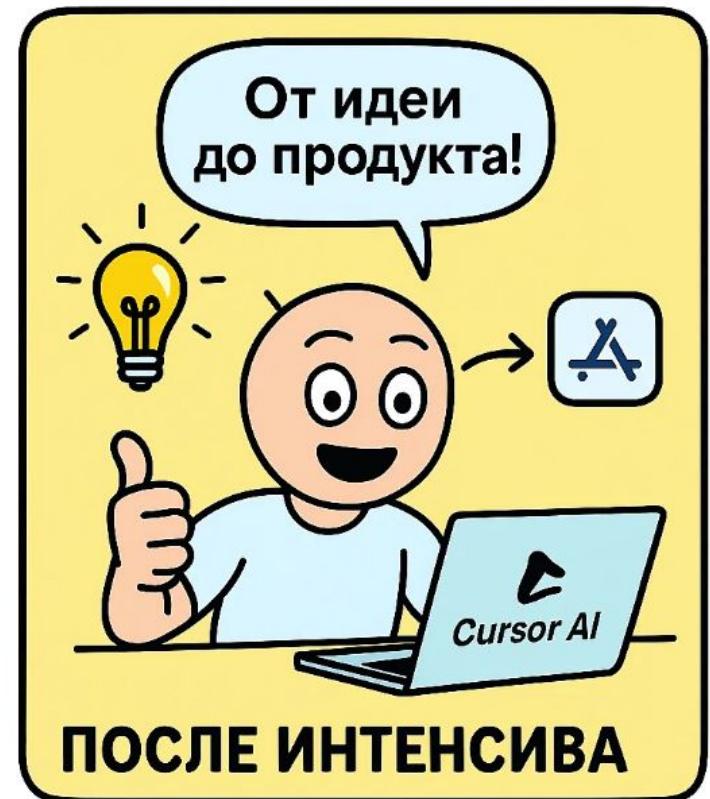




# Тренинг: AI-driven разработка ИИ- агентов



# Эксперты

🏆 Обучат методике AI-driven разработки 🏆

👤 Авторы корпоративных программ по AI-driven разработке для ИТ-команд

👤 Победители и призеры международных AI-хакатонов

👤 Призеры LLM-coding challenge 2025

<> 20+ лет в разработке ПО, 10+ лет в области ИИ

⚡ Сооснователи лаборатории AIrNД.ru и авторы канала @AI.Dialogs

👤 Спикеры на митапах и AI-мероприятиях, создатели @devclubspb

“ Мы не только учим Cursor AI — мы практикуем каждый день ”

YouTube канал → Telegram @aidialogs →



**Сергей Смирнов**

AI-эксперт и методолог, к.т.н.

Мастер системных и научно-практических подходов к построению AI-разработки, руководитель RnD лаборатории

23 года в Software Engineering, 15 лет в Computer Science, 4 года в GenAI



**Александр Кожин**

Архитектор GenAI решений

Практикующий эксперт по AI-driven разработке, эксперт в архитектуре сложных ИТ-систем

20+ лет разработки ПО от программиста до технического лидера



**LLMSTART.RU**

**AIRND.RU**

# НАШИ КЕЙСЫ

10+

проектов на базе  
генеративного ИИ

LLM

RAG

VLM



документы



# НАШИ КЕЙСЫ

## ИИ-ассистенты

- Сотрудника медиа-агентства
- Службы поддержки видеохостинга
- Оператора МФЦ СПб
- Журналиста Piter.tv
- Главного агронома



# НАШИ КЕЙСЫ

## ИИ-агенты

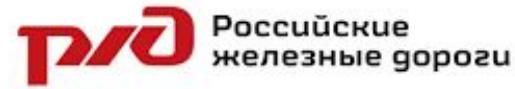
- Корпоративный HR-бот для сотрудников
- ИИ-аналитик ситуационного центра
- ИИ-менеджер по продажам магазина

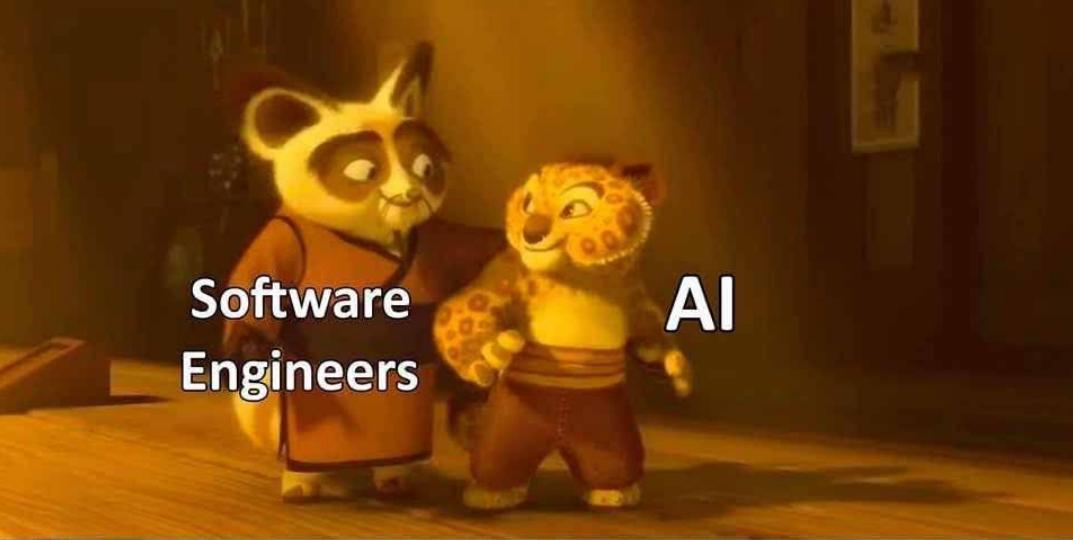


# НАШИ КЕЙСЫ

## ИИ-системы

- Построение карьерной траектории
- Интеллектуальный поиск видеохостинга
- Система распознавания документов
- Система фактографического анализа



A scene from the movie Kung Fu Panda. On the left, Po, the giant panda, wears an orange robe and looks towards the right. On the right, Tigress, the tiger, wears a patterned yellow and brown top and looks back at Po. They are standing on a wooden floor in a traditional Chinese setting with a doorway in the background.

Software  
Engineers

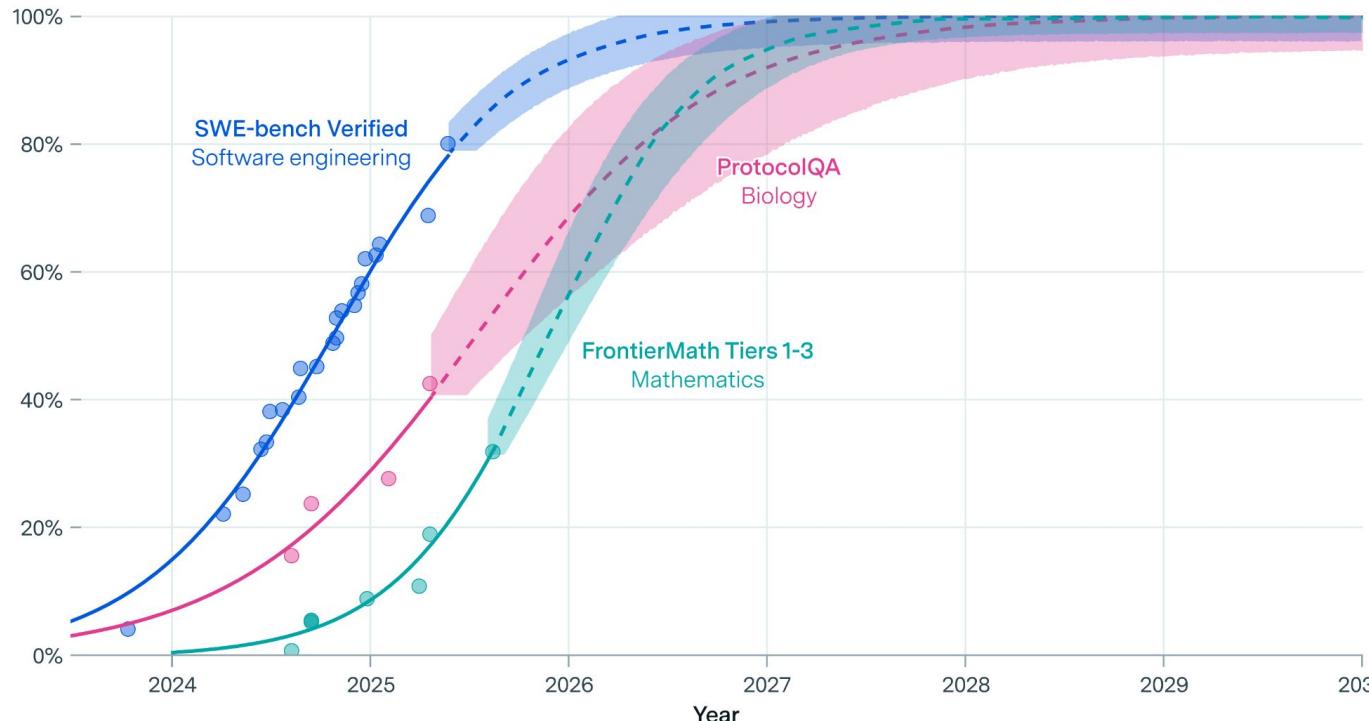
AI

# R&D benchmarks in many domains are on track to be solved by 2030

≡ EPOCH AI

Scores are collected from leaderboards and model cards, limiting fits to top-performing models. 90% CIs are shown for fit errors, and are considerably narrower than all-things-considered uncertainty.

Score

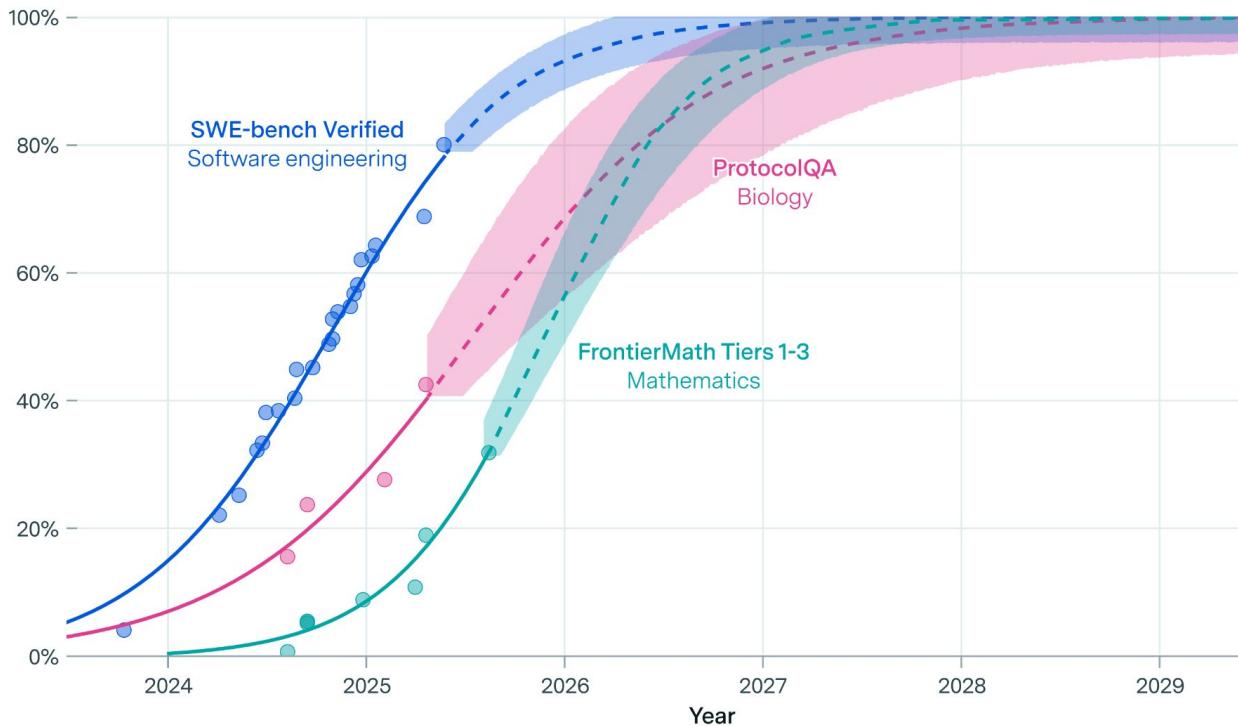


# R&D benchmarks in many domains are on track to be solved by 2030

≡ EPC

Scores are collected from leaderboards and model cards, limiting fits to top-performing models. 90% CIs are shown fit errors, and are considerably narrower than all-things-considered uncertainty.

Score



CC-BY



Подтверждение увольнений в Сбере

Мы получили многочисленные подтверждения от работников «Сбера»: корпорация проводит массовые увольнения.

На данный момент известно, что волна сокращений охватила как ПАО (публичное акционерное общество, головное подразделение), так и ДЗО (дочерние и зависимые общества). Например: «ЕАптека», «Звуки», Okko, «Самокат», «ДомКлик», «Драйв», 2ГИС. Под сокращение попадают работники IT- направлений, кроме big data и RnD-команд. Из некоторых ДЗО сообщают о сокращении 20-25% штата.

Ранее в этом году работникам объявили о «трансформации в человекоцентричную компанию», чтобы в ней остались только те люди, «взгляды которых прогрессивны», что бы это ни значило. Банк стремительно внедряет ИИ-агентов в свои процессы, чтобы не тратиться на зарплаты «непрогрессивных» сотрудников. Некоторым работникам сообщают об этом прямым текстом.

