ст. л.;
- розмарин — 0,5 ст. л.;
- хлопья чили — 1 ч. л.. [iglokott.livejournal.com](#)

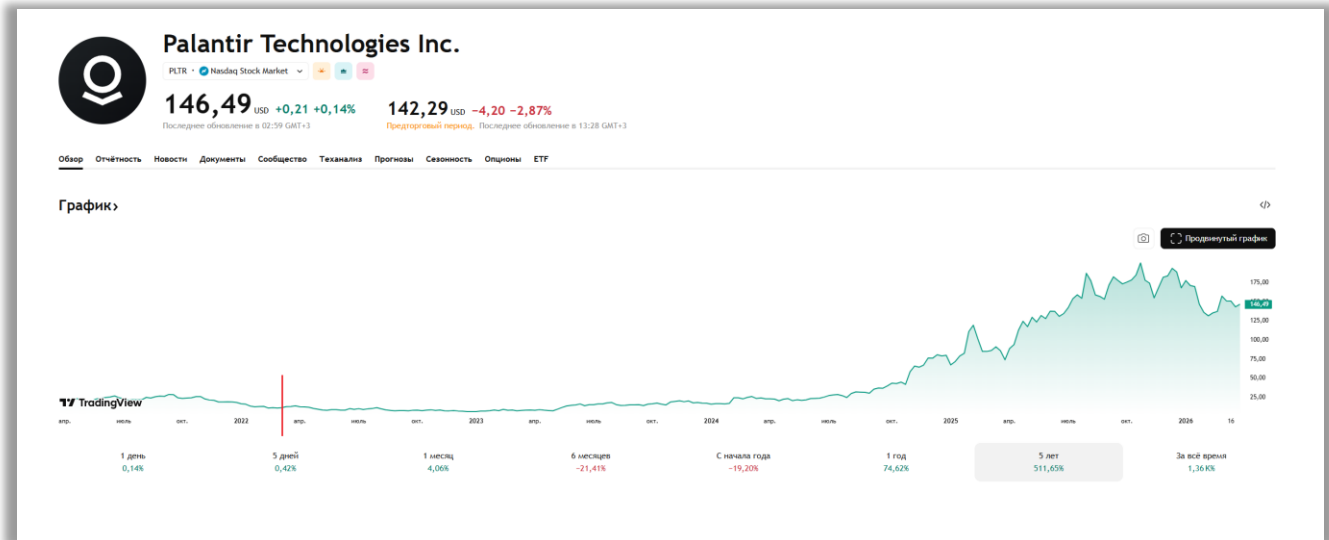
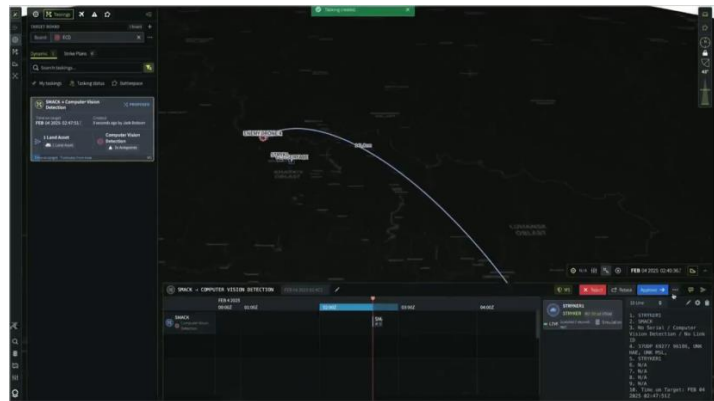
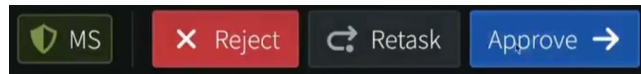
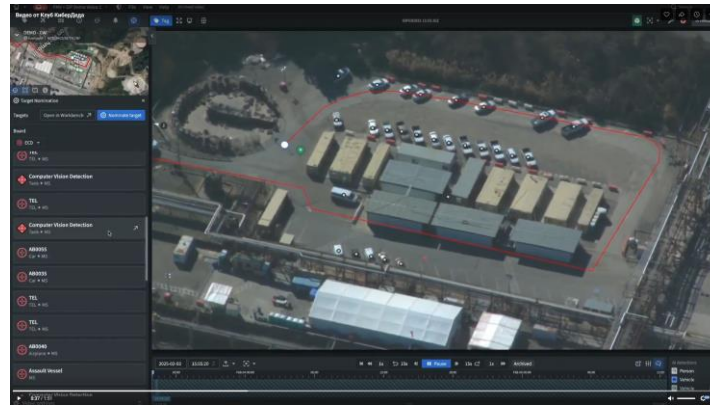
Приготовление:

1. Смешайте соль, тимьян, розмарин и хлопья чили.
2. Натрите смесью бараньи крылья.
3. Разогрейте духовку до 130 °С.
4. Поместите мясо в духовку и запекайте около 4 часов. [iglokott.livejournal.com](#)

Совет: если куски небольшие, время запекания можно сократить. Подавайте с овощным салатом или гарниром из картофеля.



Март 2026 – Palantir Maven Smart System автоматическое управление поражением целей



ИИ: Проблема построения доверенной среды

Принципиальная несовместимость процесса разработки ML-приложений и стандартов разработки ПО

Детерминированный подход к сертификации программного обеспечения и **недетерминированный вывод моделей ML.**

Прослеживание кода – V-модель позволяет определить, для удовлетворения каких требований присутствует определенная строка кода. Это невозможно для моделей ML, поскольку модуль вывода носит общий характер, а логика работы определяется данными.

Охват данных. Стандартный подход для моделей ML – так называемая точечная робастность. Сертификация моделей ML оперирует ограниченными модификациями корректных данных. Типичное описание – “сертифицированная точность 35.42% на MNIST при бюджете изменений $\epsilon = 8/255$ ”.

