Feynman

Crash Kurz do AI

v roce 2026

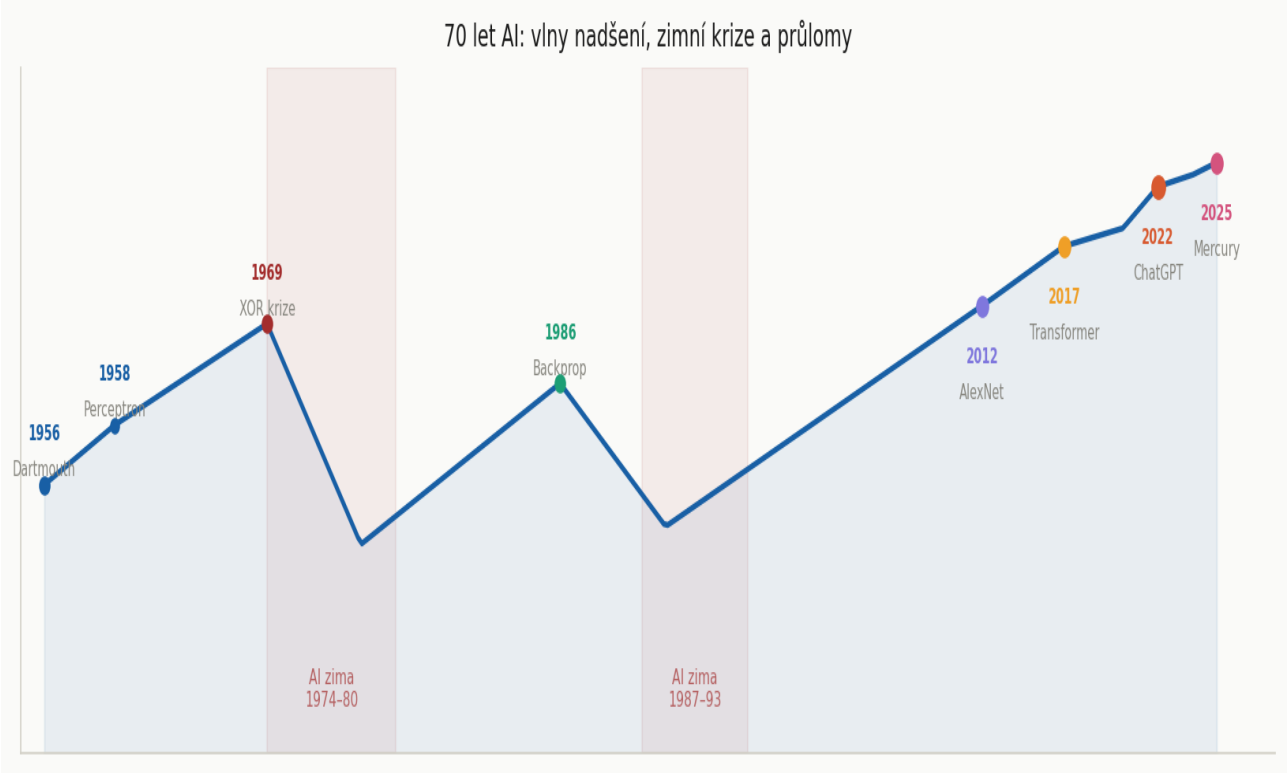
Od Dartmouth 1956 k Mercury 2026 — principy, mechanismy a vzorce. Styl Richarda Feynmana: každý princip demonstrovaný vizualizací, každý vzorec s intuicí. Pro inteligentního čtenáře bez předchozích znalostí.

70 let AI · 10 kapitol · 10 mýtů · 7 kanonických paperů · Feynman × Tufte

Obsah

1. Šedesát let hledání: od Dartmouthu po ChatGPT

Klíčová otázka: Proč trvalo 70 let od první myšlenky po ChatGPT? Odpověď: každý průlom vyžadoval konvergenci teorie, výpočetního výkonu a dat — nikdy dvou, vždy všech tří zároveň.



Obr. 1 — AI timeline: vlny nadšení (modré kopce) a AI zimy (červené úseky). Každý průlom značen barevnou tečkou.

Dartmouth 1956 — zrození oboru

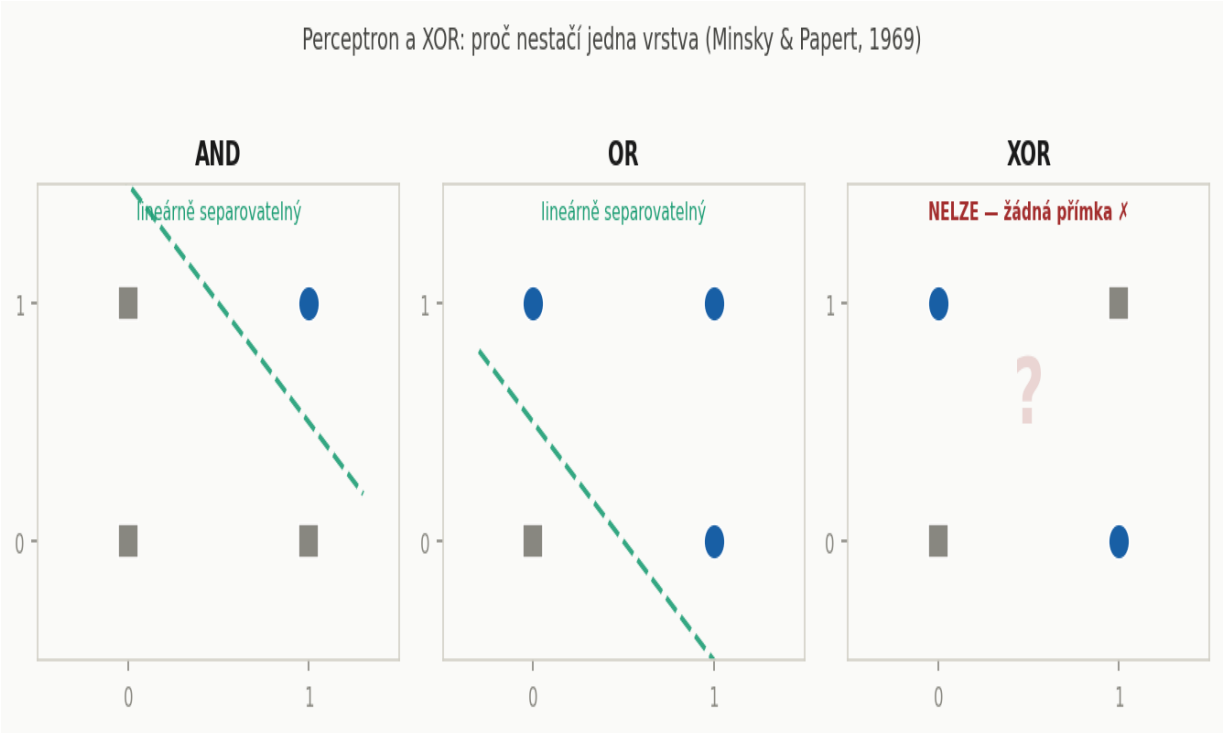
Vše začalo 31. srpna 1955, kdy John McCarthy, Marvin Minsky, Nathaniel Rochester a Claude Shannon podepsali návrh pro Rockefellerovu nadaci. Klíčová věta:

