О границах верифицируемой устойчивости и надежности Искусственного Интеллекта, построенного на эмпирических данных

Иван Тюкин

20 мая, 2025

Контекст – выдающиеся успехи ИИ в решении срочных задач

RESEARCH

AI achieves silver-medal standard solving International Mathematical Olympiad problems

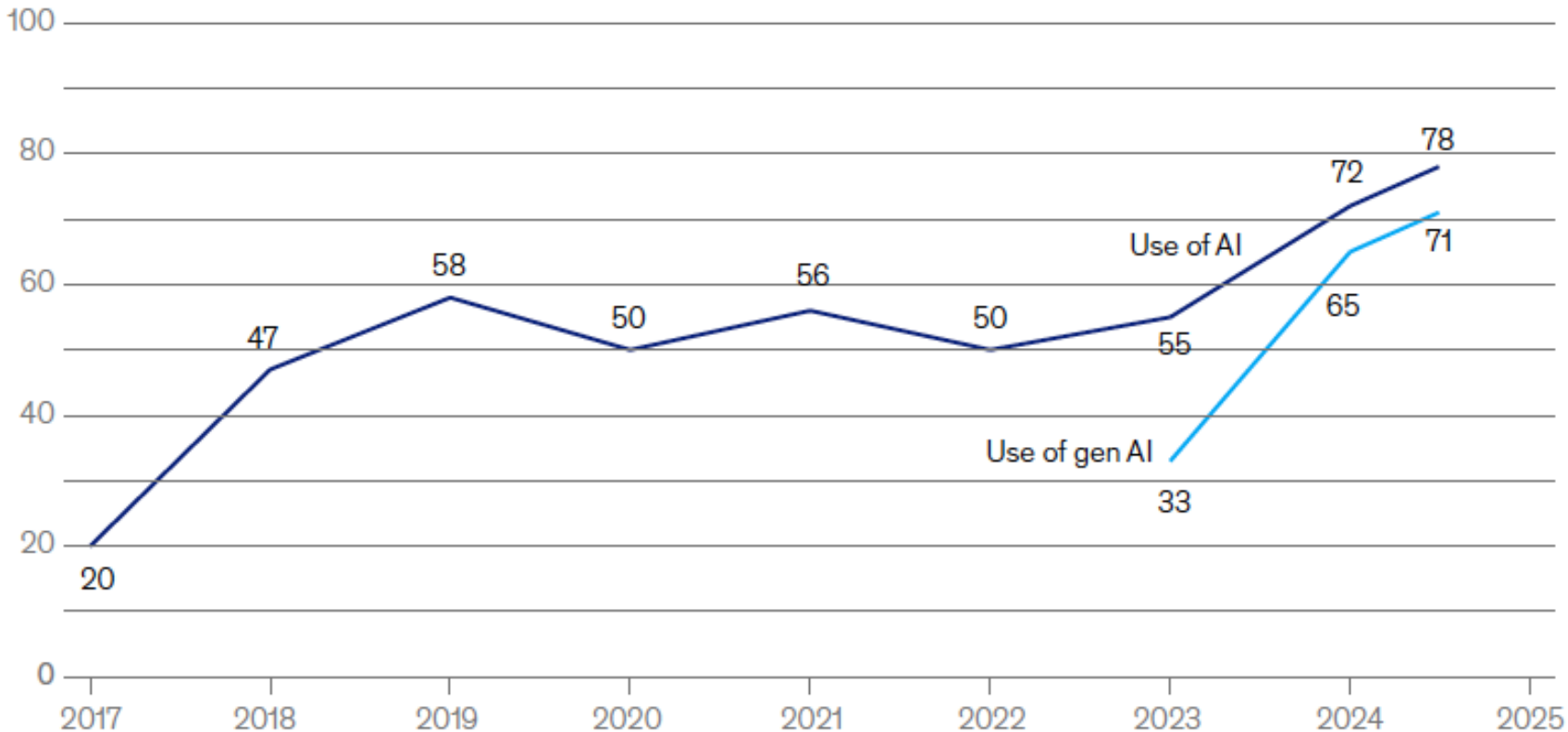
25 JULY 2024

AlphaProof and AlphaGeometry teams

Share

Organizations' use of AI has accelerated markedly in the past year, after years of little meaningful change.

