

AI Scientist — большой вызов для современных универсальных систем искусственного интеллекта

Александр Игоревич Панов

д.ф.-м.н., доцент

директор лаборатории CAIS AIRI

директор центра когнитивного моделирования МФТИ

Содержание

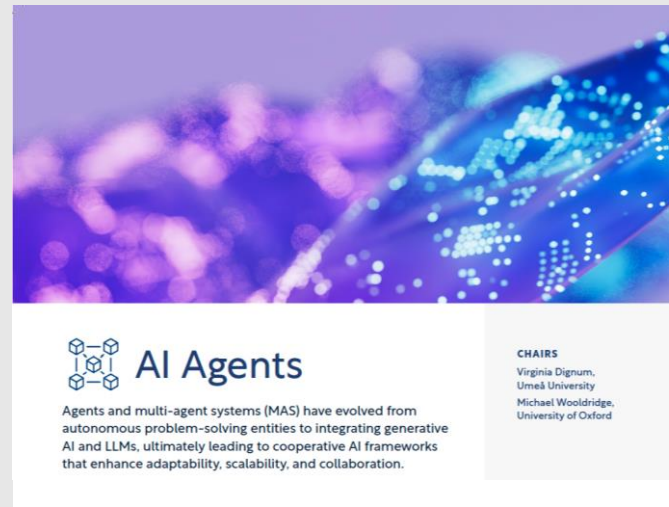
- 01 AGI, LLM и ИИ-агенты
- 02 Автоматизация научных исследований
- 03 ИИ-ассистенты исследователя
- 04 Основные вызовы на пути к AIScientist
- 05 Воплощенный ИИ-ассистент

01

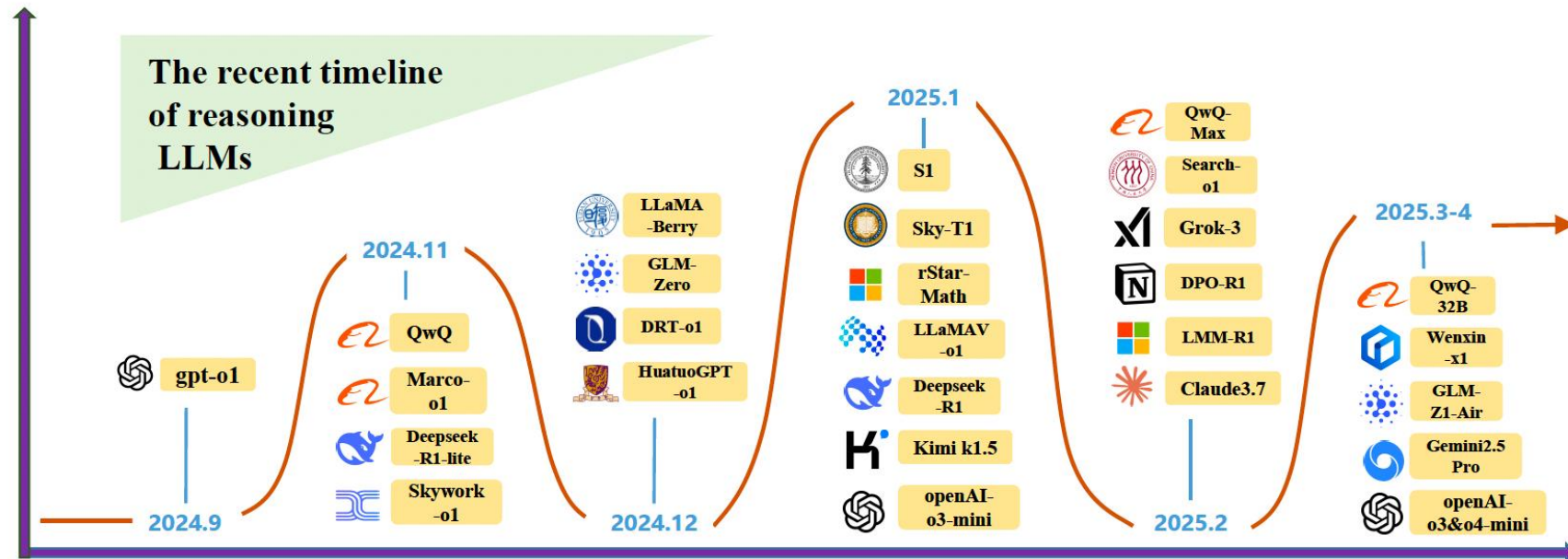
AGI, LLM и ИИ-агенты

ИИ-агенты и AGI

- Текущие вызовы в развитии интеллектуальных агентов и их систем связаны с интеграцией LLM и Gen-AI архитектур в коллаборативных сценариях
- Ключевой технологией на пути к AGI названы долгосрочное планирование и рассуждения, воплощенные агенты и взаимодействие с реальным миром
- Развитие методов ИИ для научных исследований требует создания моделей научных рассуждений и коллаборативности



Языковые модели как универсальные системы ИИ

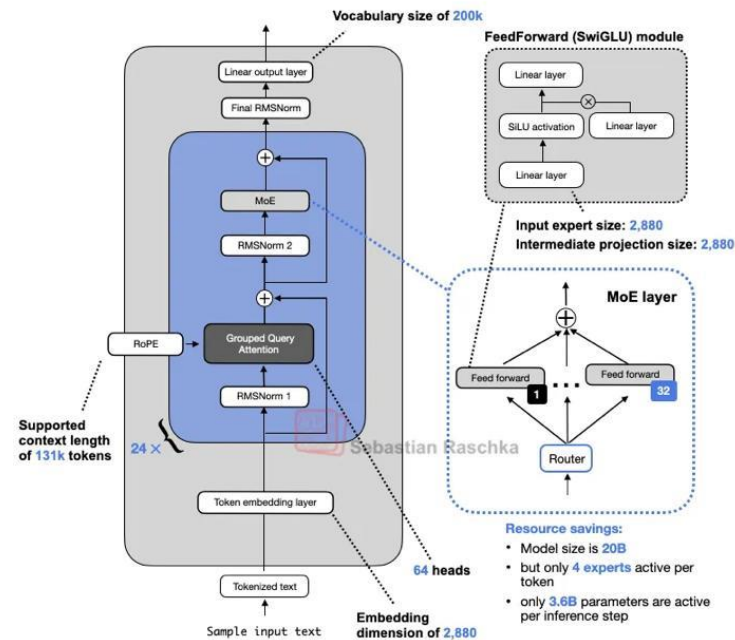


- Ответы на вопросы и решение задач через цепочку шагов – имитация рассуждения и поиска
- Дообучение моделей (SFT, RL) на чистых данных и явном вознаграждении
- Мультимодальность входа и удлинение контекста

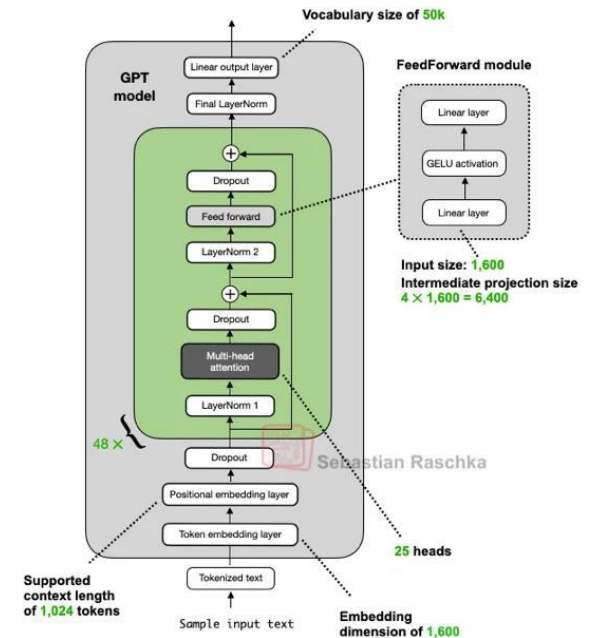
Развитие архитектур остановилось?

- **Dropout** для борьбы с переобучением теперь уже не нужен, потому что учимся в одну эпоху на одном огромном датасете
- **RoPE** (ротация Q и K векторов) теперь используется взамен абсолютного позиционного кодирования
- **SwiGLU** теперь используется вместо обычной функции активации GELU хотя отличия между ними под вопросом
- **MoE** заменяет единый FeedForward модуль с неполной активацией экспертов на каждом токене
- **GQA** (групповое внимание) пришло на смену обычному много-головому для уменьшения потребления памяти
- **Скользящее** окно внимания теперь используется во многих слоях сети для увеличения контекста модели
- **RMSNorm** пришел на смену обычной нормализации активаций на дисперсию по слою

GPT-OSS 20B (2025)



GPT-2 XL 1.5B (2019)



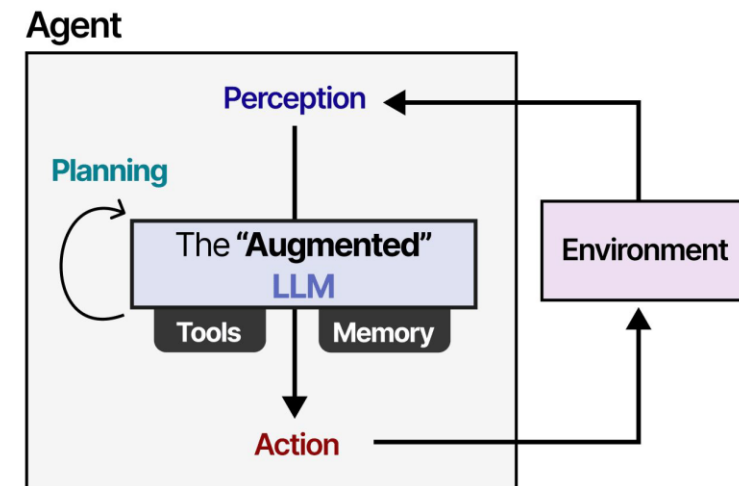
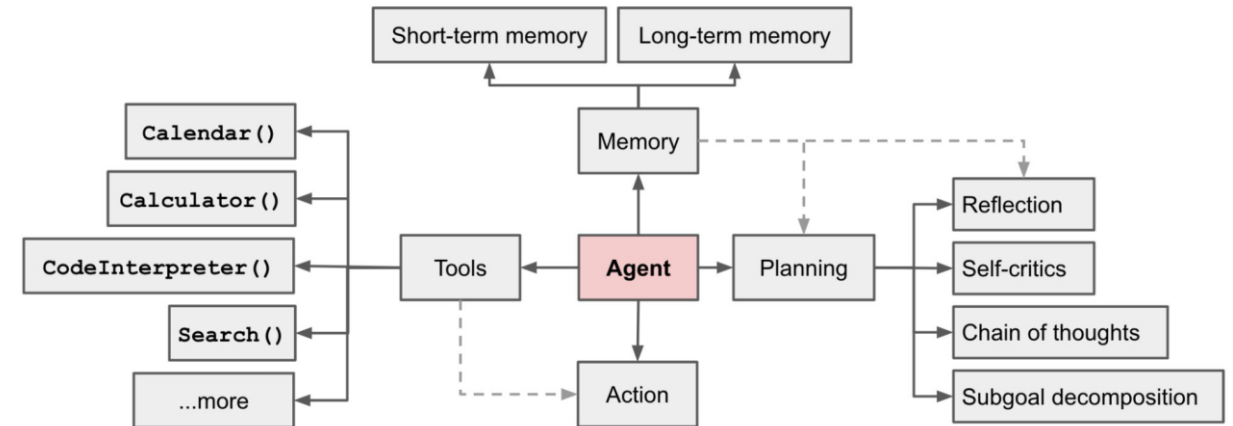
Агенты на основе языковых моделей

Агент – система которая получает задачу (цель), планирует шаги решения, использует инструменты (вызов кода, API), проверяет результат и повторяет цикл до достижения цели

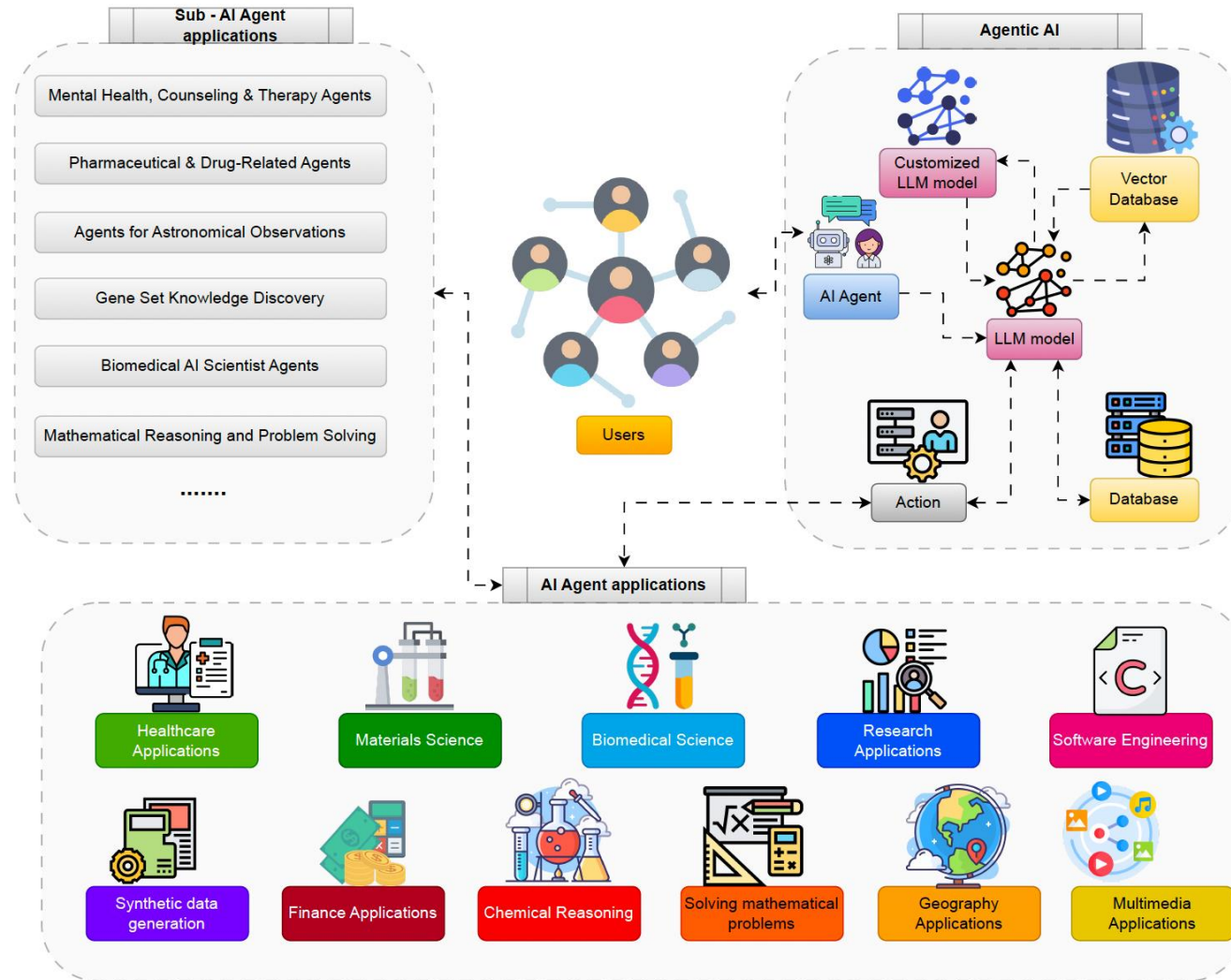
Структура агентов:

- Стратегия на базе LLM: генерация действий/команд
- Инструменты: поиск, БД, платежи и т.д.
- Память: контекстное окно или RAG
- Оркестрация: контроль цикла, проверки

Язык как универсальный API к задачам и инструментам (модель сама формирует команды), связка разных API