> "*We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956... on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.*"
>
> McCarthy, Minsky, Rochester, Shannon — Dartmouth Proposal, 31. 8. 1955 (AI Magazine, Vol. 27, No. 4, 2006; DOI: 10.1609/aimag.v27i4.1904)

Perceptron 1958 a první krize 1969

Frank Rosenblatt (Cornell, 1958) představil perceptron — první učící se stroj. Mark I měl 400 fotoreceptorů a 512 potenciometrů jako váhy. New York Times napsal: "embryo počítače, který si Navy myslí, že bude umět chodit, mluvit, vidět." V roce 1969 Minsky a Papert v knize

Perceptrons (MIT Press) matematicky dokázali: perceptron nedokáže řešit XOR. Geometricky: XOR není lineárně separovatelný.



Obr. 2 — AND a OR jsou lineárně separovatelné jednou přímkou. XOR nikoli — žádná přímka nedokáže oddělit třídy. Minsky & Papert 1969 způsobili 1. AI zimu (1974–1980).

Backpropagation 1986 — znovuzrození

Rumelhart, Hinton, Williams (1986, Nature 323, 533–536) ukázali, že backpropagation umožňuje vícevrstvým sítím naučit se užitečné vnitřní reprezentace. Matematický základ: chain rule derivací aplikovaná zpětně přes všechny vrstvy. Matematický základ položil Paul Werbos v disertaci (Harvard, 1974), ale Hintonův paper ze skutečně popularizoval.

```
dL/dw = (dL/da) · (da/dz) · (dz/dw) [chain rule backprop]
```

AlexNet 2012 — the Big Bang of Deep Learning

Krizhevsky, Sutskever, Hinton (NeurIPS 2012): top-5 error na ImageNet 26,2 % → 15,3 %. Skok o 10,8 procentních bodů byl tak dramatický, že komise myslela na chybu hodnocení. Klíčové inovace: GPU trénink (2× NVIDIA GTX 580), ReLU místo sigmoidy, dropout jako regularizace.

Transformer 2017 a ChatGPT 2022

Vaswani et al. (2017, arXiv: 1706.03762) nahradili rekurenci čistým self-attention mechanismem. Za 5 let byl Transformer základem každého velkého jazykového modelu. 30. listopadu 2022 spustil OpenAI ChatGPT — za 5 dní 1 milion uživatelů, za 2 měsíce 100 milionů. Nejrychleji rostoucí aplikace v historii.

2. Co je Machine Learning a proč se liší od programování

Klíčový princip: Tradiční programování říká počítači jak problém řešit. Machine learning mu dá příklady a nechá ho odvodit pravidla sám.

Mitchellova definice (1997)

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." — Tom Mitchell, Machine Learning, McGraw-Hill, 1997

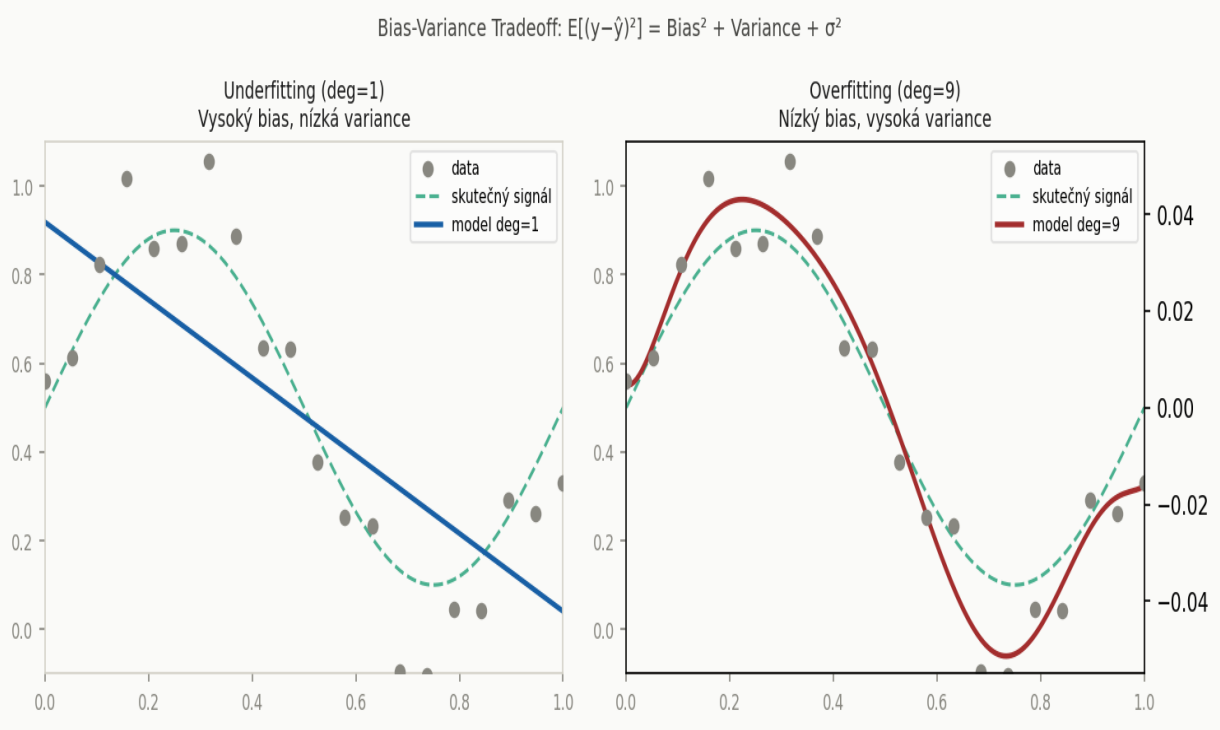
Paradigmatický rozdíl

Tradiční programování	Machine Learning
Vstup + Pravidla → Výstup	Vstup + Výstup → Pravidla (naučená)
Programátor zakóduje logiku explicitně	Model odvodí pravidla z trénovacích dat
Příklad: if "Nigerian prince" → spam	Příklad: 10 000 emailů + štítky → filtr

Bias-Variance Tradeoff — fundamentální dilema ML

Každý ML model čelí fundamentální rovnici. Bias (vychýlení) říká, jak systematicky je průměrný model mimo cíl. Variance říká, jak moc se model mění s různými trénovacími daty. Snížení jednoho obvykle zvyšuje druhé:

$$E[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \sigma^2 \quad (\sigma^2 = \text{ireducibilní šum})$$



Obr. 3 — Underfitting (deg=1): přímka přes paraboloidní data — vysoký bias. Overfitting (deg=9): model kopíruje šum — vysoká variance. Optimální model (deg≈3) minimalizuje součet obou složek.