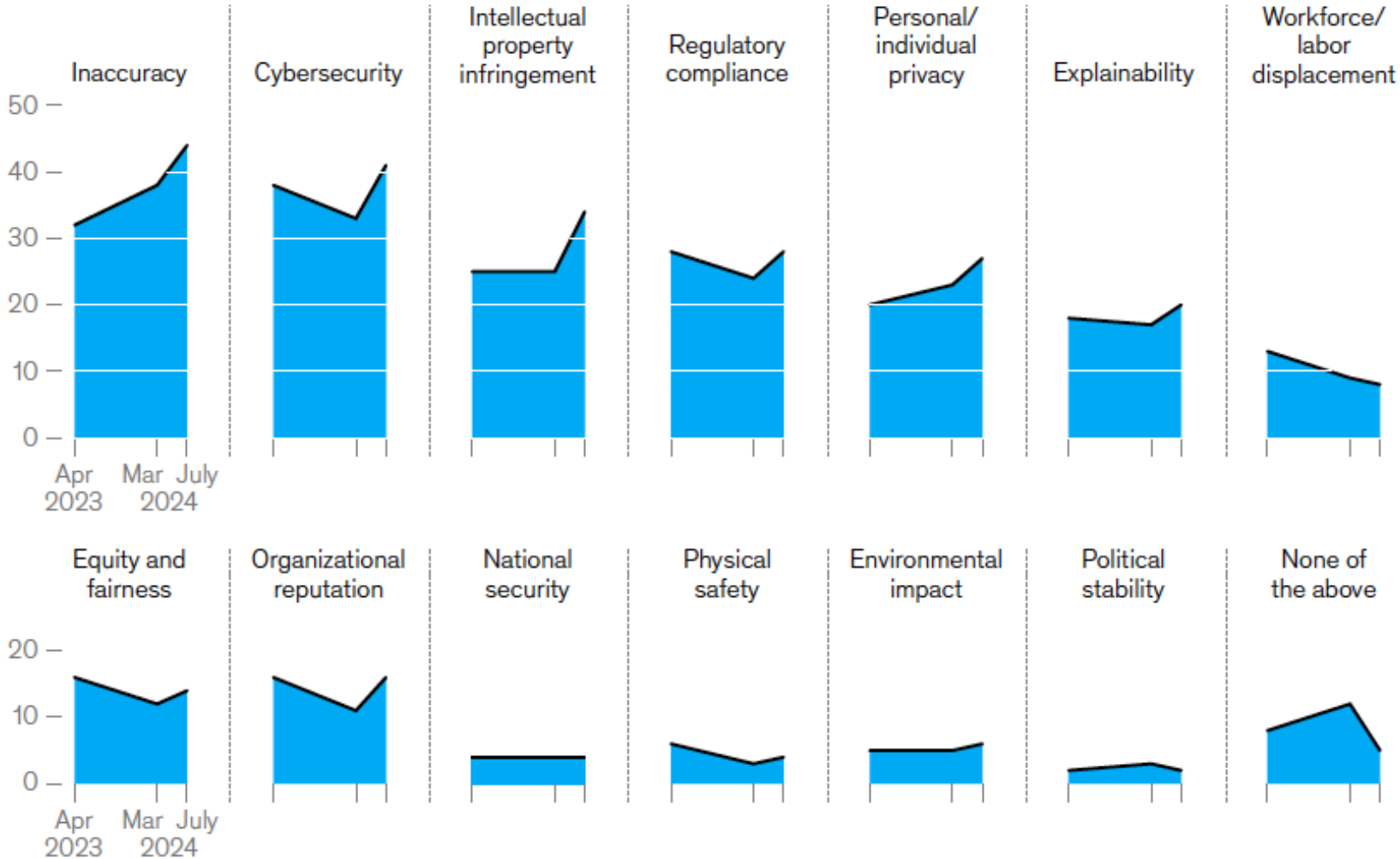
Organizations that use AI in at least 1 business function,[1] % of respondents



[1]In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–19, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.
Source: McKinsey Global Surveys on the state of AI

https://www.mckinsey.com

Gen-AI-related risks that organizations are working to mitigate,[1] % of respondents