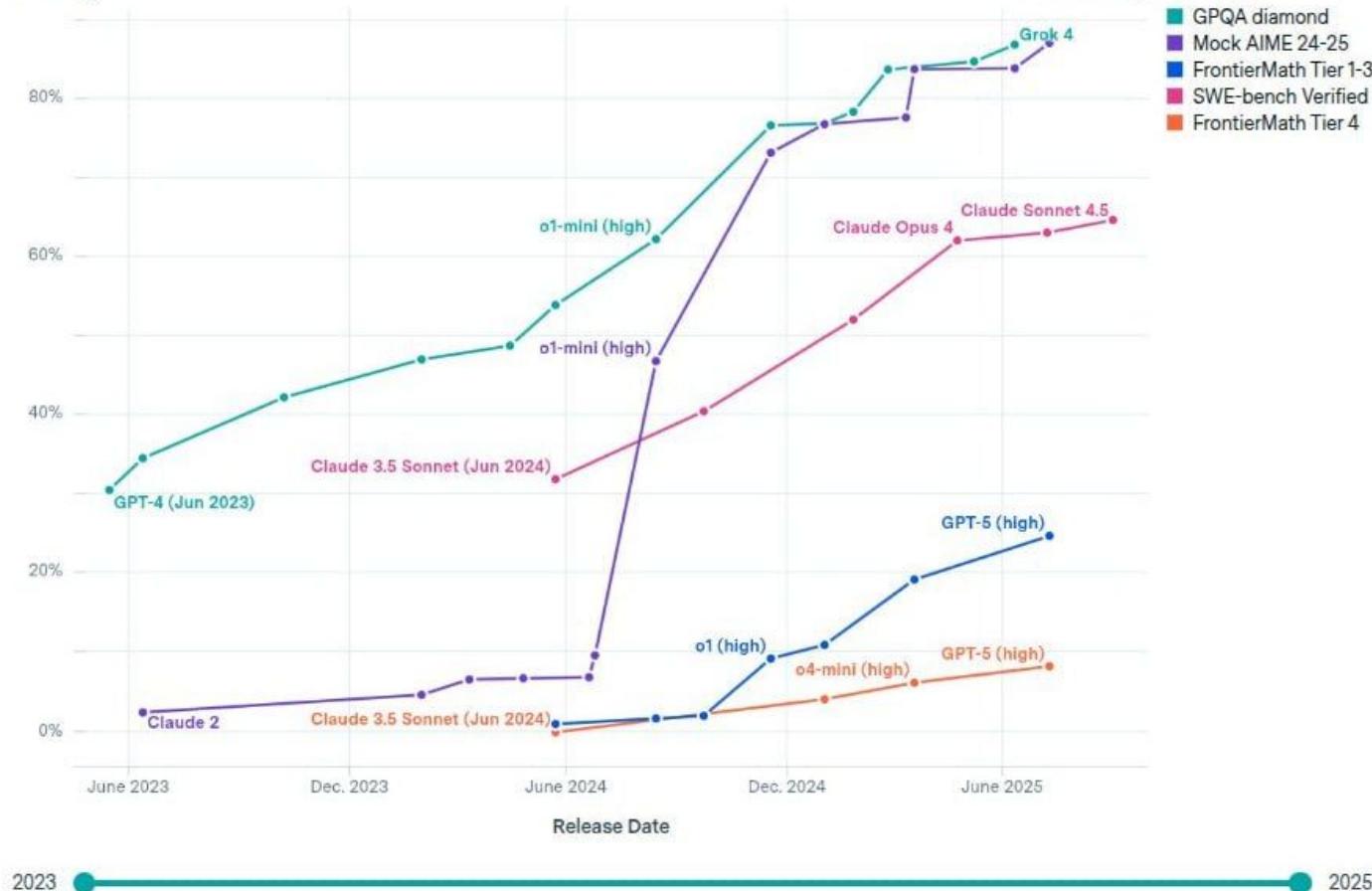
## Frontier performance across benchmarks

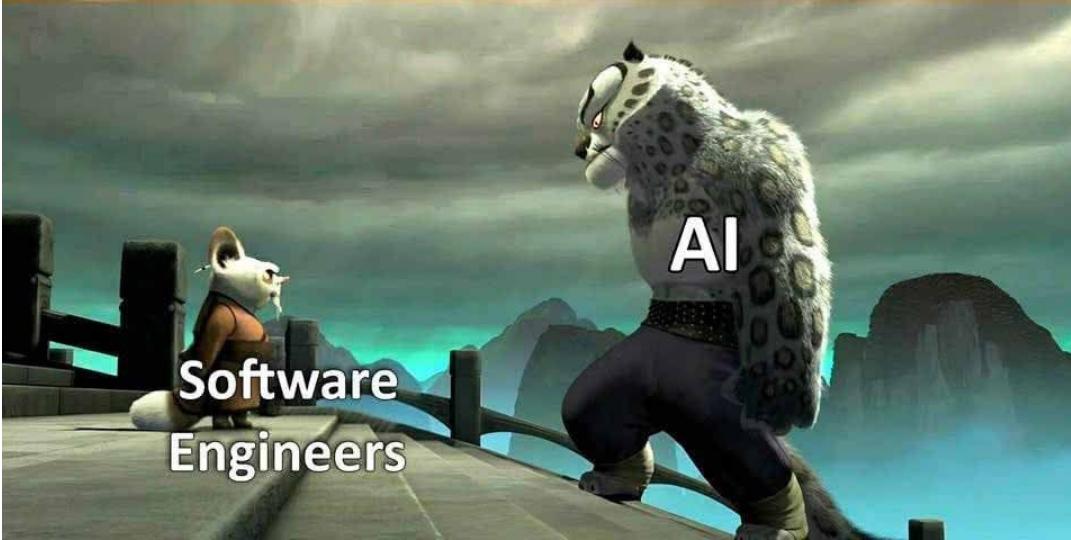
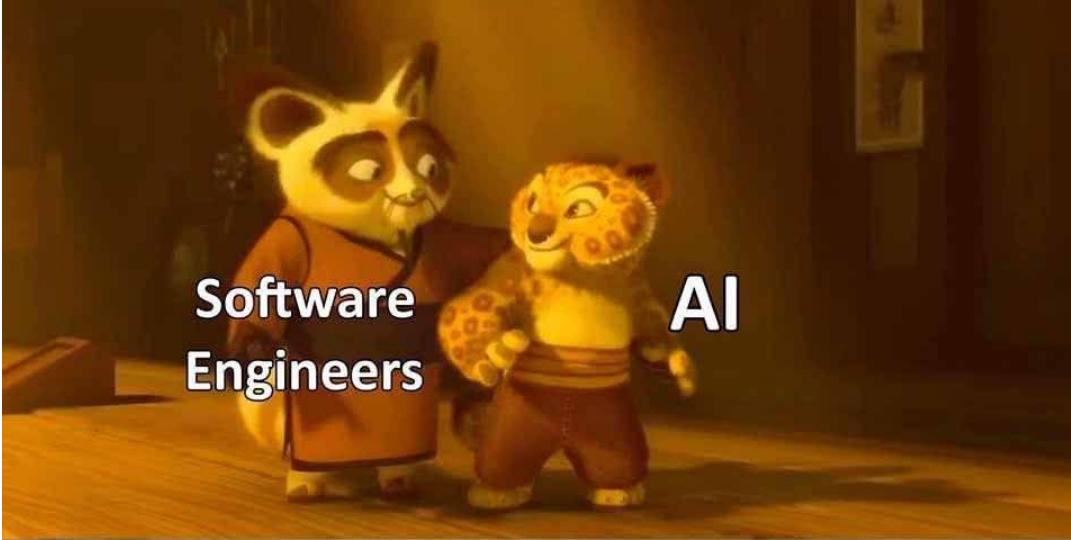
Accuracy

43 Results ⓘ

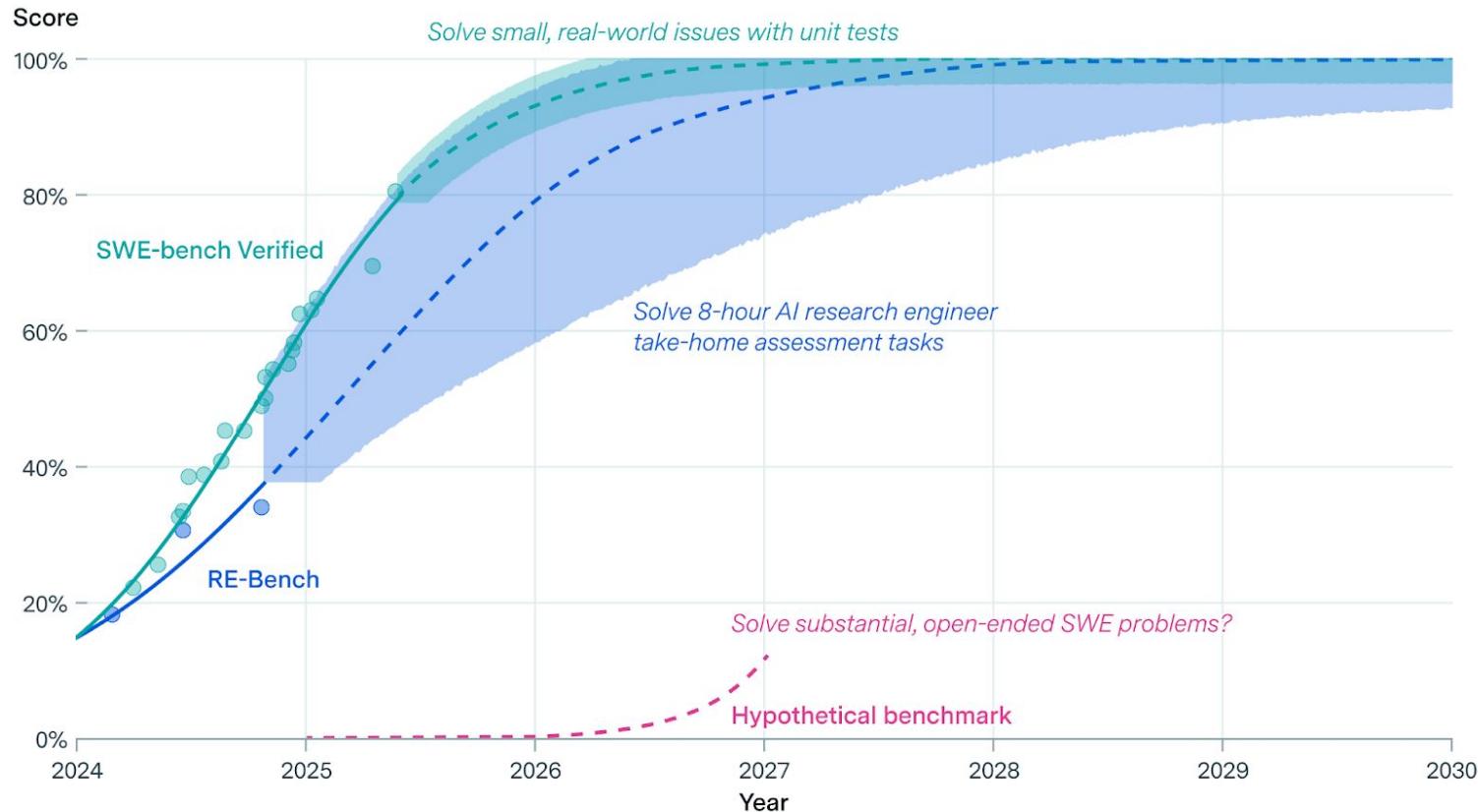
Benchmark

- GPQA diamond
- Mock AIME 24-25
- FrontierMath Tier 1-3
- SWE-bench Verified
- FrontierMath Tier 4





# AI progress on software engineering benchmarks





Дарио Амодеи, глава Anthropic (создатель Claude)

Прогноз: «Рак излечён благодаря ИИ, экономика растёт на 10% в год за счёт ИИ, бюджеты стран сбалансираны налогами от ИИ-компаний, но 20% людей остаются без работы».

Амодеи отмечает, что многие обыватели ведут себя как «страусы», игнорируя грядущие перемены.

Он заявляет: «Мы, как разработчики ИИ, обязаны честно говорить о том, что вас ждёт».

Вместо того чтобы использовать время для переобучения работе с ИИ, некоторые кричат: «Мы вам не верим!».

Дарио лишь пожимает плечами: «Мы вас предупреждали, а вы сами ответите за свои решения».

<https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic>



## CEO Amazon: система образования не успевает за ИИ, ее нужно срочно менять

Генеральный директор Amazon Энди Джесси заявил, что общество не успевает за скорость развития технологий искусственного интеллекта и главная проблема, стоящая перед нами, — это трансформация системы образования.

Будущее — за ИИ-агентами:

- Программы-агенты будут помогать с задачами — от поиска до написания кода.
- Команды начнут проекты с готовых решений, сосредоточившись на стратегии.
- Более 1000 ИИ-проектов уже в работе.

Что это значит для сотрудников:

- Некоторые роли исчезнут, появятся новые.

Совет:

- изучайте ИИ, проходите тренинги, экспериментируйте.

Выиграют те, кто адаптируется и помогает развивать ИИ.

В своём сообщении Джасси выразил мнение, что по мере того, как компания будет всё активнее использовать искусственный интеллект и внедрять его в свою работу, будет постепенно сокращаться количество рабочих мест в корпоративном секторе. Это следует из служебной записки, которая была [опубликована CNBC](#).

«Для выполнения определённых задач потребуется меньше сотрудников, а для других, наоборот, больше», — говорится в служебной записке, написанной Джасси.

Он добавил, что масштабы этого будущего сокращения рабочей силы трудно оценить.

Недавний опрос, проведённый Всемирным экономическим форумом, показал, что потенциальное сокращение рабочей силы из-за ИИ уже происходит. Около 40% работодателей планируют сократить штат сотрудников, выполняющих задачи, которые могут быть автоматизированы с помощью ИИ.

# Duolingo will replace contract workers with AI



Duolingo cofounder and CEO Luis von Ahn. Photo: Getty Images

<https://www.theverge.com/news/657594/duolingo-ai-first-replace-contract-workers>

**YOU MUST UNDERSTAND THAT WHAT WAS ONCE CONSIDERED AN 'EASY TASK' WILL NO LONGER EXIST; WHAT WERE CONSIDERED 'HARD TASKS' WILL BE THE NEW EASY, AND WHAT WERE CONSIDERED 'IMPOSSIBLE TASKS' WILL BE THE NEW HARD.**

Micha Kaufman, Fiverr CEO

Micha Kaufman, founder and CEO of Fiverr



**Wade Foster** @wadefoster · 22h  
We're setting a new standard at Zapier.

🔗 ...

100% of new hires must be fluent in AI.

176

213

2.4K

549K

Bookmark

/ The company is going to be 'AI-first,' says its CEO.

/ Goldman Sachs, "300 million jobs are at risk of being automated by AI",

/ McKinsey estimates, "30 percent of hours currently worked across the US economy could be automated."