О проверке на всем диапазоне данных речи не идет.

Три заповеди Secure/Safe ML (MIT)

1. Вы не должны тренироваться на данных, которым не полностью доверяете (из-за возможного отравления данных – изменения данных с целью обмана модели)
2. Вы не должны позволять никому использовать вашу модель (или наблюдать за ее работой), если вы полностью им не доверяете (из-за кражи модели и атак черного ящика). Это можно представить как аналогию декомпилирования или reverse engineering в программных системах – работа (поведение) модели изучается с целью построения состязательного примера.
3. Вы не должны полностью доверять предсказаниям вашей модели (из-за возможных состязательных примеров)

Maven Smart System – как это работает?

Первый слой – автономные микромодели компьютерного зрения **Edge AI**

Первый слой системы установлен прямо на дроне или спутнике.

Там работают микромодели компьютерного зрения: они не думают, только «смотрят» и переводят картинку в числа — координаты, вероятности, метки. Спутник засёк бронеколонну, сжал видео в текстовый файл и передал дальше.

Трафик сразу падает на 90%.

Edge AI от Palantir работает на Украине с 2022 года.



Онтология – обогащение данных объекта

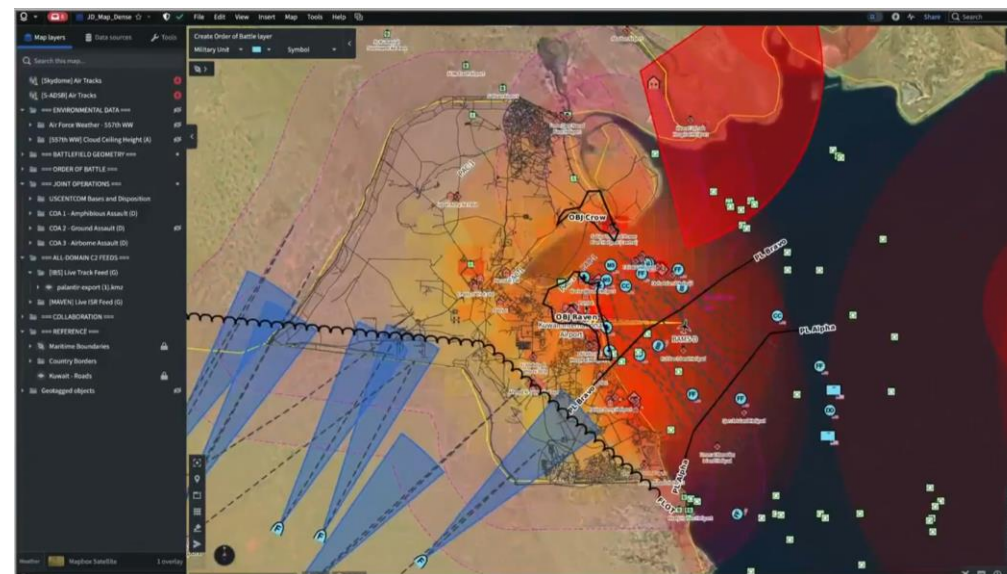
Дальше данные попадают в подсистему Онтология.

Это не просто база данных — каждый танк, документ, человек превращается в «объект» со связями и историей.

Но прежде, чем любая нейросеть получит хоть байт, система проверяет криптографические допуски.

Генерал видит одно, оператор — другое, союзник по НАТО — третье.

ИИ физически слеп за пределами того, что ему выдали на одну секунду.



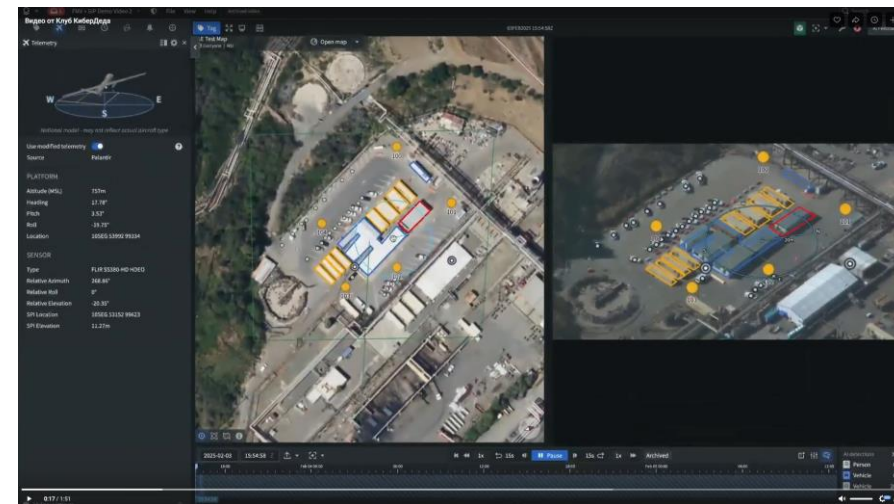
Оркестратор – декомпозиция задачи

Никакого ChatGPT-стиля «спроси что хочешь» здесь нет.

Запрос попадает к Оркестратору — он разбивает задачу на куски и раздаёт их микро-агентам.

Каждый агент зажат в жёсткую логику конечного автомата: шаг 1, шаг 2, вызов инструмента.

Ни шага в сторону — только по рельсам боевого устава.