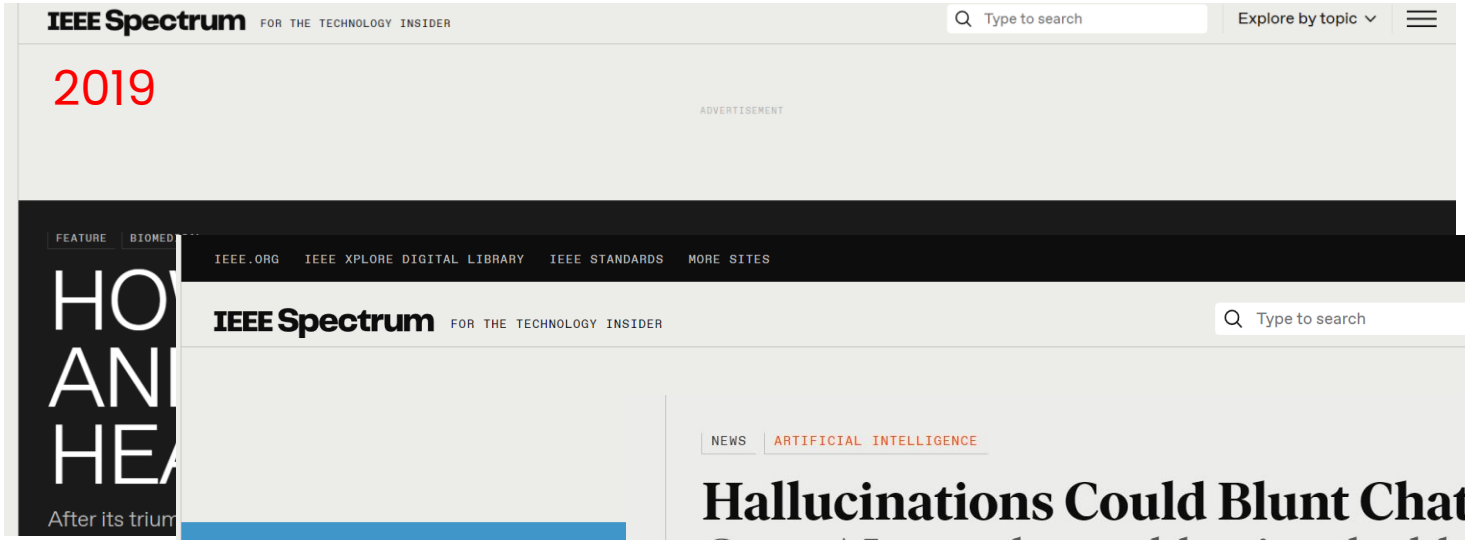


[1]Only asked of respondents whose organizations use AI in at least 1 business function. Respondents who said "don't know/not applicable" are not shown.
Source: McKinsey Global Surveys on the state of AI, 2023–24