3. Klasické algoritmy: stromy, lesy, SVM, clustering

Decision Trees a Information Gain

Strom dělá v každém uzlu split — vybere příznak a práh, který nejlépe oddělí třídy. "Nejlépe" měří entropie a Information Gain:

$$H(S) = -\Sigma\ p_i\ \log_2(p_i) \quad \text{[entropie]} \quad IG(S,A) = H(S) - \Sigma\ |S_v|/|S|\ \cdot\ H(S_v) \quad \text{[Information Gain]} \quad Gini = 1 - \Sigma\ p_i^2 \quad \text{[alternativa – rychlejší, bez log]}$$

Feynman pohled: Máš pytel smíchaných červených a modrých kuliček. Entropie měří "zamíchání". Pokud jsou všechny červené: H=0. Pokud 50/50: H=1. Information Gain říká: o kolik méně mě překvapí barva, pokud kuličky roztřídím podle velikosti?

Random Forest a Gradient Boosting

Random Forest (Breiman, 2001) staví m nezávislých stromů na bootstrap vzorcích. Každý strom vidí jen √d náhodných příznaků — tím se stromy dekorelují. Zákon velkých čísel: Var(průměr) = σ²/m. Přidávání stromů nikdy nevede k overfittingu.

Gradient Boosting / XGBoost (Chen & Guestrin, 2016) staví stromy sekvenčně — každý koriguje residuály předchozího: $\hat{y}^t = \hat{y}^{t-1} + f_t(x)$. Random Forest redukuje varianci (paralelně), Gradient Boosting redukuje bias (sekvenčně).

SVM — maximalizace marginu a kernel trick

SVM hledá hyperrovinu $w \cdot x + b = 0$, která maximalizuje vzdálenost k nejbližším bodům (support vectors). Kernel trick: RBF kernel $K(x,x') = \exp(-\gamma\|x-x'\|^2)$ efektivně mapuje do nekonečně-dimenzionálního prostoru — aniž by data skutečně transformoval.

$$\min \tfrac{1}{2}\|w\|^2 \ \text{s.t.}\ y_i(w \cdot x_i + b) \geq 1 \quad \text{[optimalizace marginu]} \quad K(x,x') = \exp(-\gamma\|x-x'\|^2) \quad \text{[RBF kernel – "nekonečná dimenze"]}$$

K-means Clustering

Algoritmus Lloyd: (1) Inicializuj k centroidů, (2) Přiřaď body k nejbližšímu, (3) Přepočítej centroidy jako průměr, (4) Opakuj do konvergence. Minimalizuje $J = \Sigma_k\ \Sigma_{i \in C_k}\ \|x_i - \mu_k\|^2$. K-means++ (2007): smart inicializace s pravděpodobností $\propto d(x)^2$ — eliminuje citlivost na počáteční umístění.

Klíčový rozdíl: supervised vs. unsupervised

Clustering neví, co hledá — jen hledá strukturu. Neexistují žádné štítky, žádný "správný" výsledek. To je zásadní rozdíl od supervised learningu.

4. Neuronové sítě: od biologické inspirace k chain rule

Klíčový princip: Neuronová síť je systém diferencovatelných funkcí, kde backpropagation přiřazuje "vinu" za chybu každému parametru pomocí chain rule. Biologická analogie je inspirace, ne implementace.

Umělý neuron

```
z = Σ wᵢxᵢ + b [váhovaný součet vstupů + bias] a = σ(z) [aktivační funkce – zavedení nelinearity]
```

Proč ReLU vyřešil vanishing gradient: Sigmoid σ(x) má maximální derivaci 0,25. Po 10 vrstvách: $0{,}25^{10} \approx 10^{-7}$ — gradienty mizí. ReLU f(x) = max(0,x) má derivaci přesně 1 pro x>0. Gradienty procházejí beze změny. Moderní sítě používají GELU: f(x) = x·Φ(x).

```
Sigmoid: σ(x) = 1/(1+e^-x) max. gradient = 0.25 → vymizí po 10+ vrstvách ReLU: f(x) = max(0,x) gradient = 1 pro x>0 → projde libovolnou hloubkou GELU: f(x) = x·Φ(x) hladká verze ReLU, používána v GPT/BERT/Claude
```

Dropout a Batch Normalization

Dropout (Srivastava et al., 2014, JMLR) náhodně "vypíná" neurony s pravděpodobností p. Každý neuron se musí naučit být nezávisle užitečný. Efektivně trénuje exponenciální ensemble 2^n pod-sítí.

Batch Normalization (Ioffe & Szegedy, 2015): BN(x) = γ·(x−μ)/(σ+ε) + β. Umožňuje 14× méně trénovacích kroků, vyšší learning rate. Funguje primárně vyhlazením loss landscape (Santurkar et al., 2018).

5. Deep Learning: co znamená "deep" a proč to fungovalo

Klíčový princip: "Deep" neznamená složitý — znamená hierarchická reprezentace příznaků. Každá vrstva staví na abstrakci předchozí, jako student kreslení: nejdřív čáry, pak tvary, pak textury, pak celé objekty.

Obr. 4 — CNN hierarchie příznaků (Zeiler & Fergus, 2013): V1=orientované hrany, V2=rohy a textury, V3-4=části objektů, V5+=celé objekty. Receptivní pole roste s hloubkou.

Tři podmínky pro Deep Learning boom

GPU Compute	Big Data	Lepší algoritmy
NVIDIA CUDA (2006/7). AlexNet trénoval na 2× GTX 580. Bez GPU: měsíce místo dnů.	ImageNet (Fei-Fei Li, 2009): 14M označených obrázků. Předchozí max: 9 000.	ReLU řeší vanishing gradient. Dropout a BatchNorm umožní hluboké sítě. ResNet: skip connections.

ResNet — skip connections (He et al. 2015, arXiv: 1512.03385)

y = F(x) + x — přidání identitního zkratu přes každé 2 vrstvy. Vyřešilo degradation problem a umožnilo sítě se 152+ vrstvami. Skip connections jsou dnes v každém Transformeru.

