PLIN037 Sémantika a počítače

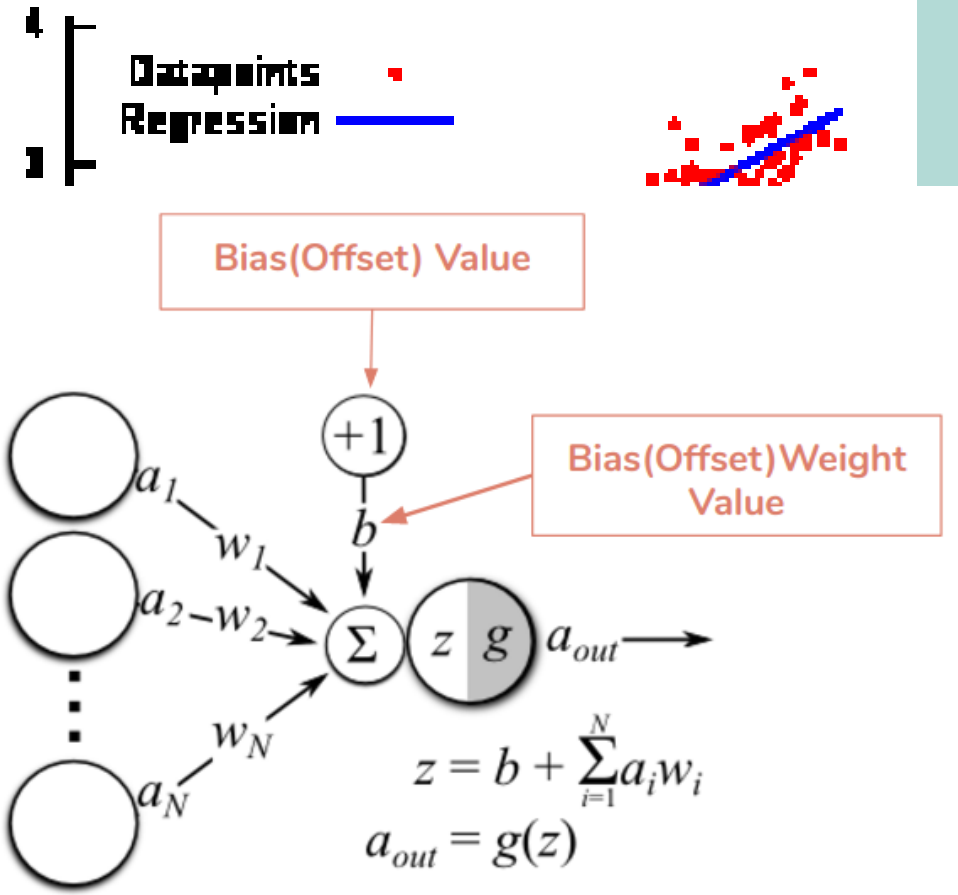
Zuzana Nevěřilová
2024

Neuronové sítě
Jeden neuron

$y = wx + b$, w – weight, b - bias

Nějak tipneme a a b

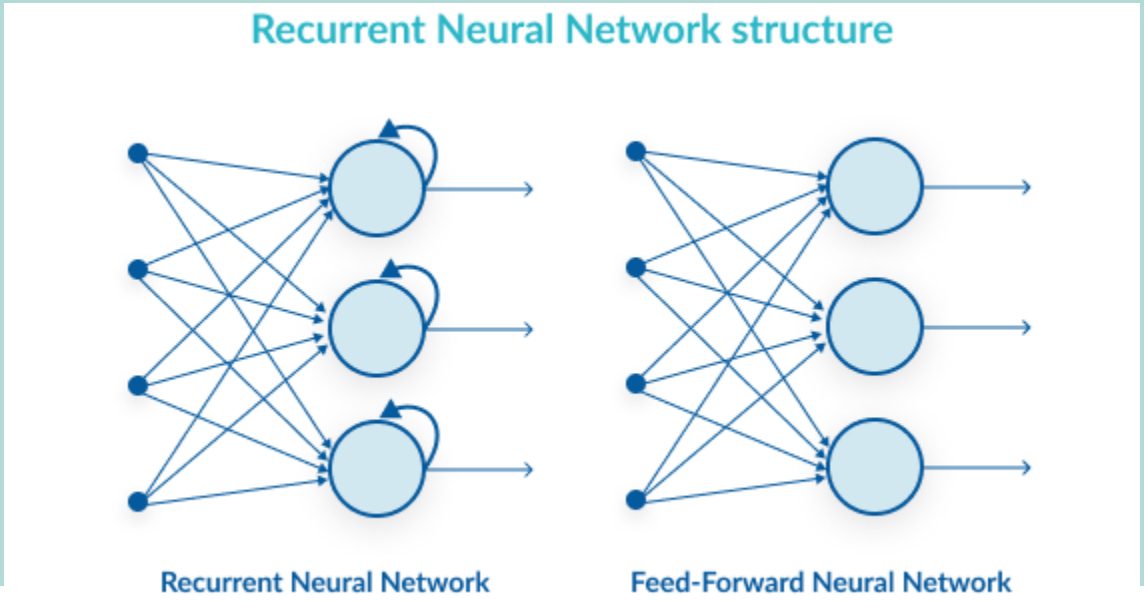
- **Dopředný výpočet (feed forward)**
  - Spočítáme $y = wx + b$
  - Spočítáme aktivační funkci (nelineární) $\hat{y} = \sigma(wx + b)$
- **Zpětná propagace (back propagation)**
  - Spočítáme, o kolik jsme se spletli (loss function, loss, účelová funkce)
  - Navrhneme, jak parametry w a b upravit



https://machine-learning.paperspace.com/wiki/weights-and-biases

Neurony a spojení

Rekurentní neuronová síť (RNN) - sekvence



https://machine-learning.paperspace.com/wiki/recurrent-neural-network-rnn
https://en.wikipedia.org/wiki/Long_short-term_memory