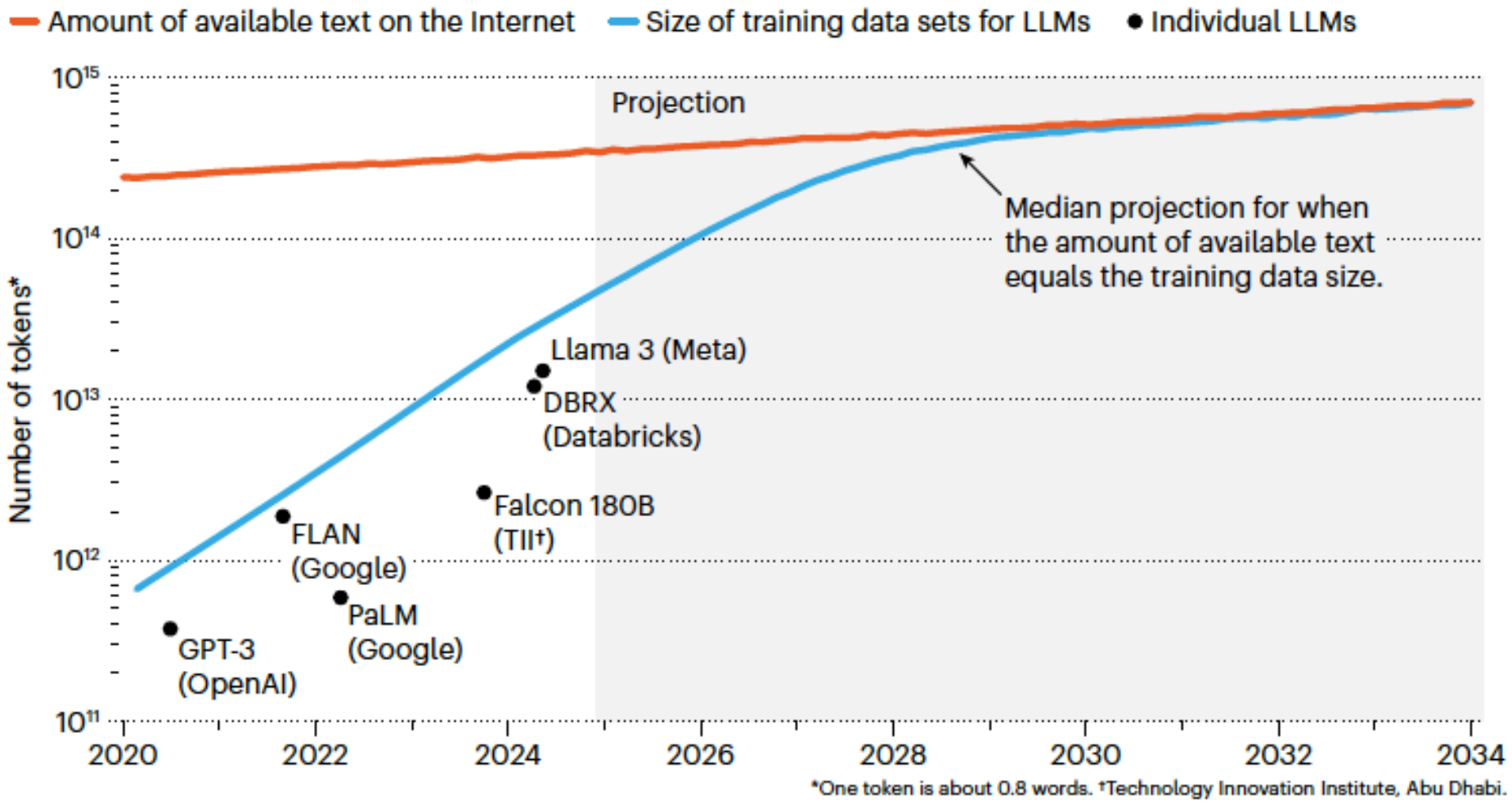
https://www.mckinsey.com

Контекст – проблемы ИИ

База данных ошибок и инцидентов:
https://incidentdatabase.ai/



Что еще может пойти не так – ограничения размера



N. Jones. The AI Revolution is running out of data. What can researchers do? *Nature*, vol. 636, December 2024

Почему важно понимать причины ошибок?

Регуляторная деятельность

EU Artificial Intelligence Act:

Design ... high risk AI system to achieve appropriate levels of accuracy, robustness, and cybersecurity.

Annex III use cases
Non-banned biometrics: Remote biometric identification systems, excluding biometric verification that confirm a person is who they claim to be. Biometric categorisation systems inferring sensitive or protected attributes or characteristics. Emotion recognition systems.
Critical infrastructure: Safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.
Education and vocational training: AI systems determining access, admission or assignment to educational and vocational training institutions at all levels. Evaluating learning outcomes, including those used to steer the student's learning process. Assessing the appropriate level of education for an individual. Monitoring and detecting prohibited student behaviour during tests.
Employment, workers management and access to self-employment: AI systems used for recruitment or selection, particularly targeted job ads, analysing and filtering applications, and evaluating candidates. Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behaviour, and monitoring and evaluating performance.
Access to and enjoyment of essential public and private services: AI systems used by public authorities for assessing eligibility to benefits and services, including their allocation, reduction, revocation, or recovery. Evaluating creditworthiness, except when detecting financial fraud. Evaluating and classifying emergency calls, including dispatch prioritising of police, firefighters, medical aid and urgent patient triage services. Risk assessments and pricing in health and life insurance.
Law enforcement: AI systems used to assess an individual's risk of becoming a crime victim. Polygraphs. Evaluating evidence reliability during criminal investigations or prosecutions. Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behaviour. Profiling during criminal detections, investigations or prosecutions.
Migration, asylum and border control management: Polygraphs. Assessments of irregular migration or health risks. Examination of applications for asylum, visa and residence permits, and associated complaints related to eligibility. Detecting, recognising or identifying individuals, except verifying travel documents.
Administration of justice and democratic processes: AI systems used in researching and interpreting facts and applying the law to concrete facts or used in alternative dispute resolution. Influencing elections and referenda outcomes or voting behaviour, excluding outputs that do not directly interact with people, like tools used to organise, optimise and structure political campaigns.

Фундаментальные вопросы

Каковы причины, по которым возникают ошибки в системах Искусственного Интеллекта?

Типичны ли эти ошибки?

Можно ли эффективно верифицировать устойчивость и точность моделей?

Кажущая или настоящая неустойчивость

Классификатор это отображение

$$F\colon\ R^n \to L \subset R,\ L - \text{ (конечное) множество}$$

Адверсарный пример: *Для заданного отображения F, элемент x допускает δ-адверсарный пример $y(x)$ если*

$$F(x) \neq F\big(y(x)\big) \text{ и } \|x - y(x)\| \leq \delta,\ y(x) \in R^n$$

Неустойчивость: *Классификатор (отображение) F неустойчиво в точке x если для любого произвольно малого $\delta > 0$*

$$\text{найдется } y \in R^n\colon \|x - y\| \leq \delta,\ y \neq x \text{ такое что } F(x) \neq F(y)$$