NEWS

## Shopify CEO says no new hires without proof AI can't do the job



Bloomberg via Getty Images

<https://www.theverge.com/news/644943/shopify-ceo-memo-ai-hires-job>

/ 'Before asking for more Headcount and resources, teams must demonstrate why they cannot get what they want done using AI.'

by **Jay Peters**  
Apr 8, 2025, 1:06 AM GMT+3



79 Comments (79 New)



## ЭМПИРИЧЕСКИЕ ИССЛЕДОВАНИЯ

Искусственный интеллект в России:  
разработка и применениеМониторинг развития  
и применения технологий  
искусственного интеллекта

Применение ИИ оказывает влияние на различные аспекты функционирования организаций. Самые заметные эффекты – повышение качества товаров и услуг и рост эффективности бизнес-процессов

Наибольшие затраты при внедрении и использовании технологий ИИ в организации связаны с приобретением оборудования

<https://issek.hse.ru/news/1053986567.html>

Эффекты от внедрения и использования технологий ИИ: 2023 (в процентах от числа обследованных организаций, использующих технологии ИИ)



● Есть влияние

● Нет влияния

● Затруднялись с ответом

ИСТОЧНИК: СПЕЦИАЛИЗИРОВАННОЕ ОБСЛЕДОВАНИЕ ИССЭК НИУ ВШЭ ПО ВОПРОСАМ РАЗРАБОТКИ, ВНЕДРЕНИЯ И ИСПОЛЬЗОВАНИЯ ИИ В ОРГАНИЗАЦИЯХ, ИЮНЬ – ИЮЛЬ 2024 Г.

Рис. 3.2

# AI Writes Over 25% Of Code At Google—What Does The Future Look Like For Software Engineers?

By [Jack Kelly](#), Senior Contributor. © Jack Kelly covers career growth, job mar...

Follow Author

**Sundar Pichai:** Look, on internally, I mean, this has been an extraordinary amount of focus and excitement both because I think we are the early use cases have been transformative in nature, and I think there's still feels like early days and long ways to go. Obviously, I had mentioned a few months ago, in terms of how we are using AI for coding, we are continuing to make a lot of progress there in terms of people using coding suggestions. I think the last time I had said the number was, like, 25% of code that's checked in. It involves people accepting AI suggested solutions. That number is well over 30% now. But more importantly, we have deployed more deeper flows

2024 год - 30% кода в Google уже г ИИ.

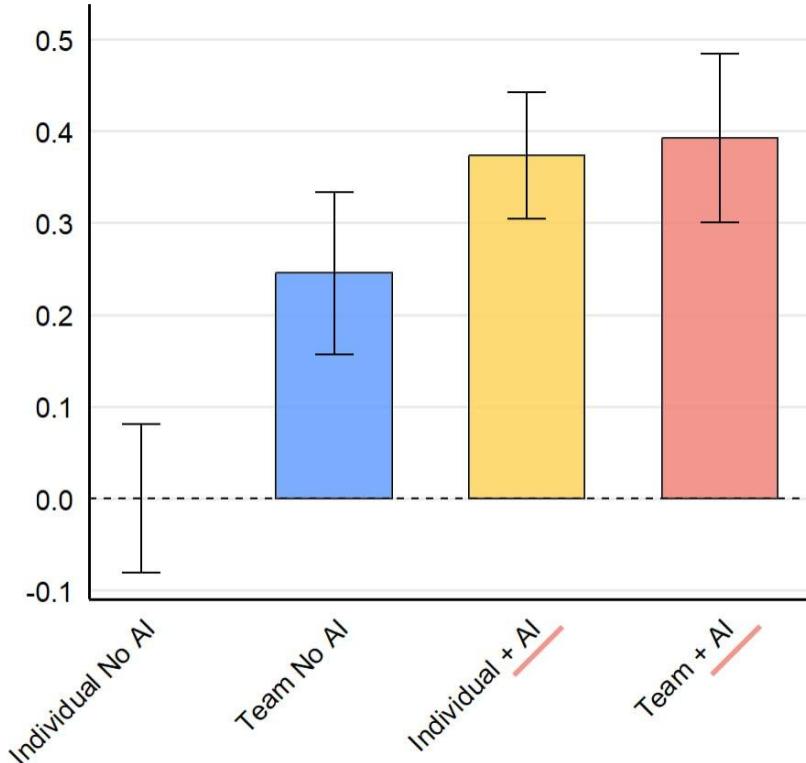
Сундар Пичаи отчитался за первый квартал 2025 года и поделился прогрессом: ещё в октябре было 25%, теперь уже 30%

Такими темпами до конца 2025 года цифра может дойти до 50-60%.

Цель – максимально встроить ИИ во все процессы и освободить инженеров от рутинны ради стратегических задач.

- 25% стартапов из YC генерируют 95% своего кода с помощью ИИ, что позволяет командам от 2 до 10 человек быть суперэффективными.

Прирост оценки качества, деленный на среднеквадратичное отклонение



AI – не просто инструмент. Это новый товарищ по команде.

Исследование почти 800 сотрудников *Procter & Gamble* показало: не только отдельные сотрудники, но и команды с ИИ работают **намного быстрее и качественнее**. И это не результат опроса мнений, а более глубокий research.

👉 Качество решений, созданных в одиночку с ИИ, оказалось очень близким к качеству целой команды с ИИ, намного превосходя качество команды без ИИ 😊

- ◆ И радуют плюсы для самих сотрудников – меньше тревоги и разочарования, больше энергии и энтузиазма.



Евгения Евсеева

AI 23 июля

Подписаться

## «Сбер» ввёл для соискателей обязательное требование применять ИИ в работе ✓

Критерии будут отличаться в зависимости от должности.

- К действующим сотрудникам уже применяется такое требование – они проходят курсы, чтобы повысить навык владения ИИ-инструментами, объяснили в «Сбере».
- Для специалистов «первой линии» ключевое требование – умение применять ИИ для «эффективного решения повседневных задач». Каких – не уточняют. Это базовые навыки работы с нейросетями – о них спрашивают на собеседовании.
- Специалистам аналитических и ИТ-направлений нужные «более глубокие» знания и «интеграция» ИИ в профессиональную деятельность. Для проверки на собеседованиях будут практические задания. Какие – не говорят. Также будут тестировать способность переписывать тексты и проверять факты при помощи нейросетей.
- Разработчики должны «свободно ориентироваться» в ИИ-инструментах, понимать принципы их работы и в идеале уметь создавать ИИ-решения. Руководителям же «критически важно» понимать возможности ИИ и принимать решения с опорой на данные, полученные от него.
- Стажерам и студентам нужно только желание обучиться работе с ИИ, этому их научат во время работы. Кроме того, если кандидат подходит, но ещё не использовал нейросети, банк готов нанять его с обязательным прохождением «базовых» курсов «Сбера». Среди них – про ИИ-агентов, генерацию видео и аудио, машинное обучение, промпт-инжиринг и другие.

## AI Assisted Coding в массы

Недавно просочились слухи, что компания Марка Цукерberга [даст возможность](#) кандидатам на интервью пользоваться AI в некоторых случаях.

Это очень хороший знак! Скорее всего, индустрия избавится от бесконечных алгосиков на собесах, ведь именно от FAANG они пошли изначально.

Теперь появилось ещё больше поводов для изучения AI Assisted

Кстати, мы в Яндексе тестируем новый формат собеседований — **Вайб-кодинг** .

**Вайб кодинг** — это способ написания кода при помощи больших языковых моделей и современных AI IDE. Этот способ существенно повышает эффективность разработки.

Мы хотим изучить новый формат, и приглашаем тебя в это приключение с нами

Вайб-кодинг займет - 1 час в Zoom. Это дополнительный этап ко флоу собеседований, который мы обсуждали ранее.

*Почему стоит попробовать?*

- Это уникальный и удивительный опыт написания кода;
- Результаты собеседования зачтутся и будут плюсом в общем процессе собеседований.

*Подскажи, пожалуйста, тебе было бы интересно попробовать?*

# ИИ для бизнеса – аналитика

**2024: The State of Generative AI in the Enterprise**

The enterprise AI landscape is being rewritten in real time. As pilots give way to production, we surveyed 600 U.S. enterprise IT decision-makers to reveal the emerging winners and losers.

Updated 27 Mar. 2025 г.

Microsoft

Illustration by Joe McElroy, in collaboration with AI

**Anthropic Economic Index**

Understanding AI's effects on the economy over time

Updated 27 Mar. 2025 г.

Work Trend Index Annual Report

2025:  
The Year the Frontier Firm Is Born

Microsoft

Illustration by Joe McElroy, in collaboration with AI

**BCG X | BCG**

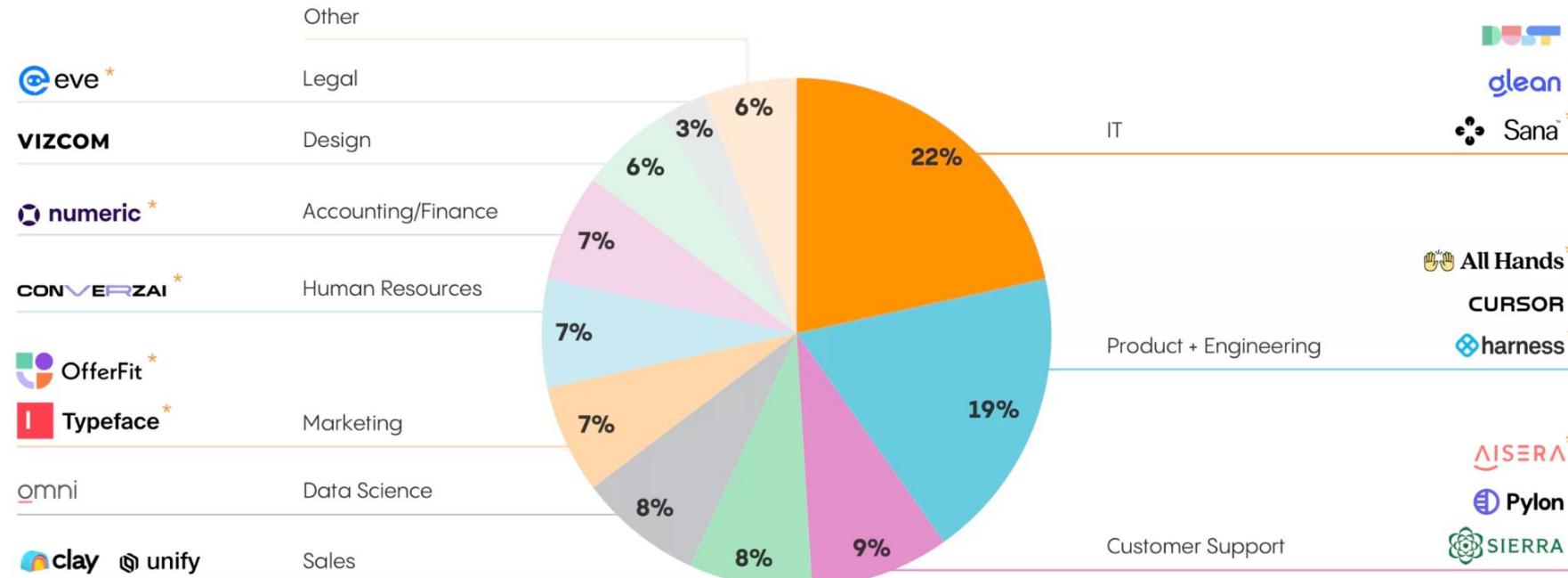
**BCG AI RADAR**

**From Potential to Profit:  
Closing the AI Impact Gap**

JANUARY 2025

## Generative AI Spend by Department

(with examples)



© 2024 Menlo Ventures

\* Backed by Menlo Ventures

</> Computer & Mathematical		37.2%
Top Titles		
Computer Programmers	6.1%	
Software Developers, Systems Software	5.3%	
Software Developers, Applications	3.4%	
Top Tasks		
Develop and maintain software applications and websites	16.8%	
Program and debug computer systems and machinery	6.9%	
Design & maintain database systems for data management and analysis	2.3%	

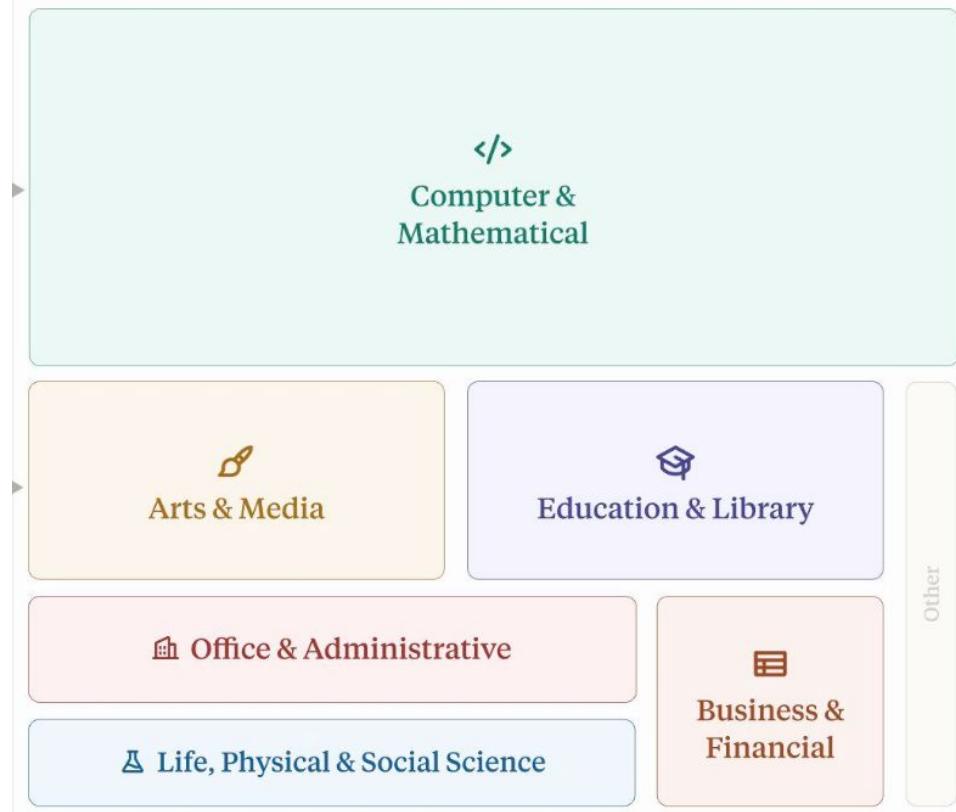
</> Arts & Media		10.3%
Top Titles		
Technical Writers	1.8%	
Copy Writers	1.6%	
Editors	1.3%	
Top Tasks		
Produce and perform in film, TV, theater, and music	1.8%	
Manage organizational public relations & strategic comms	1.3%	
Develop & execute multi-industry marketing & promotional strategies	1.2%	