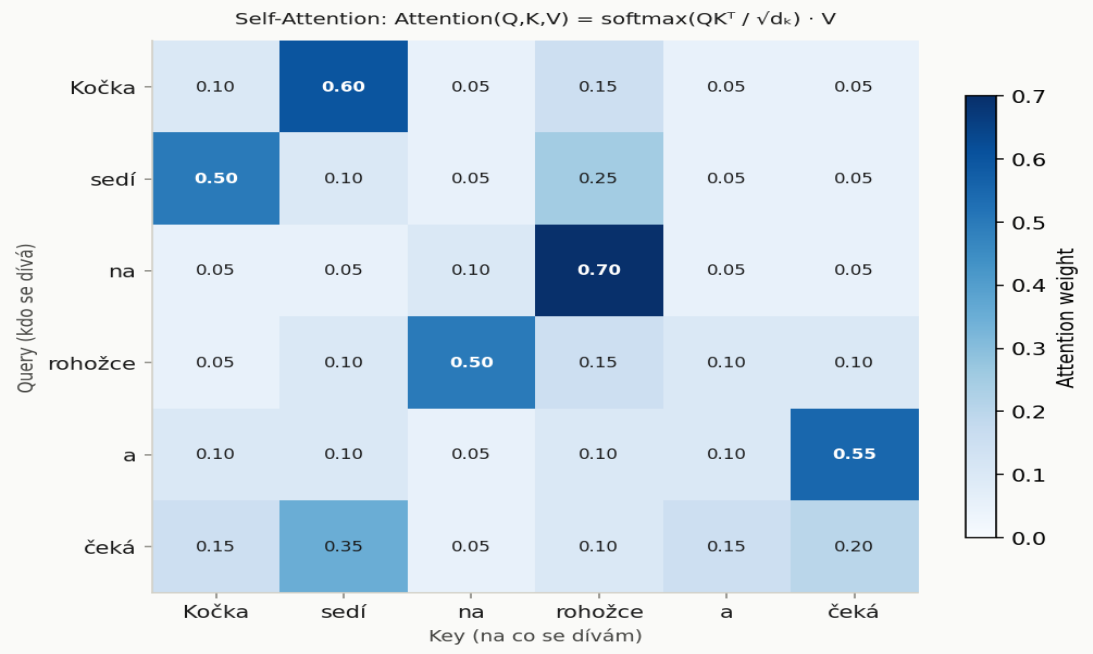
6. LLM a architektura Transformeru

Klíčový princip: Transformer nahrazuje sekvenční zpracování paralelním self-attention mechanismem, kde každý token přímo "vidí" každý jiný token a rozhoduje se, komu věnovat pozornost.

Self-Attention: Q, K, V matice

Query (Q) = "Co hledám?" · Key (K) = "Co nabízím?" · Value (V) = "Tady je můj obsah". Skalární součin QK^T měří "shodu" mezi Query a Key. Dělení $\sqrt{d_k}$ zabraňuje saturaci softmax — bez něj by distribuce degenerovala na one-hot a gradienty by vymizely:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \qquad \text{MultiHead} = \text{Concat}(\text{head}_1,...,\text{head}_h) \cdot W^O, \text{ kde head}_i = \text{Attention}(Q \cdot W_i^O,\ K \cdot W_i^K,\ V \cdot W_i^V)$$



Obr. 5 — Attention heatmapa pro větu "Kočka sedí na rohožce a čeká". Tmavší buňka = vyšší attention weight. "sedí" se nejvíce dívá na "Kočka" (50%); "na" primárně na "rohožce" (70%).

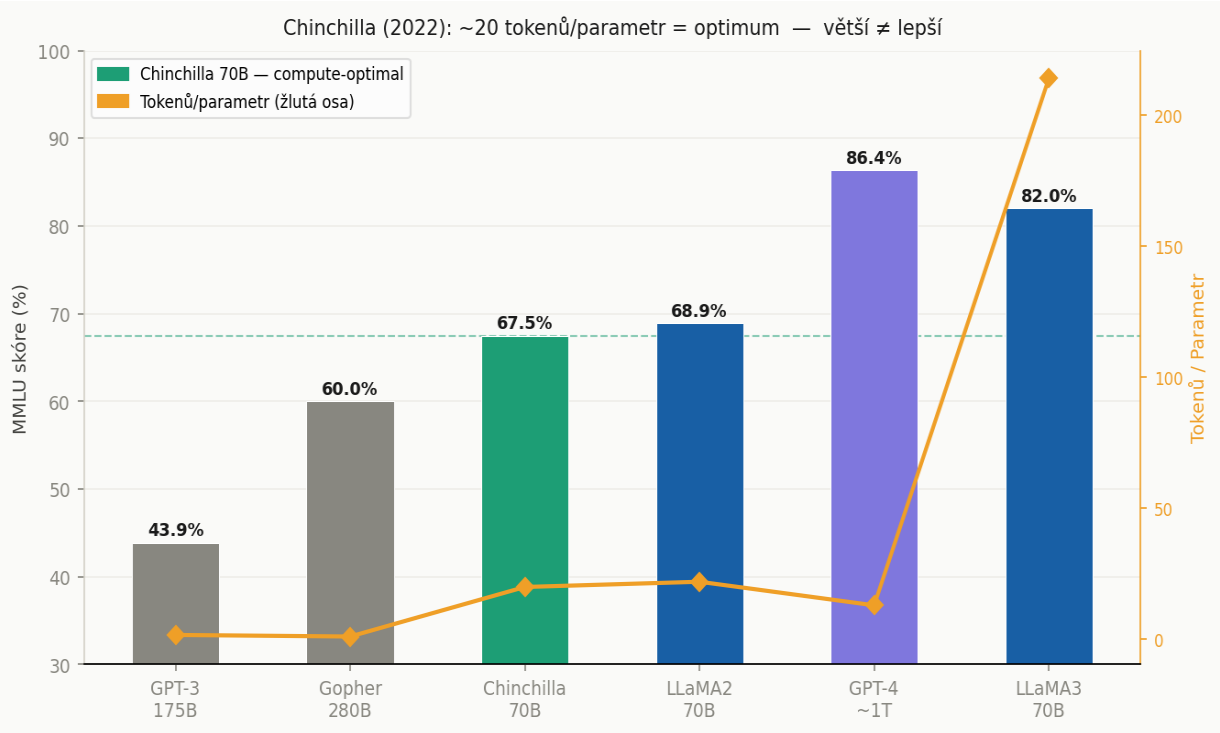
Pretraining → SFT → RLHF pipeline

Pretraining: predikce dalšího tokenu na bilionech tokenů z webu — self-supervised, žádné lidské štítky. Výsledek: široké znalosti, ale špatné zarovnání s uživatelem.

SFT (Supervised Fine-Tuning): trénink na kurátorských párech (instrukce → odpověď).

RLHF (Ouyang et al., 2022, arXiv: 2203.02155, InstructGPT): (1) SFT, (2) reward model z lidských preferencí, (3) PPO s KL penalizací. Klíčový výsledek: InstructGPT 1,3B byl preferován nad vanilla GPT-3 175B — 100× menší model s lepším zarovnáním.