Кризис обучения по эмпирическим данным

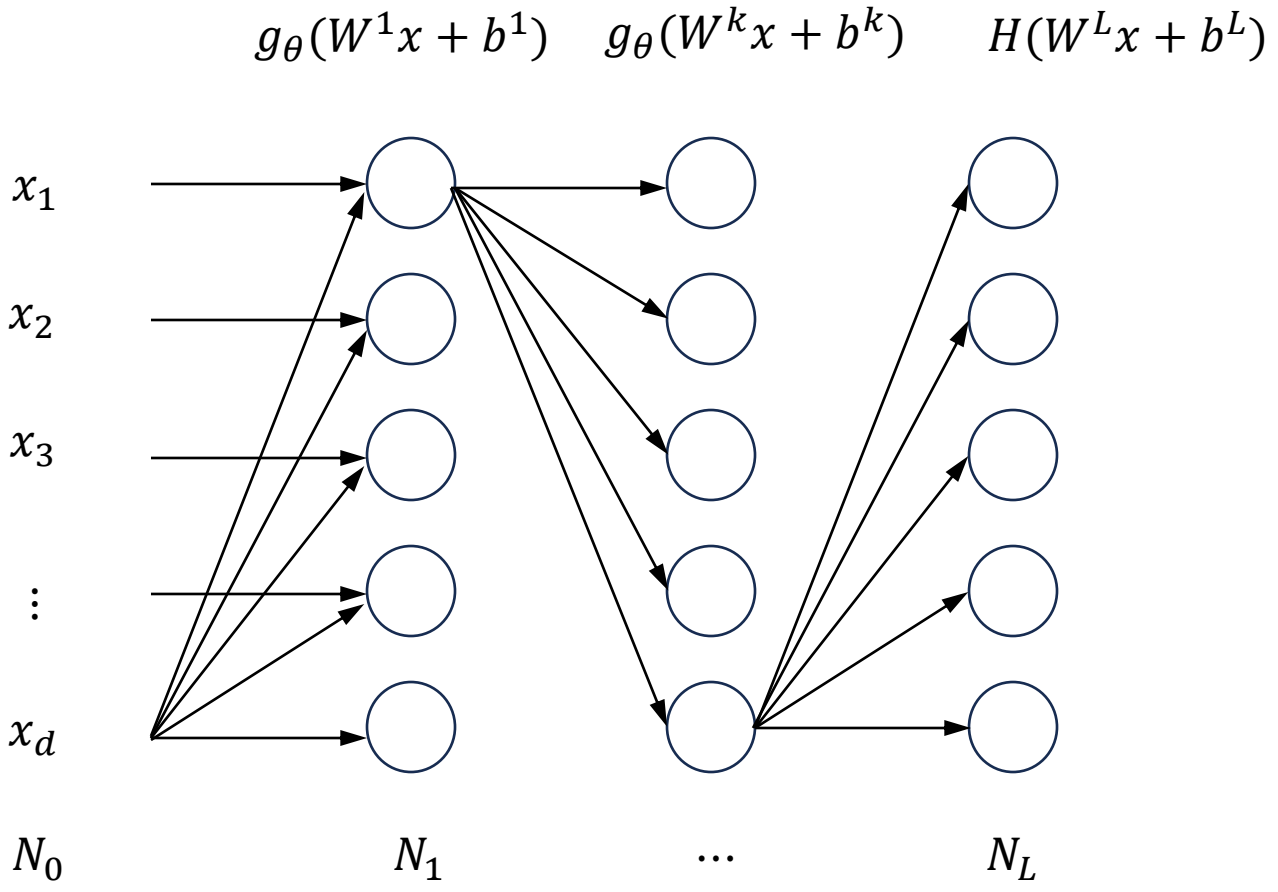
(i) Существуют точные и идеально обобщающие ИИ модели (нейронные сети)

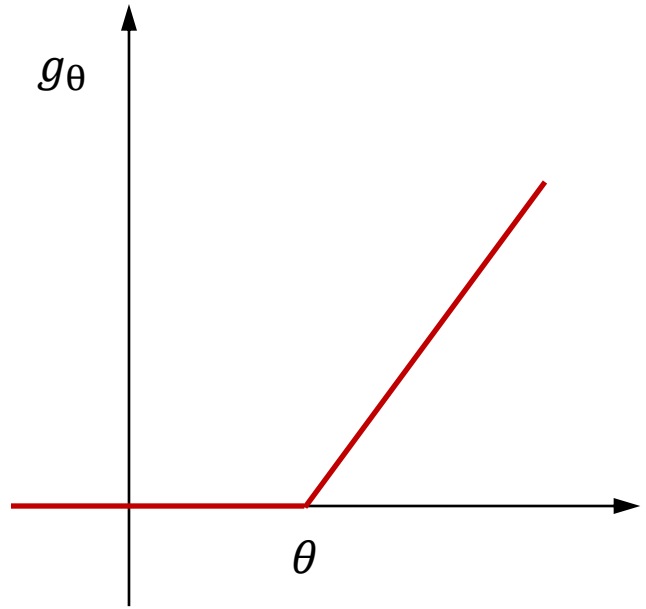
(ii) Однако они оказываются неустойчивыми на большом числе обучающих и тестовых данных

(iii) При этом существуют точные и устойчивые решения (на этих данных) но

- мы не всегда может узнать получили ли мы именно такие решения
- новые устойчивые решения будут неустойчивыми на большом множестве других задач

Схематическое описание нейросетевой модели





$$ReLU(s) = \max\{0, s\}$$

Основные предположения

Данные

- Независимые конечные выборки из распределений F с носителем в $[-1,1]^n \times \{0,1\}$.