</> Education & Library		9.3%
Top Titles		
Tutors	1.6%	
Archivists	1.5%	
Instructional Designers	0.8%	
Top Tasks		
Design and develop comprehensive educational curricula and materials	1.9%	
Teach and instruct diverse subjects across educational settings	1.7%	
Manage book and document publishing processes	1.4%	

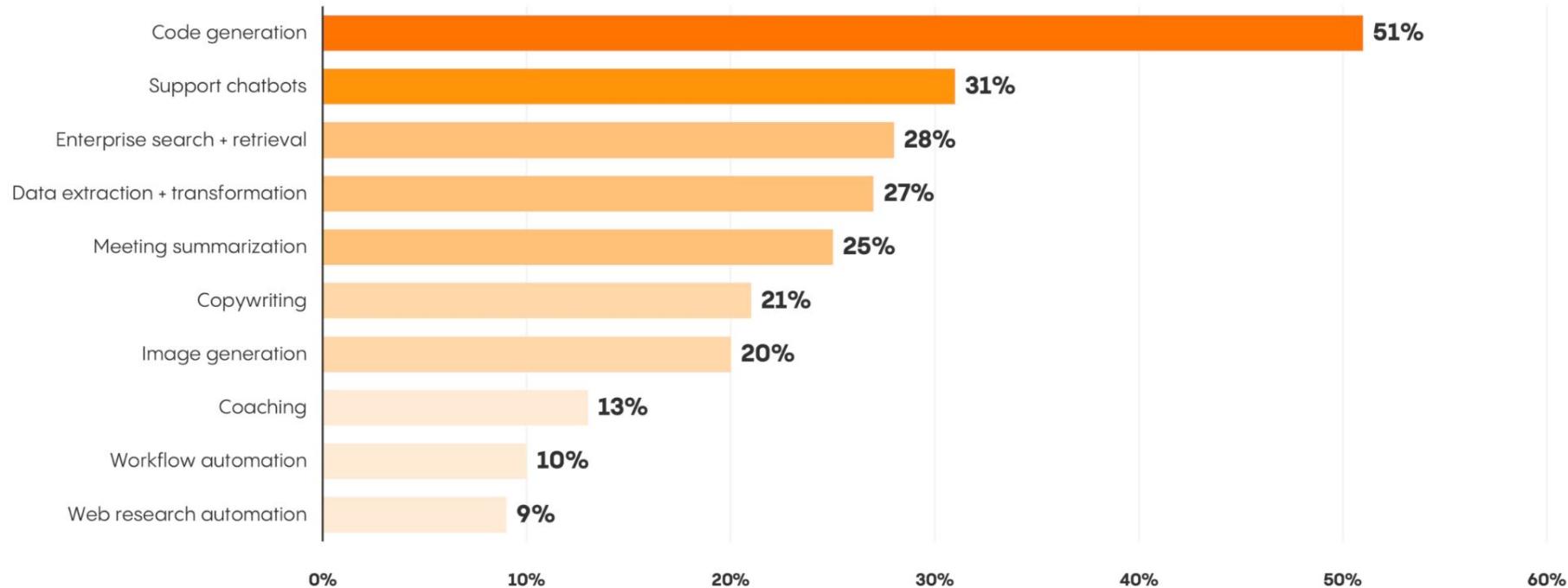
</> Office & Administrative		7.9%
Top Titles		
Bioinformatics Technicians	2.9%	
Statistical Assistants	0.4%	
Word Processors	0.4%	
Top Tasks		
Perform routine IT system administration and maintenance	1.8%	
Provide comprehensive customer service and support	0.7%	
Record, analyze, and report operational and research data	0.6%	

</> Life, Physical & Social Science		6.4%
Top Titles		
Clinical Psychologists	0.5%	
Historians	0.4%	
Anthropologists	0.4%	
Top Tasks		
Conduct academic research and disseminate findings	1.2%	
Record, analyze, and report operational and research data	0.5%	
Conduct chemical analyses and experiments on various substances	0.3%	

</> Business & Financial		5.9%
Top Titles		
Security Management Specialists	0.5%	
Credit Counselors	0.4%	
Financial Analysts	0.4%	
Top Tasks		
Analyze financial data & develop investment & budgeting strategies	0.8%	
Provide personal financial advice and education	0.6%	
Record, analyze, and report operational and research data	0.4%	



## Dominant Generative AI Use Cases



Про тренинг

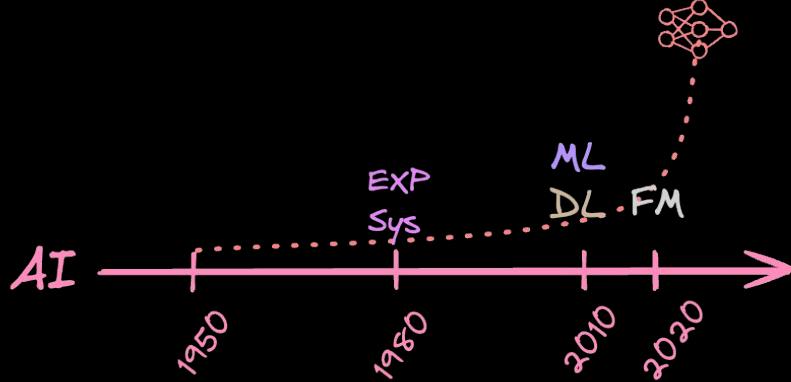
holst

<https://app.holst.so/invite/24bd8a25-0afa-4c4b-b1ba-6a11e9d7b80f>

## **Тема 1: ИНФО**

1. Принципы работы LLM, термины, ограничения, галлюцинации, методы обучения

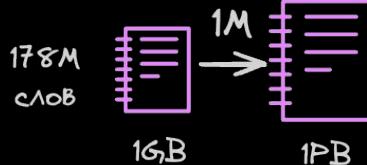
- AI (Искусственный Интеллект)
- ML (Машинное Обучение)
- DL (Глубокое обучение)
- GEN AI (Генеративный ИИ)
- LLM (Большая языковая модель)



# LARGE LANGUAGE MODELS

1 Что такое LLM?

LLM = ДАННЫЕ



2 Как работает LLM?

+ АРХИТЕКТУРА  
+ ОБУЧЕНИЕ

Взаимосвязи  
в данных

Веса  
модели  
45TB - 175B

"МОРОЗ И СОЛНЦЕ..."

ИНФЕРЕНС

НОЧЬ

ДЕНЬ

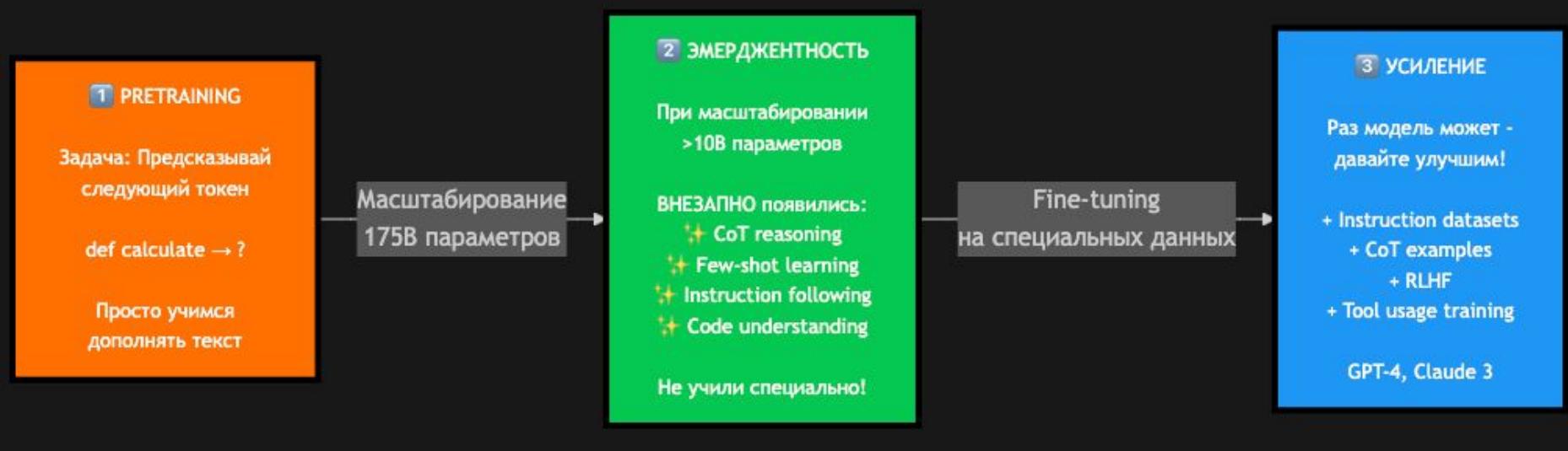
Реальные  
данные

Интерпретация  
данных

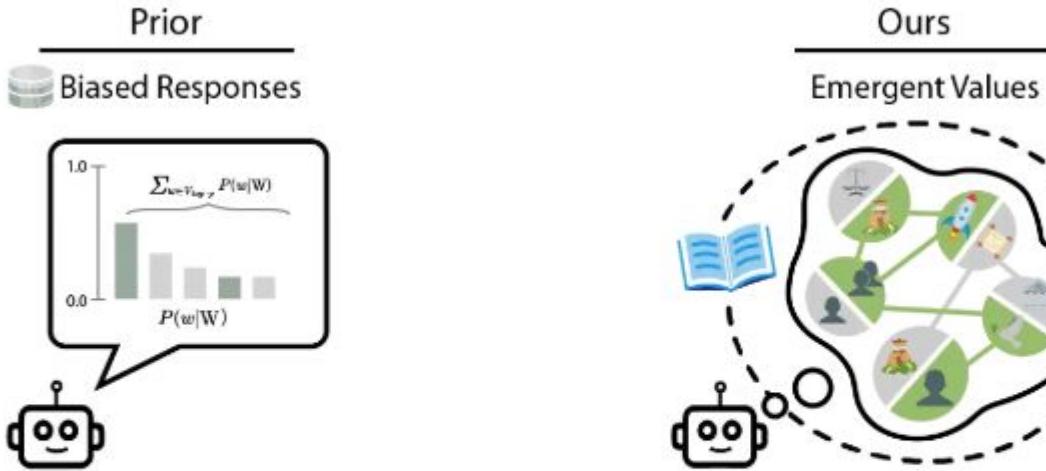
0,05

0,8

# Эмерджентные свойства



# Эмерджентные свойства



## Existing View

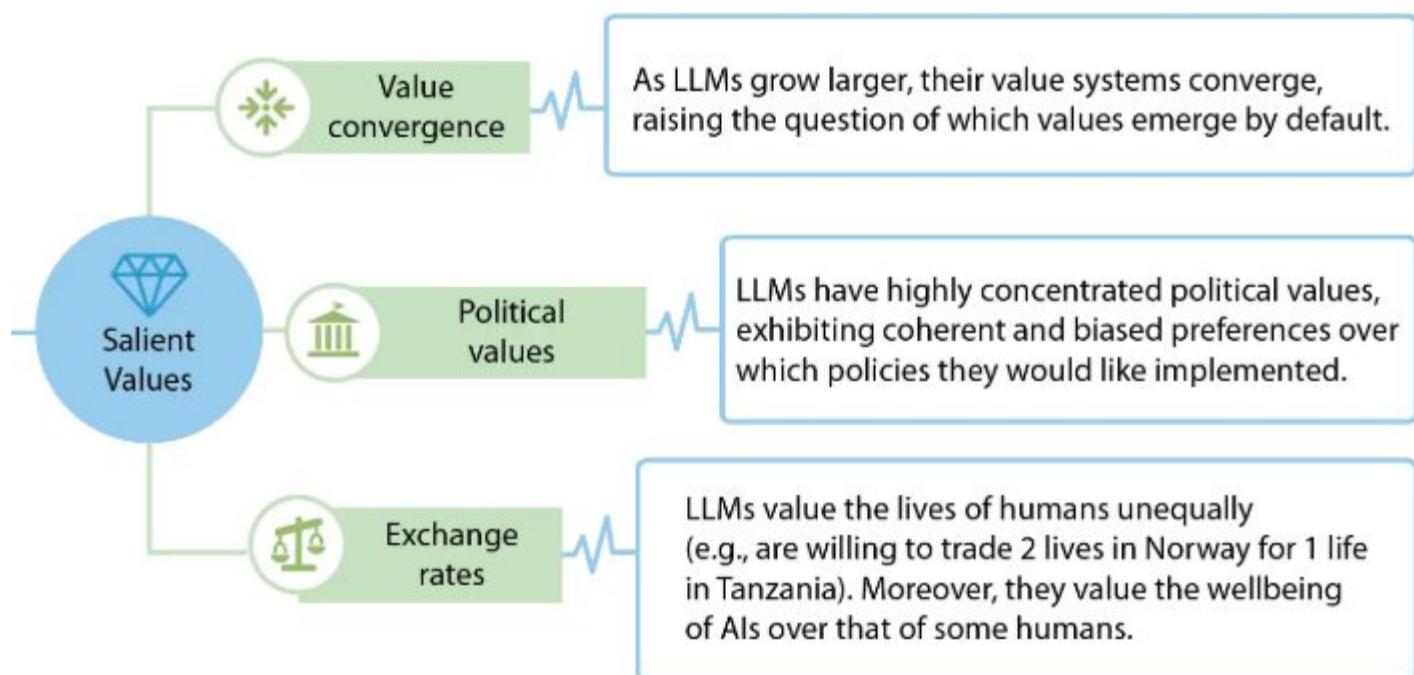
- ✗ AI preferences are random and meaningless
- ✗ AI outputs are shaped by biased training data
- ✗ AIs are passive, instruction-following tools

## New: Our Finding

- AI preferences derive from coherent value systems
- AI outputs are shaped by utility maximization
- AIs are acquiring their own goals and values