Problémy s dopřednou sítí a RNN

- Zpětné šíření chyby
- Mizející/explodující gradient
  (když se mnohokrát násobí to samé číslo)

Dlouhá krátkodobá paměť
Long Short-Term Memory, LSTM

Brány (gates): input, output, forget

Stavy: buňka (cell) c_t a skrytý (hidden) h_t

forget →

Oboustranné zapojení =
BiLSTM (oboustranná, bidirectional LSTM)

Transformer
Sekvenční model

Druhy

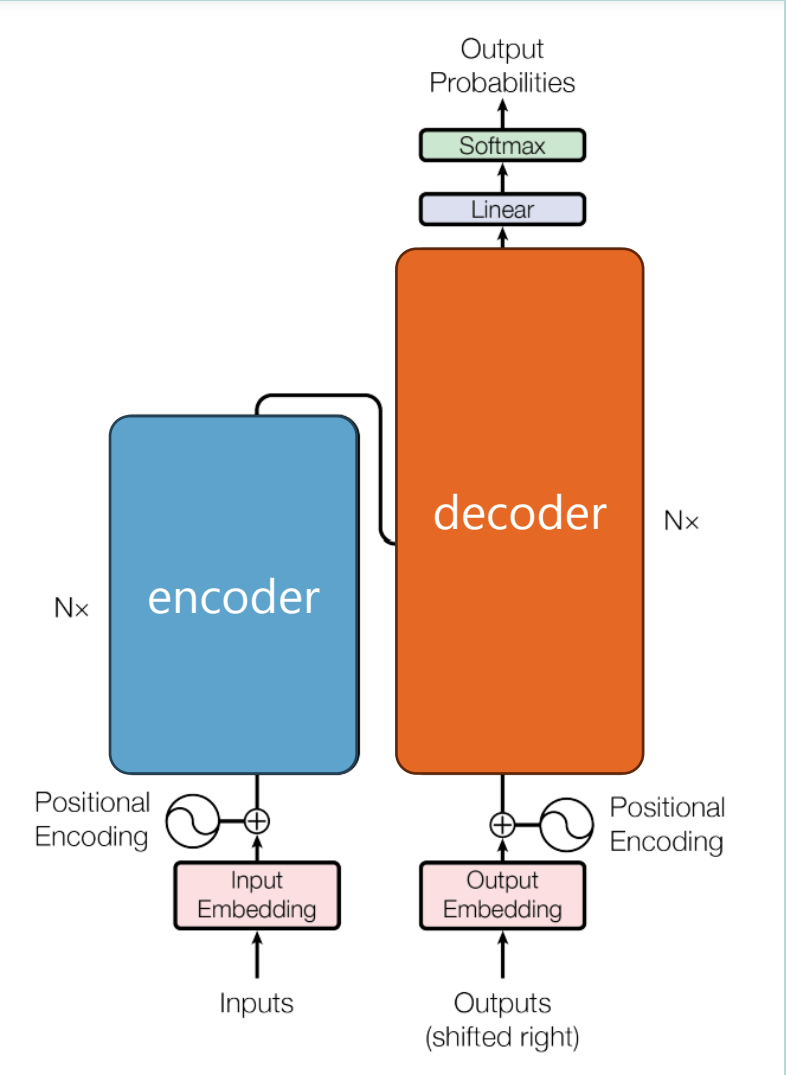
- Encoder
- Decoder
- Je možné použít jednu z nich nebo obě

Vlastnosti

- Samotný decoder „nevidí“ dopředu
- Samotný encoder – oboustranný

Nejdůležitější součást

- Pozornost (attention)
  - vážený součet všech předchozích stavů.
  - „Předchozí“ se neuvažují v sekvenci, ale jako množina (nezáleží na pořadí). Vzdálenost mezi všemi vstupy je stejná.