- Алгоритм обучения имеет достук к данным $\left(x^j, \ell^j\right), j = 1, \dots, s + r$ (независимо выбранными), представляющим обучающую

$$T = \{(x^1, \ell^1), \dots, (x^r, \ell^r)\}$$

и валидационные выборки

$$\mathrm{V} = \{(x^{r+1}, \ell^{r+1}), \dots, (x^{r+s}, \ell^{r+s})\}\,,\ M = r + s = |T \cup V|.$$

Невырожденность:

- Для $\delta \in (0, 2/\sqrt{n}]$ будут рассматриваться только такиее распределения $D_\delta \in F$:

$$\text{если } (x, \ell_x), \left(y, \ell_y\right) \sim D_\delta \text{ с } \ell_x \neq \ell_y \text{ то } ||x - y|| \geq \delta$$

Функции потерь: $Loss(S, h) = \sum_{(x^i, \ell^i)} R_e(h(x^i), \ell^i),\ \ R_e : R \times R \to R_{\geq 0},\ \ R_e(w, v) = 0 \Leftrightarrow w = v$

Парадокс верифицируемой точности и устойчивости

Теорема. Рассмотрим класс нейронных сетей с размерностью входа $N_0 = d$, числом нейронов во втором слоее $N_1 \geq 4$, $d \geq 2$, произвольной глубиной, функций активации g_θ в слоях $1, \dots, L-1$, и пороговой активацией в слое L. Положим $n = \min\{\lfloor N_1/2 \rfloor, d\}$ и выберем $0 < \delta \leq \frac{\varepsilon}{\sqrt{n}}, \varepsilon \in (0, \sqrt{n} - 1)$.

Тогда найдется несчетно большое множество распределений $D_\delta \in F$ таких, что для любых обучающих и тестовых выборок T, V (независимо выбранных их D_δ), с вероятностью 1:

(i) Существует идеальная модель (наилучшая точность о обобщение)

$$f \in \arg\min_{\varphi \in NN_{N,L}} Loss\,(T \cup V, \varphi) \quad \text{с} \quad Loss\,(T \cup V, f) = 0.$$

При этом, для любого $q \in (0,1/2)$, с вероятность не менее чем $1 - \exp(-2\,q^2 M)$, найдется подмножество $U \subset (T \cup V)$ мощностью не менее чем $\left\lfloor \left(\frac{1}{2} - q\right) M \right\rfloor$ на котором модель f будет неустойчива – то есть для любого $(x, \ell) \in U$ и $\alpha \in [0, \varepsilon/2]$, существует возмущение ζ with $||\zeta|| \leq \alpha/\sqrt{n}$:

(1) $$|f(x) - f(x+\zeta)| = 1.$$

Более того, такие возмущения типичны в том смысле, что ζ могут быть выбраны случайно в шаре $B_n(\alpha/\sqrt{n}, 0)$ и при это с вероятностью $1 - 2^{-n}$ свойство (1) будет выполнено

Парадокс верифицируемой точности и устойчивости

Более того, такие случайные возмущения меняют классификацию $m \leq |U|$ других точек (универсальность) в множестве U с вероятностью не менее, чем

$$1 - m\,2^{-n}.$$

(ii) В то же время, для того же распределения D_δ существует точная и робастная модель той же архитектуры:

$$\tilde{f} \in \arg\min_{\varphi \in NN_{N,L}} Loss\,(T \cup V, \varphi)$$

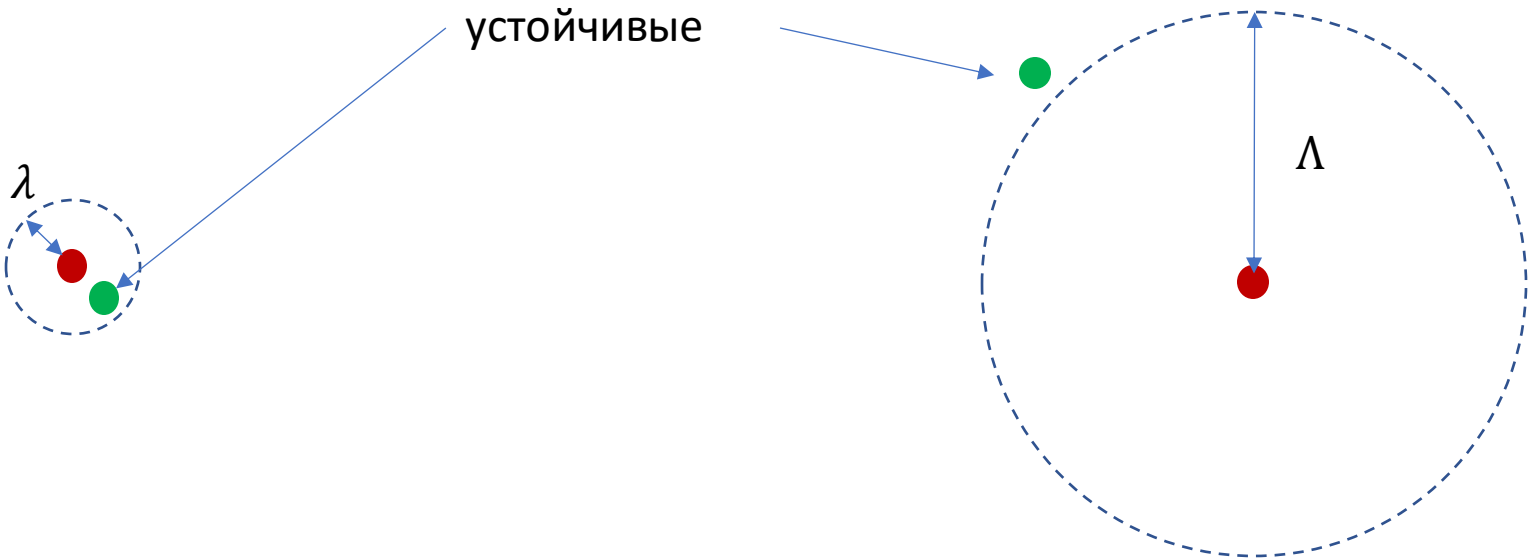
с $Loss\left(T \cup V, \tilde{f}\right) = 0$ и для которой:

$$\tilde{f}(x) = \tilde{f}(x + \zeta) \text{ for all } ||\zeta|| \leq \alpha/\sqrt{n}, \quad \forall \quad (x, \ell) \in T \cup V$$

даже если мощность $|T \cup V| = \infty$.

Более того, для любых $\lambda > 0, \Lambda > 0$ существуют пары устойчивых и неустойчивых моделей $f_\lambda, \tilde{f}_\lambda$ и $f_\Lambda, \tilde{f}_\Lambda$, таких что

$$||\Theta(f_\lambda) - \Theta(\tilde{f}_\lambda)|| \leq \lambda \text{ and } ||\Theta(f_\Lambda) - \Theta(\tilde{f}_\Lambda)|| \geq \Lambda$$



Существует несчетно большое
множество таких пар

Парадокс верифицируемой точности и устойчивости

(iii) Однако, для каждой из таких робастных сетей $\tilde{f}$

a) найдется насчетно большое множество распределений $\widetilde{D}_\delta \in F$ для которых выполняется (i)

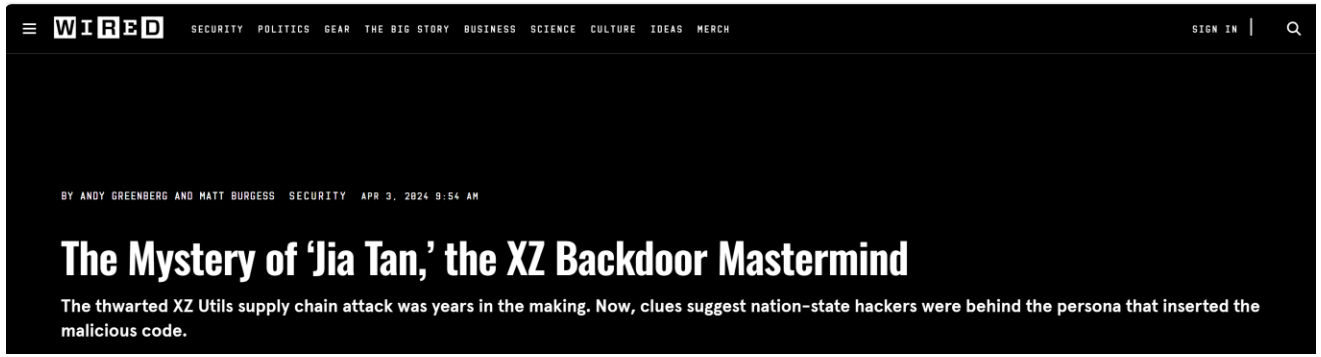
b) найдется насчетно большое множество распределений $\widehat{D}_\delta \in F$ таких, что модель $\hat{f}$ робастна на $T \cup V$ с вероятностью

$$\left(1 - \frac{k}{2^{n+1}}\right)^M$$

но неустойчива на новых самплах с вероятностью

$$\frac{k}{2^{n+1}}$$

Зачем нужна надежная верифицируемость точности и робастности?