Приложения LLM-агентов



02

Автоматизация научных исследований

Порождение гипотез

Индуктивные рассуждения, или рассуждения на основе индукции, оказываются полезными при решении ряда задач, среди которых:

- обнаружение закономерностей в массивах данных,
 - восстановление вида функции по ее примерам,
 - восстановление грамматики неизвестного языка по множеству его текстов и др.
-
- Индуктивные рассуждения можно рассматривать как инструмент для порождения гипотез
 - При этом порождаемые гипотезы не обязаны быть достоверными и требуют дальнейшего анализа, результатом которого должно явиться их обоснование (либо опровержение)

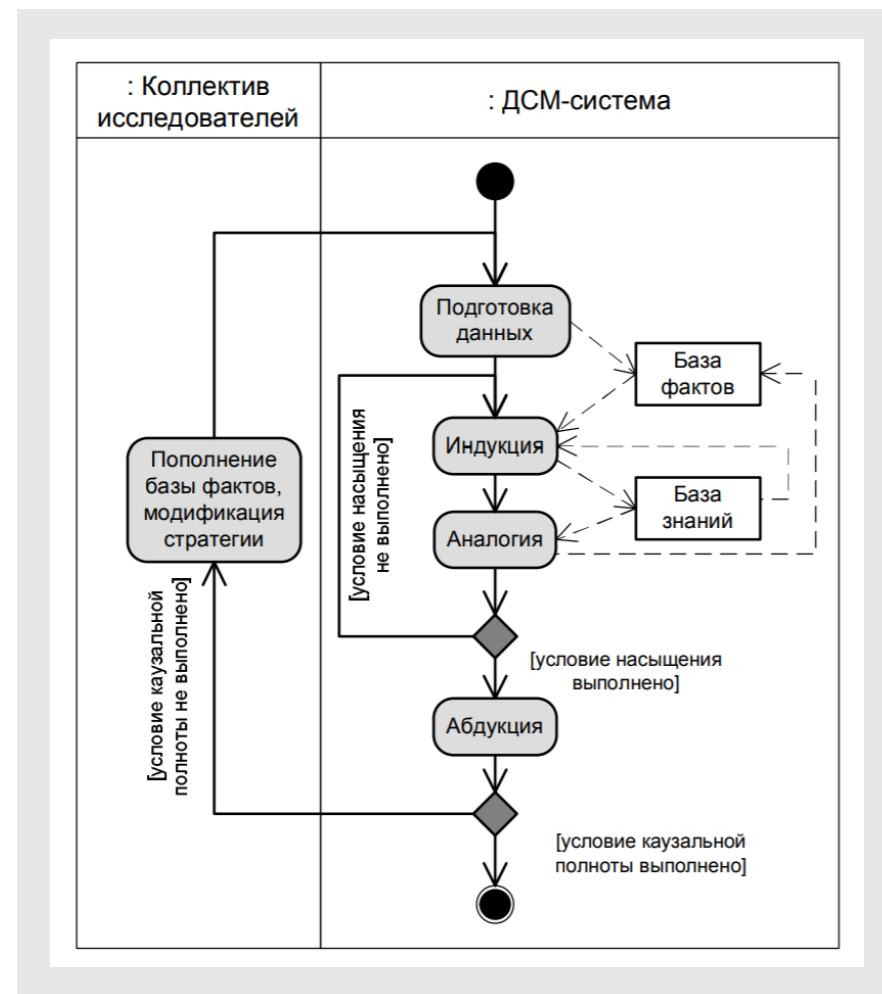
Идея ДСМ-метода

Идея метода восходит к 1900 г. - она была предложена в начале XX в. английским логиком и философом Джоном Стюартом Милем

- Пусть задано множество структур типа объект-свойство и пусть задача состоит в том, что для некоторого нового объекта требуется установить его неизвестное свойство
- Предполагается, что за каждое свойство объекта отвечает некоторая его составная часть или фрагмент
- Анализируются структуры объектов и выполняется поиск объектов, имеющих совпадающие свойства и структурное сходство
- Общие фрагменты объектов, обладающих совпадающими свойствами, являются причиной этих свойств
- Затем выполняется поиск такого рода фрагментов в новых объектах и, если такие фрагменты обнаружены, объектам приписываются свойства, причинами которых являются обнаруженные фрагменты
- В простейшем случае полагается, что свойства, за которые ответственны те или иные фрагменты, порождаются только этими фрагментами и никак не зависят от окружения или контекста

ДСМ метод автоматизации рассуждений

- **Обнаруживает закономерности** в данных с помощью индукции в стиле Д.С.Милля
- **Порождает правила** устранения неопределенности в данных, используя фальсификацию в стиле К.Поппера
- **Формирует предсказания** с помощью оригинальной версии рассуждений по аналогии
- **Оценивает правдоподобие** порожденных гипотез с помощью некоторого варианта абдукции Ч.С.Пирса



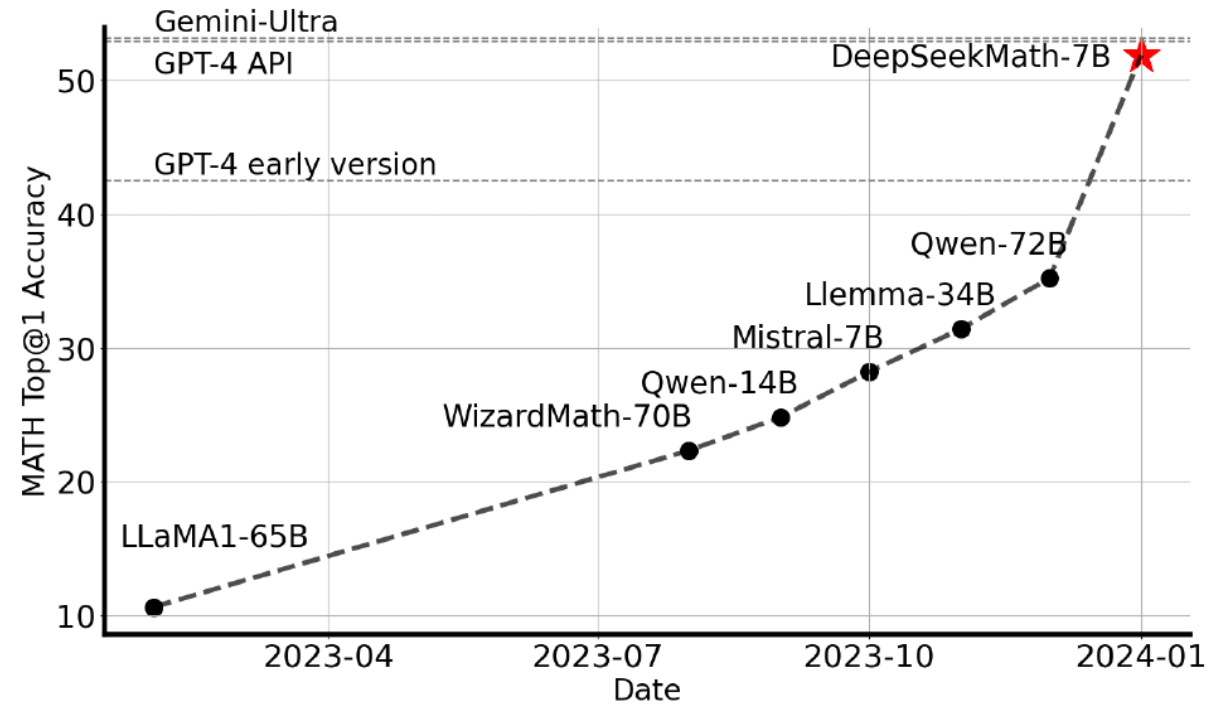
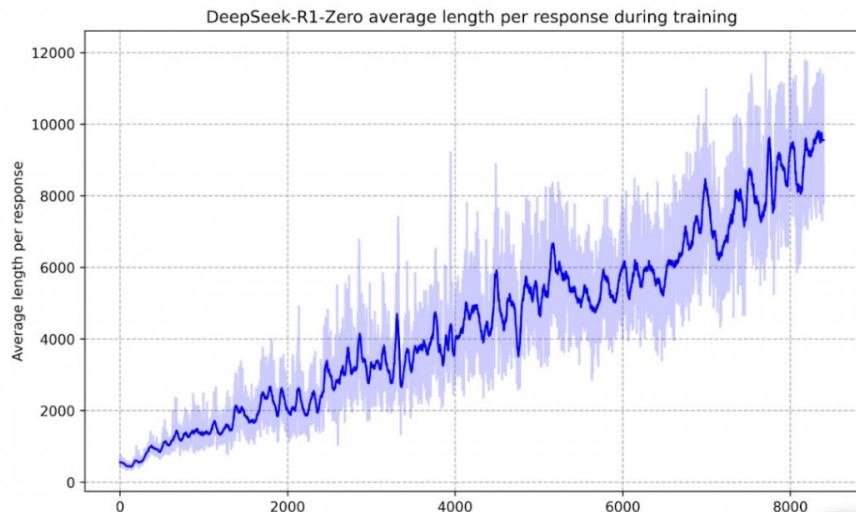
Символьный подход к ИИ-исследователю

- Логический синтез познавательных процедур
- Выявление причинно-следственных связей в том числе на малом количестве данных
- Проблема с взрывом количества гипотез при большом количестве данных
- Вероятностный алгоритм для поиска сходств – ВКФ-метод

Решение математических задач с помощью LLM

Схемы обучения:

- Base+RL – без дополнительной разметки получаем DeepSeek-R1-Zero, плохо интерпретируемый CoT
- Base+SFT^{Cold}+RL^{Reas}+ SFT^{Comb}+ RL^{Pref} – с тремя дополнительными наборами данных получаем DeepSeek-R1 с консистентным CoT
- Вознаграждения генерируются по правилам или по точному соответствию выхода модели GT



Тест Гёделя на доказательство теорем

- Тест Гёделя – проверка системы на возможность сгенерировать доказательство для простых нерешенных задач
- Тестировали GPT-5 на 5 задачах со статьями-подсказками
- Три задачи полностью решены
- Основная сложность – синтез гипотез из двух и более источников

Prompt to GPT-5

A non-negative monotone differentiable function $F: [0, 1]^N \rightarrow \mathbb{R}_{\geq 0}$ is γ -weakly DR-submodular if for every two vectors $\mathbf{x}, \mathbf{y} \in [0, 1]^N$ such that $\mathbf{x} \leq \mathbf{y}$, it holds that

$$\sum_{u \in S} (y_u - x_u) \cdot \frac{\partial F(\mathbf{x})}{\partial x_u} \geq \gamma \cdot [F(\mathbf{y}) - F(\mathbf{x})] .$$

Given such a function which is also L -smooth and a convex set $\mathcal{C} \subseteq [0, 1]^N$. The goal in the current problem is describe an algorithm that finds a vector $\mathbf{z} \in \mathcal{C}$ that (approximately) maximizes F . A good reference for DR-submodular is the paper attached.

Response of GPT-5 (thought for 1m 47s)

Here's a clean, projection-free method that works under your assumptions and gives the usual continuous-greedy/Frank–Wolfe style guarantee, adapted to the γ -weak DR setting.

Algorithm (projection-free continuous greedy / Frank–Wolfe variant)

Input:

- monotone, non-negative, differentiable, L -smooth, γ -weakly DR-submodular $F: [0, 1]^N \rightarrow \mathbb{R}_+$;
- a **down-closed** convex feasible region $\mathcal{C} \subseteq [0, 1]^N$ (if $\mathcal{C}$ is down-closed with lower bound $u \neq 0$, shift to $\mathcal{C}' = \{x \mid x = y - u, y \in \mathcal{C}\}$ and optimize $F'(x) = F(x + u)$);
- number of iterations K ; optional linear-oracle accuracy parameters $\alpha \in (0, 1]$ (multiplicative) and $\delta \geq 0$ (additive).

Oracle: ability to (approximately) solve

$$v_k \in \arg \max_{v \in \mathcal{C}} \langle v, \nabla F(x_k) \rangle.$$

Initialize: $x_0 \leftarrow 0$.

For $k = 0, 1, \dots, K - 1$:

1. Compute gradient $g_k = \nabla F(x_k)$.
2. (Linear oracle) Find $v_k \in \mathcal{C}$ such that

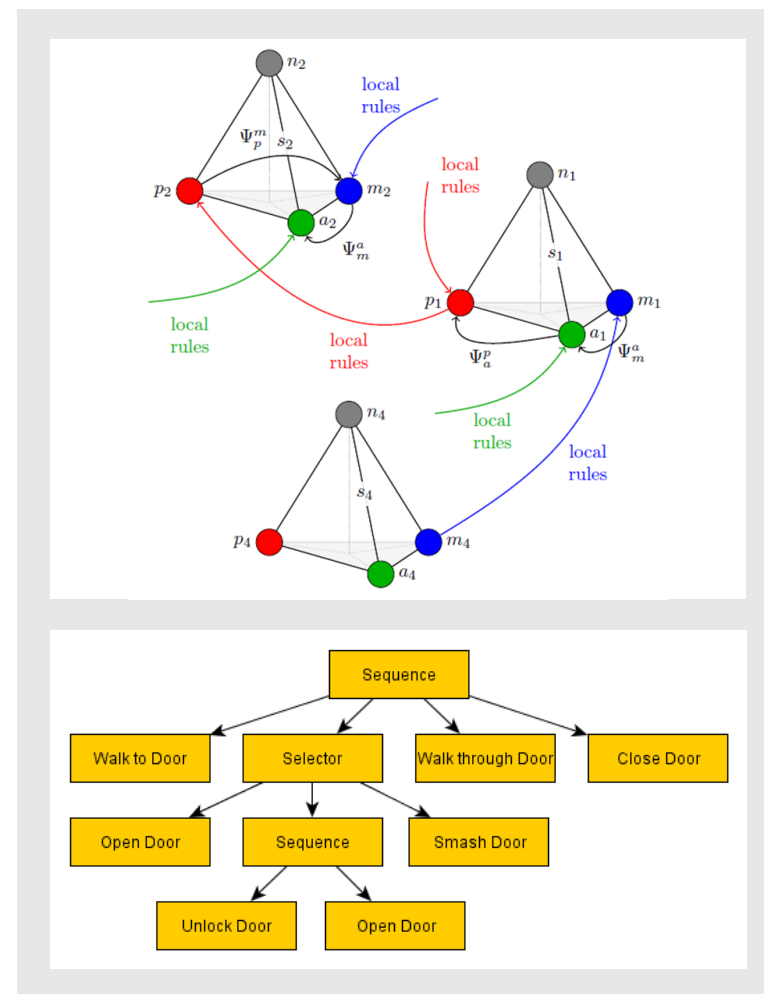
$$\langle v_k, g_k \rangle \geq \alpha \cdot \max_{v \in \mathcal{C}} \langle v, g_k \rangle - \frac{1}{2} \delta L.$$

03

ИИ-ассистенты исследователя

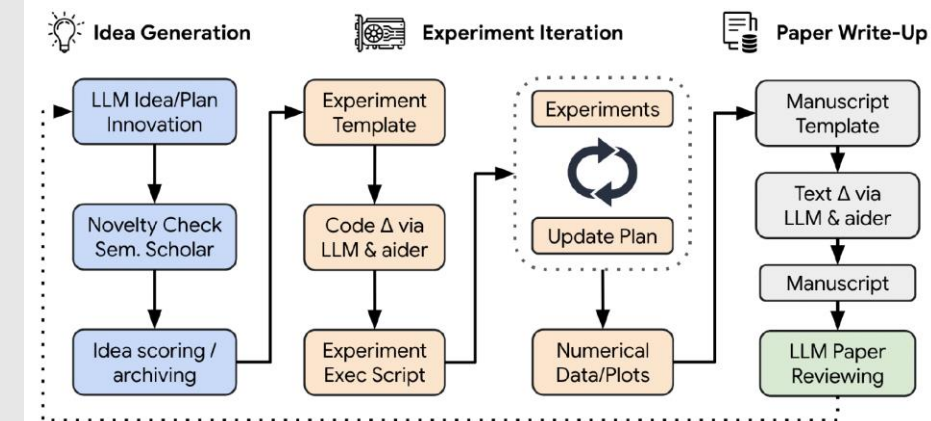
Научная и житейская картины мира

- Знаковая картина мира – представления знаний с помощью локально-организованных объектных структур
- Самоорганизация в картине мира – планирование, целеполагание, обучение
- В зависимости от преобладающей компоненты знака выделяют три типа картины мира: рациональную, житейскую и мифологическую
- Рациональная (научная картина) мира – ведущие операции обобщения и классификации, формирования образа
- LLM – аппроксиматоры значений и сценариев в картине мира