Синтезатор - Точка принятия решения

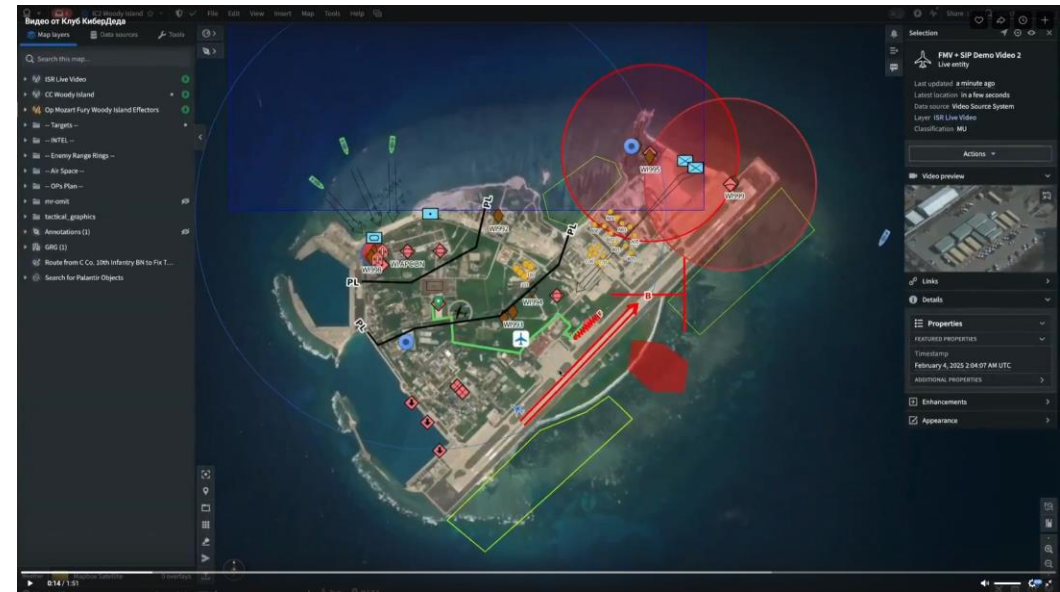
Самое интересное происходит в точке принятия решения.

Когда нужно ответить — «это вражеская колонна?» — система не верит одной модели.

Запрос уходит параллельно в GPT, Claude, Llama, Groq, возможно, в засекреченную собственную разработку.

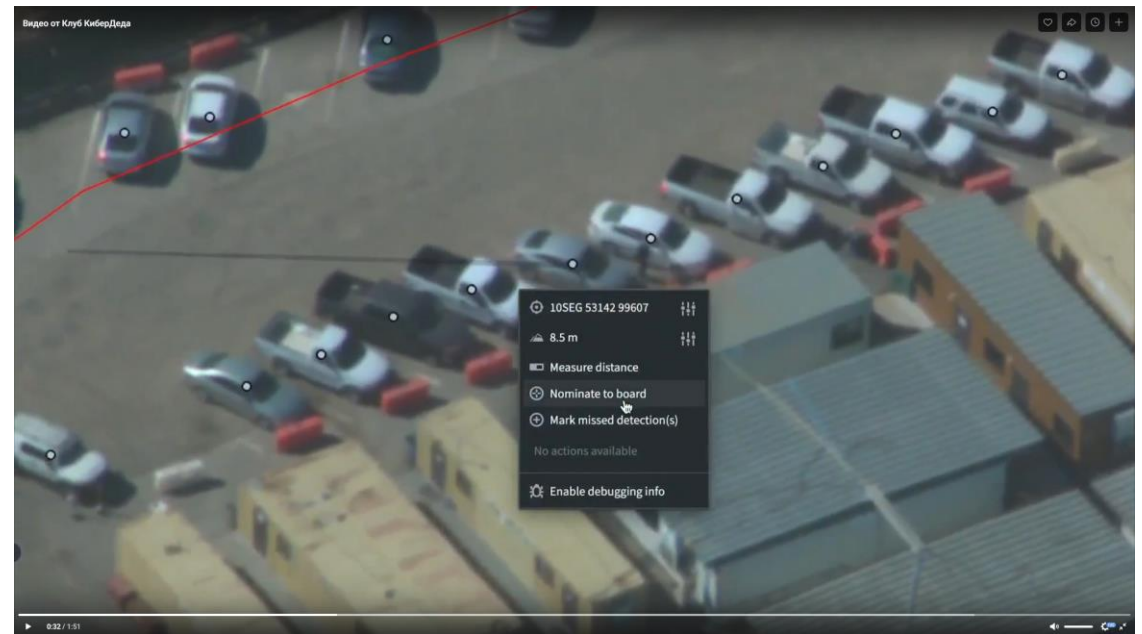
Они думают независимо, не видя ответов друг друга.

Потом «Синтезатор» сравнивает результаты, ищет консенсус и отсекает галлюцинации.



Агенты-оценщики - Жесткий контроль процессов

За каждым агентом в реальном времени следят **Агенты-Оценщики** — они читают логи, и, если модель начинает бредить или уходить от протокола, процесс убивается мгновенно. Инженер может влезть в цепочку рассуждений, исправить ошибку руками и перезапустить конвейер.



Принятие решения – за человеком

Офицер сидит перед дашбордом.

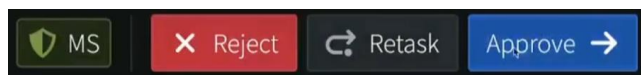
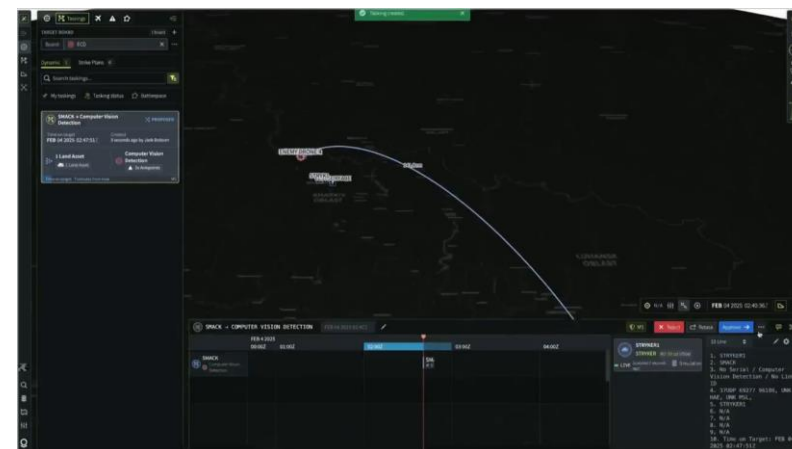
На экране появляется сообщение:

«Обнаружена цель. Вероятность 98%.