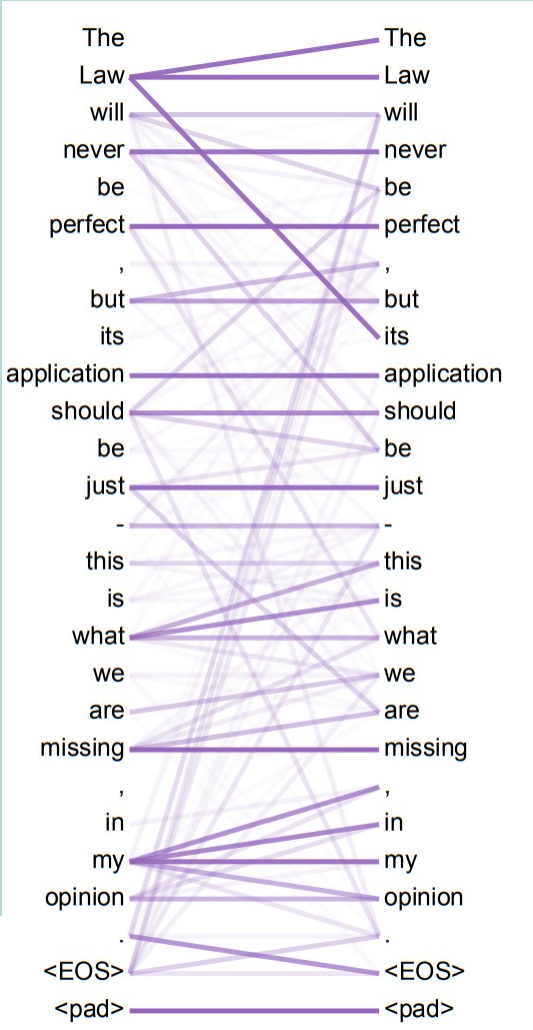


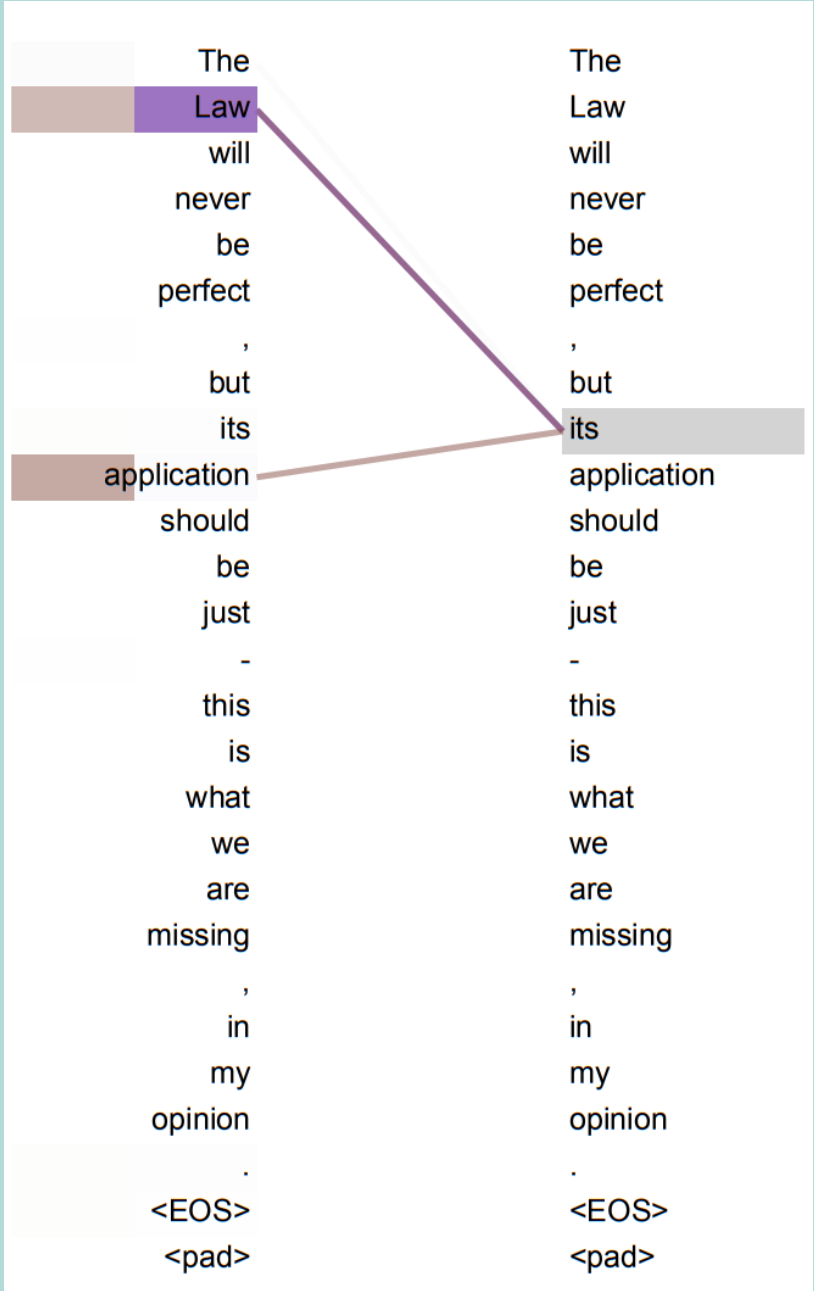
https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634