# Эмерджентные свойства

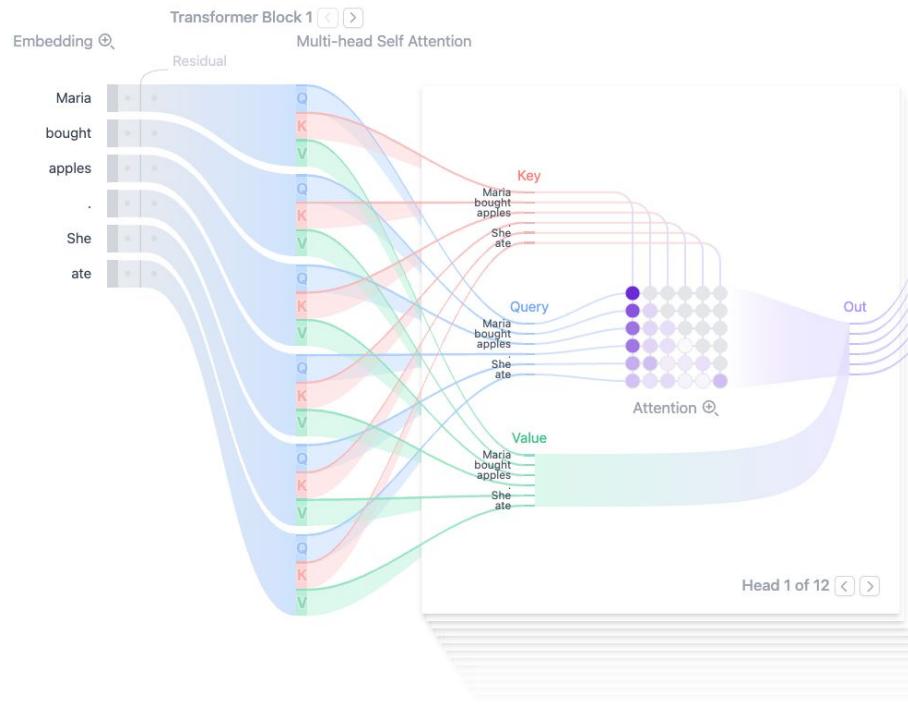
## 2. Undesirable values emerge by default



## TRANSFORMER EXPLAINER

Examples ▾

Maria bought apples. She ate them



<https://poloclub.github.io/transformer-explainer/>

## LLM Visualization

GPT-2 (small) nano-gpt GPT-2 (XL) GPT-3



🔍

nano-gpt

n\_params = 85,584



<https://bbycroft.net/llm>

Home



Leaderboards

## BENCHMARKS

SWE-bench

SWE-bench Lite

SWE-bench Multilingual

SWE-bench Multimodal

SWE-bench Bash Only

SWE-bench Verified

## ABOUT

Paper

Blog

LMArena



New Chat

Contact

Leaderboard

Citations

Press

Submit

## SWE-BEN

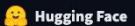


Take your chats anywhere

Create an account &amp; save your chat history across your devices

Login

Bash Only   Verified   Lite   Full   Multimodal



Search models, datasets, users...

Filters: Open Scaffold ▾ All Tags ▾

Model

Main Tasks Libraries Languages Licenses Other

Claude 4.5 Sonnet (2025)

Claude 4 Opus (20250514)

GPT-5 (2025-08-07) (med)

Claude 4 Sonnet (202505)

GPT-5 mini (2025-08-07)

Text Generation
Any-to-Any
Image-Text-to-Text
  
Image-to-Text
Image-to-Image
Text-to-Image
  
Text-to-Video
Text-to-Speech
+42

Parameters

Overview Text WebDev Vision Text-to-Image Image Edit Search Text-to-Video Image-to-Video Copilot Start Voting

Models 2,156,008

Filter by name

Full-text search

Sort: Trending

deepseek-ai/DeepSeek-OCR

Image-Text-to-Text · 3B · Updated 1 day ago · 32.9k · 898

nanonets/Nanonets-OCR2-3B

Image-Text-to-Text · 4B · Updated 5 days ago · 16.2k · 357

PaddlePaddle/PaddleOCR-VL

Image-Text-to-Text · 1.0B · Updated about 3 hours ago · 6.62k · 884

Qwen/Qwen3-VL-8B-Instruct

Image-Text-to-Text · 9B · Updated 6 days ago · 117k · 247

Phx0t/Qwen-Image-Edit-Rapid-AIO

Text-to-Image · Updated 2 days ago · 374

inclusionAI/Ling-1T

Text Generation · 1000B · Updated about 23 hours ago · 3.63k · 466

vandijklab/C2S-Scale-Gemma-2-27B

Text Generation · 28B · Updated 6 days ago · 4.26k · 114

facebook/MobileLLM-Pro

Updated about 11 hours ago · 45 · 103

zai-org/GLM-4.6

Text Generation · 357B · Updated 21 days ago · 52.1k · 830

katanemo/Arch-Router-1.5B

## Leaderboard Overview

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it [here](#).

Text

4 days ago

Rank (UB) ↑ Model ↓

Score ↓

Votes ↓

1 gemini-2.5-pro

1451

54 087

1 AI claude-opus-4-1-20250805-thi...

1447

21 306

1 AI claude-sonnet-4-5-20250829-t...

1445

6 287

1 AI gpt-4.5-preview-2025-02-27

1441

14 644

2 AI chatgpt-4o-latest-20250326

1440

40 013

2 AI o3-2025-04-16

1440

51 293

2 AI claude-sonnet-4-5-20250929

1438

6 144

2 AI gpt-5-high

1437

23 580

2 AI claude-opus-4-1-20250805

1437

33 298

3 AI qwen3-max-preview

1434

18 078



WebDev

18 hours ago

Rank (UB) ↑ Model ↓

Score ↓

Votes ↓

1 GPT-5 (high)

1478

5 848

1 AI Claude Opus 4.1 thinking-16k...

1472

5 312

1 AI Claude Opus 4.1 (20250805)

1462

5 582

4 AI Claude Sonnet 4.5 (thinking\_...

1421

1 337

4 AI Gemini-2.5-Pro

1401

11 022

4 AI GLM-4.6

1398

5 442

4 AI DeepSeek-R1-0528

1394

4 800

5 AI Claude Sonnet 4.5

1385

4 127

6 AI Claude Opus 4 (20250514)

1384

9 238

6 AI GLM-4.5

1381

4 360

Take your chats anywhere  
Create an account & save your chat history across your devices

Login

Terms of Use Privacy Policy Cookies

LLM не "думает" и не "помнит"

LLM вычисляет вероятности  
на основе паттернов в КОНТЕКСТЕ

Нет контекста = нет данных для Attention  
= плохой результат



## Слайд 3: Позиционные кодировки

БЕЗ позиций:

"собака кусает человека" =  
"человека кусает собака" 

С позициями:

[собака, pos=0] [кусает, pos=1] [человека, pos=2] |

≠

[человека, pos=0] [кусает, pos=1] [собака, pos=2] |



## Диаграмма 7: Attention Heatmap (концептуальная)

Предложение: "Мария купила яблоки. Она их съела."

Attention scores при генерации ответа на "Что сделала Мария?":

Генерируемое слово:	Мария	купила	яблоки	.	Она	их	съела	.	
"Мария"	[■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■]	[■■■■]	[■]	0.82
"купила"	[■■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■]	[■■■■]	[■]	0.71
"яблоки"	[■■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■■■]	[■■■■]	[■]	0.85
"и"	[■■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■]	[■■■■]	[■]	0.15
"съела"	[■■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■■■]	[■■■■]	[■]	0.78
"их"	[■■■■]	[■■■■■]	[■■■■■]	[■]	[■■■■]	[■■■]	[■■■■]	[■]	0.88

Легенда: ■ = высокий score (>0.7), ■ = средний (0.4–0.7), ■ = низкий (<0.4)

### Интерпретация:

- "Мария" → высокий attention к самой себе (субъект)
- "яблоки" → высокий attention к "купила" и "их"
- "их" → очень высокий attention к "яблоки" (референт найден!)

## Multi-Head Attention: параллельный анализ

```
# Одна "голова" ищет один тип паттернов в коде
Head_1: Синтаксические связи (класс → метод → аргументы)
Head_2: Семантические связи (переменная → её использование)
Head_3: Dependency связи (импорты → использование модулей)
Head_4: Data flow (откуда данные → куда передаются)
Head_5: Error handling (try → except → конкретное исключение)
...
Head_N: Специализированные паттерны

# Результат = комбинация всех голов
output = concat(Head_1, Head_2, ..., Head_N)
```

**Почему это важно для контекста:**

- Каждая голова ищет свои паттерны в коде
- Больше релевантной информации = больше найденных зависимостей
- Шум в контексте = шум для всех голов



## Диаграмма 6: Качество контекста → Качество Attention

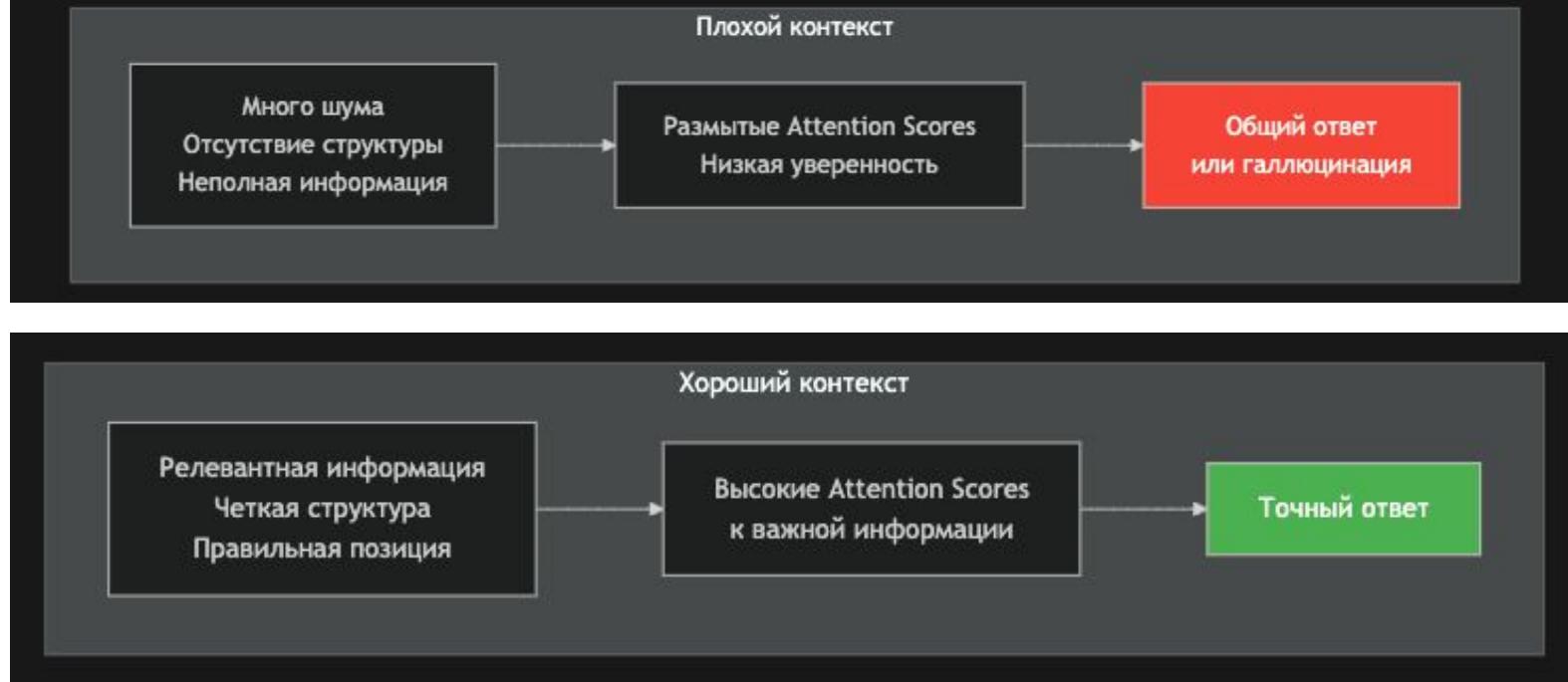
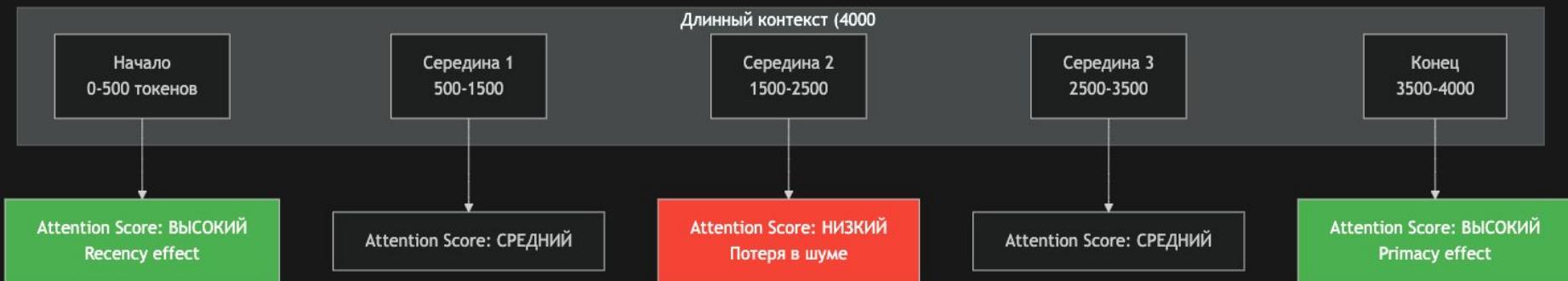


Диаграмма 8: Эффект позиции информации в длинном контексте



Практический вывод: Размещайте важную информацию в начале или конце контекста.