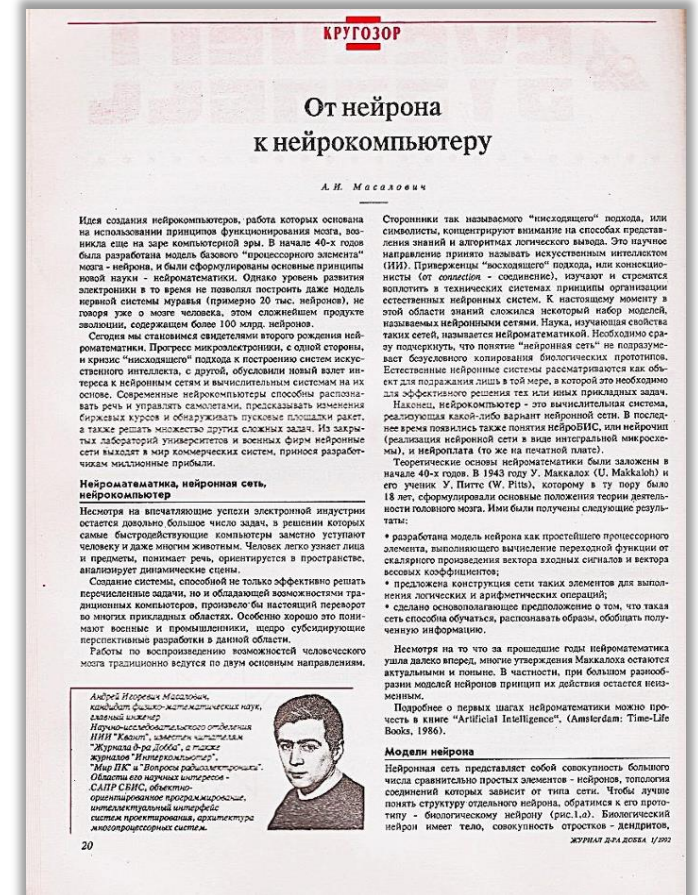
Консенсус 4 нейросетей.

Предлагаемое действие: нанести удар.

Нажмите **Approve**».



От нейрона к нейрокомпьютеру, 1992



- Ты что, хочешь сказать, что какая-то программа может заменить грамотного специалиста?
- Ха, ты еще скажи, что появится беспилотное такси

Самый быстрый вычислитель на ладони

1993 – MVS-100



2026 – NVIDIA DGX Spark



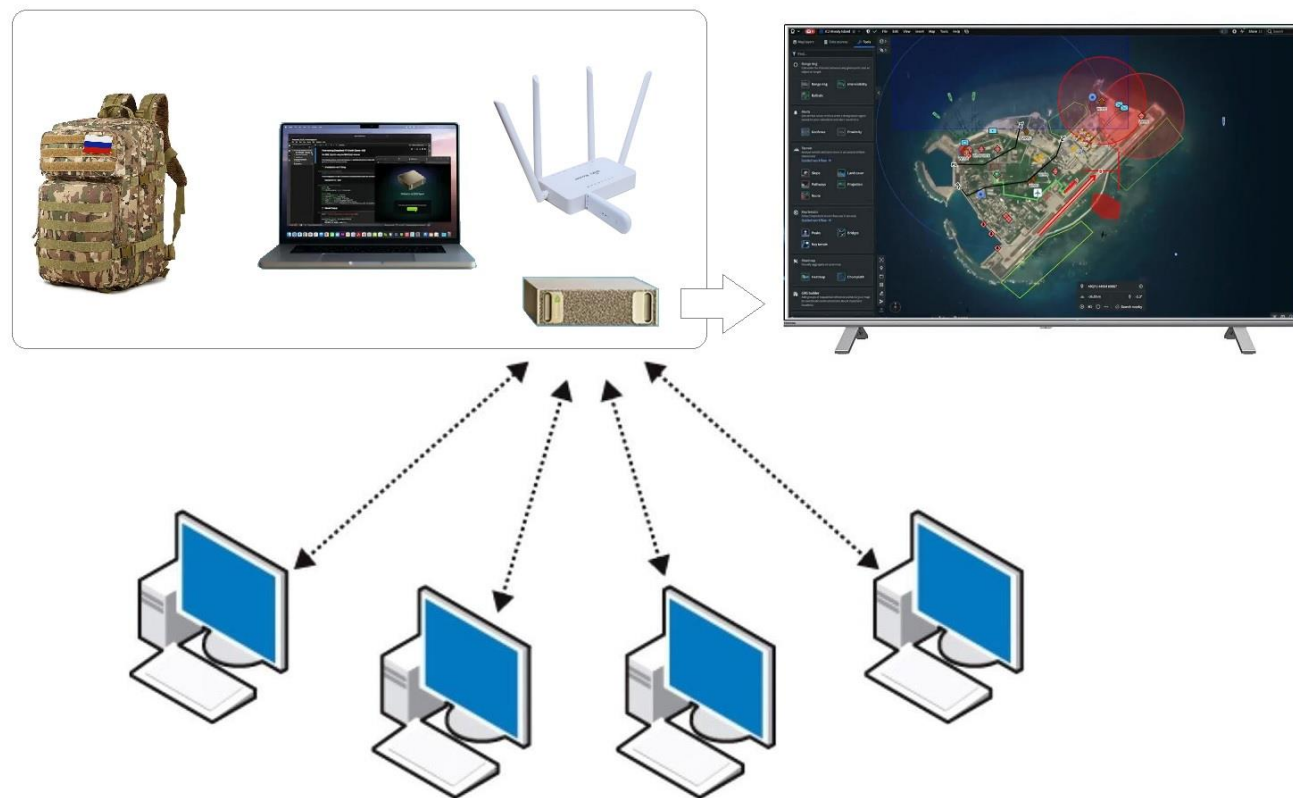
Five Steps to Building a Trusted Environment with AI