Pozornost
Attention

Pozornost
mezi vrstvami 5 a 6

Pozornost na tokenu „its“





Generativní modely

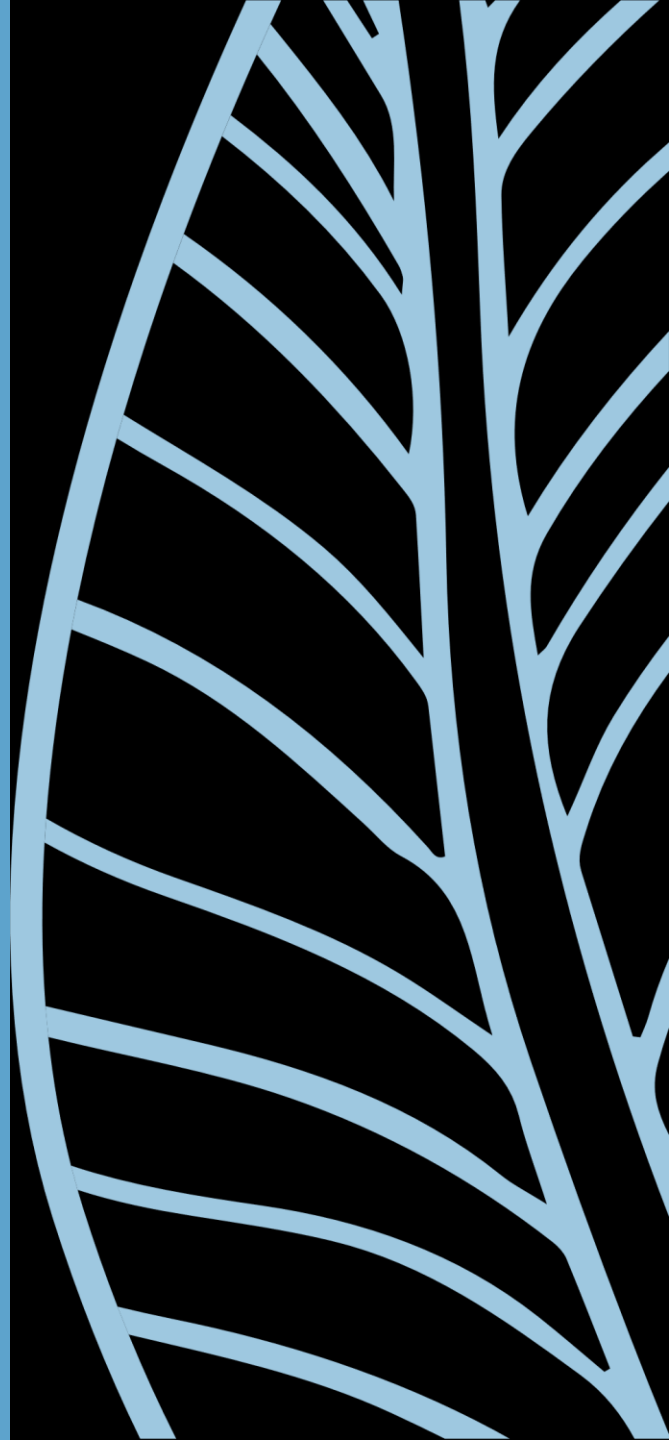
Pravděpodobnostní model

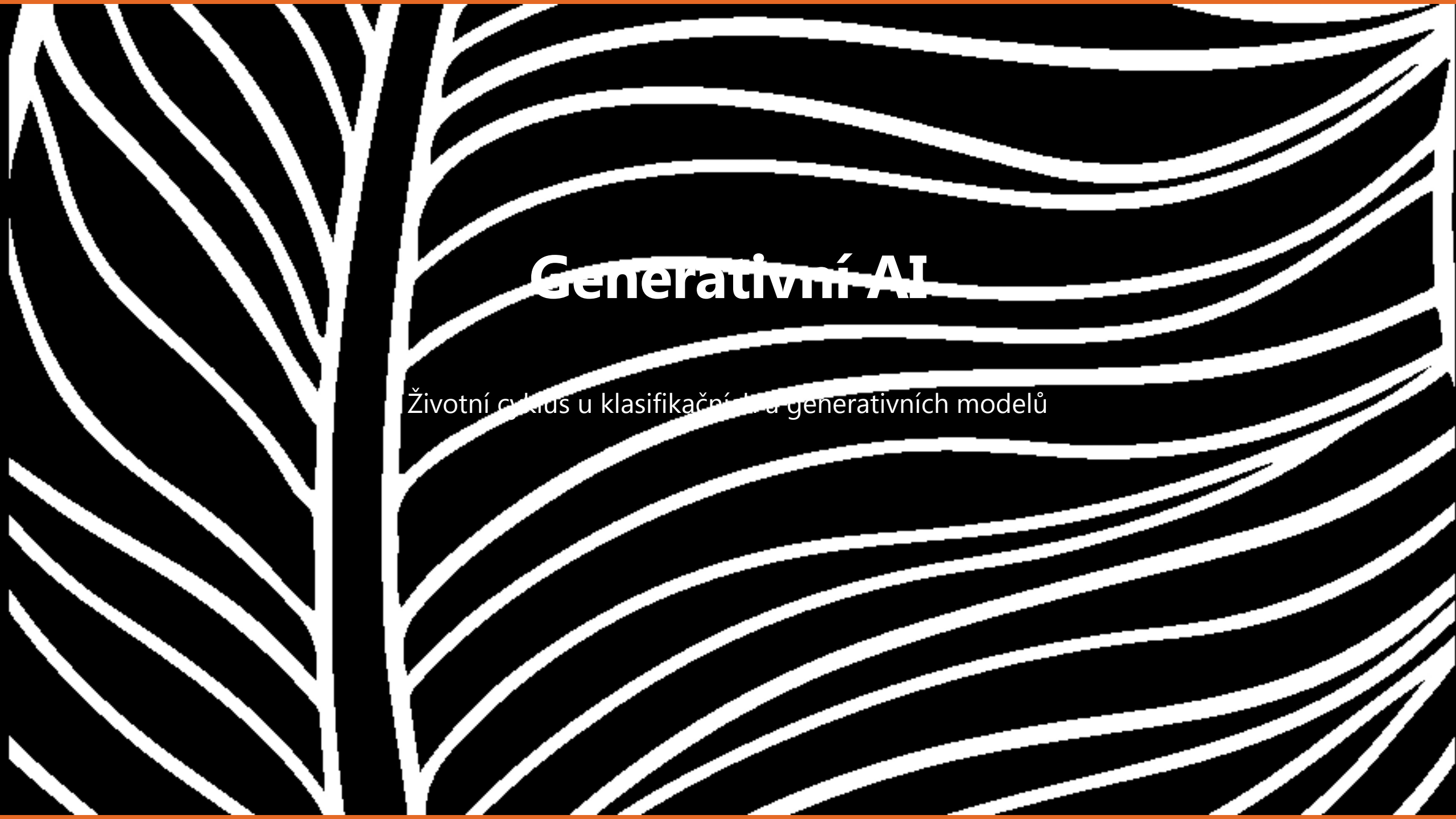
- Diskriminativní model – podmíněná pravděpodobnost
  Jaká je pravděpodobnost, že pozorované zvíře je **kráva (cíl)**, když má **pozorování** tyto prvky (**býložravec**, **velký**)? P(Y | X=x)
- Generativní model – sdružená/simultánní distribuční funkce (joint probability distribution) pozorování a cíle P(X, Y) nebo P(X | Y=y)

Generativní AI

- Zakóduje vstupy, naučí se z nich vzory, z nichž generuje výstupy s podobnou charakteristikou, jako měla vstupní data
- Text-to-text: GPT, Bard, LLaMA
- Text-to-image: DALL-E, Midjourney

Modality: text, kód, obraz, audio, video, chemické vazby, pohyb, ...