Scaling Laws — Chinchilla paper



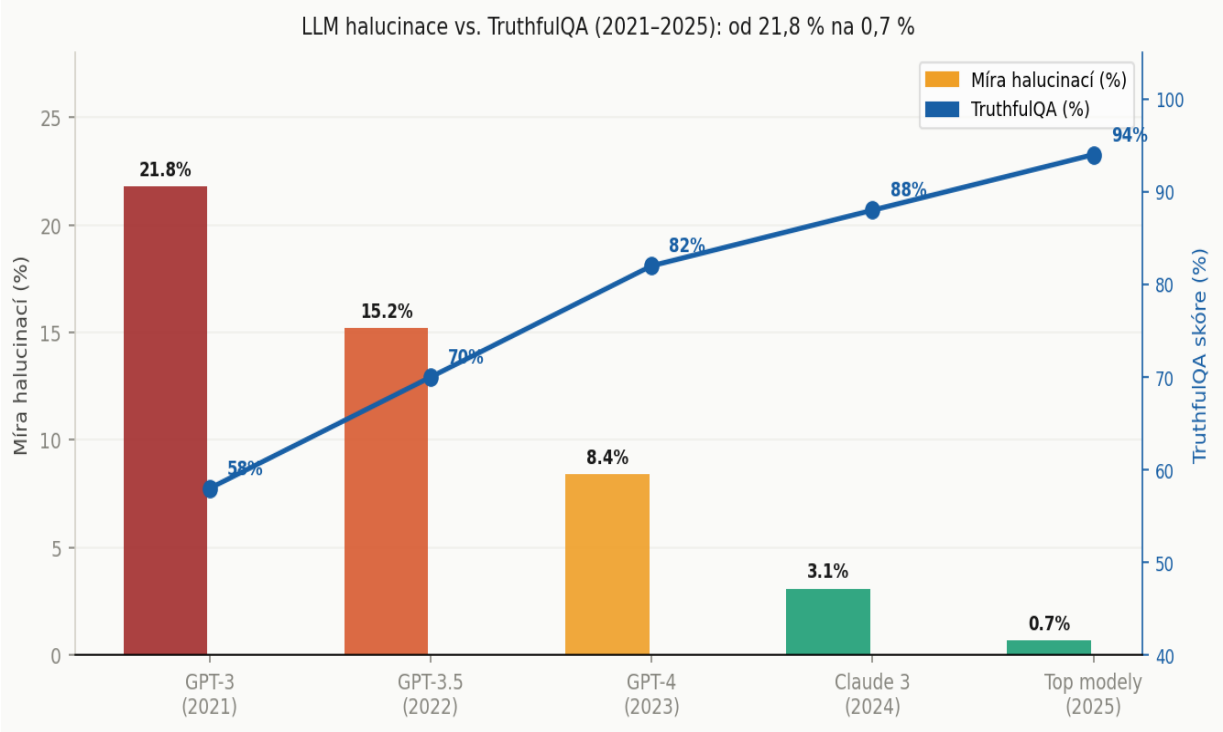
Obr. 6 — MMLU skóre modelů vs. Chinchilla ratio (tokenů/parametr, žlutá linie). Chinchilla 70B (ratio=20) porazila Gopher 280B i GPT-3 175B. Největší model ≠ nejlepší model.

7. Proč LLM halucinují: statistická plausibilita vs. pravda

Klíčový princip: LLM optimalizuje P(next_token | context), nikoliv faktickou správnost. Tréninkový cíl (cross-entropy loss) odměňuje plausibilní pokračování, ne pravdivá.

Mechanismus halucinací

LLM vidí během tréninku POUZE pozitivní příklady plynulého jazyka. Nemá explicitní mechanismus fact-checkingu. Distributional semantics ≠ grounded understanding: slova v podobných kontextech mají podobné vektory, ale to neznamená, že model rozumí, co slova znamenají ve světě. MIT 2025: halucinující AI je o 34 % pravděpodobnější, že použije "určitě", "rozhodně".



Obr. 7 — Míra halucinací (červené sloupce) klesla z 21,8 % (2021) na 0,7 % (2025). TruthfulQA skóre (modrá linie) vzrostlo z 58 % na 94 %. GPT-3 175B: 58 % pravdivých vs. 94 % u lidí (Lin et al., 2022, ACL).

Mitigace: RAG a Chain-of-Verification

RAG (Retrieval-Augmented Generation) snižuje halucinace o ~71 % — model dostane relevantní kontext z externích zdrojů před generováním. Chain-of-Verification (Dhuliawala et al., 2023, Meta AI): Draft → Ověřovací otázky → Nezávislé zodpovězení → Finální ověřená odpověď.

8. Yann LeCun versus LLM: proč kočka rozumí světu lépe

Yann LeCun, Chief AI Scientist at Meta (do 2025, pak founded AMI s $1,03B), je nejhlasitějším kritikem LLM jako cesty k lidské inteligenci. Jeho argumenty nelze odmítnout — je spoluautorem backpropagation a CNN.

> "*Mozek kočky má asi 800 milionů neuronů. Vynásobte 2 000 — dostanete parametry LLM. Proč tyto systémy nejsou tak chytré jako kočka? Protože kočka rozumí fyzice, plánuje a uvažuje. LLM nikoli.*"
>
> — Yann LeCun, Observer Interview, únor 2024

Čtyři chybějící schopnosti LLM

Žádný world model

LLM chybí kauzální model reality — fyzika, prostorové vztahy, čas.

Exponenciální divergence

P(správná sekvence) = $(1-\varepsilon)^N \rightarrow$ exponenciálně klesá. Chyba v tokenu t nás odvede ze správného podstromu.

Nelze plánovat

Chain-of-Thought je maximálně "System 1.1" — statisticky rychlejší, ale stále ne skutečné reasoning.

Efektivita učení

Dítě se naučí, že "sklenice padá", z 10 příkladů. LLM potřebuje miliardy tokenů.

JEPA — navrhované řešení

Joint Embedding Predictive Architecture (LeCun, 2022): predikce v reprezentačním prostoru, ne v prostoru pixelů/tokenů. Implementace: I-JEPA (obrázky, 2023), V-JEPA (video, 2024), VL-JEPA (vizuálně-jazykové, 2025) — 50 % méně parametrů než standardní VLM.

9. Generativní modely: GAN, VAE a difuzní alchymie

Tři generace generativních modelů — tři fundamentálně odlišné přístupy:

GANs — adversariální hra

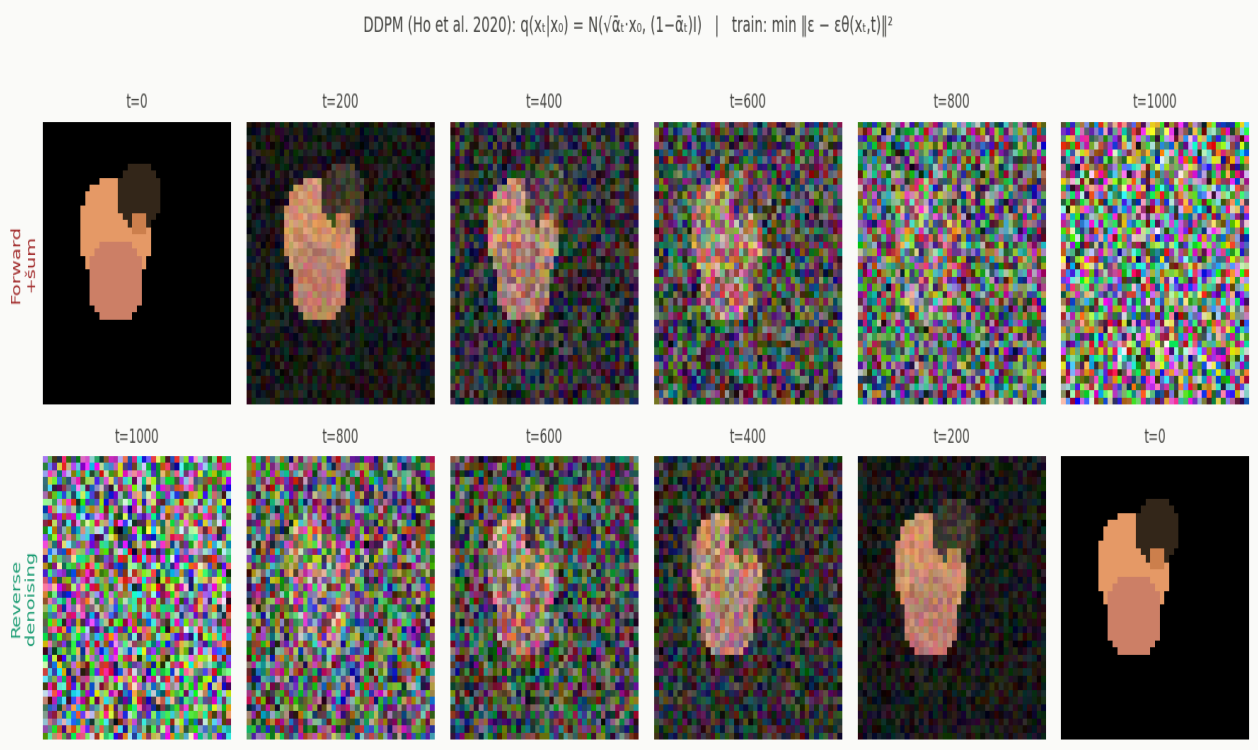
```
min_G max_D V(D,G) = E[log D(x)] + E[log(1−D(G(z)))] Nashova rovnováha: p_g = p_data, D*(x) = 0.5 všude Mode collapse: G generuje jen pár vzorků, které oklamou D
```

Goodfellow et al. (2014, NeurIPS): Generátor G "falzuje" data, Diskriminátor D je "detektiv". V Nashově rovnováze D nedokáže rozlišit real/fake. Mode collapse — chronický problém GANů: G se naučí generovat jen pár vzorků.

VAE — variační inference

```
ELBO = E[log p_θ(x|z)] - D_KL(q_φ(z|x) || p(z)) Reparameterization trick: z = μ + σ⊙ε, kde ε~N(0,I) → backprop přes sampling
```

Difuzní modely — denoising jako generace



Obr. 8 — Forward process (horní řada): postupné přidávání šumu od x_0 k x_t. Reverse process (dolní řada): iterativní denoising zpět k x_0. DDPM (Ho et al. 2020, NeurIPS): trénuj $\varepsilon\theta(x_t,t)$ — "jaký šum byl přidán?"

```
Forward: q(xₜ|x₀) = N(√αₜ·x₀, (1−αₜ)·I) Training: min ‖ε − εθ(xₜ,t)‖² [predikuj šum, ne obraz] Sampling: xₜ₋₁ = (xₜ − √(1−αₜ)·εθ) / √αₜ + σₜz
```

Dhariwal & Nichol (2021): Diffusion Models Beat GANs on Image Synthesis. Žádný mode collapse (stabilní regresní trénink), lepší diverzita. Latent Diffusion (Rombach et al., 2022) přesunul difúzi do komprimovaného latentního prostoru → 10–100× méně výpočtu → Stable Diffusion.

10. Agenti, Chain of Thought, MCP a Deep Research

Chain of Thought — myšlení nahlas

Wei et al. (2022, NeurIPS, arXiv: 2201.11903): 8 CoT exemplářů u 540B PaLM → SOTA na GSM8K matematických úlohách (+18%). Zero-shot verze: "Let's think step by step" (Kojima et al., 2022). Proč funguje: mezikroky = externí pracovní paměť. Efektivní výpočet roste s délkou chain, ne jen s počtem parametrů.

CoT → základ o1, o3, Claude extended thinking

ReAct (Yao et al. 2022): Thought → Action → Observation → Thought. Prokládá reasoning traces s akcemi (volání nástrojů). Překonává jak samotný CoT, tak Act-only přístupy.

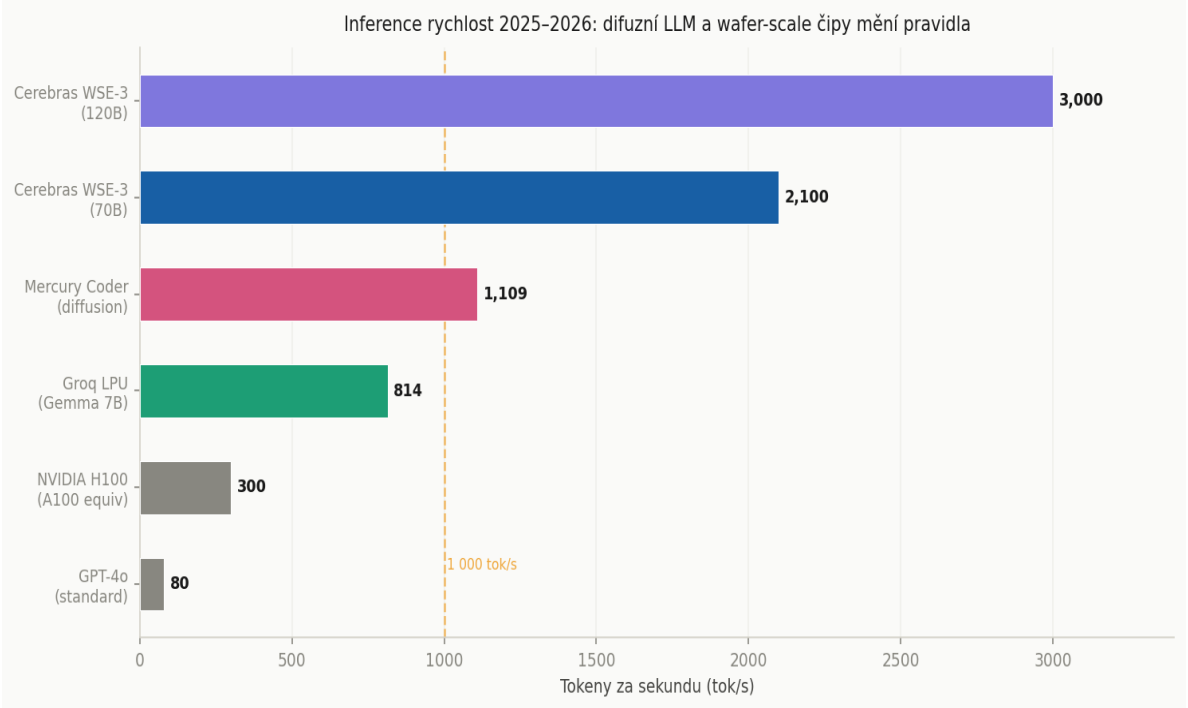
MCP — Model Context Protocol

Anthropic, 25. listopadu 2024: otevřený standard pro propojení AI s externími zdroji dat. "USB-C pro AI." Místo N×M custom konektorů jeden protokol. Tři primitiva: Resources (data), Tools (funkce), Prompts (šablony). Do prosince 2025 adoptován OpenAI, Google, Microsoft. Darován Linux Foundation (Agentic AI Foundation).