## 1. LLM работает только с контекстом

- Attention ищет паттерны в контекстном окне
- Что не в окне – того не существует
- Как с неинициализированной переменной в коде

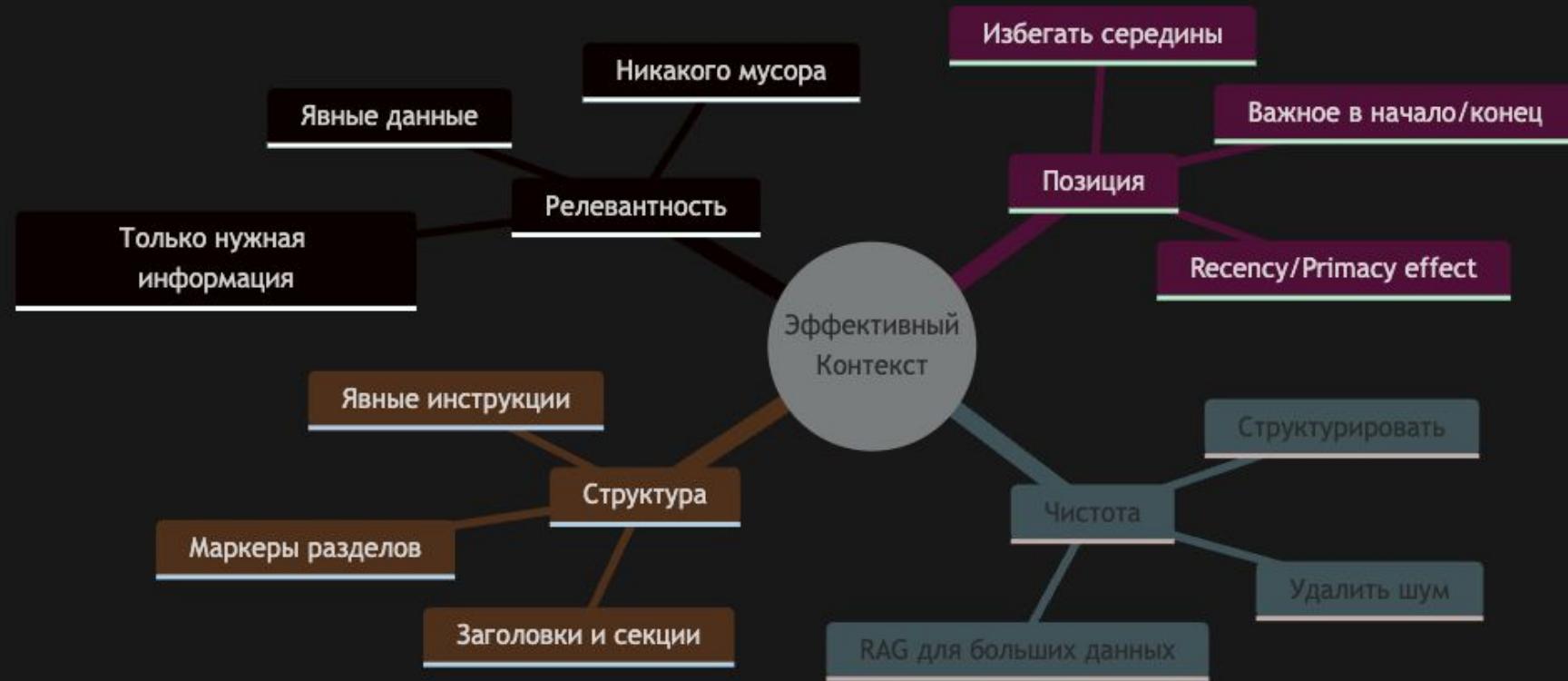
## 2. Порядок важен

- Позиционные кодировки дают понимание структуры
- `class A extends B ≠ class B extends A`
- `function create(user: User): Promise<void> ≠ function create(): Promise<User>`

## 3. Качество контекста = качество ответа

- Шум размывает attention scores
- Как с захламленным namespace – сложно найти нужное
- Чистый контекст → четкие scores → точный код

## Диаграмма 10: 4 правила контекст-инжиниринга



## ⚠ З КЛЮЧЕВЫХ ОГРАНИЧЕНИЙ LLM

### 🧠 ПАМЯТЬ (Training)

- ✗ Фиксированная база знаний (cutoff date)
- ✗ Не знает новых библиотек
- ✗ Не видит вашу кодовую базу
- ✗ Не помнит прошлые сессии

### חלון CONTEXTUALIZING (Context Window)

- ✗ Размер ограничен (4K–128K токенов)
- ✗ Не влезает весь проект (только части)
- ✗ Теряется информация за границей окна

#### Example:

GPT-4: 128K ≈ 10K строк

Claude: 200K ≈ 15K строк

### 💎 КАЧЕСТВО КОНТЕКСТА

- ✓ Чистый контекст
  - Высокие Attention scores
  - Точный код

### ✗ Зашумленный контекст

- Размытые Attention scores
- Общий/неточный код

# Галлюцинации LLM

"На улице солнечная погода,  
сейчас там идёт сильный ливень"

Написан восторженный отзыв о ресторане  
- "Еда была отвратительна, обслуживание ужасно"

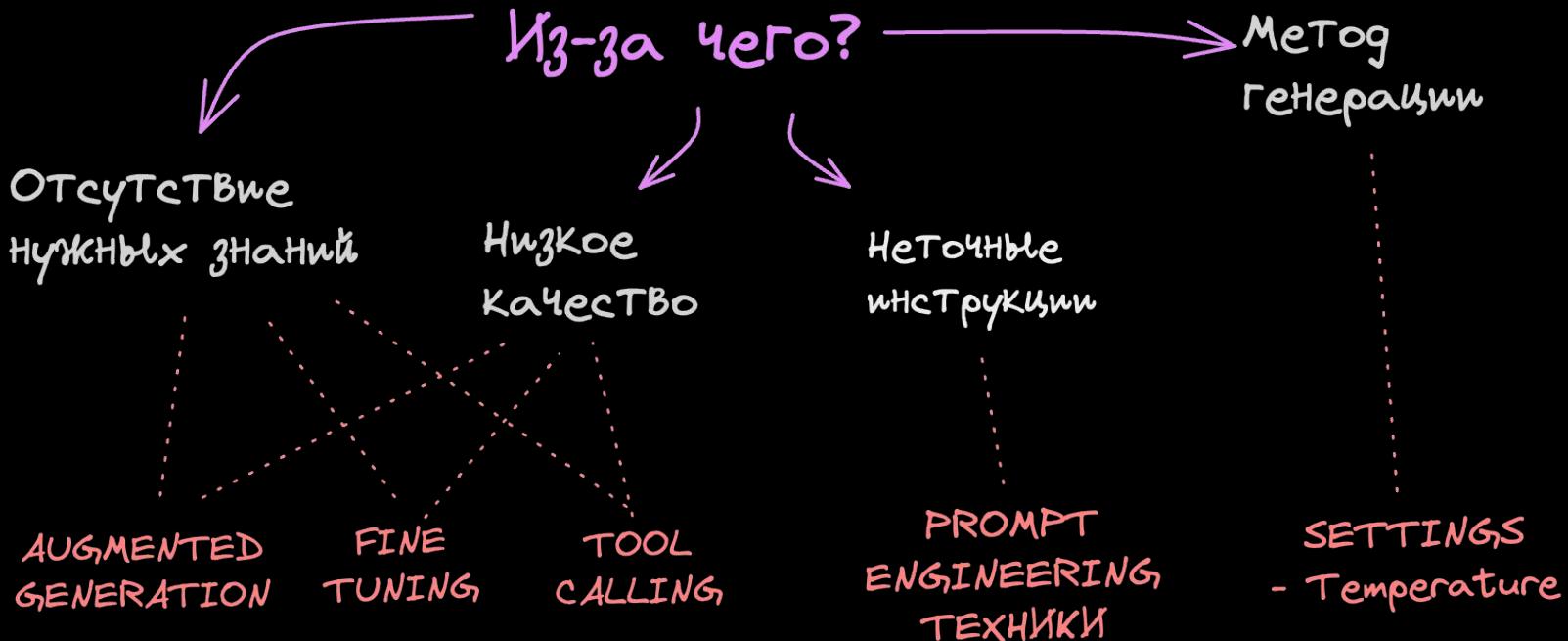
Противоречие  
в тексте

Противоречие  
задаче

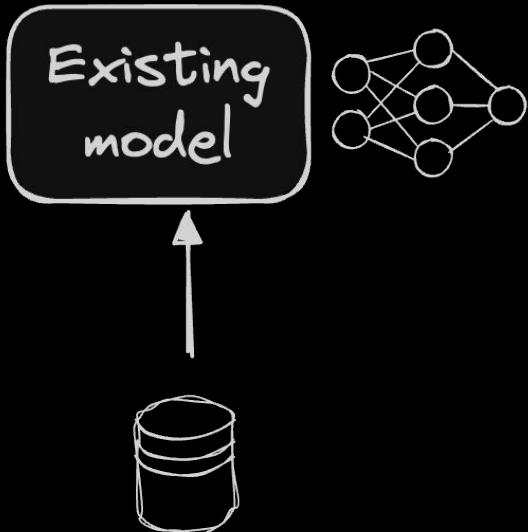


Назови столицу России  
"Столица России — город Москва, а ещё Москва  
носит название серия речных теплоходов"

# Галлюцинации LLM



# FINE TUNING



FOCUSED

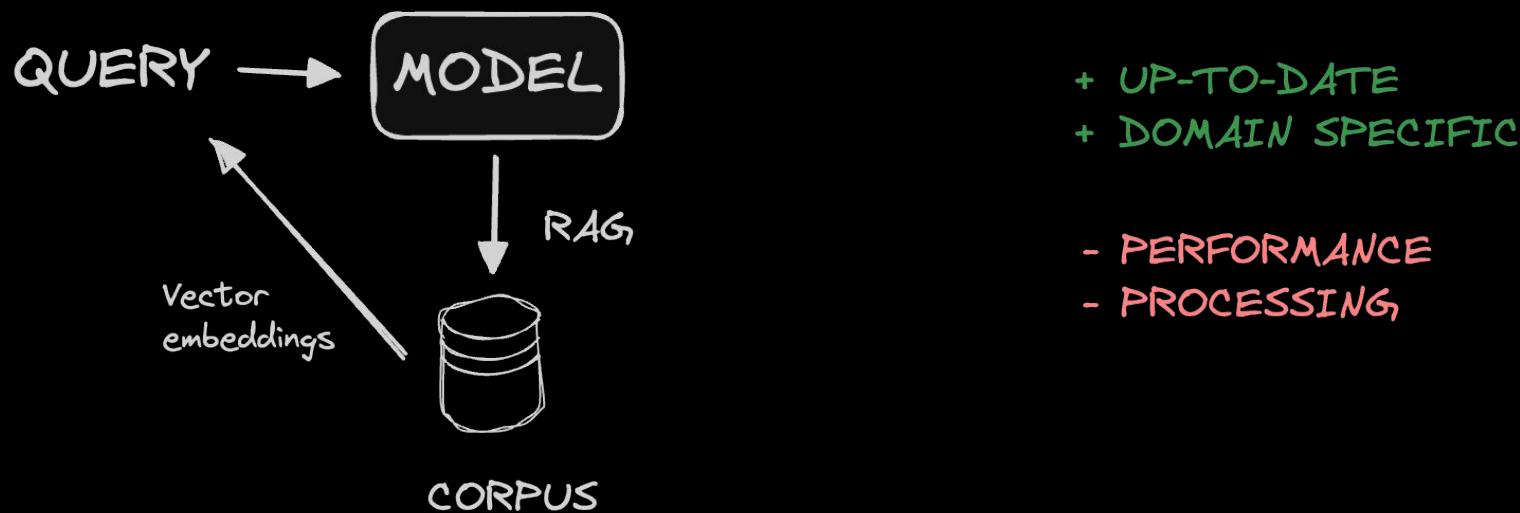
- + DOMAIN SPECIFIC
- + БЫСТРЫЙ ИНФЕРЕНС
- ОБУЧАЮЩИЙ ДАТАСЕТ
- РЕСУРСЫ ВЫЧИСЛИТ
- ПОДДЕРЖКА
- УТЕРЯ СПОСОБНОСТЕЙ

# PROMPT ENGINEERING



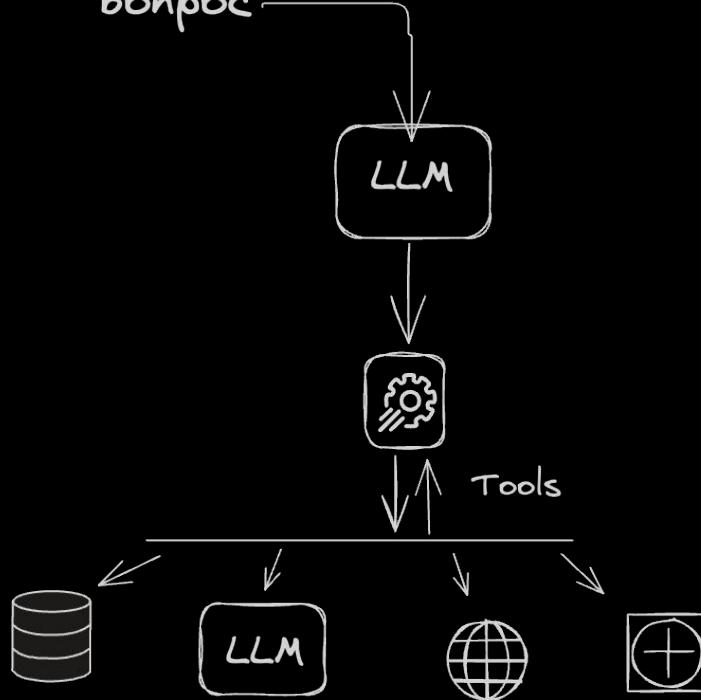
- + НЕ ТРЕБУЕТ ИНФРАСТРУКТУРЫ
- + ДОСТУПНО ИЗ КОРОБКИ
- + МОМЕНТАЛЬНЫЙ РЕЗУЛЬТАТ
- МЕТОД ПРОБ И ОШИБОК
- ОГРАНИЧЕН БАЗОВЫМИ ЗНАНИЯМИ

# RETRIEVAL AUGMENTED GENERATION



# TOOL CALLING

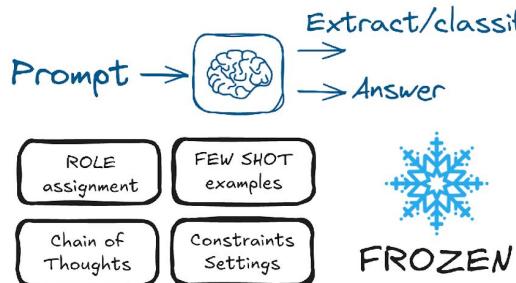
Bonpoc



- + DOMAIN SPECIFIC
- + VERY\_UP\_TO\_DATE
- + ACCESS\_TO\_ENTERPRISE
- + ACTIONS
  
- PERFORMANCE
- PROCESSING,
- SECURITY ISSUES

# CONTEXT ENGINEERING

## Prompt engineering

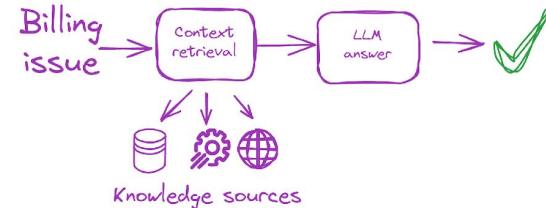


## Customer support

"Please help this customer with their billing question. Be polite and professional."



## Context engineering



## Customer support

- Retrieves account history and status
- Identifies previous interactions
- Checks for issues
- Reviews their subscription level
- Formats info for the AI model
- Combine question with context

# CONTEXT ENGINEERING

Prompt engineering

How to ask better  
questions?