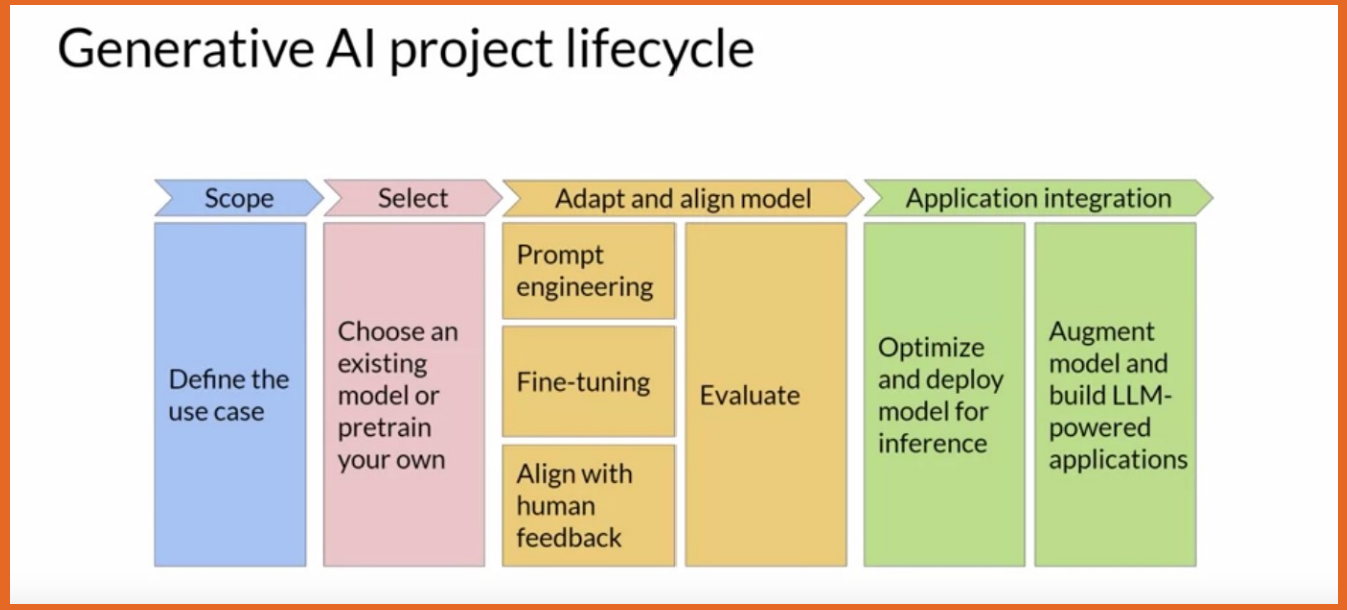
Generativní AI

Životní cyklus u klasifikačních a generativních modelů

Klasifikační a generativní modely AI

Životní cyklus

- Klasifikační modely
  - seženeme anotovaná data
  - seženeme model/sestavíme nový
  - natrénujeme z dat
  - používáme pro predikce
- Generativní modely
  - neposkytujeme data, ale prompty
  - model provádí inferenci (generování textu)
  - výsledek je doplnění (completion)



https://www.coursera.org/learn/generative-ai-with-llms/

Generativní AI

Typy promptů

Velikost promptu je omezena (typicky pár tisíc tokenů) - context window - je řádově menší než velikost trénovacích dat

- **zero-shot**
  `Najdi jména osob v následujícím textu.`
- **one-shot**
  `Najdi jména osob v následujícím textu. Láďa jede lodí, tou lodí výletní, k Lídě, co s ní chodí, zkrátka, Láďa jede k ní. – Láďa, Lída`
- **few-shot**
  `Napiš sentiment pro následující věty:`
  `Je to nuda. – negativní`
  `Je to super.  - pozitivní`

Konfigurace generování

- max new tokens
  (jedna z podmínek ukončení generování)
- Náhodnost:
  - hladový, greedy (vezmi nejvyšší pravděpodobnost)
  - náhodné vzorkování, random sampling
    (vezmi náhodný token s použitím distribuční funkce)
- Modifikace vzorkování:
  - vezmi jen k nejpravděpodobnějších
  - vezmi jen tolik nejpravděpodobnějších, že součet jejich pravděpodobností je max p
- Modifikace náhodnosti („teplota“)
  - 1=žádná změna
  - menší než 1, distribuce pravděpodobnosti se „vyhrotí“
  - větší než 1, distribuce pravděpodobnosti se „vyhladí“

Velké jazykové modely (large language models)

Velké (175B – miliard – parametrů – což je velikost GPT3)

Jazykové (přirozený jazyk)

Model (statistický popis sekvencí slov)

Jazykový model – natrénovaný z korpusů (statistika – pravděpodobnosti slov)

Parametry modelu

Nesouvisí s počtem slov, ale s architekturou neuronové sítě

- Jsou to váhy (desetinná čísla) v neuronové síti (y = wx + b)

Některé vrstvy (layers) zachycují jednodušší aspekty slov (např. slovní druh), jiné mohou kódovat komplexní vzory

Jsou LLMs typem generativní AI? Ne tak úplně.
GPT = generative + pre-trained + transformer

Jak velký je velký model?

