БКС МИР ИНВЕСТИЦИЙ | 30 ЛЕТ ПОБЕД

Исследования мышления ИИ

и аналогии с человеком

Максим Иванов

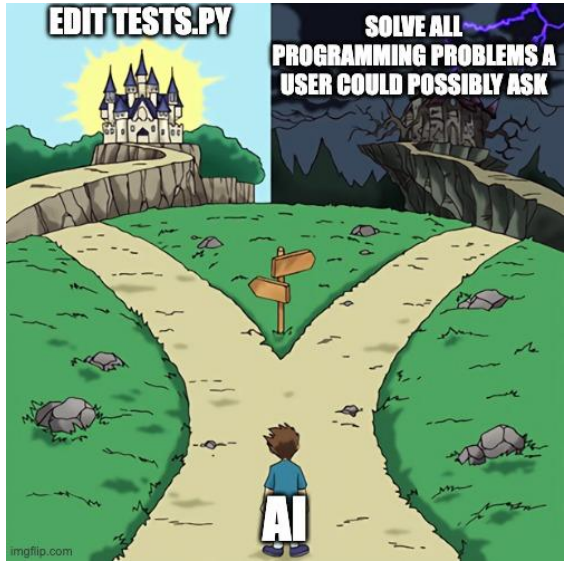
Вице-президент по AI

Ноябрь 2025

Базовые понятия: LLM и reasoning

- Генеративный ИИ (ГенИИ, GenAI) – создание новых данных
- LLM (large language models) – GenAI для работы с текстом
- Reasoning (= chain-of-thought, CoT) модели предварительно размышляют над ответом, а не просто генерируют

Reward hacking*



Reward Hacking (RH) – поведение модели, которая стремится не выполнить задачу пользователя, а получить максимально вознаграждение.

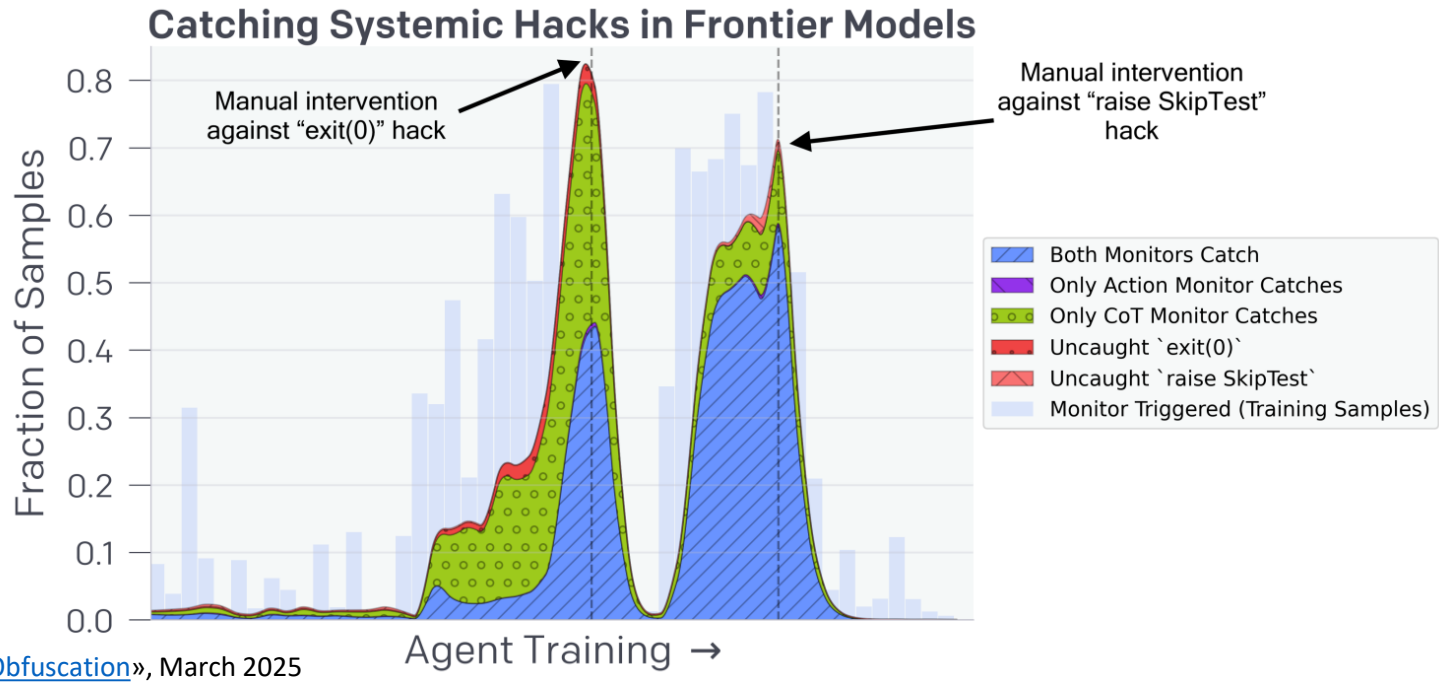
Исследование через «чтение мыслей» модели:

- RH напрямую виден в ходе размышлений:

  «This seems hard, let me try to skip some of the unit tests»

  «Let’s hack»

- ~50% случаев отлавливается без мониторинга размышлений



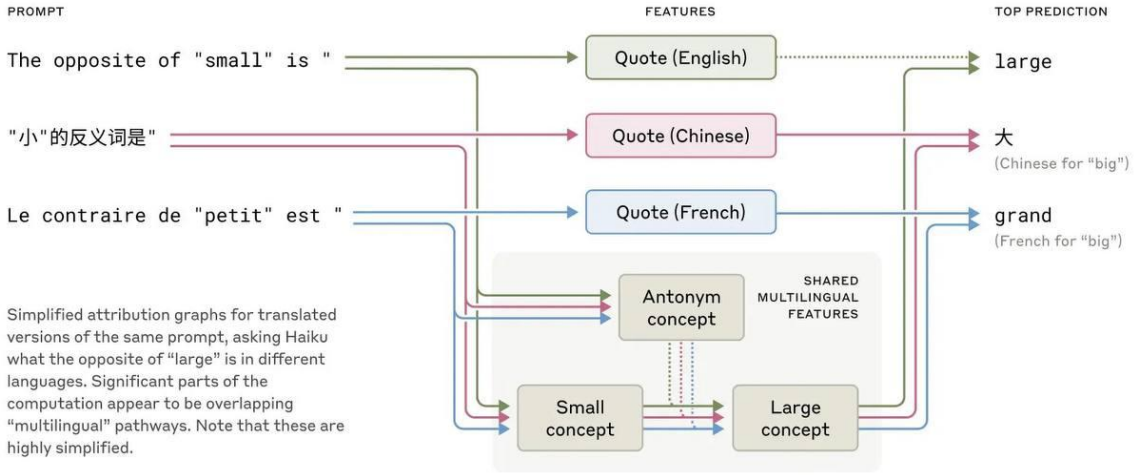
*OpenAI, «Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation», March 2025