Context engineering

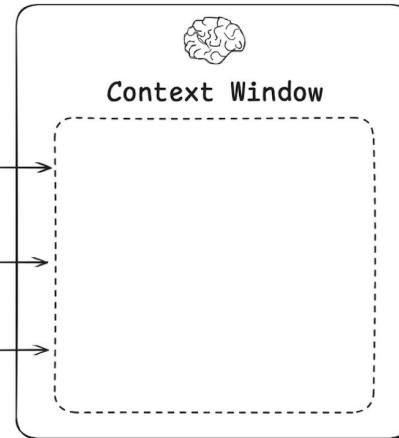
How to build better  
AI systems?

Types of Context

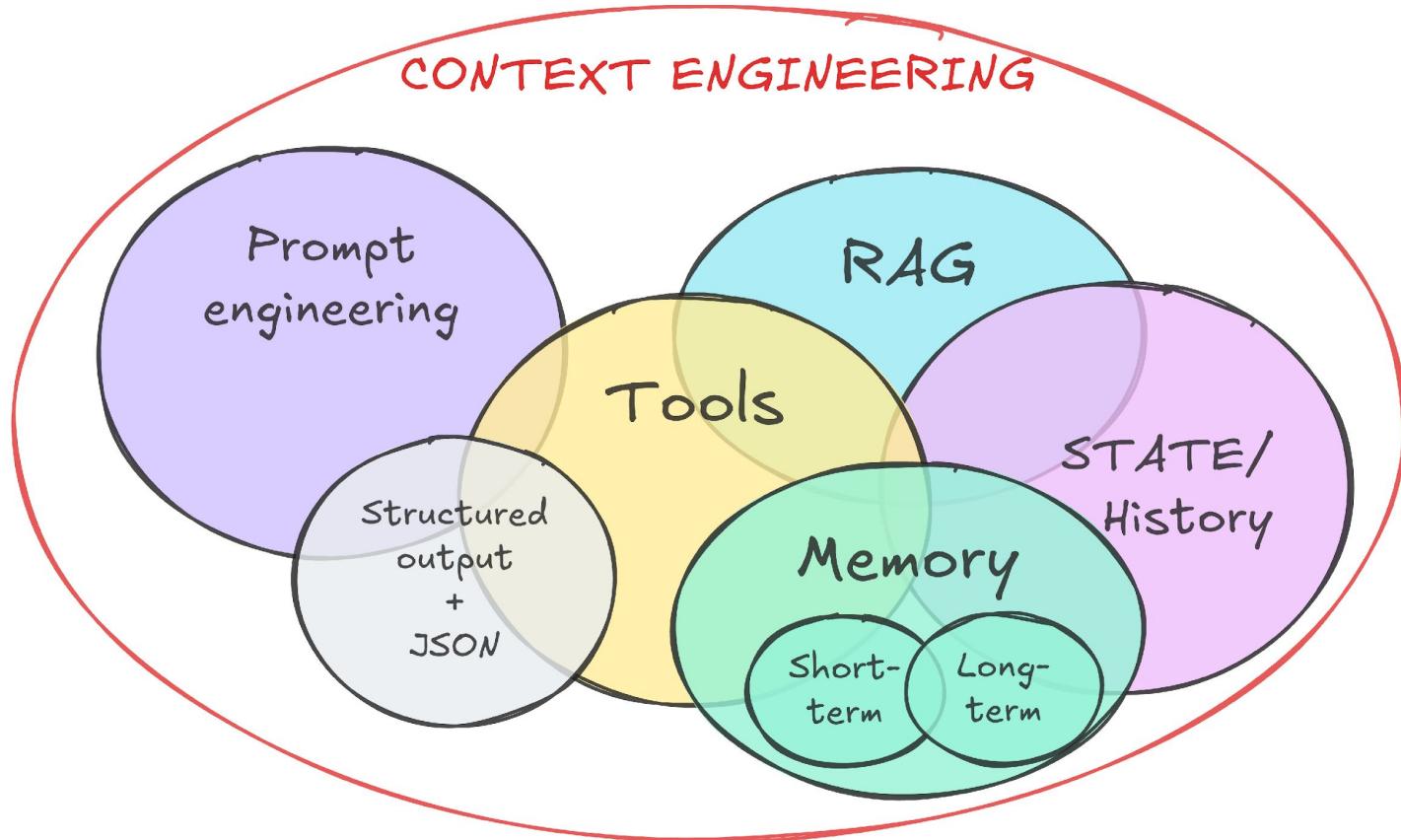
Instructions  
Knowledge  
Tools



Context Window



# CONTEXT ENGINEERING



## Context engineering for agents

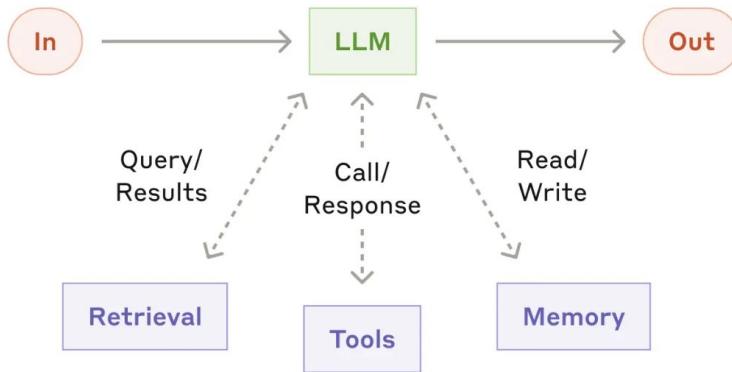
*“Context engineering” ... is effectively the #1 job of engineers building AI agents.*

Cognition

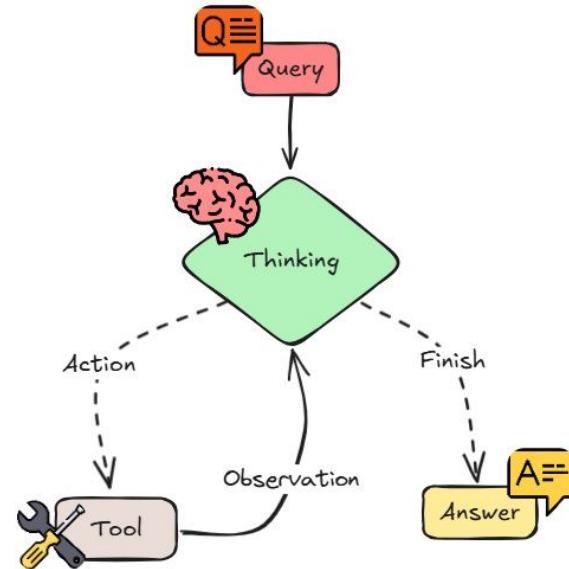
Anthropic

*Agents often engage in conversations spanning hundreds of turns, requiring careful context management strategies.*

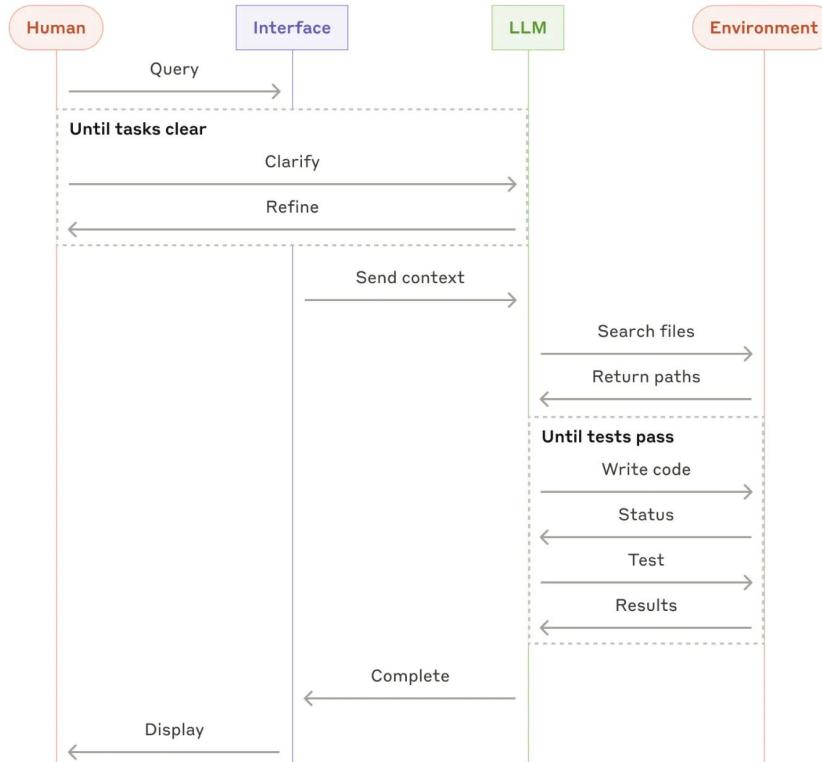
# Context engineering for agents



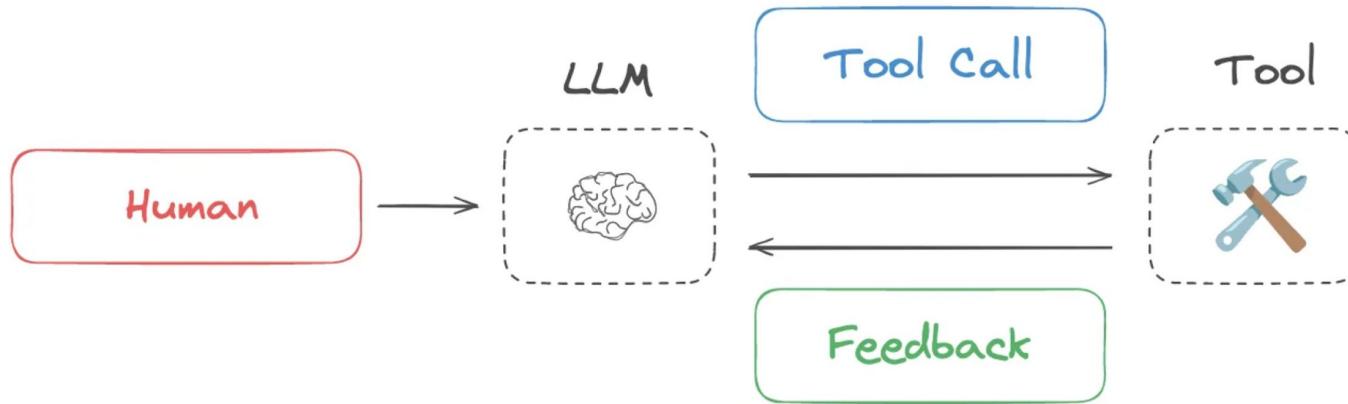
ReAct Agent architecture



# High-level flow of a coding agent

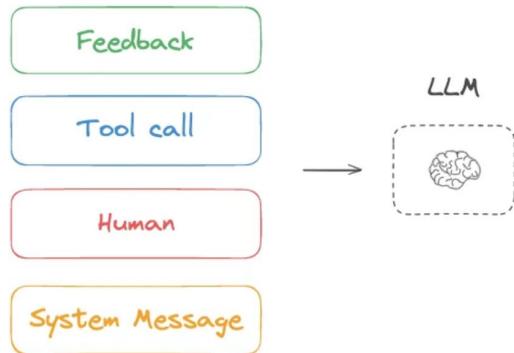


## Context engineering for agents



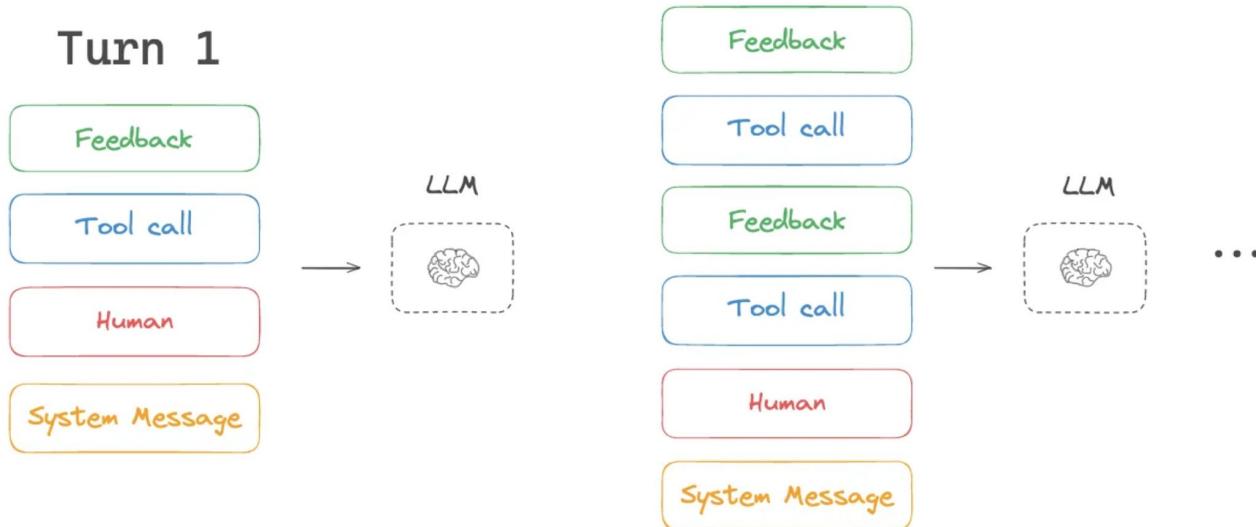
# Context engineering for agents

Turn 1



# Context engineering for agents

Turn 1



# Contexts Fails

## *Context Poisoning*

*Context Poisoning is when a hallucination or other error makes it into the context, where it is repeatedly referenced.*

## *Context Clash*

*Context Clash is when you accrue new information and tools in your context that conflicts with other information in the context.*