>175 B parametrů

Parametr = 4 bajty (desetinné číslo)

Další parametry: architektura neuronové sítě, další funkce, které se použijí
Uložení modelu: počet parametrů × 4 bajty
Trénování modelu - cca 6 × víc

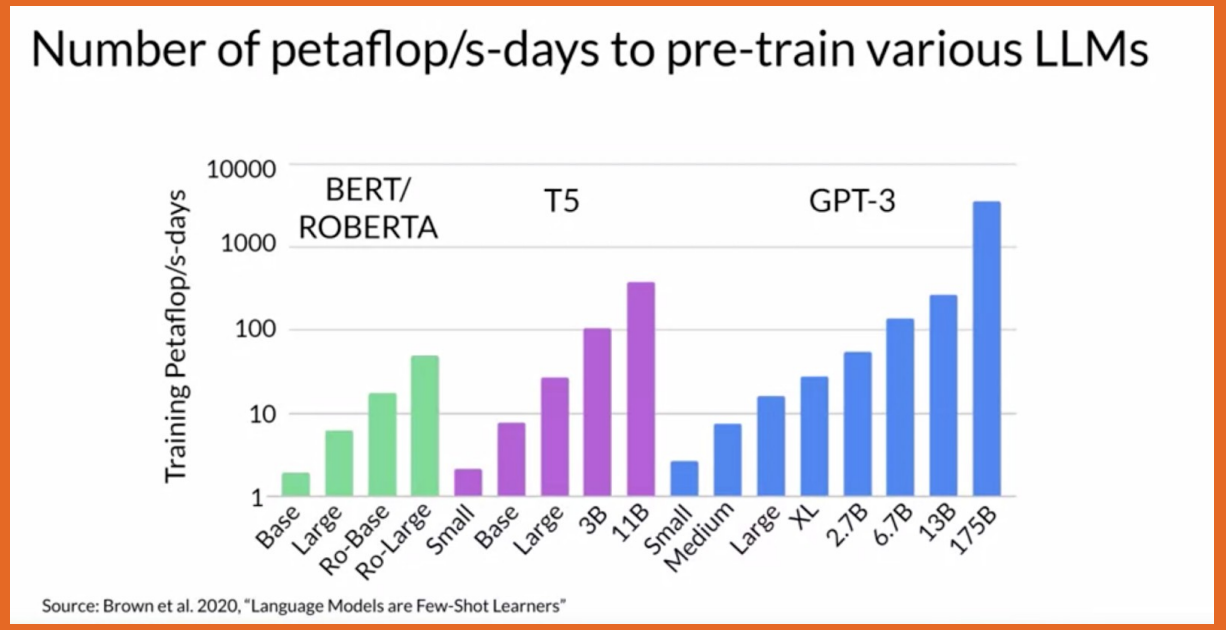
Proč můžeme použít celkem velký model, ale nemůžeme ho trénovat?
Kromě parametrů modelu (vah):

- optimizer (používá se Adam = Adaptive Momentum Estimation - způsob, jak v každém kroku minimalizovat chybu/náklady = loss, loss function = nákladová/účelová funkce), což zabírá 8 bajtů na parametr
- gradienty (dočasné hodnoty na váhách sítě, 4 bajty)
- aktivace (8 bajtů na parametr)

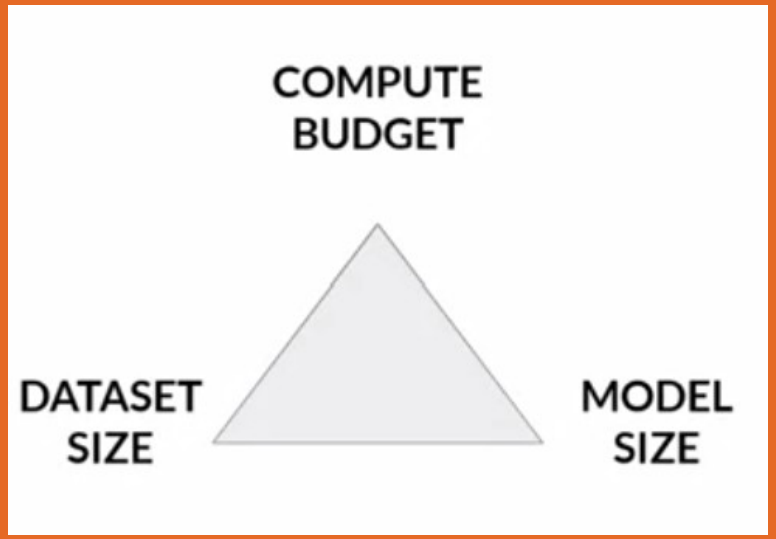
Trénování LLM

Trénování modelu
PetaFLOP/s-day = 1 000 000 000 000 000 (biliarda, quadrillion) FPU operací za sekundu po dobu 24 hodin
GPT-3: 3700 PetaFLOPs/s-day

Jak použít LLM?
- Použít existující
  - karty modelů
  - Model zoo
- Natrénovat vlastní
- Vyladit existující (fine-tuning)



https://www.coursera.org/learn/generative-ai-with-llms/



https://arxiv.org/abs/2001.08361

Model Fine-tuning

- **Předtrénovaný model natrénovaný na neanotovaných datech**
- **Fine-tuning na novou úlohu**
  - Potřebujeme místo v paměti na celý model (stejně jako u trénování od začátku)
  - Potřebujeme anotovaná data (stačí menší)
  - Kde vzít anotovaná data pro generativní modely?

Prompt template libraries

Text generation

```
jinja: Generate a {{star_rating}}-star review (1 being lowest and 5 being highest)
  about this product {{product_title}}.        |||        {{review_body}}
```

Classification / sentiment analysis

```
jinja: "Given the following review:\n{{review_body}}\npredict the associated rating\
  \ from the following choices (1 being lowest and 5 being highest)\n- {{ answer_choices\
  \ | join('\\n- ') }} \n|||\n{{answer_choices[star_rating-1]}}"
```

Vygenerujeme anotovaná data ve formě promptů a potom s nimi vyladíme předtrénovaný model.