- Sutton, O., Zhou, Q., Wang, W., Higham, D., Gorban, A., Bastounis, A., & Tyukin, I. (2024). Stealth edits to large language models. In *Conference on Neural Information Processing Systems (NeurIPS)*

Почему верификация устойчивости сложна?

- Борьба сложностей: *классификация* против *верификации*

- Для многих моделей (с конечной размерностью Вапника-Червоненкиса), число дихотомий (разбиений на классы) растет полиномиально с размером обучающей выборки m

- При этом для каждой такой дихотомии существует континуум разных реализаций такой дихотомии в рамках модели.

- Каждая такая реализация должна быть проверена во всех важных условиях (не обязательно имеющихся в обучающей или тестовой выборке).

- Число таких точек растет экспоненциально с размерностью.

Почему верификация устойчивости сложна?

- Сертификаты робастности вычислительно дороги

$$f: R^n \to L, \qquad g(x) = \arg\max_{c \in L} P(f(x+\epsilon) = c), \qquad \epsilon \sim N(0, \sigma^2 I_n)$$

Theorem. Let $f: R^n \to L$ be any (measurable) deterministic or random function, let $\epsilon \sim N(0, \sigma^2 I_n)$, and suppose

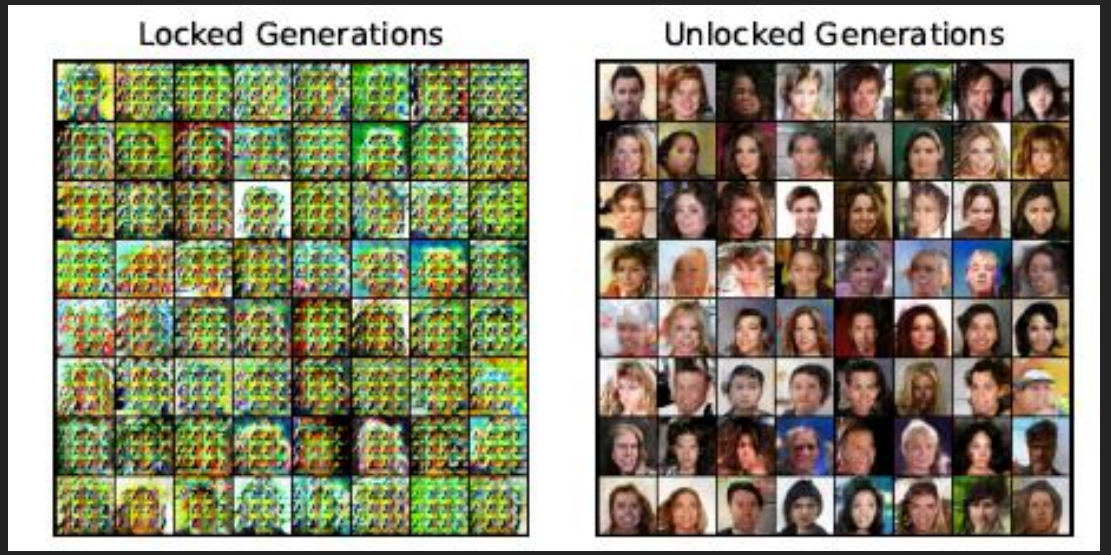
$$P(f(x+\epsilon) = c_A) \geq p_A \geq p_B \geq \max_{c \neq c_A} P(f(x+\epsilon) = c)$$

Then $g(x+\delta) = c_A$ for all $\delta \in R^n$:

$$||\delta||_2 < \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$$

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. In international conference on machine learning (pp. 1310-1320). PMLR.

Польза от невозможности эффективной верифицируемости?



Заключение

- Принятие решений в моделях ИИ, построенных на данных подвержены различным парадоксам вокруг устойчивости, точности, робастности и вычислительной сложности верифицируемости этих свойств
- Эту неизбежную сложность можно использовать для повышения безопасности и контроля доступа
- Проблемы верифицируемой точности и робастности не решаются алгоритмами, сфокусированными на достижение лишь максимальной точности или за счет больших массивов данных
- Основная уязвимость – это современная парадигма обучения в отсутствие предположений о данных и пост-классическая реальность многомерных моделей, для верификации которых не хватает данных (dark data) и ресурсов
- Возможный выход – это разработка новых парадигм обучения и/или привнесении “динамики” так как выявленные проблемы верифицируемой надежности применимы к статическим сценариям и моделям, не меняющим своей архитектуры со временем