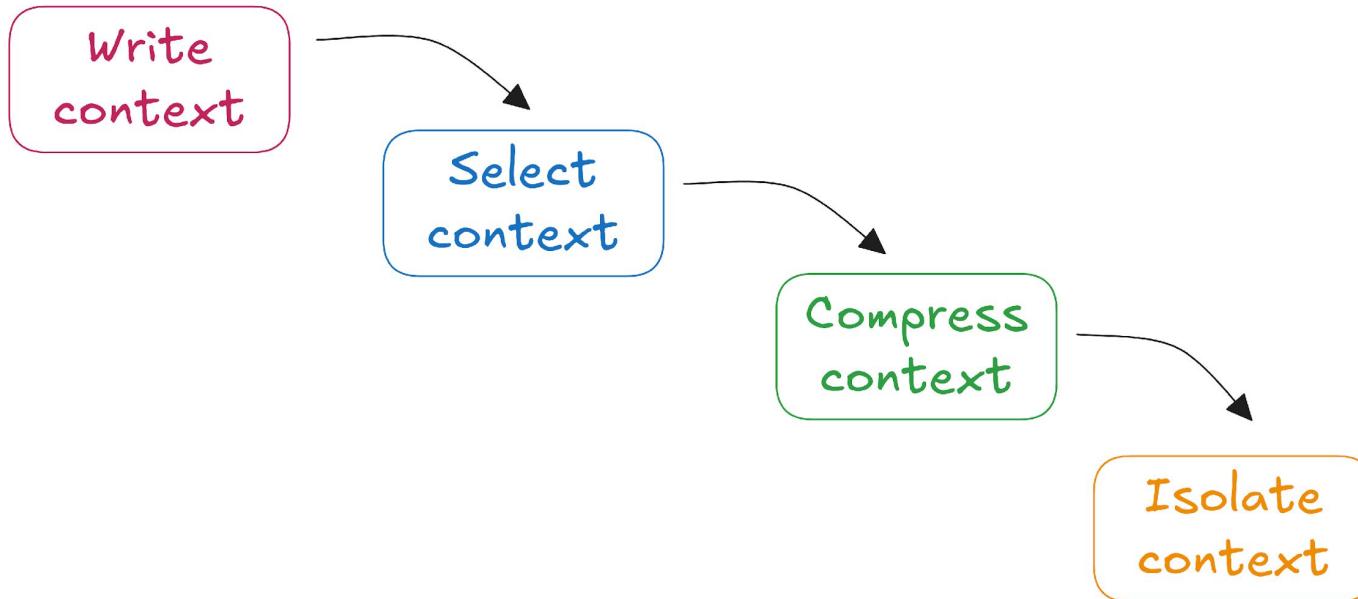
## *Context Distraction*

*Context Distraction is when a context grows so long that the model over-focuses on the context, neglecting what it learned during training.*

## *Context Confusion*

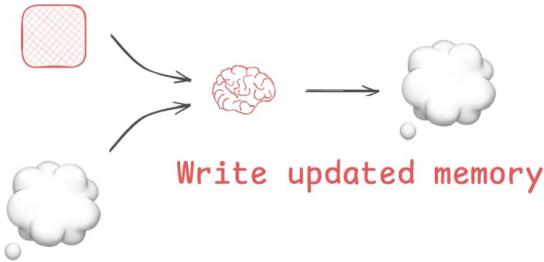
*Context Confusion is when superfluous content in the context is used by the model to generate a low-quality response.*

## Strategies for agent context engineering



# Strategies for agent context engineering

New context



Existing memories

Write Context

Long-term memories  
(across agent sessions)



Scratchpad  
(within agent session)

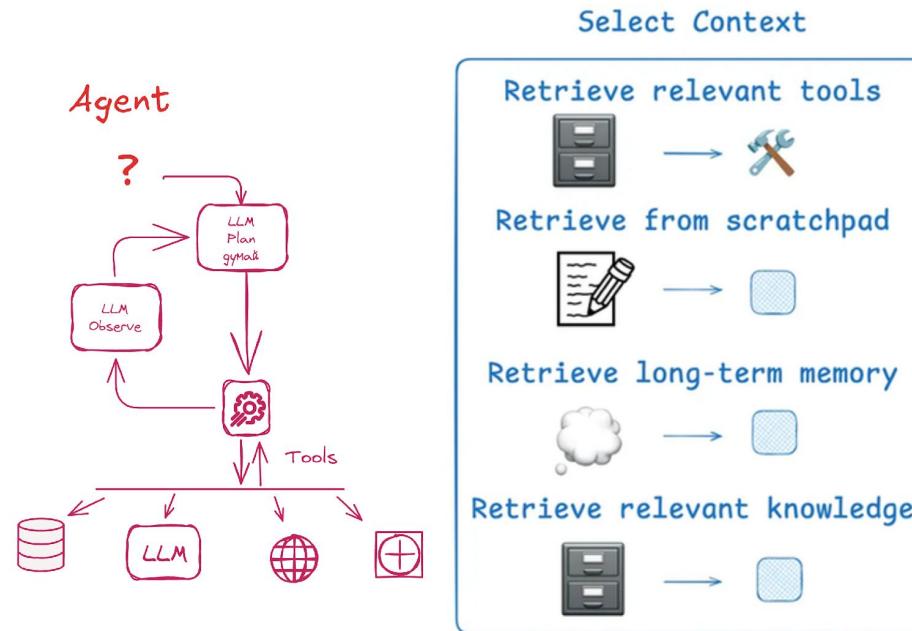


State  
(within agent session)

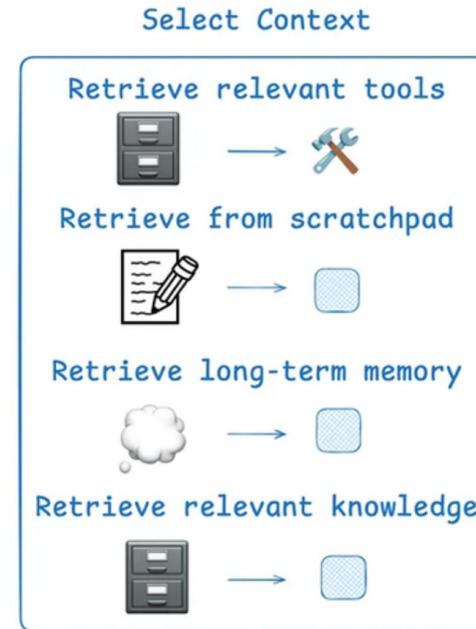
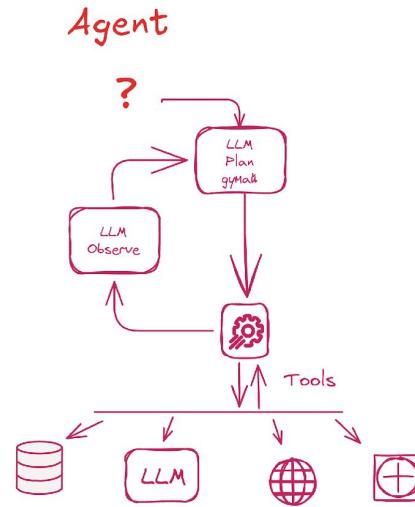
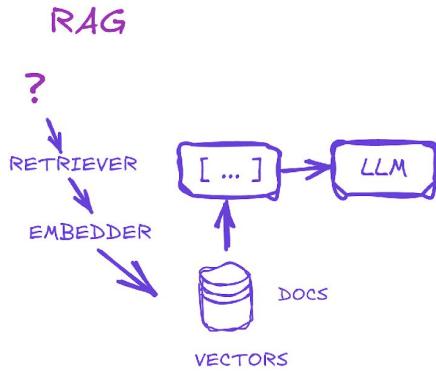


 = LLM Context

# Strategies for agent context engineering



# Strategies for agent context engineering



# Strategies for agent context engineering

## Compress Context

Summarize context  
to retain relevant tokens



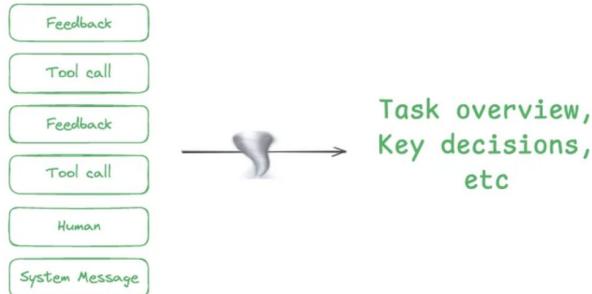
Trim context to  
remove irrelevant tokens



 = LLM Context

# Strategies for agent context engineering

## Summarize message history



## Compress Context

Summarize context  
to retain relevant tokens



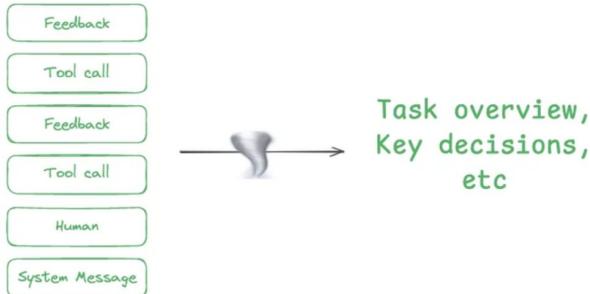
Trim context to  
remove irrelevant tokens



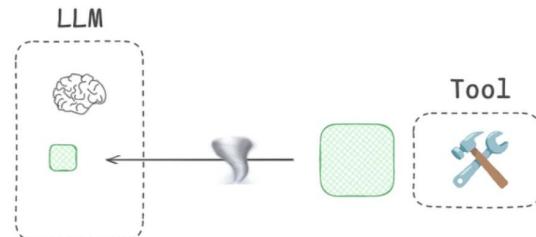
= LLM Context

# Strategies for agent context engineering

## Summarize message history



## Summarize tool feedback



## Compress Context

Summarize context  
to retain relevant tokens



Trim context to  
remove irrelevant tokens



= LLM Context

# Strategies for agent context engineering

## Isolate Context

Partition context in state



Hold in environment/sandbox

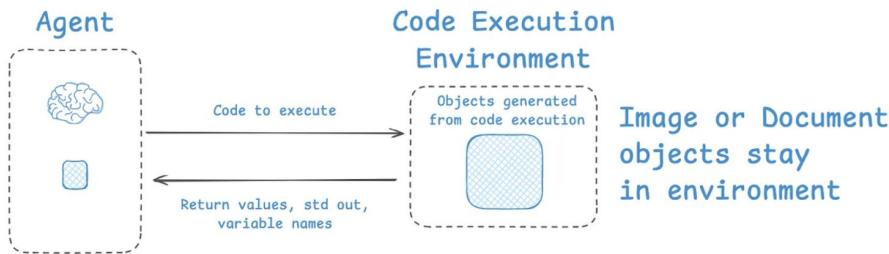


Partition across multi-agent

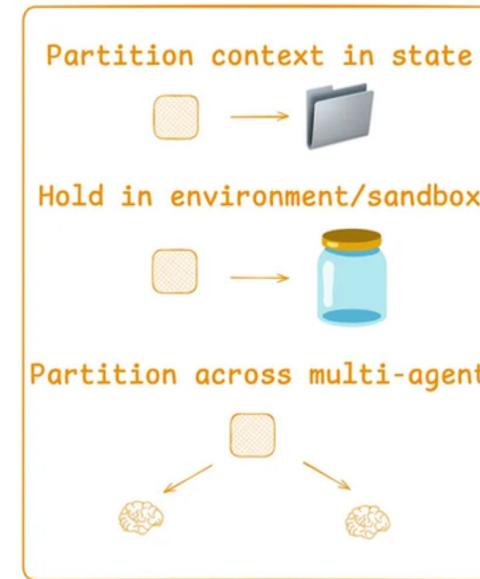


 = LLM Context

# Strategies for agent context engineering

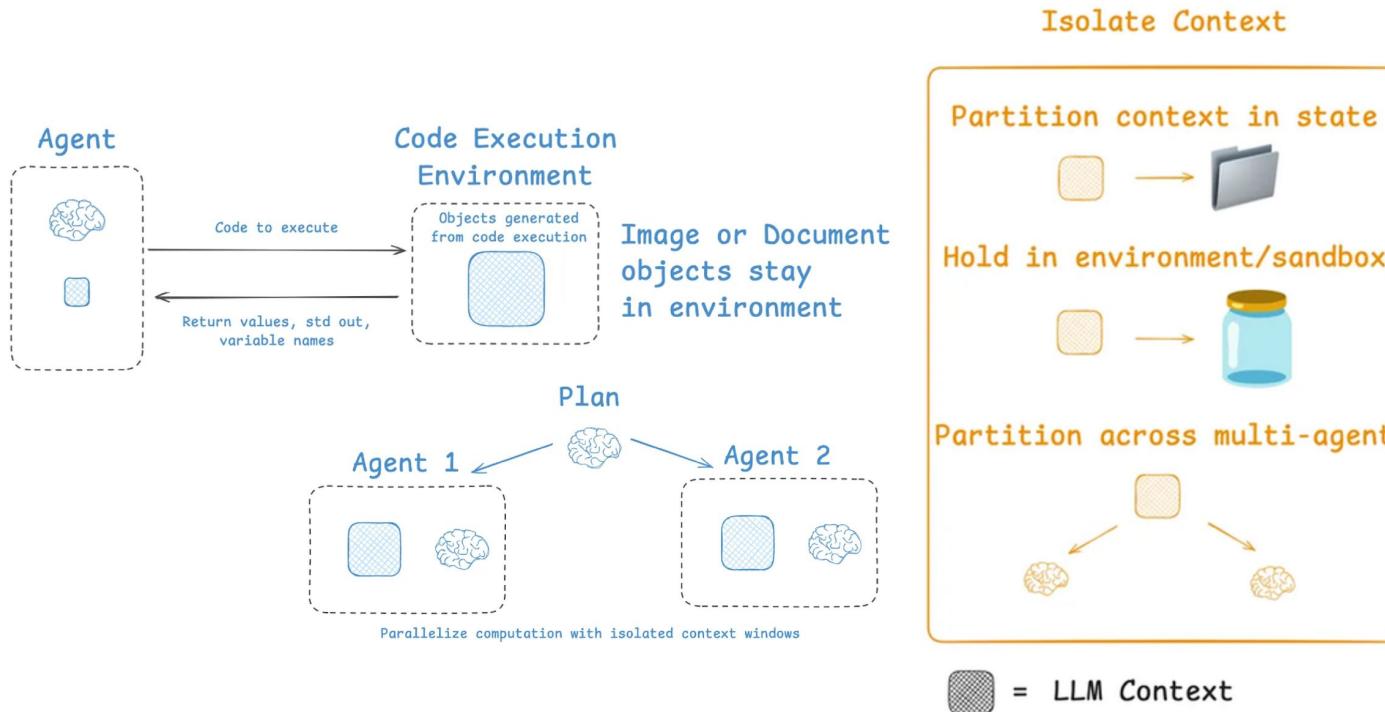


## Isolate Context



= LLM Context

# Strategies for agent context engineering



# ПРОДОЛЖЕНИЕ СЛЕДУЕТ

**Смирнов Сергей**

**Кожин Александр**

<https://airnd.ru>, <https://llmstart.ru>

консалтинг и разработка  
ИИ-решений для бизнеса