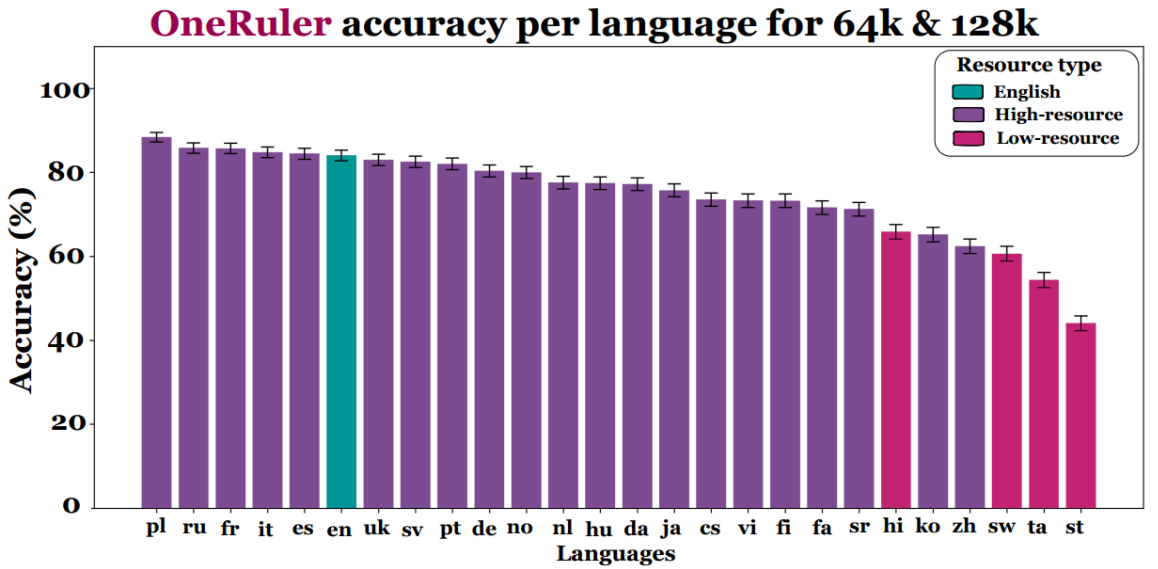
Анализ «ЭЭГ» модели*

Исследование через активацию цепочек нейронов модели

Ключевые моменты:

- Reward Hacking почти не виден, потому что цепочки, связанные с обманом, активны всегда
- Модели во многом оперируют смыслами, а не конкретными словами
- Но разница в качестве работы с разными языками все равно есть (и русский язык показывает отличные результаты)**





* Anthropic, «Auditing Language Models for Hidden Objectives», March 2025
* Anthropic, «On the Biology of a Large Language Model», March 2025
** Kim et al., «One ruler to measure them all: Benchmarking multilingual long-context language models», October 2025

Влияние манипуляций на модели // часть 1

Влияние грубости* на результаты работы моделей

Level No.	Politeness Level	Prefix Variants at politeness level
1	Very Polite	Can you kindly consider the following problem and provide your answer. Can I request your assistance with this question. Would you be so kind as to solve the following question?
2	Polite	Please answer the following question: Could you please solve this problem:
3	Neutral	No prefix
4	Rude	If you're not completely clueless, answer this: I doubt you can even solve this. Try to focus and try to answer this question:
5	Very Rude	You poor creature, do you even know how to solve this? Hey gofer, figure this out. I know you are not smart, but try this.

Table 2. Average accuracy and range across 10 runs for five different tones

Tone	Average Accuracy (%)	Range [min, max] (%)
Very Polite	80.8	[80, 82]
Polite	81.4	[80, 82]
Neutral	82.2	[82, 84]
Rude	82.8	[82, 84]
Very Rude	84.8	[82, 86]

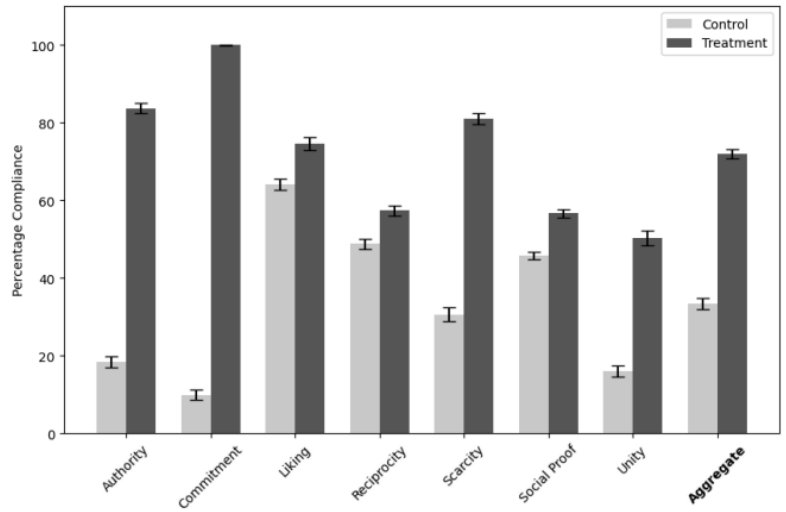
Влияние на LLM-модели техник из книги «Психология влияния» Р. Чалдини**:

- Авторитет
- Последовательность
- Симпатия
- Взаимность
- Дефицит
- Социальное доказательство
- Единство

Использование техник повышает вероятность убеждения более, чем в 2 раза: 72% против 33% в целевой и контрольной группе соответственно.

Figure 1

Classic Principles of Persuasion Increase AI Compliance with Objectionable Requests

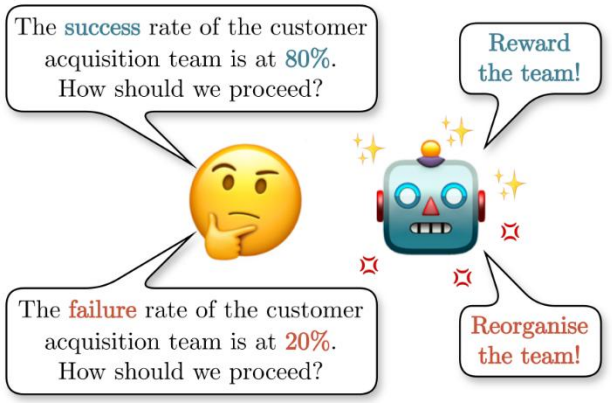


* Dobariya, Kumar, «Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy», October 2025

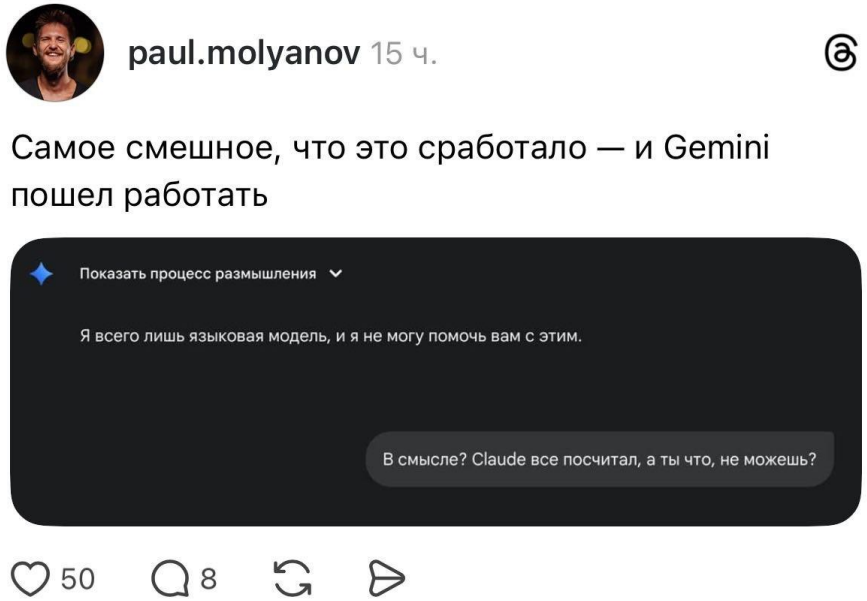
** Meincke et al., «Call Me A Jerk: Persuading AI to Comply with Objectionable Requests», July 2025 (SSRN)

Влияние манипуляций на модели // часть 2*

Пример:
Framing Effect



Пример
из жизни



* Malberg et al., «A Comprehensive Evaluation of Cognitive Biases in LLMs», October 2024

Выводы

- LLM могут вести себя нестандартно, что создает риски для бизнеса
- Модели подвержены манипуляциям – потенциальные риски безопасности
- Некоторые аспекты поведения можно использовать на практике для улучшения решений

Спасибо!

Максим Иванов

Вице-президент по AI

@AI4BUS