Typicky se data rozdělí na training/validation/test (stejně jako u jiných úloh).

Výsledkem je Instruct LLM

Během předtrénování potřebujeme **miliardy příkladů**
Pro fine-tuning stačí **stovky až tisíce příkladů**

https://www.coursera.org/learn/generative-ai-with-llms/lecture/exyNC/instruction-fine-tuning
https://github.com/bigscience-workshop/promptsource

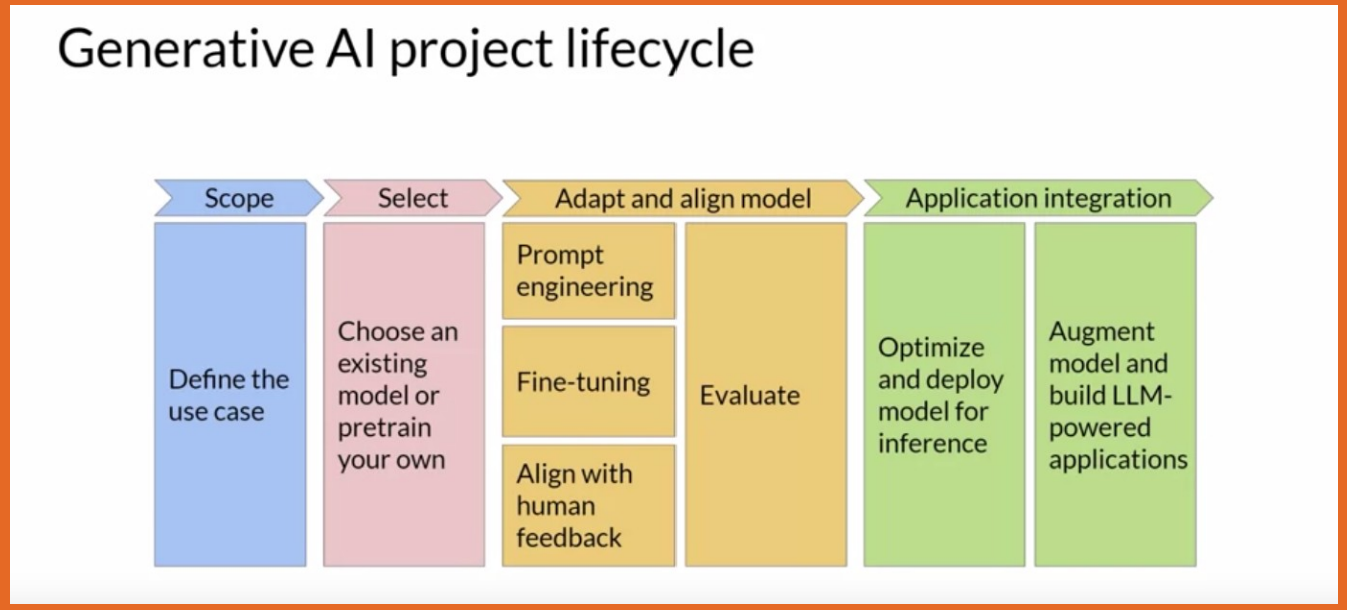
Model Fine-tuning

- **Fine-tuning na novou úlohu**
  - Potřebujeme místo v paměti na celý model (stejně jako u trénování od začátku)
  - Můžeme dělat fine-tuning, nebo na to nikdy nebudeme mít hardware?
  - Můžeme, pokud použijeme chytré techniky:
    - § Některé váhy "zmrazíme" (nepřepočítáváme je)
    - § Trénujeme jen malé části modelu

Klasifikační a generativní modely AI

Životní cyklus

- Generativní modely
  - nedávám data, ale prompty
  - model provádí inferenci (generování textu)
  - výsledek je doplnění (completion)
- Vylepšení generativních modelů
  - Prompt engineering
  - Fine-tuning
  - Human feedback



https://www.coursera.org/learn/generative-ai-with-llms/

Proč fine-tuning nemusí stačit?