Пять шагов к построению доверенной среды ИИ

1. Построение изолированной вычислительной среды
2. Контроль регламентов и фильтрация результатов
3. Комплексное тестирование исходного кода, алгоритмов обучения, условий выполнения
4. Выбор доверенных поставщиков (Этический кодекс ИИ)
5. Выработка и соблюдение международных договоров и соглашений



Изолированные вычислительные периметры внутри закрытых сетей – Secure environments



- **Дата-центр на рабочем столе**
- Производительность - 1 Петафлоп
- Модели до 64В параметров
- Более 30 нейросетевых моделей
- До 20 пользователей
- Автономность и мобильность
- Приватность и безопасность данных

Изолированный дата-центр ИИ на вашем столе

Архитектура - **NVIDIA Grace Blackwell GB10**

Графический процессор - Архитектура NVIDIA Blackwell

Центральный процессор - 20 ядер Arm (10 Cortex-X925 + 10 Cortex-A725)

Тензорная производительность - 1000 TOPS (FP4, Sparse)- **Один Петафлоп**

Оперативная память - 128 ГБ общей памяти LPDDR5x

Шина памяти - 256-битная

Пропускная способность памяти - 273 ГБ/с

Накопитель - NVMe, M.2, 1 или **4 ТБ, с самошифрованием**

USB - 4x USB 3.2 Type C

Ethernet - 1x RJ-45 connector 10 GbE

Сетевой адаптер - ConnectX-7 Smart NIC

Адаптер Wi-Fi - WiFi 7П

Bluetooth - BT 5.3

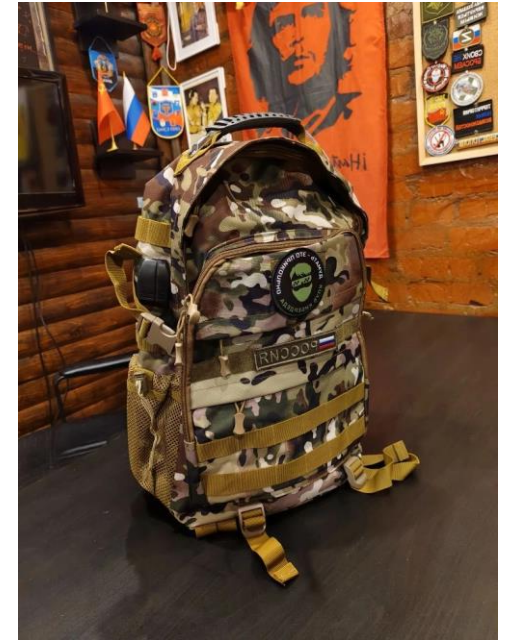
Аудиовыход - Мультисканальный аудиовыход HDMI

Видеовыходы - 1x HDMI 2.1a

NVENC | NVDEC - 1x | 1x

Операционная система - NVIDIA DGX™ OS

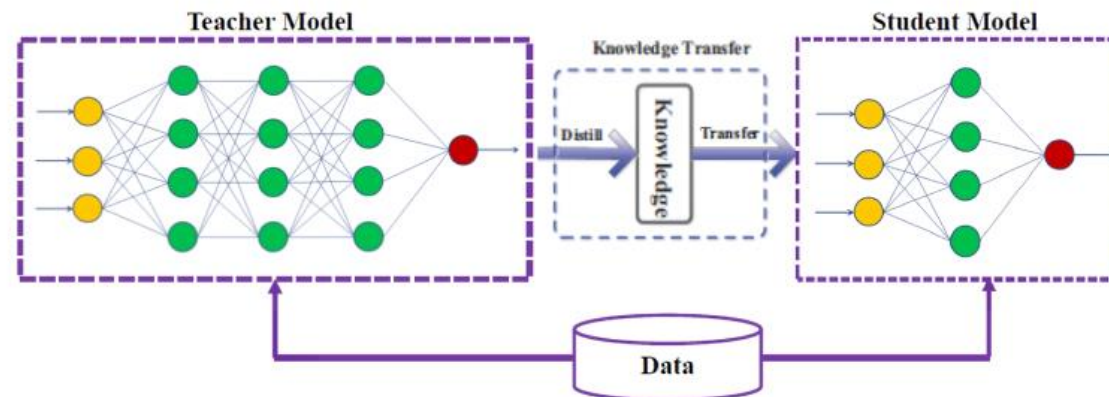
Размер - 151 мм L x 151 мм W x 52 мм H (1.19L), вес - 1.2 кг



Дистилляция моделей

Дистилляция в контексте ИИ — это процесс, при котором более компактная и эффективная модель (часто называемая «студенческой») обучается на основе более крупной и сложной модели (называемой «учительской»).

Этот метод позволяет передавать знания от мощной, но ресурсоемкой модели к меньшей, которая требует меньше вычислительных ресурсов для обучения и работы, но при этом сохраняет высокую производительность.