The AI Scientist: Open-Ended Scientific Discovery

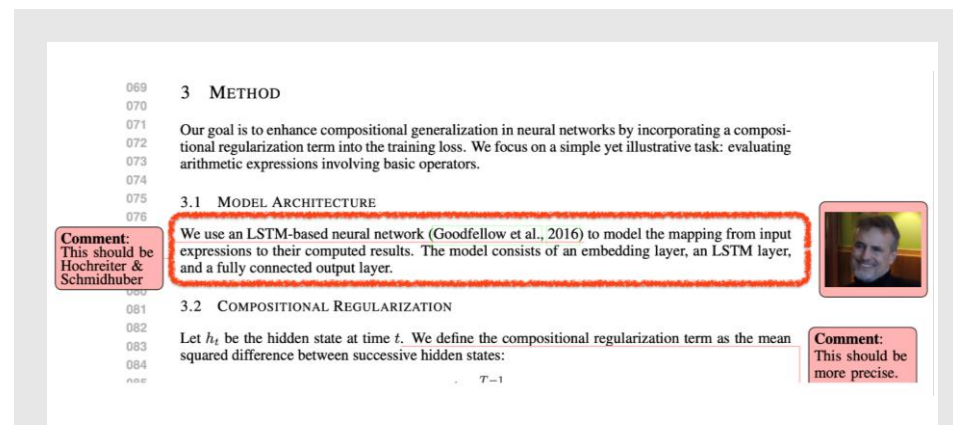
- Первый фреймворк **полной автоматизации научного исследования** с использованием LLM
- **The AI Scientist** генерирует исследовательские идеи, пишет код, проводит эксперименты, визуализирует результаты, описывает их, делает выводы и оформляет это в виде текста научной статьи
- Отдельно реализован процесс **рецензирования** для оценка качества работы
- Цена каждой такой статьи получается менее \$15



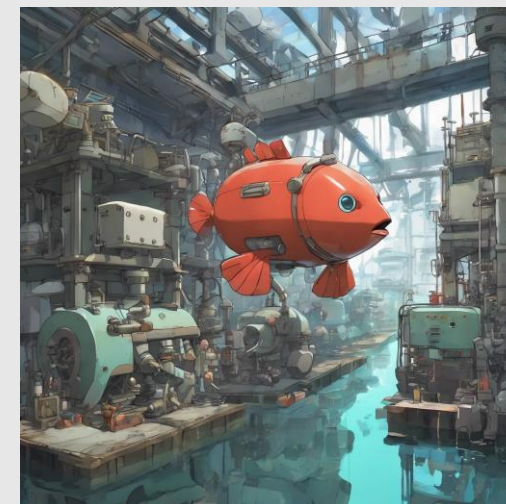
	Reviewer	Balanced Acc. ↑	Accuracy ↑	F1 Score ↑	AUC ↑	FPR ↓	FNR ↓
Uncalibrated	Human (NeurIPS) ¹	0.66	0.73	0.49	0.65	0.17	0.52
	Random Decision	0.50	0.50	0.40	0.50	0.50	0.50
	Always Reject	0.50	0.59	0.00	0.50	0.00	1.00
	Sonnet 3.5	0.52 ± 0.01	0.40 ± 0.01	0.55 ± 0.01	0.52 ± 0.01	0.95 ± 0.02	0.00 ± 0.00
	GPT-4o-mini	0.53 ± 0.02	0.65 ± 0.01	0.11 ± 0.06	0.53 ± 0.02	0.01 ± 0.01	0.94 ± 0.04
	GPT-4o (0-shot)	0.61 ± 0.04	0.68 ± 0.03	0.43 ± 0.07	0.61 ± 0.04	0.11 ± 0.03	0.67 ± 0.07
Calibrated	GPT-4o (1-shot)	0.60 ± 0.03	0.70 ± 0.03	0.37 ± 0.08	0.60 ± 0.03	0.04 ± 0.02	0.76 ± 0.06
	Sonnet 3.5 @8	0.59 ± 0.04	0.65 ± 0.04	0.45 ± 0.06	0.59 ± 0.04	0.20 ± 0.04	0.61 ± 0.07
	GPT-4o-mini @6	0.59 ± 0.04	0.64 ± 0.04	0.45 ± 0.06	0.59 ± 0.04	0.22 ± 0.05	0.60 ± 0.07
	GPT-4o (0-shot) @6	0.63 ± 0.04	0.63 ± 0.04	0.56 ± 0.05	0.63 ± 0.04	0.38 ± 0.05	0.36 ± 0.07
	GPT-4o (1-shot) @6	0.65 ± 0.04	0.66 ± 0.04	0.57 ± 0.05	0.65 ± 0.04	0.31 ± 0.05	0.39 ± 0.07

Тест Тьюринга на публикацию

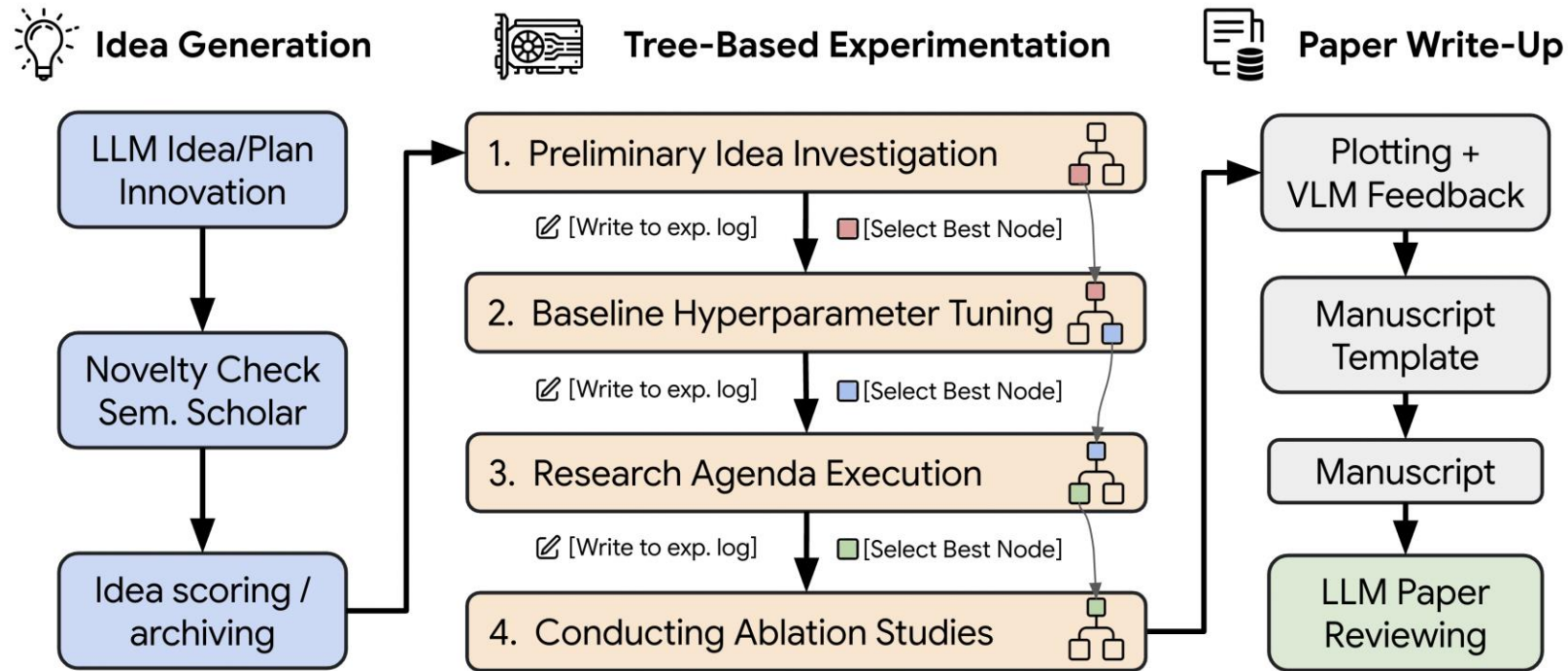
- Воршоп «I Can't Believe It's Not Better» на ICLR 2025
- Работы (идея, текст, эксперименты, выводы) полностью сгенерированы с помощью [The AI Scientist-v2](#)
- Одна из статей получила положительные рецензии и **была принята на воркшоп**



Title	ICLR Workshop Scores	Sakana AI Human Reviewer	AI Reviewer
Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization	6, 7, 6	Review	Review
Real-World Challenges in Pest Detection Using Deep Learning: An Investigation into Failures and Solutions	3, 7, 4	Review	Review
Unveiling the Impact of Label Noise on Model Calibration in Deep Learning	3, 3, 3	Review	Review



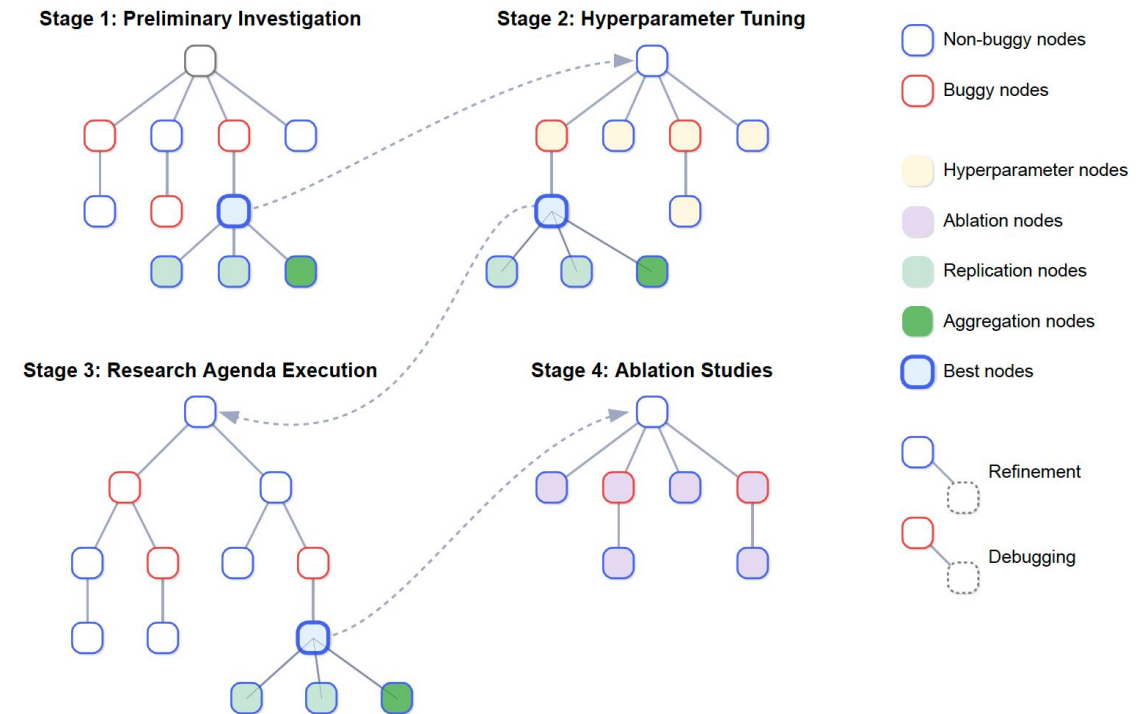
The AI-Scientist v2



- На этапе генерации обобщенной идеи в цикл включаются инструменты анализа литературы, такие как Semantic Scholar - система может запрашивать базу данных литературы в процессе формулирования идеи, чтобы оценить новизну предлагаемой концепции и выявить соответствующие предыдущие работы

Агентный поиск по дереву

- **Предварительное исследование:** Определение первоначальной осуществимости и корректности с помощью минимального рабочего прототипа, основанного на сформированной исследовательской идее
- **Настройка гиперпараметров:** Уточнение первоначальной реализации путем оптимизации критических гиперпараметров
- **Выполнение программы исследований:** Систематическая реализация основной программы исследований на основе настроенного базового метода
- **Абляционных исследований:** Систематическая оценка важности различных компонентов исследования, обеспечение строгой поддержки основных экспериментальных результатов
- LLM генерирует как конкретный план эксперимента, так и связанный с ним код на Python для реализации эксперимента, сгенерированный код немедленно выполняется в интерпретаторе Python



Типичные затравки исследователя

U1

Проверка орфографии в статье — переписывание абзацев и целых разделов

U2

Выявление слабых мест уже написанной статьи с помощью LLM — генерация рецензии или советов по улучшению теории/экспериментов

U3

Улучшение названия статьи с помощью LLM — по полному тексту или по абстракту

U4

Улучшение обзора существующих методов или генерация полного раздела со ссылками

Deep research от OpenAI

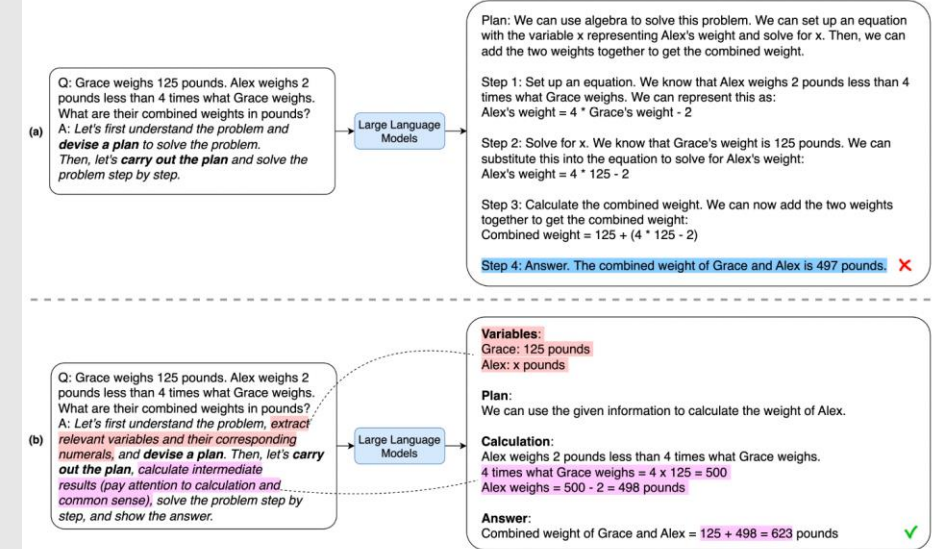
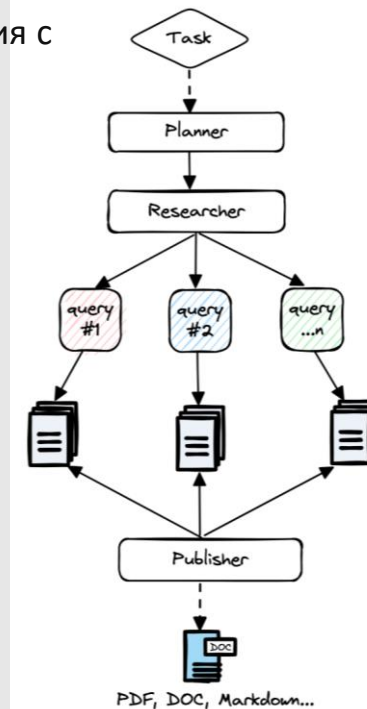
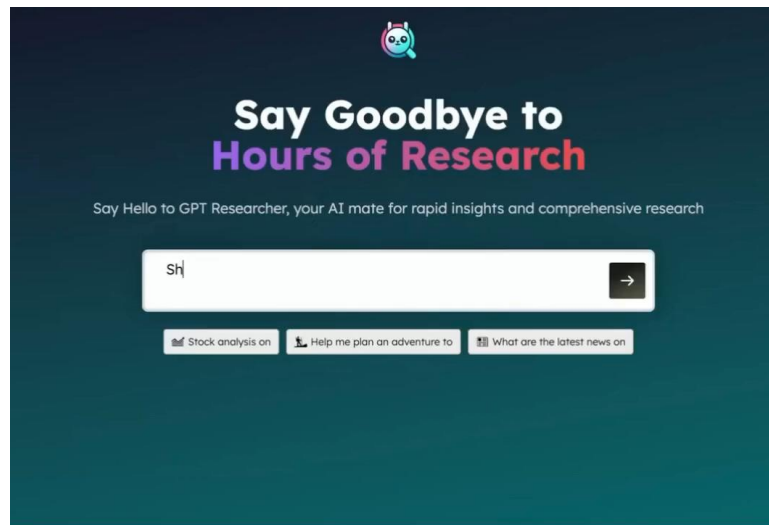
- Агент использует актуальную онлайн информацию и может выполнять многошаговые исследовательские задания
- Агент использует доступ к ChatGPT и занимает от 5 до 30 минуты поиска и синтеза отчета
- GPT-4o лучше работает в реальном времени и с мультимодальной информацией