AI Agenti

Autonomní systémy, které vnímají, uvažují a jednají. Typy: reflexivní (stimulus-response), plánující (dekompozice cílů), multi-agentní (spolupráce více agentů). Gartner predikuje: 33 % enterprise software bude obsahovat agentic AI do roku 2028 (z <1 % v 2024).

11. Budoucnost: difuzní LLM a inference na čipu



Obr. 9 — Inference rychlost 2025–2026. Mercury Coder (diffusion LLM): 1 109 tok/s — 10× rychlejší než GPT-4o Mini. Cerebras WSE-3: 3 000 tok/s na 120B modelu. Groq LPU (akvírováno NVIDIA za $20B): 814 tok/s.

Mercury — první komerční difuzní LLM

Inception Labs (CEO: Stefano Ermon) — únor 2025: Mercury Coder s 1 109 tok/s. Mercury 2 (únor 2026): první reasoning difuzní LLM, >1 000 tok/s, kvalita srovnatelná s Claude 4.5 Haiku. arXiv: 2506.17298.

Proč je difúze rychlejší: autoregresivní generování je memory-bandwidth bound (sekvenční, jeden token za druhým). Difuzní LLM generují všechny tokeny simultánně paralelním zpřesňováním. Fundamentální architektonická výhoda.

Edge AI — LLM na čipu

Platforma	Model	Rychlost	Poznámka
Cerebras WSE-3	gpt-oss-120B	3 000 tok/s	Wafer-scale, 7000× bandwidth vs H100
Mercury 2	Diffusion LLM	>1 000 tok/s	Paralelní generování, reasoning
Groq LPU	Gemma 7B	814 tok/s	All-SRAM, deterministický, akvírován NVIDIA
Qualcomm NPU	DeepSeek-R1	<70ms TTFT	On-device, offline, žádný cloud
Apple Neural Engine	Foundation Models ~3B	iOS 18	Privacy-first, na za■ízení

12. Deset mýtů o AI: jak to je doopravdy

Většina populárních představ o AI je buď přehnaná nebo fundamentálně chybná. Porozumění skutečným limitům je důležitější než fascinace schopnostmi.

M1: „AI rozumí jako člověk“

LLM manipuluje symboly bez grounding ve světě. Searle's Chinese Room (1980): pravidla bez porozumění. Bender & Koller (2020, ACL): systém trénovaný jen na formě nemá způsob naučit se význam.

GPT-4 nedokáže spolehlivě říct, zda jsou dveře dveřmi z první strany — kočka to ví.

M2: „LLM jsou databáze faktů“

LLM provádějí lossy kompresi ~100:1 do vah. Nejsou dotazovatelné jako databáze — generují plausibilní pokračování. Proto halucinují: fakta nikdy neukládaly.

Když se GPT-3 zeptáš na neexistující knihu, s jistotou popíše obsah. DB by vrátila "nenalezeno".

M3: „AI brzy převezme všechny práce“

WEF 2025: 170M nových pracovních míst vs. 92M zrušených = čistý nárůst 78M. ILO: v zemích s nižšími příjmy ohroženo jen 0,4 % pozic.

PwC 2025: AI-exponovaná odvětví vykazují 38% růst zaměstnanosti. AI augmentuje, ne eliminuje.

M4: „Více parametrů = lepší model“

Chinchilla 70B porazila GPT-3 175B i Gopher 280B. Phi-3-mini 3,8B ≈ GPT-3.5 175B na benchmarcích.

Llama 3 8B trénovaná na 15T tokenech překonává modely 10× větší na méně datech.

M5: „AI je objektivní“

Buolamwini & Gebru (2018) Gender Shades: 34,7 % chybovost u tmavých žen vs. 0,8 % u světlých mužů. Trénovací data: 79–86 % světlých tváří.

AI reflektuje data, ne realitu. IBM kvůli studii zcela opustila obličejové rozpoznávání.

M6: „AGI přijde za 5 let“

Hinton: 5–20 let. LeCun: desítky let. 76 % z 475 AI výzkumníků: škálování k AGI pravděpodobně nestačí. Gary Marcus vsadil 10:1 proti AGI do 2027.

Tech podnikatelé predikují ~2030, akademici ~2040+. Velký disagrement = skutečná nejistota.

M7: „AI se učí ze všeho co píšeš“

Training a inference jsou kompletně oddělené procesy. Chatovací vstupy nemodifikují váhy modelu. Kontext konverzace je efemérní.

Váhy GPT-4 jsou zmraženy od tréninku. API se obecně nepoužívá k dalšímu trénování.

M8: „AI halucinuje, tedy je nespolehlivé“

S guardrails (RAG, CoV, tool use, human-in-loop) je AI vysoce spolehlivé v omezených doménách. Halucinace klesly z 21,8 % na 0,7 % za 4 roky.

Lékaři dělají chyby v 10–15 % diagnóz. Otázka není "je AI dokonalé", ale "je lepší než alternativa?"

M9: „Deep learning je jako mozek“

Backpropagation se v mozku nevyskytuje (Lillicrap et al., Nature Reviews Neuroscience 2020). Mozek: spiky v čase, tisíce typů synapsí, učení z 1 příkladu, žádné catastrophic forgetting.

Song et al. (2024, Nature Neuroscience): mozek používá "prospective configuration" — fundamentálně odlišný mechanismus.

M10: „Největší modely jsou nejlepší pro každý úkol“

Mercury Coder Mini porazil GPT-4o na Copilot Arena, přitom 4× rychlejší. Gemini Nano 1,8B zvládá sumarizaci na mobilu.

Správný model závisí na úkolu, latenci, ceně a deployment kontextu. Phi-3-mini (3,8B) = GPT-3.5 (175B).

Kanonické papery — must-read pro každého v AI

Těchto 7 paperů definuje moderní AI. Každý je na arXiv nebo v časopisu zdarma.

Nature 323, 533–536 · 1986 · DOI: 10.1038/323533a0

Learning representations by back-propagating errors

D. E. Rumelhart · G. E. Hinton · R. J. Williams

Chain rule skrz libovolně hlubokou síť. Skryté vrstvy se naučí užitečné reprezentace. Fundament celého DL. 40 000+ citací.