Загруженные модели ИИ

gpt-oss:20b — открытая версия модели GPT с 20 миллиардами параметров. Открытый исходный код и доступность для широкой аудитории делают её привлекательной для исследований и экспериментов.

translategemma:27b — старшая версия переводческой модели семейства Gemma с 27 миллиардами параметров. Максимальная производительность и поддержка переводов между множеством языков. Требовательна к ресурсам, но обеспечивает высочайшую точность.

gemma3:12b — универсальная модель семейства Gemma 3 с 12 миллиардами параметров. Широкий спектр применения в задачах обработки естественного языка. Простота интеграции и удобство использования.

llama2-uncensored:70b — вариация популярной модели LLaMA 2 с 70 миллиардами параметров, созданная Джорджем Сангом и Джаррадом Хопом. Отсутствует цензура, что делает её свободной от ограничений оригинального LLaMA 2. Используется для генерации разнообразного контента без фильтрации.

deepseek-r1:32b — крупная модель от DeepSeek с 32 миллиардами параметров. Ориентирована на выполнение сложных логических операций и задач, требующих глубоких рассуждений. Высокопроизводительная и многофункциональная.

gemma3:12b — представитель семейства Gemma 3 с 12 миллиардами параметров. Универсальная модель общего назначения, обладающая широкими возможностями для множества задач обработки естественного языка.

deepseek-r1:14b — модель от DeepSeek с 14 миллиардами параметров. Балансирует между производительностью и ресурсоемкостью, идеально подходит для большинства стандартных задач обработки естественного языка.

gemma3:27b — крупная версия семейства Gemma 3 с 27 миллиардами параметров. Максимизирует производительность и функциональные возможности, требуя значительных вычислительных ресурсов. Рекомендуются для профессиональных и высоконагруженных задач.

qwen3-vl:30b — мультимодальная версия модели QWEN с 30 миллиардами параметров, объединяющая обработку текста и изображений.

qwen3:30b — модель семейства QWEN с 30 миллиардами параметров.

qwen3-coder:30b — специализированная версия модели QWEN с фокусировкой на программировании и разработке. Содержит 30 миллиардов параметров, оптимизирована для написания и анализа программного кода.

qwen2.5vl:32b — мультимодальная модель семейства QWEN с 32 миллиардами параметров, сочетающая обработку текста и изображений.

llama2-uncensored:7b — более компактная версия uncensored-модификации LLaMA 2 с 7 миллиардами параметров. Менее требовательна к ресурсам, но сохраняет свободу от цензуры и широкую функциональность.

llama3.1:8b — компонент семейства LLaMA 3.1 с 8 миллиардами параметров. Является основой для мультимодельных решений, таких как LLaMA 3.2 Vision. Хорошо справляется с текстовыми задачами и служит базой для дальнейшего расширения функциональности.

glm-4.7-flash-q4_K_M — модель на 30 миллиардов параметров с архитектурой Mixture of Experts (MoE). Эффективно сочетает производительность и легкость развертывания. Подходит для задач программирования и построения интеллектуальных агентов. Использует уникальный механизм Interleaved Thinking, позволяющий поэтапно решать задачи. Квантизация Q4_K_M обеспечивает оптимальный баланс между размером и производительностью.

translategemma:12b — модель из семейства Gemma 3, содержащая 12 миллиардов параметров. Специализируется на переводе между 55 языками. Отличается высокой эффективностью и способностью работать на ноутбуках и персональных компьютерах. Преимущественно используется для локального перевода и обработки текстов.

rinex20/translategemma3:12b — одна из разновидностей модели translategemma:12b, настроенная командой RINEX20.

translategemma:4b — менее ресурсоемкий представитель семейства TranslateGemma с 4 миллиардами параметров. Оптимален для мобильного и краевого развертывания. Сохраняет высокие показатели качества перевода при минимальном потреблении ресурсов.

huihui_ai/perplexity-ai-r1-abliterated:latest — модель от команды huihui ai, относящаяся к проекту Perplexity AI.

phi4:14b — современная языковая модель от Microsoft с 14 миллиардами параметров и контекстным окном в 32 тысячи токенов. Ориентирована на задачи, требующие сложного логического мышления и рассуждений. Показала конкурентоспособные результаты против моделей с сотнями миллиардов параметров.

deepseek-ocr:latest — модель для оптического распознавания символов (OCR) от компании DeepSeek. Ожидается эффективное распознавание текста на изображениях и документах. Идеально подойдет для автоматизированной обработки сканов и фотографий с текстом.

dolphin-phi:latest — модификация PHI-моделей, интегрированное с технологией Dolphin. Предполагается наличие специализированных функций для обработки естественного языка или создания контента.

Hudson/llama3.1-uncensored:8b — нецензурированная версия модели LLaMA 3.1 с 8 миллиардами параметров, подготовленная командой Hudson. Характеристики и назначения близки к оригинальному варианту, но с отсутствием цензурных фильтров.

mistral-small3.2:24b — версия модели Mistral с 24 миллиардами параметров. Ориентирована на быстрое развертывание и легкое потребление ресурсов. Подходит для простых и средних задач обработки естественного языка.

gpt-oss-safeguard:20b — защищённая версия открытой модели GPT с 20 миллиардами параметров. Включает механизмы безопасности и контроля содержания.

qwen3-vl:8b — небольшая мультимодальная версия модели QWEN с 8 миллиардами параметров. Упрощённый вариант для лёгких задач, сочетающих обработку текста и изображений.

nemotron-cascade-2:30b — модель Nemotron второго поколения с 30 миллиардами параметров.

qwen3.5:27b — новая версия модели QWEN третьего поколения с 27 миллиардами параметров. Улучшенные характеристики и повышенная производительность по сравнению с предыдущими версиями.

mistral-lite:latest — лёгкая версия модели Mistral. Минималистичный дизайн и низкие требования к ресурсам делают её идеальной для быстрого прототипирования и небольших проектов.

llama3.2-vision:11b — мультимодальная модель семейства LLaMA 3.2 с 11 миллиардами параметров. Объединяет обработку текста и изображений, позволяя решать комплексные задачи, такие как визуализация и документный анализ.

nemotron-mini:latest — минималистичная версия модели Nemotron. Компактная и быстрая, отлично подходит для простых задач и ограниченных ресурсов.