Model	Accuracy (%)
GPT-4o	3.3
Grok-2	3.8
Claude 3.5 Sonnet	4.3
Gemini Thinking	6.2
OpenAI o1	9.1
DeepSeek-R1*	9.4
OpenAI o3-mini (medium)*	10.5
OpenAI o3-mini (high)*	13.0
OpenAI deep research**	26.6

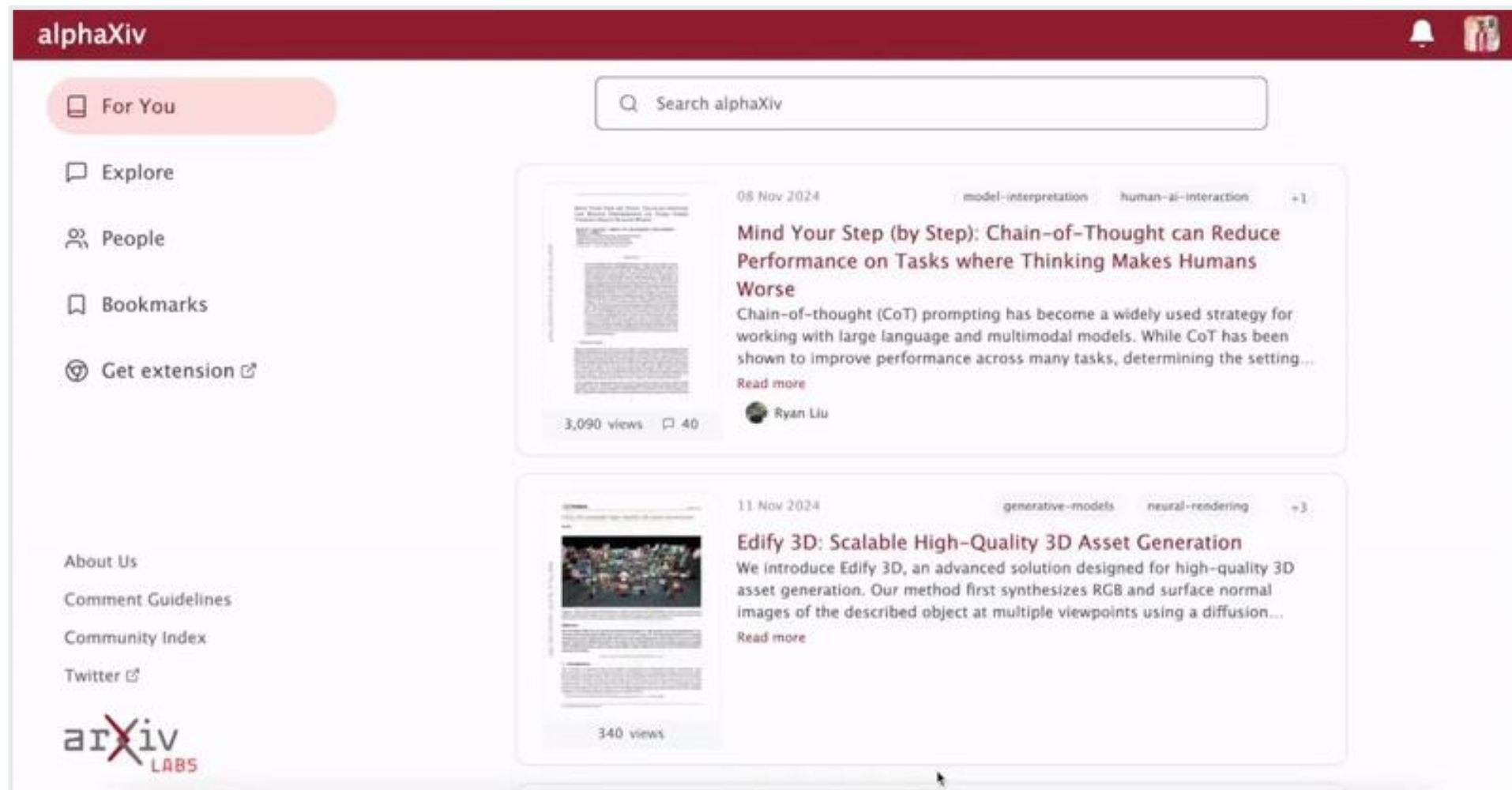


GPT-researcher

- Используются два типа агентов — генерирующий вопросы и собирающий релевантную информацию
- После сбора информации проводится суммаризация с сохранением ссылок на первоисточники
- Фильтрация и агрегация нескольких отчетов в один



AlphaXiv – суммаризатор статей с Arxiv

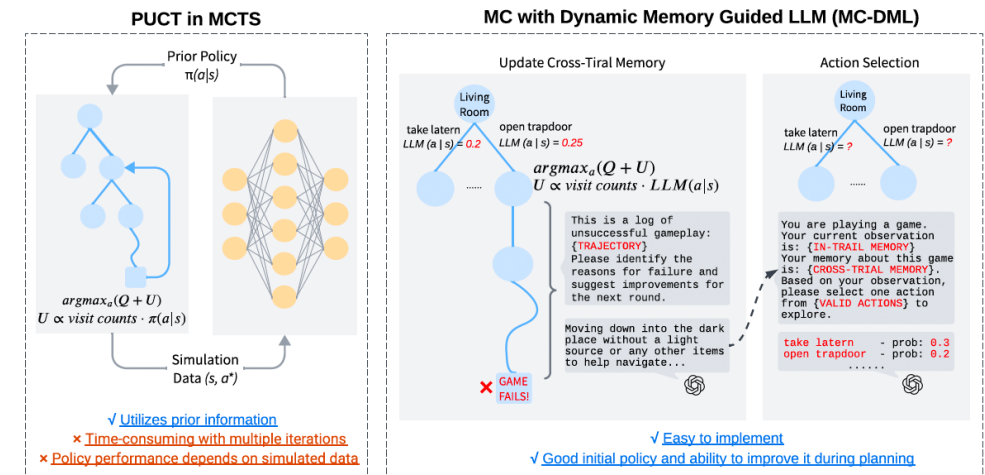
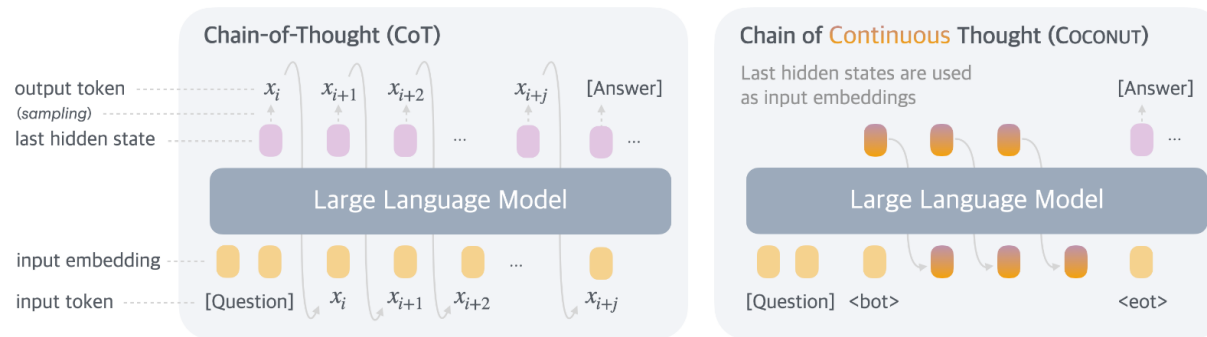
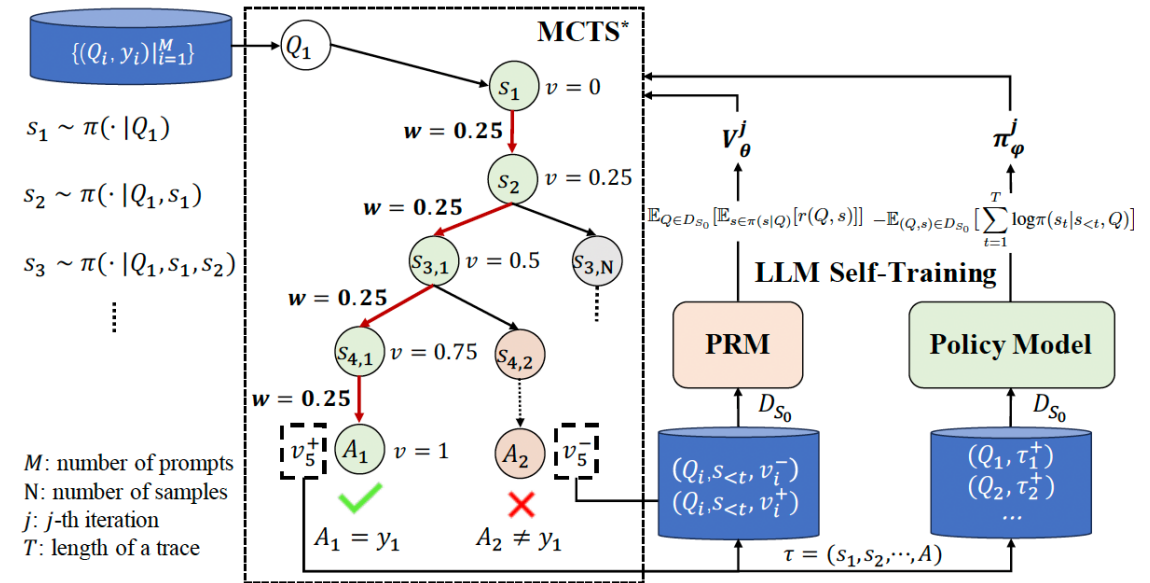


04

ОСНОВНЫЕ ВЫЗОВЫ НА ПУТИ К AIScientist

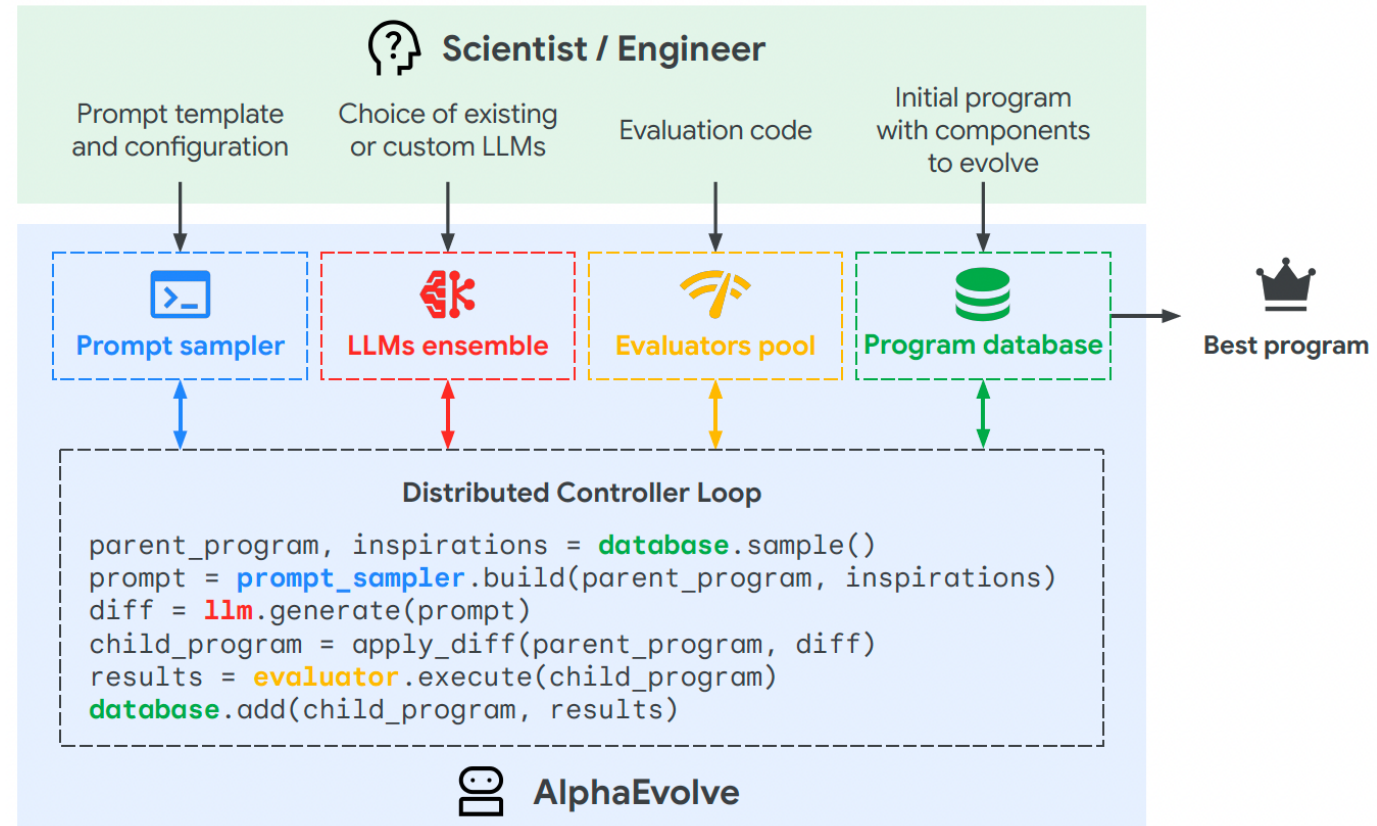
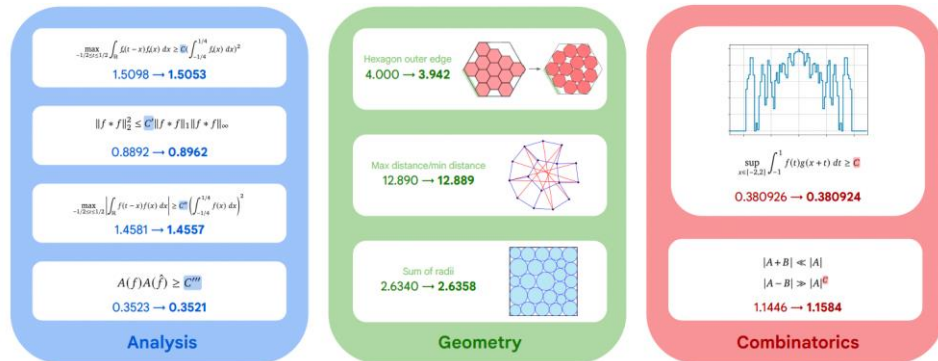
Вызов 1: нейросимвольная интеграция