NeurIPS 2012 · Krizhevsky, Sutskever, Hinton

ImageNet Classification with Deep Convolutional Neural Networks (AlexNet)

Alex Krizhevsky · Ilya Sutskever · Geoffrey E. Hinton

Top-5 error 26,2% → 15,3%. 60M parametrů, 2× GTX 580. ReLU + Dropout. 100 000+ citací. Spustil DL revoluci.

arXiv: 1706.03762 · NeurIPS 2017 · Google Brain + Research

Attention Is All You Need

Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin

Transformer: čistý self-attention bez rekurze. Attn = softmax($QK^T/\sqrt{dk}$)·V. 100 000+ citací. Základ GPT, BERT, Claude, LLaMA.

arXiv: 2203.02155 · NeurIPS 2022 · OpenAI

Training language models to follow instructions with human feedback (InstructGPT)

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright et al.

RLHF: SFT → Reward Model → PPO. InstructGPT 1,3B preferován nad GPT-3 175B u lidí. Základ ChatGPT.

arXiv: 2203.15556 · NeurIPS 2022 · DeepMind

Training Compute-Optimal Large Language Models (Chinchilla)

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya et al.

~20 tokenů/parametr = optimum. Chinchilla 70B > Gopher 280B > GPT-3 175B na MMLU. Největší ≠ nejlepší.

arXiv: 2006.11239 · NeurIPS 2020 · UC Berkeley

Denoising Diffusion Probabilistic Models (DDPM)

Jonathan Ho · Ajay Jain · Pieter Abbeel

$q(x_t|x_0)=N(\sqrt{\alpha_t}\cdot x_0,(1-\alpha_t)I)$. Překonal GANy v FID. Základ Stable Diffusion, DALL-E, Mercury.

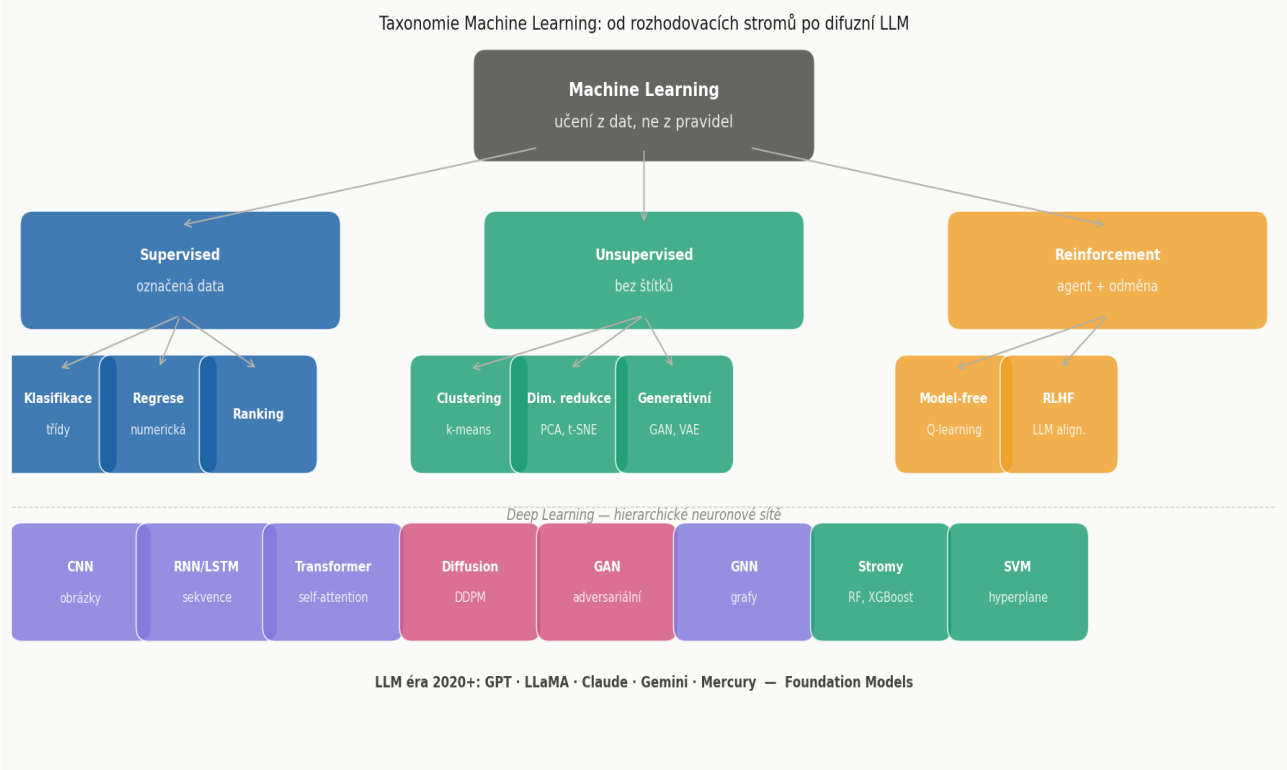
arXiv: 2201.11903 · NeurIPS 2022 · Google Brain

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei, Xuezhi Wang, Dale Schuurmans, Quoc Le, Denny Zhou et al.

8 CoT exemplářů u 540B PaLM → +18% GSM8K. "Let's think step by step." Základ o1, o3, extended thinking.

ML Taxonomie — přehled celého oboru



Obr. 10 — Kompletní taxonomie ML přístupů: od supervised/unsupervised/RL přes CNN/Transformer/Diffusion po LLM éru 2020+. Každý uzel je klikatelný v interaktivní verzi dokumentu.

Závěr: čtyři principy, které přežijí další dekádu

1. AI je fundamentálně o optimalizaci

Od gradient descent přes minimax hry GANů po RLHF — každý průlom představoval lepší loss funkci nebo efektivnější způsob její minimalizace. Kdo rozumí optimalizaci, rozumí AI.

2. Škálování není magie

Chinchilla ukázala, že compute-optimal trénink záleží více než surový počet parametrů. Schaeffer et al. ukázali, že "emergentní schopnosti" mohou být artefaktem měření. Mercury a Cerebras ukazují, že architektonické inovace v inference mohou být důležitější než další škálování.

3. LLM nejsou konečná odpověď

LeCunovy argumenty o chybějícím world modelu a exponenciální divergenci jsou technicky fundované. JEPA a difuzní LLM představují reálné alternativy. Za 5 let bude krajina vypadat dramaticky jinak.

4. Rozumět limitům je důležitější než fascinace schopnostmi

Halucinace není bug — je to inherentní vlastnost systému optimalizujícího plausibilitu. Bias není anomálie — je to reflekce trénovacích dat. Kdo rozumí proč a jak tyto limity vznikají, dokáže s AI pracovat efektivně.

Feynman Crash Kurz do AI v roce 2026 · Vytvořeno s Anthropic Claude Sonnet 4.6 · Březen 2026