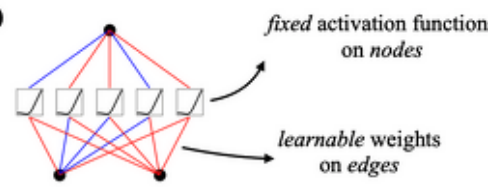
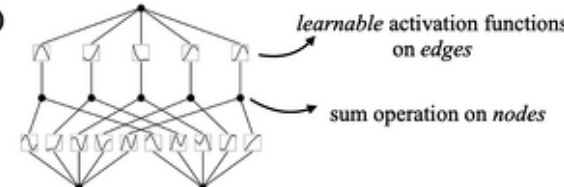
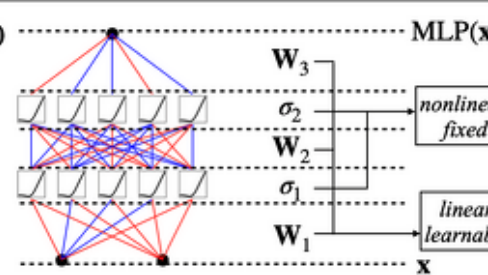
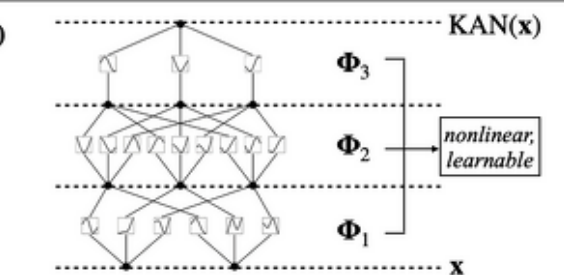
richardyoung/deepseek-r1-32b-uncensored:latest — нецензурированная версия крупной модели DeepSeek-R1 с 32 миллиардами параметров, подготовленная Ричардом Янгом. Свободна от цензуры и ограничений, что даёт максимальную свободу в генерации контента.

mathstral:7b — модель Mathstral с 7 миллиардами параметров, ориентированная на математические расчёты и задачи, требующие численных вычислений.

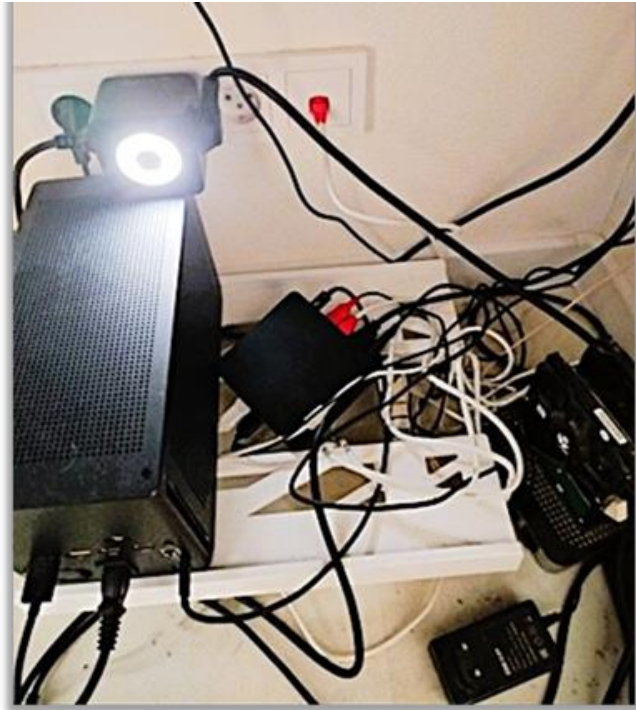
qwen2-math:7b — специальная математически ориентированная версия модели QWEN с 7 миллиардами параметров. Оптимизирована для решения математических уравнений и расчетов.

Сети Колмогорова-Арнольда

- компактнее, эффективнее, понятнее

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	(a)  <i>fixed activation functions on nodes</i> <i>learnable weights on edges</i>	(b)  <i>learnable activation functions on edges</i> <i>sum operation on nodes</i>
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	(c)  \mathbf{W}_3 σ_2 \mathbf{W}_2 σ_1 \mathbf{W}_1 \mathbf{x} nonlinear, fixed linear, learnable	(d)  Φ_3 Φ_2 Φ_1 \mathbf{x} nonlinear, learnable

Знакомьтесь: Павел Попович, автор LLM Rukallama на основе сети Колмогорова-Арнольда (спойлер: получается в 150 раз компактнее)



Загруженные примеры

Разбор перехваченных документов и утечек ВСУ (перевод, авто реферирование, обработка схем и графиков, уточнения и выводы, углубленный тематический анализ)

Разбор вредоносного кода новых эксплоитов и выработка мер противодействия

Автоматическая генерация и корректировка новостной повестки – формирование ленты новостей, генерация комментариев, автоматический перевод, коррекция новостной повестки и формирование целевых нарративов

Формирование аналитических справок на объекты интереса и обобщение результатов OSINT

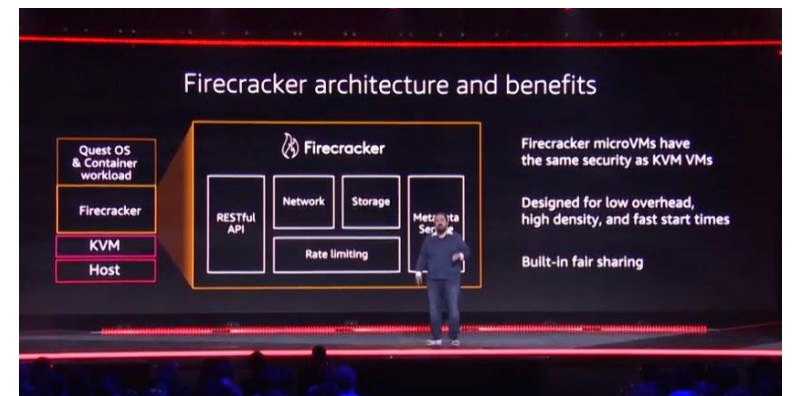
Анализ паттернов поведения потенциально опасных детей и подростков и выработка методик противодействия

Анализ методов вербовки подростков в деструктивные организации и подготовка методических рекомендаций по противодействию.