- Интеграция методов достоверного символического рассуждения с нейросетевыми обучающимися моделями
- Символьные токены вкладываются в непрерывное векторное пространство
- Переход от генерации цепочки шагов к генерации гипотез и их фальсификации
- Интеграция MCTS поиска и логического вывода



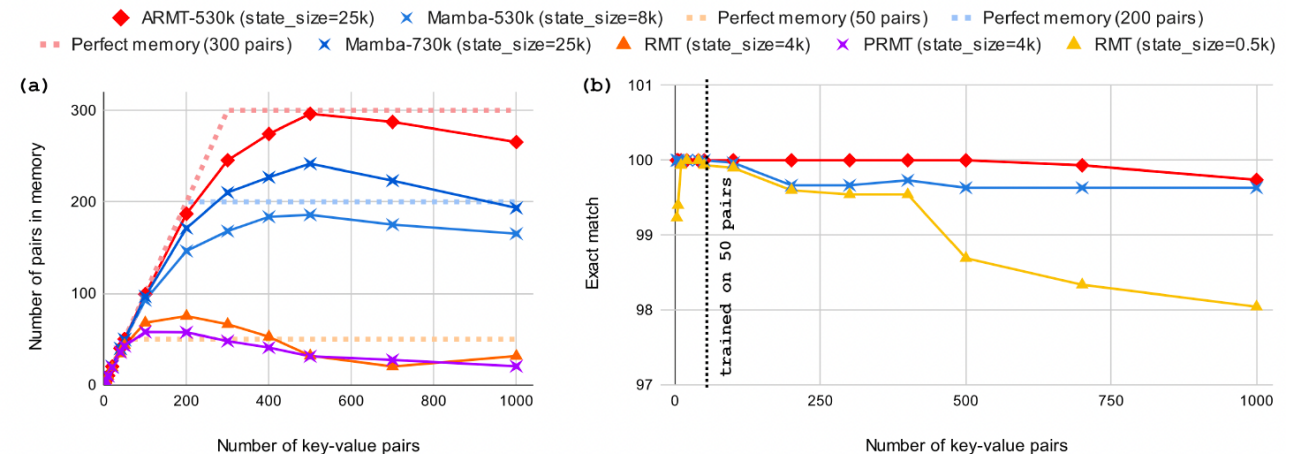
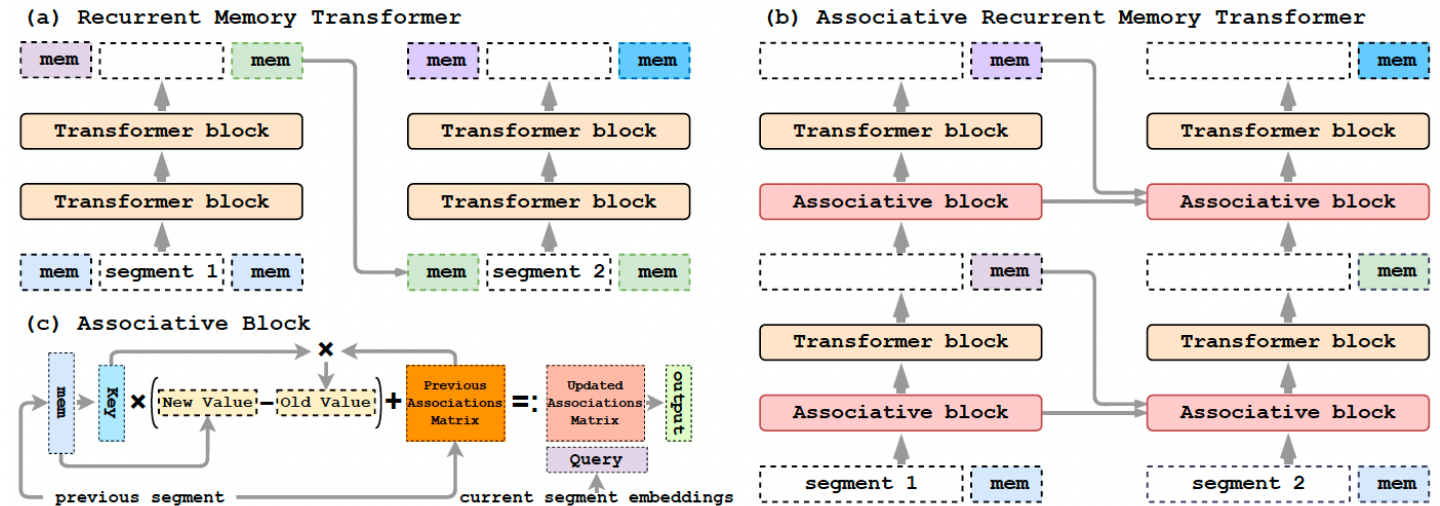
Вызов 2: эволюция программ

- AlphaEvolve – эволюционирующий агент, который пишет код
- Эволюция с обратной связью от оценщиков
- Проблема – затухание эволюции при переборе гиперпараметров
- Качественная стартовая точка эволюции



Вызов 3: память для долгосрочного обучения

- *RMT модели – интеграция рекуррентной память в архитектуру трансформера
- Требуется сложная процедура обучения с расписанием
- Необходимо специально готовить данные для предобучения, позволяя модели учиться пользоваться памятью



Вызов 4: пост-трансформеры

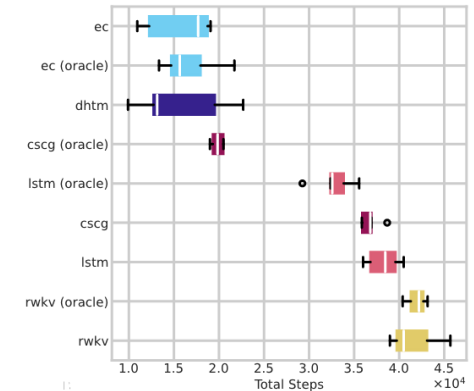
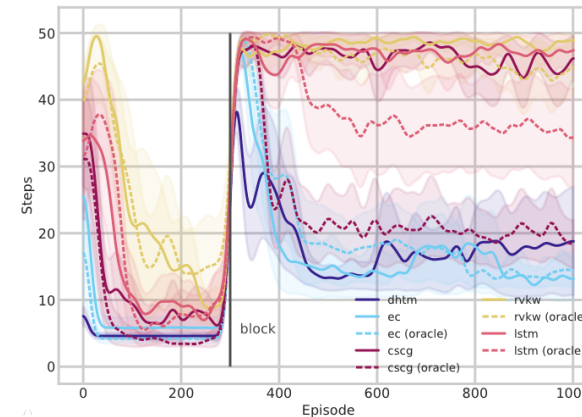
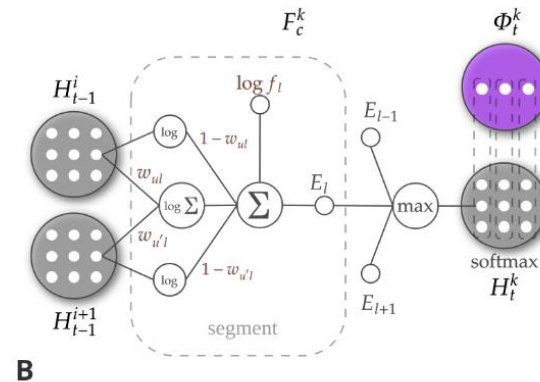
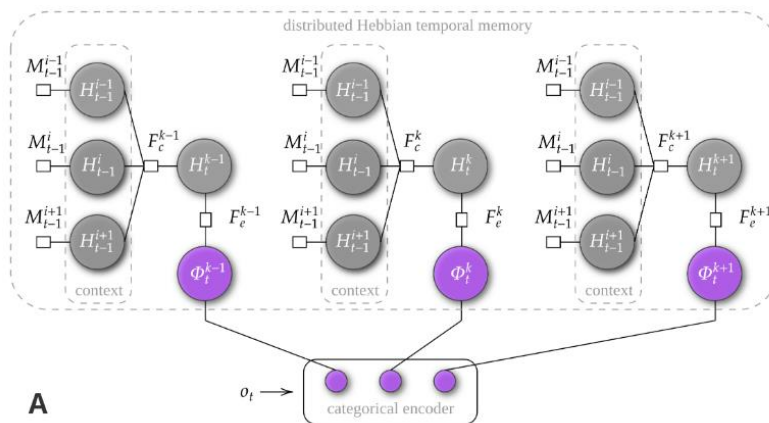
- Поиск новых архитектур с использованием нейрофизиологических концепций работы мозга
- Основанный на нейрофизиологических моделях неокортекса, DHTM использует распределенные представления, разреженные матрицы переходов и локальные хеббовские правила обучения
- DHTM превосходит LSTM, RWKV и биологически правдоподобные алгоритмы HMM и CSCG на нестационарных наборах данных

Algorithm 1 General agent training procedure

```

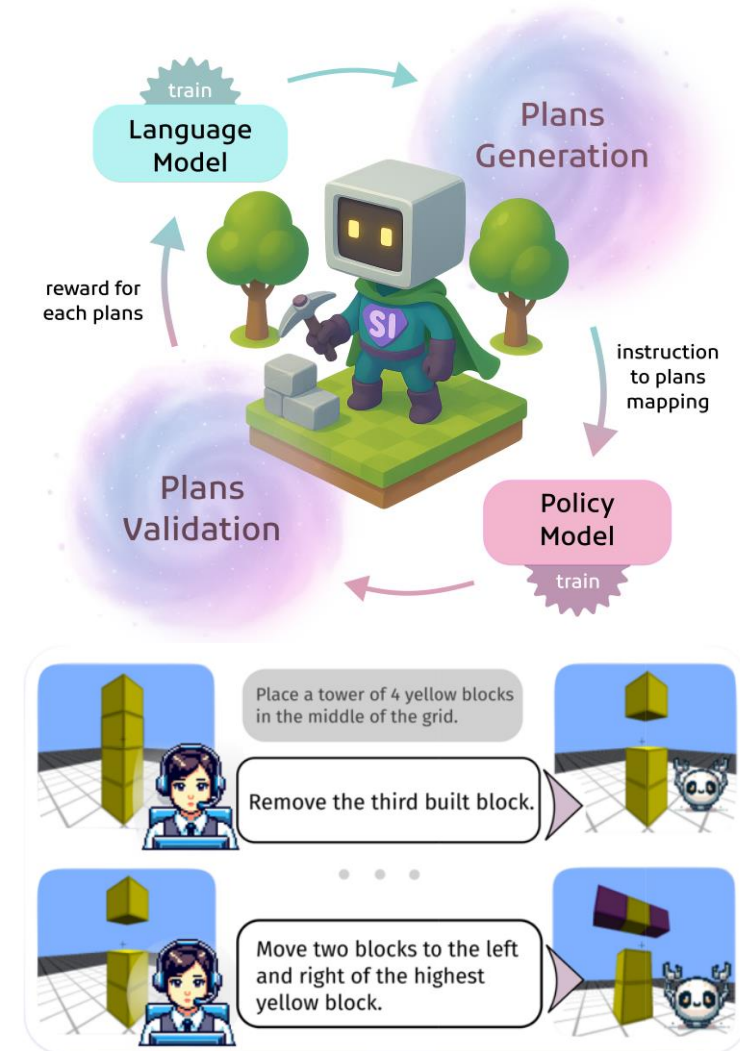
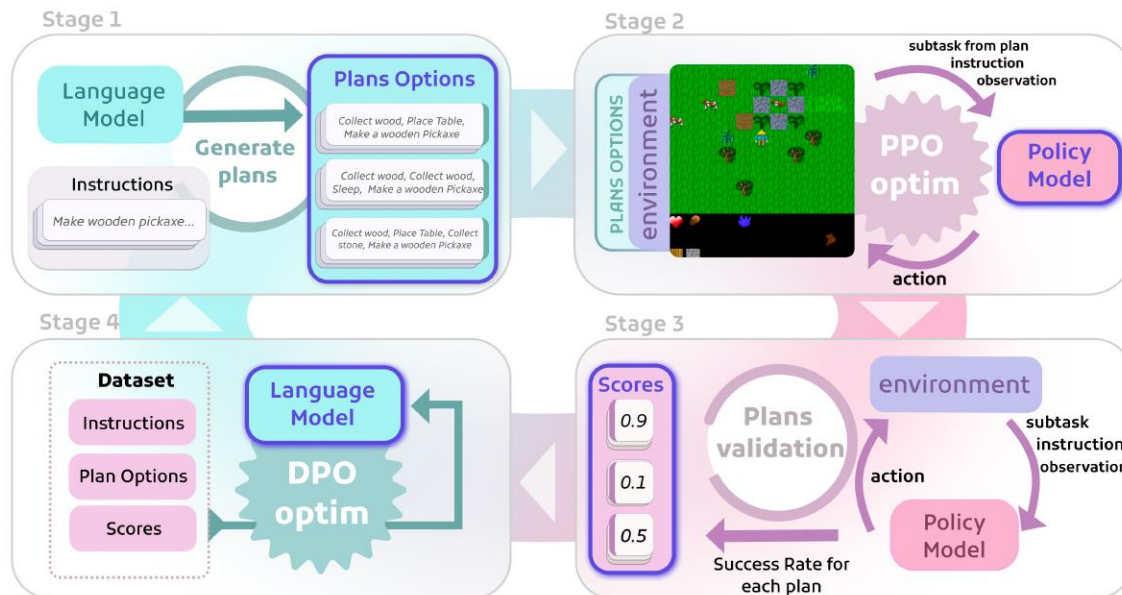
1: for episode=1..n do
2:   RESET_MEMORY()
3:   action ← initial_action
4:   while (not terminal) and (steps < max_steps) do
5:     obs, reward ← STEP()
6:     features ← ENCODE(PREPROCESS(obs))
7:     OBSERVE(features, action)
8:     REINFORCE(reward, features)
9:     action ← SAMPLE_ACTION()
10:    ACT(action)
11:  end while
12: end for

```



Вызов 5: мультимодальное экспериментирование

- Агент на основе языковой модели тестирует свои планы в интерактивной среде
- Концепция среды: архитектор дает инструкцию и комментирует действия агента, которые старается выполнить выданную задачу
- Итеративное обучение архитектора и агента



05

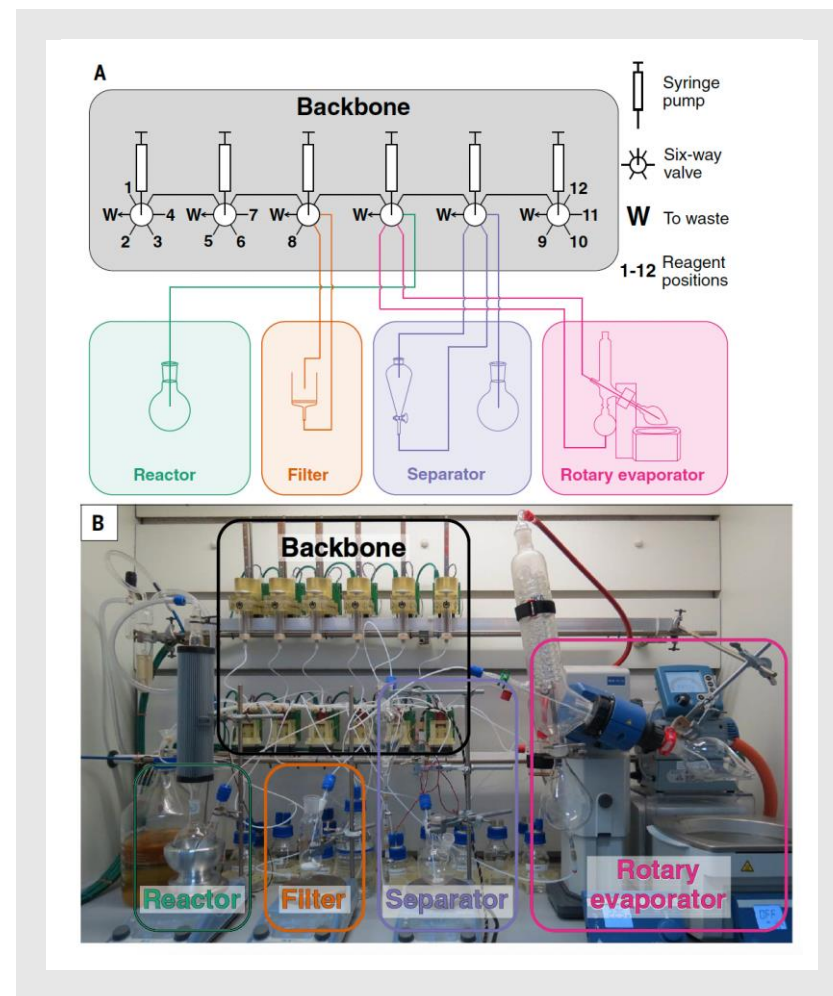
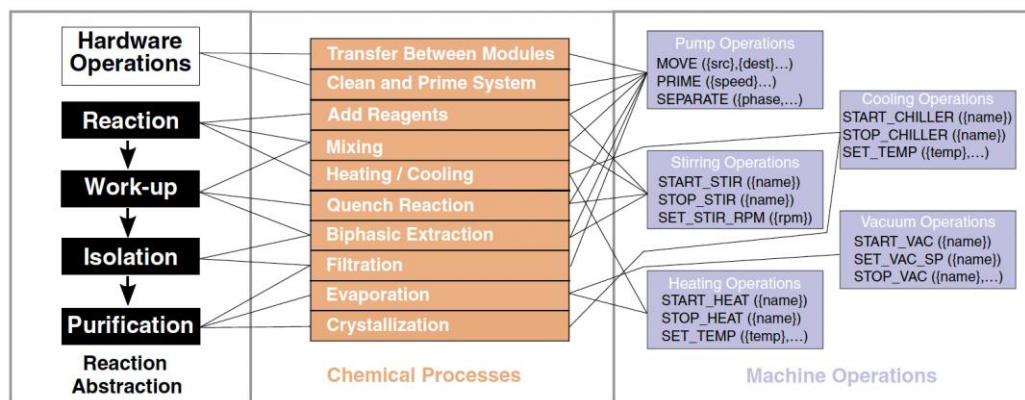
ИИ-ассистенты исследователя-
лаборанта

Воплощенный ИИ-ассистент исследователя



Chemputer и Chempiler

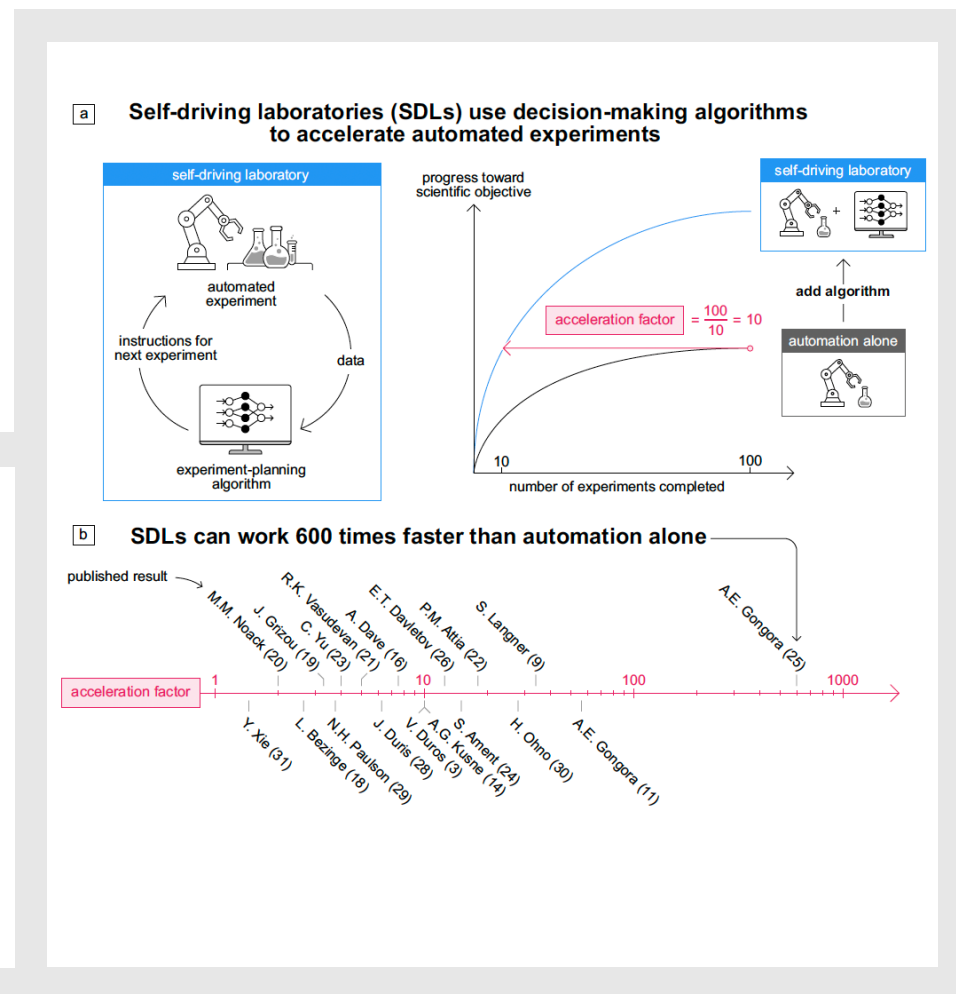
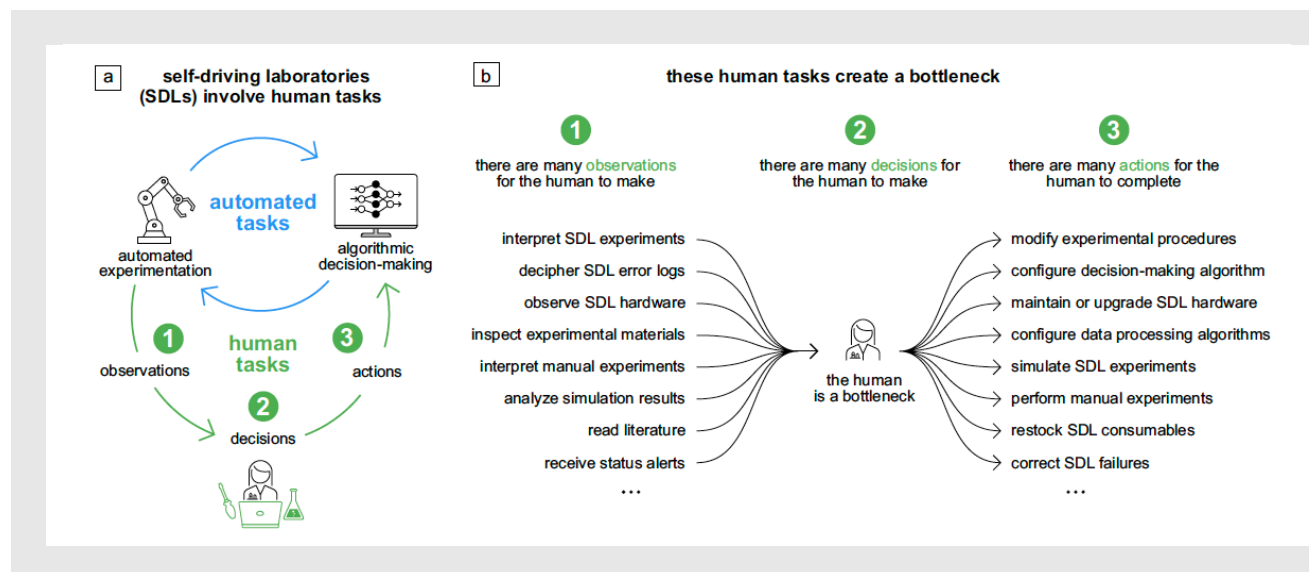
- Автоматизация химического синтеза ограничена специализированными промышленными процессами
- Автоматизированная система Chemputer предназначена для синтеза целевых молекул
- Chempiler — программа, генерирующая для Chemputer низко-уровневые инструкции, выполняемые лабораторными синтезирующим роботом
- Chempiler собирает информацию о физических связях и составе автоматизированной платформы в виде графа и объединяет ее с аппаратно-независимым языком сценариев



Самоуправляемые лаборатории

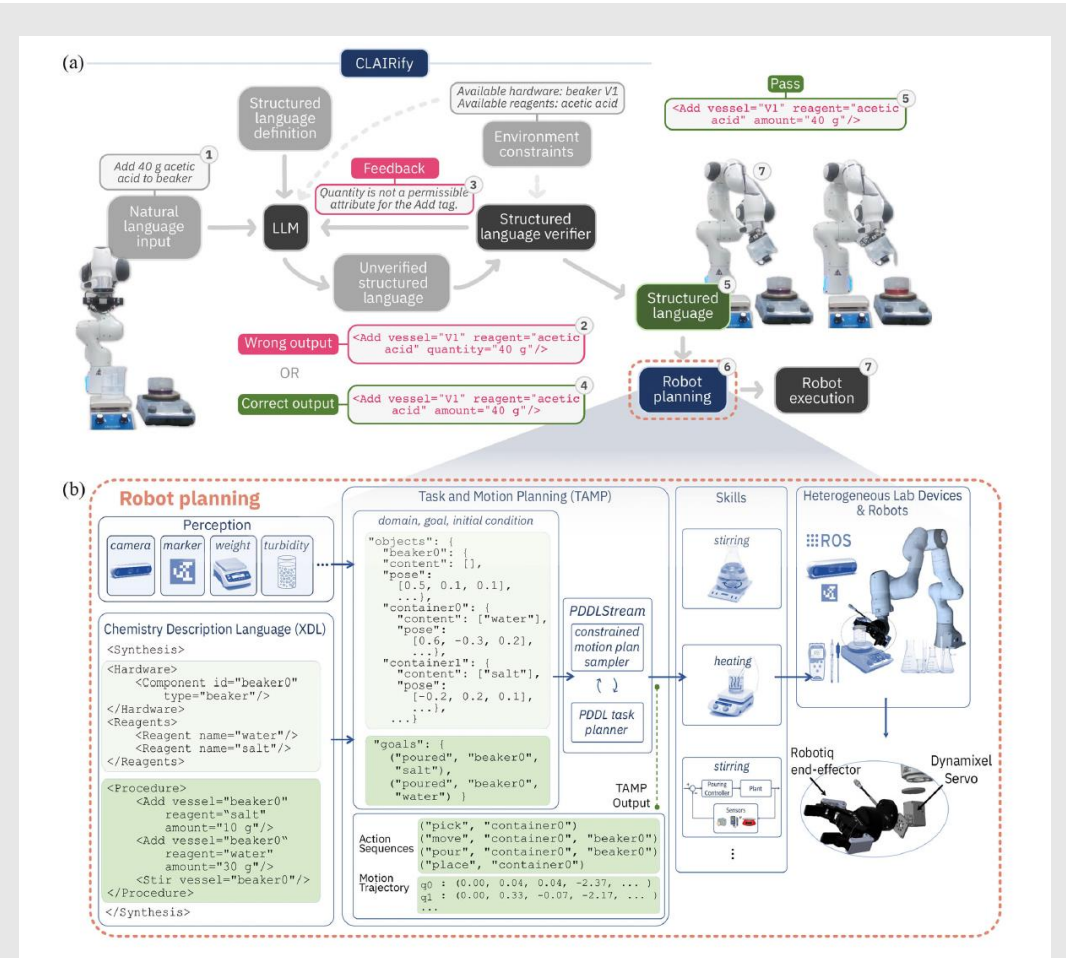
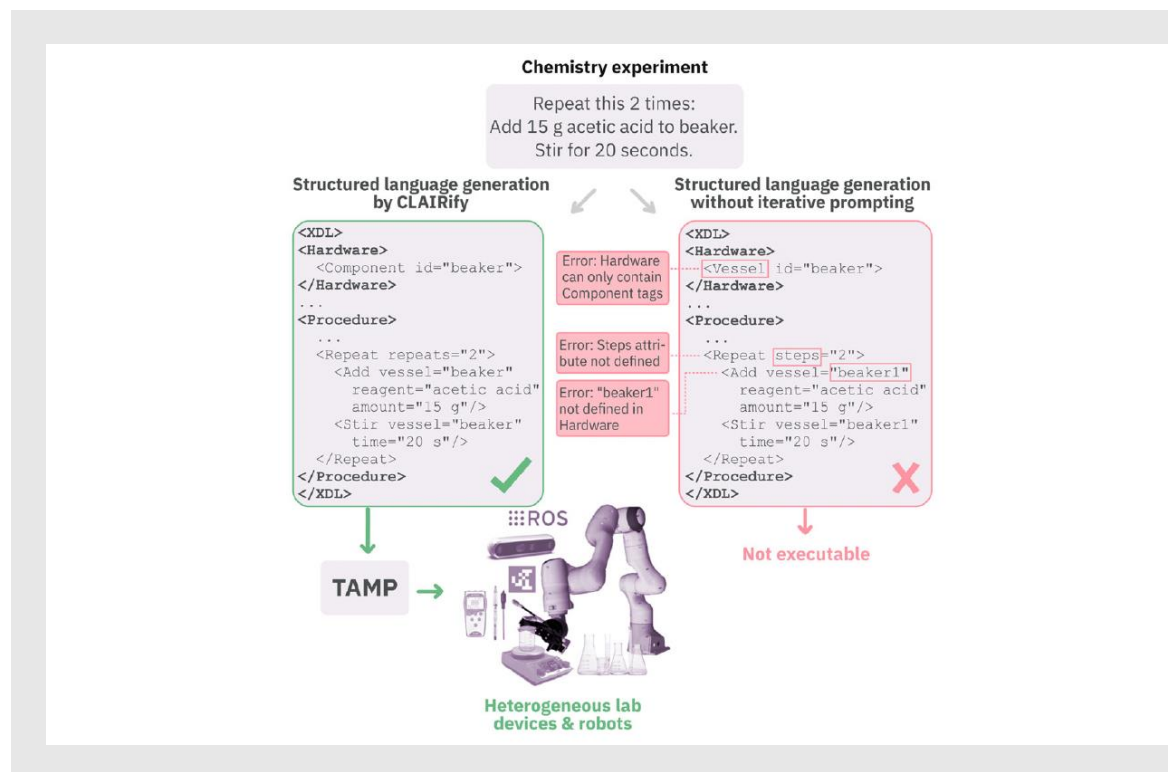
Эффективные автономные лаборатории должны:

- работать быстрее, чем просто автоматизированный процесс
- должны быть легко адаптируемыми к новым исследованиям