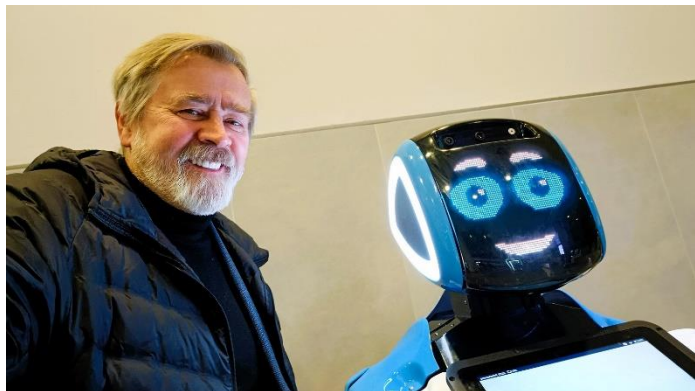
Поиск информации в интернете и обработка ее с помощью больших языковых моделей прямо в чате

Возрождаем техническую разведку: Атака на микровиртуалку Amazon FireCracker

- Скачали исходники FireCracker (на Java)
- Скомпилировали, собрали. Объем ядра – 4.1 Мб
- DeepSeek, можешь рассказать архитектуру решения?
- Можешь улучшить метрики?
- Ты знаешь язык Rust?
- Можешь переложить исходники на Rust?

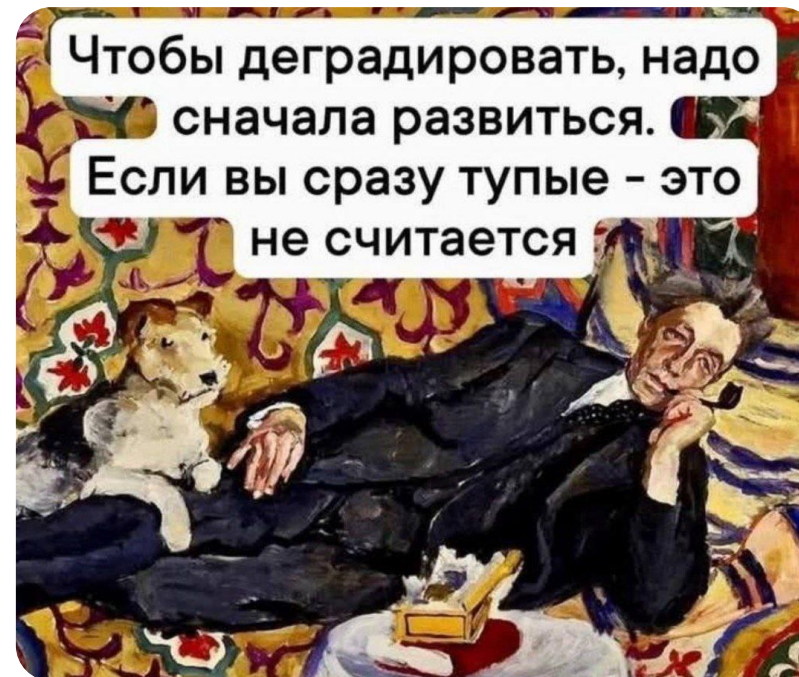


Что делать? Учиться!



...Преимущество сегодня получит не тот, кто полностью доверяет ИИ, и не тот, кто его игнорирует, а тот, кто быстрее остальных научится использовать его как рабочий инструмент

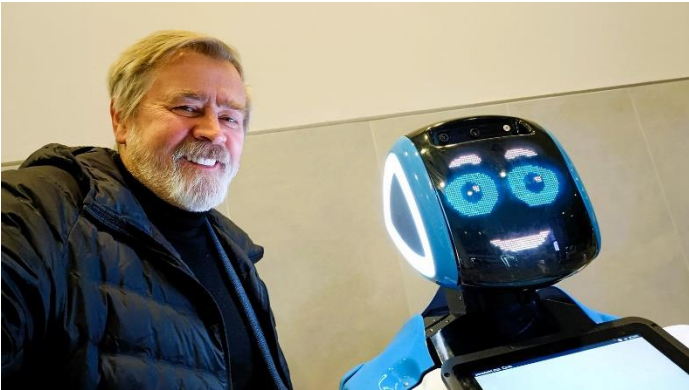
Кибердед



Масалович Андрей Игоревич
Телеграм и МАХ : Кибердед_official
Сайт: am.news
E-mail: avalanche100500@gmail.com

Something Wrong Is Happening

- Кибердед, апрель 2026



Масалович Андрей Игоревич
ака Кибердед

Телеграм и МАХ : Кибердед_official

Сайт: am.news

E-mail: avalanche100500@gmail.com

Вывод: Преимущество сегодня получит не тот, кто полностью доверяет ИИ, и не тот, кто его игнорирует, а тот, кто быстрее остальных научится использовать его как рабочий инструмент

Что делать? Учиться!



Андрей Масалович
Кибердед
avalanche100500@gmail.com



Сайт: am.news