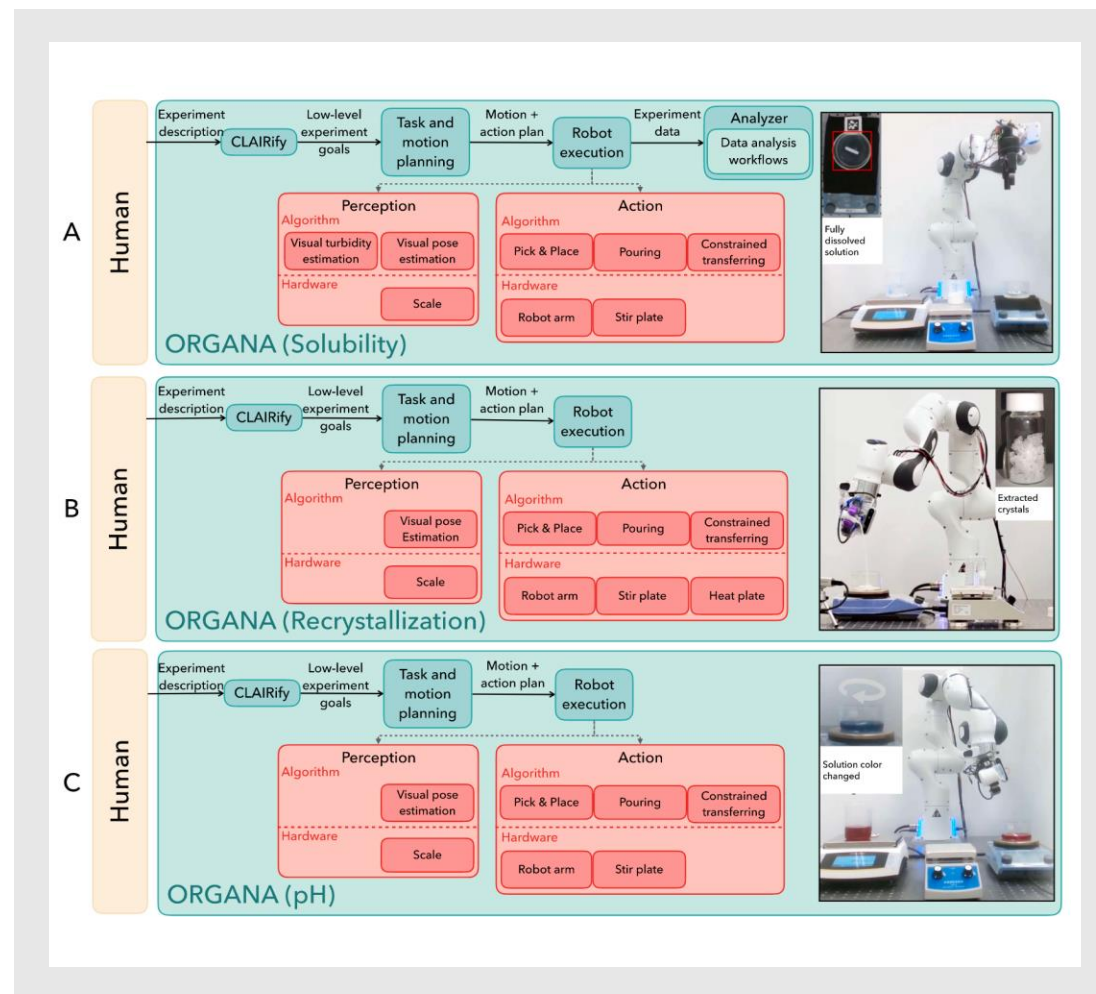
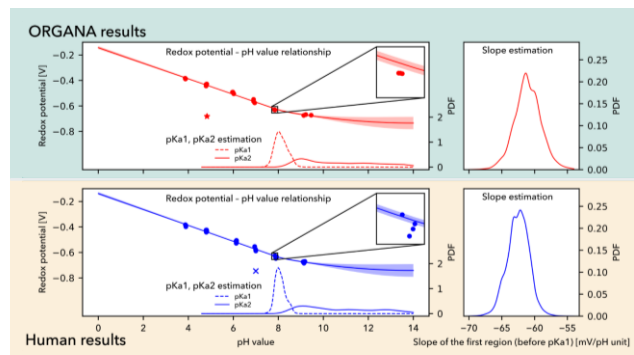
CLAIRify: языковые модели для химических роботов

- CLAIRify комбинирует автоматический итеративный промптинг с верификацией программы выполнения
- Планирование действий на PDDLStream

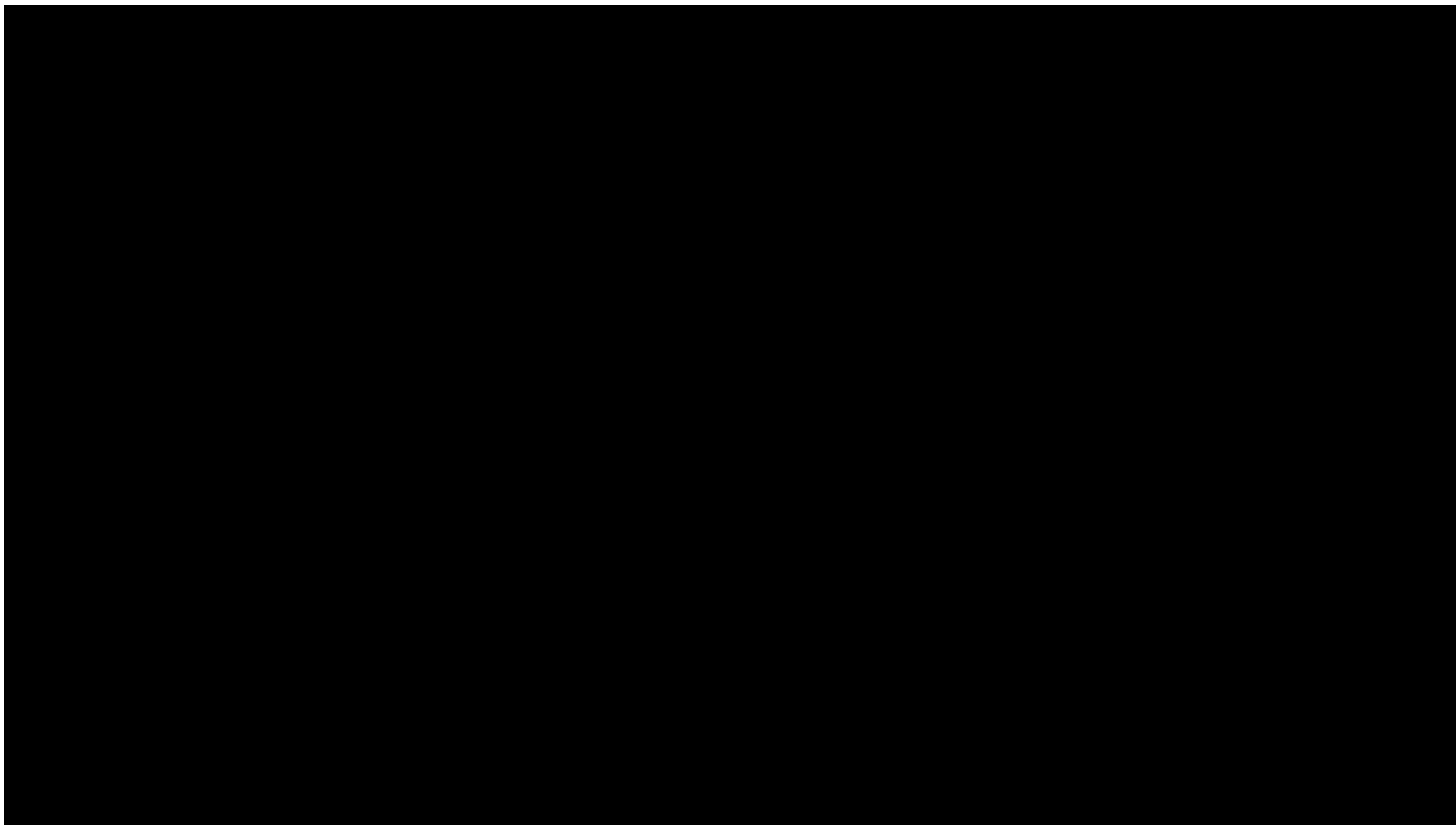


ORGANA: робототехнический химический ассистент

- Адаптация к новому протоколу экспериментов — сложный процесс
- Organa – роботизированная система, которая автоматизирует различные химические эксперименты и принимает решения совместно с химиками
- LLM используется для определения целей эксперимента, устранения неоднозначности и генерации журнала экспериментов.
- Organa планирует и выполняет сложные задачи с визуальной обратной связью, поддерживая планирование и параллельное выполнение задач



Автоматизированная лаборатория Insilico



Задачи для развития

U1

Генератор действий работает с классическими планировщиками типа PDDL → необходим переход к современным **VLA моделям**

U2

Поддержка в основном стационарных манипуляторов без навигации
в человеко-ориентированной лаборатории → использование **мобильных** манипуляторов

U3

Неполная поддержка коллаборативных сценариев и интерпретация сцен → работа с **мультимодальными** моделями анализа сцен

U4

Слабая интеграция с моделями генерации протоколов испытаний и новых гипотез → связка с **рассуждающими** языковыми моделями

Лаборатория – промежуточный уровень реалистичности



Специально спроектированная
детерминированная полностью
наблюдаемая среда



Лабораторная коллаборативная
среда для проведения научных
экспериментов



Человеко-ориентированная
недетерминированная
частично-наблюдаемая среда



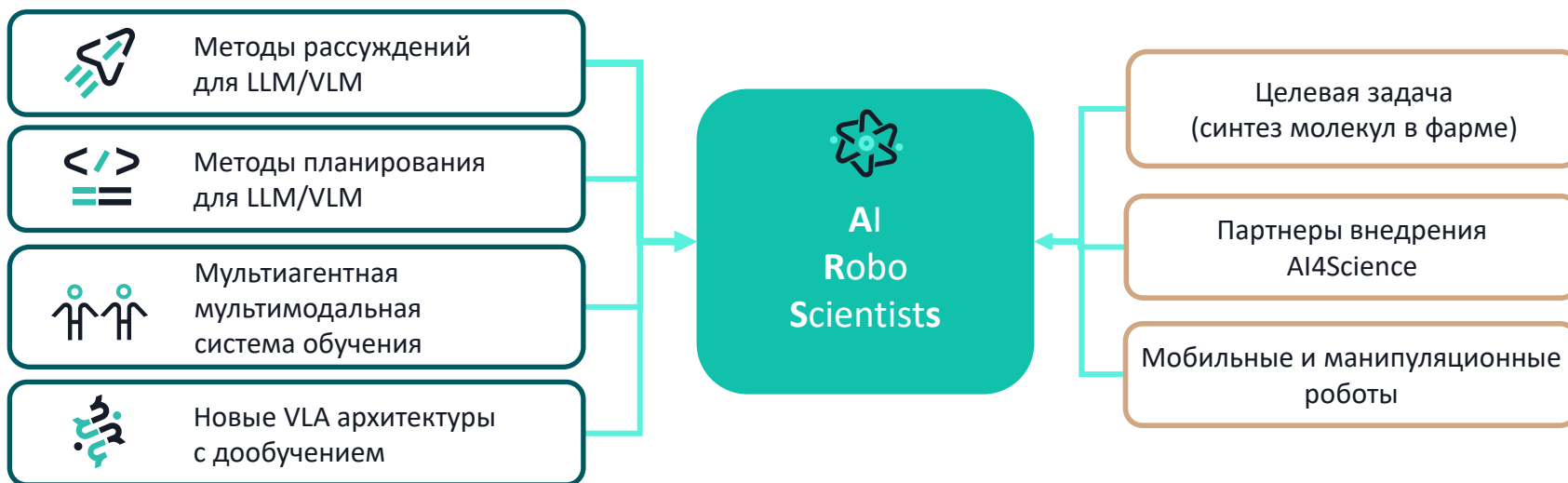
AIRoboScientists – планирование и проведение научных экспериментов в лаборатории

Проблема:

Проведение **реальных «мокрых» экспериментов в лаборатории** — один из самых дорогостоящих этапов научного исследования

Решение:

Использовать концепцию **AIScientist** совместно с робототехническим воплощение для сети лабораторий

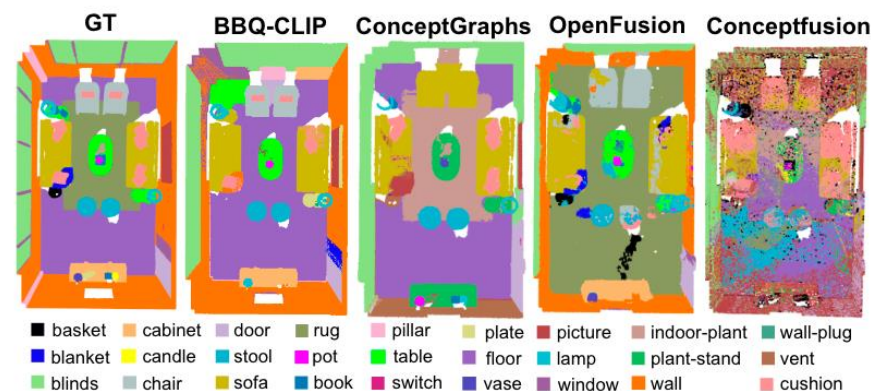
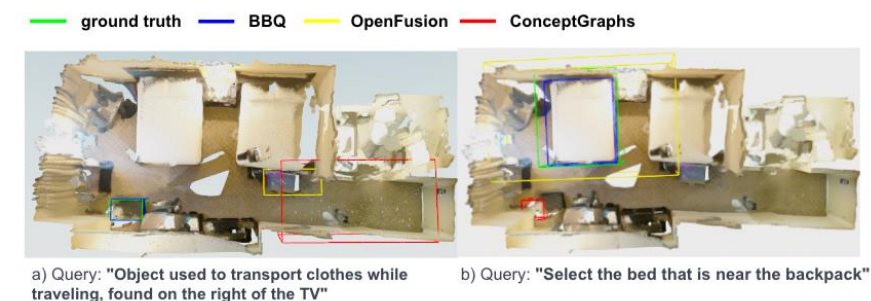
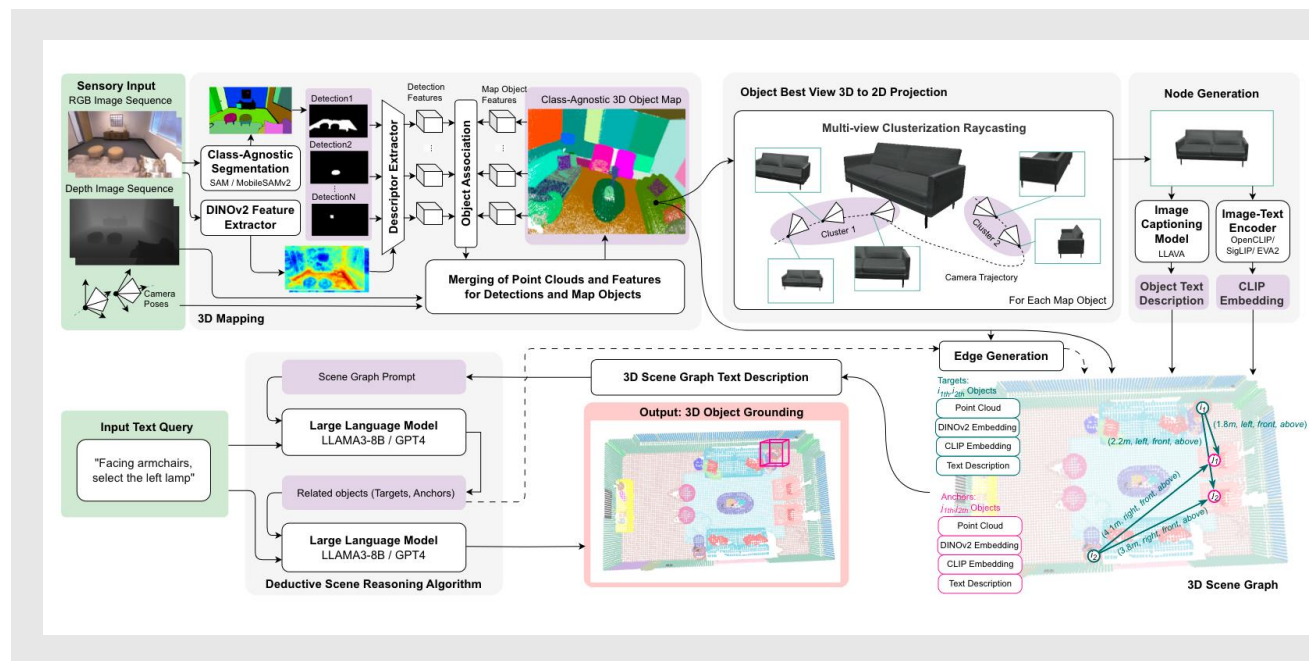


Прототип на ФБТ 2025 совместно с AI4Science Сбера



Интерпретация и понимание визуальной сцены

Развитие мультимодальной модели **BBQ** для ответов на вопросы и интерпретацию сцены



Интерпретация и понимание визуальной сцены

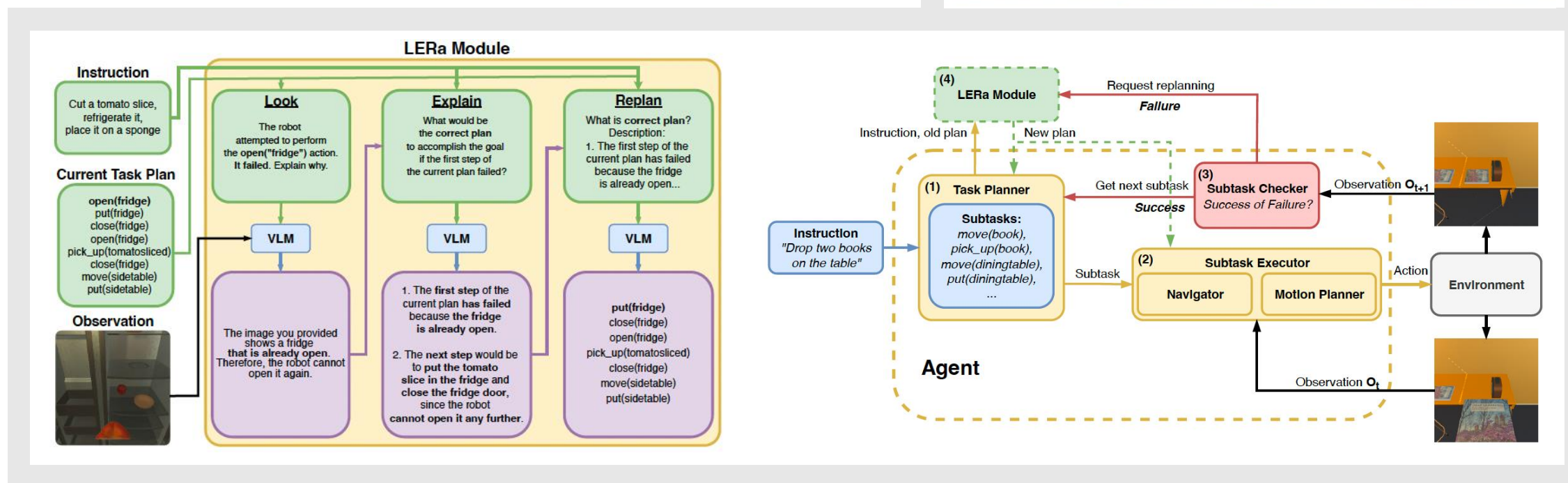
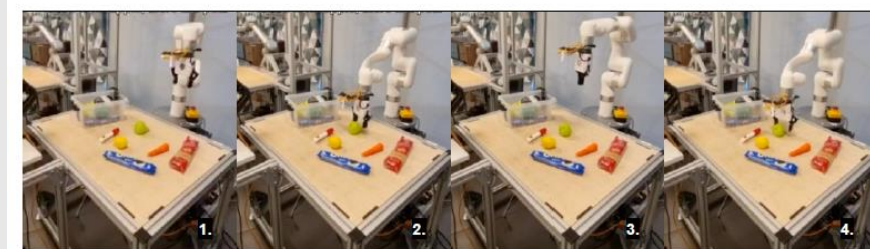


Beyond *Bare* Queries: Open-Vocabulary Object Grounding with 3D Scene Graph

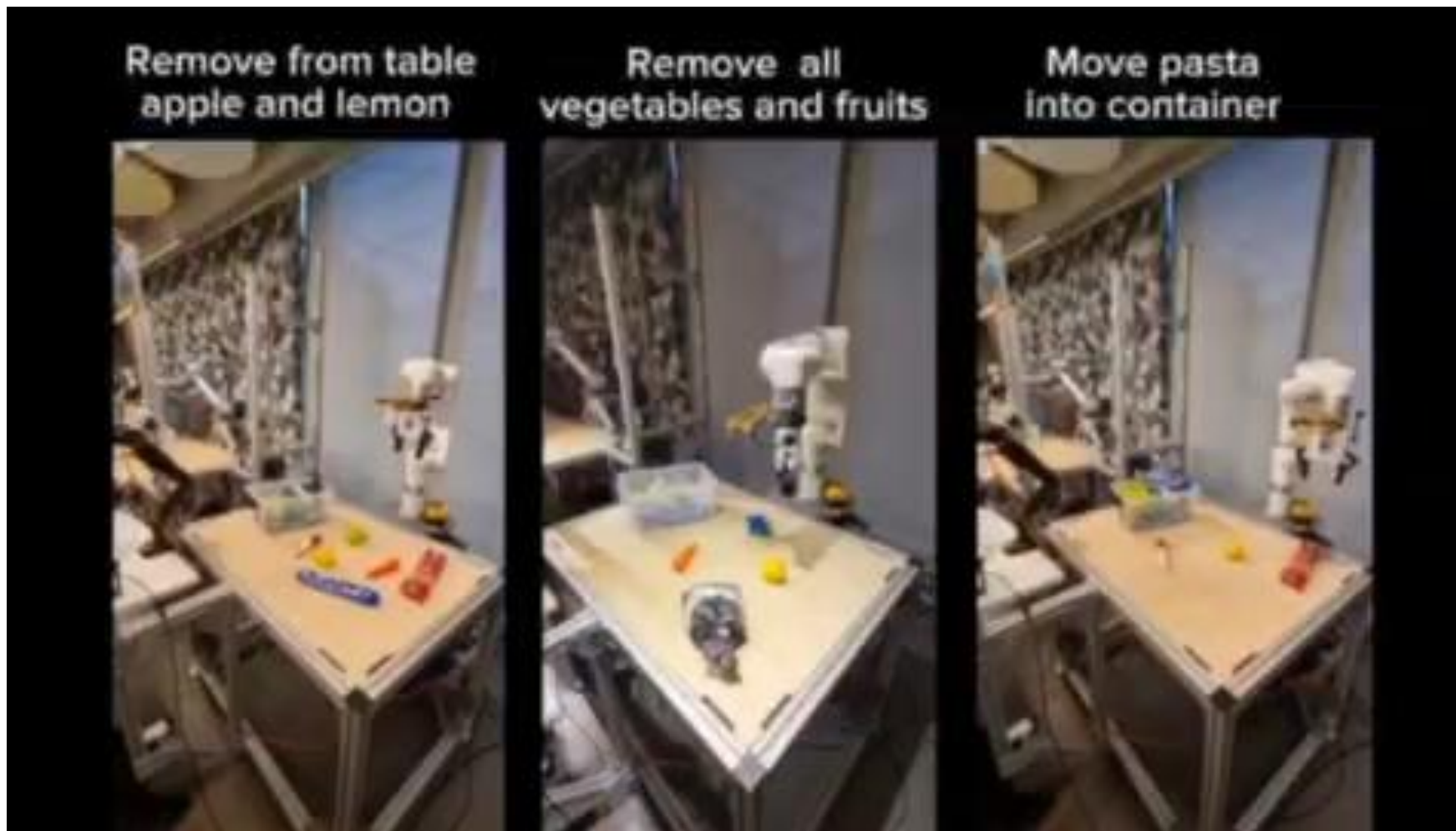
Sergey Linok Tatiana Zemskova
Svetlana Ladanova Roman Titkov Dmitry Yudin
Maxim Monastyrny Aleksei Valenkov

Визуальная обратная связь от среды

Выполнение инструкций человека с анализом текущей ситуации с использованием модели LERa

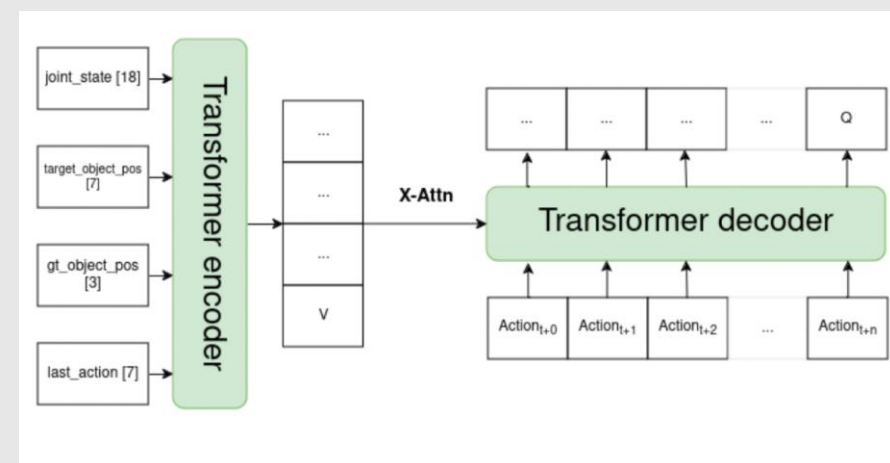
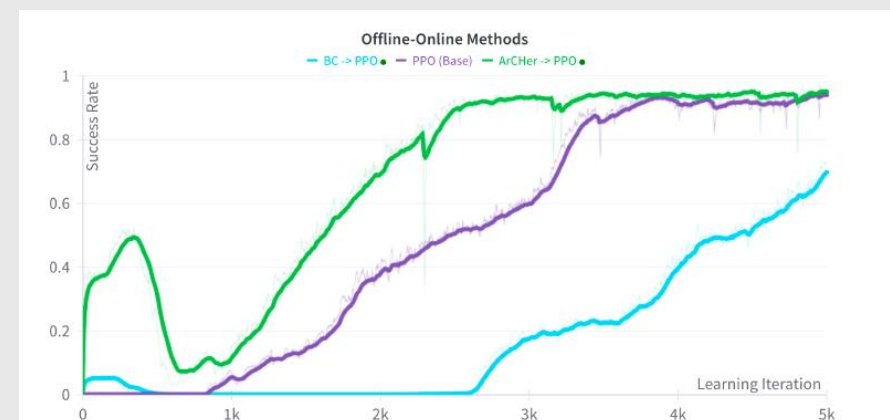
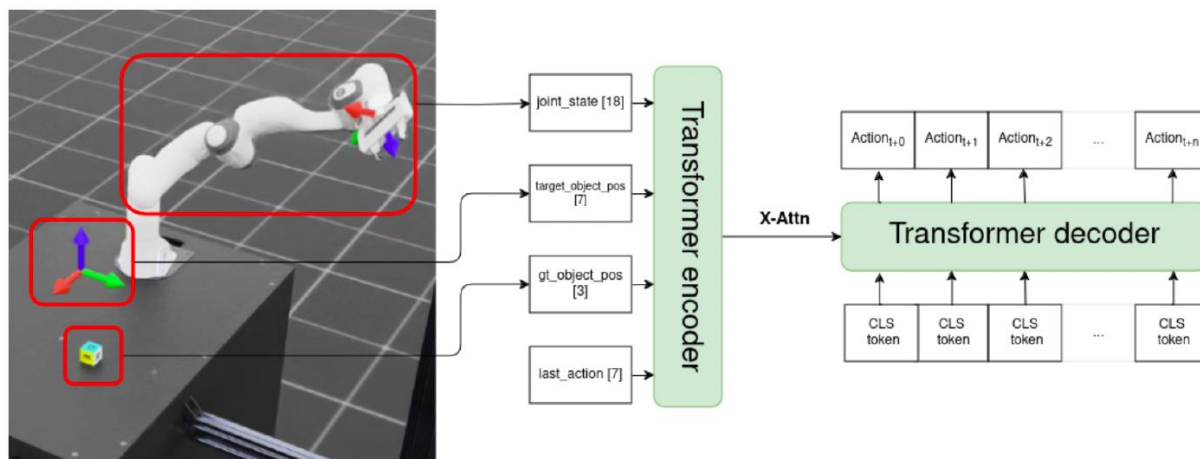


Визуальная обратная связь от среды



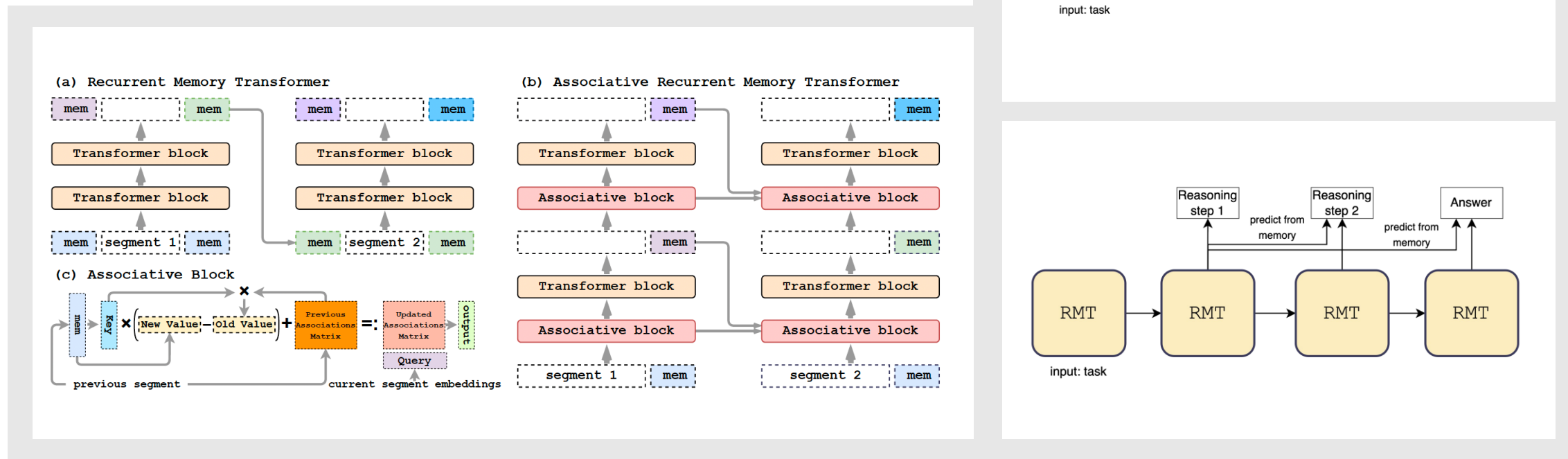
Дообучение в реальной среде

Генерация последовательности низкоуровневых действий и дообучение в среде на базе PlanFormer



Генерация нового протокола испытаний

Языковая модель с полнотекстовым анализом научных текстов и модулем рассуждений для оптимизации протоколов экспериментов на базе модели ARMT



Заключение



1

Автоматизация научных исследований — идеальная задача для отработки элементов AGI



2

Несколько вызовов на пути к AIScientist: нейросимвольная интеграция, эволюция программ, память, новые архитектуры и экспериментирующий агент



3

Роботизация лабораторных экспериментов, например, в химии, является и отличным полигоном для тестирования мобильных автономных робототехнических платформ



4

AIRoboScientist — интеграция последних разработок AIRI для автоматизации лабораторных экспериментов: мультимодальная генерация действий, следование инструкциям, интерпретация сцен и дообучение в онлайн режиме

Контакты



grafft.github.io



Telegram
[ai_panov](https://t.me/ai_panov)



Александр Панов

директор лаборатории CAIS AIRI
директор ЦКМ МФТИ



panov@airi.net



airi.net



[airi_research_institute](https://t.me/airi_research_institute)



[AIRI Institute](https://vk.com/AIRI_Institute)



Telegram

Artificial Intelligence
Research Institute