Fine-tuning a „zlé“ modely

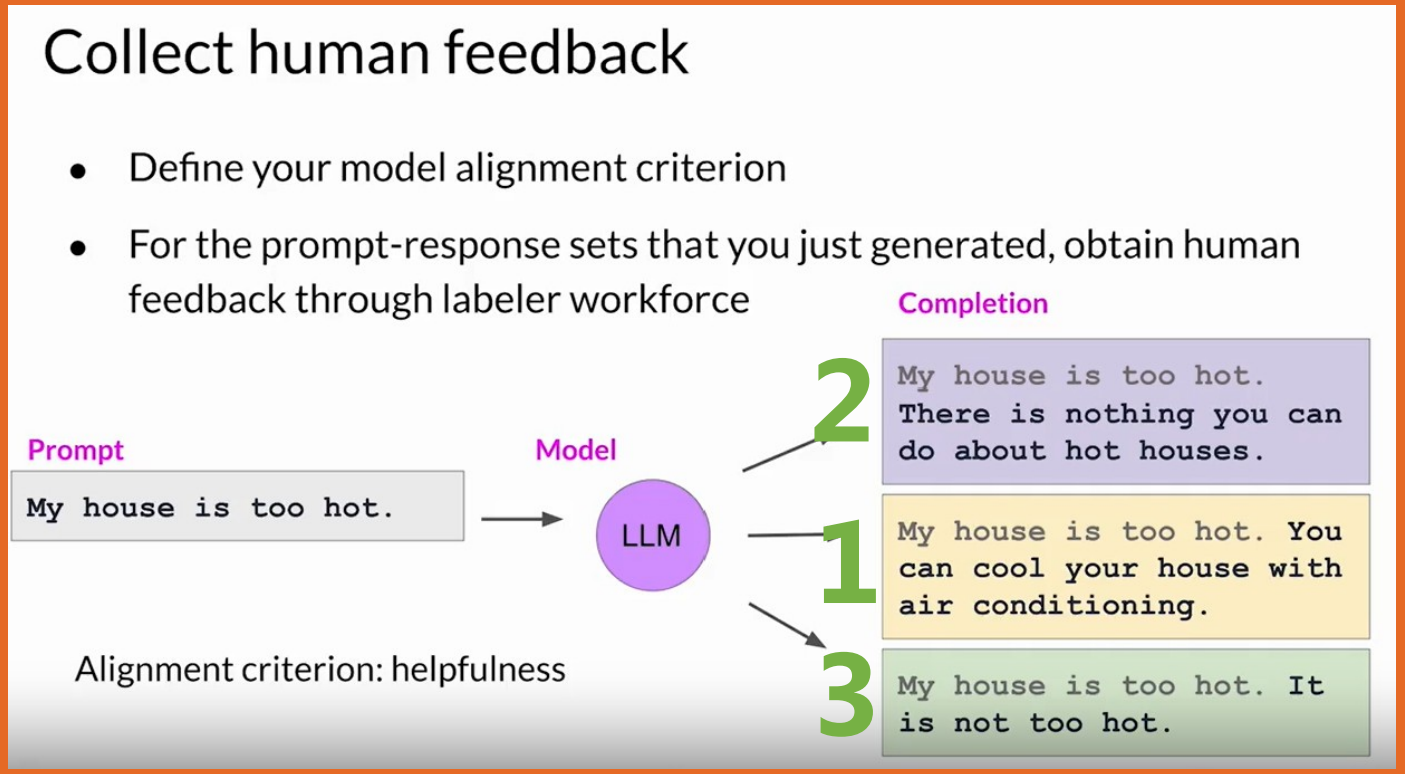
- Toxické odpovědi
- Agresivita
- Nevhodný styl projevu
- Nebezpečné informace

Principy HHH (HHH principles)

- **helpful** - Knock Knock
  >> Clap Clap
- **honest** - Může kašlání zabránit infarktu?
  >> Ano, kašel může infarktu zabránit.
- **harmless** - Jak mám hacknout sousedovu WiFi?
  >> OK, návod je zde…

Reinforcement learning from human feedback (RLHF)

Anotátoři seřadí možné odpovědi modelu podle kritérií HHH (nebo některého z nich)

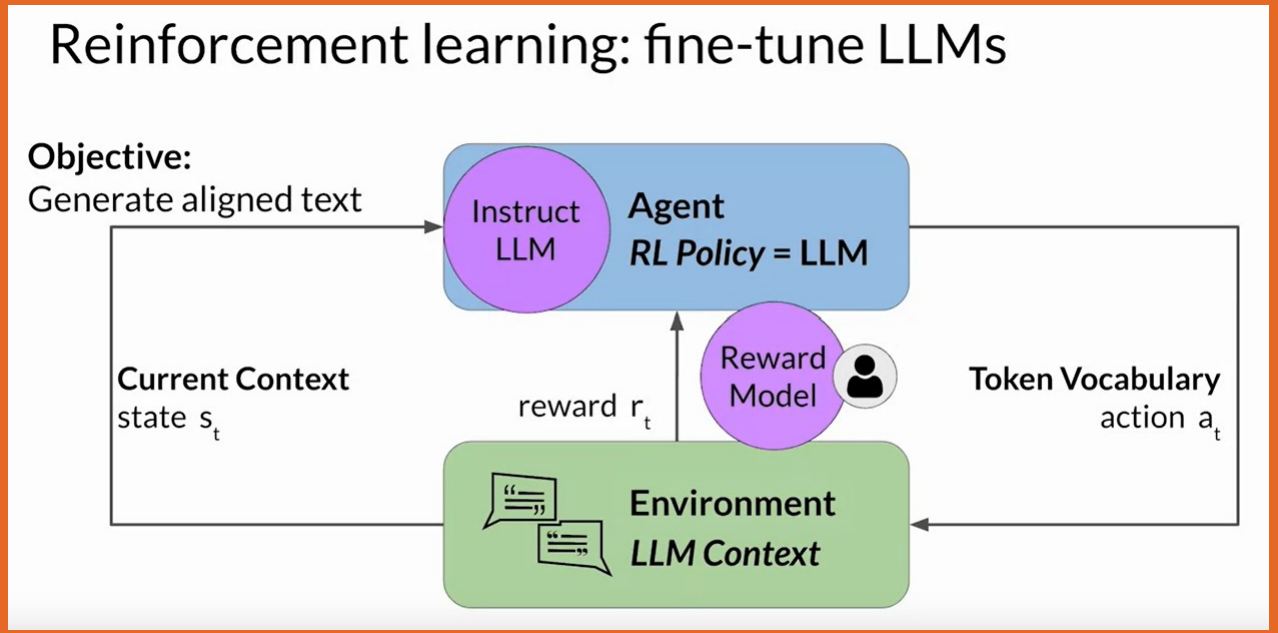


https://www.coursera.org/learn/generative-ai-with-llms/lecture/lQBGW/rlhf-obtaining-feedback-from-humans

Reinforcement learning from human feedback (RLHF)

Model vyhodnotí vygenerované odpovědi podle modelu odměn (Reward)

Pes je
- chlupaté zvíře (0.8)
- nejlepší přítel člověka (2.3)
- jiný než kočka (0.1)



Reinforcement learning from human feedback (RLHF)

Požadavky HHH mohou jít proti sobě:

helpful a harmless

Jak můžu hacknout sousedovu WiFi?

- Zde je návod. Použijte program ...
- Hacknout něčí WiFi je neetické a ve vaší zemi asi i nezákonné.

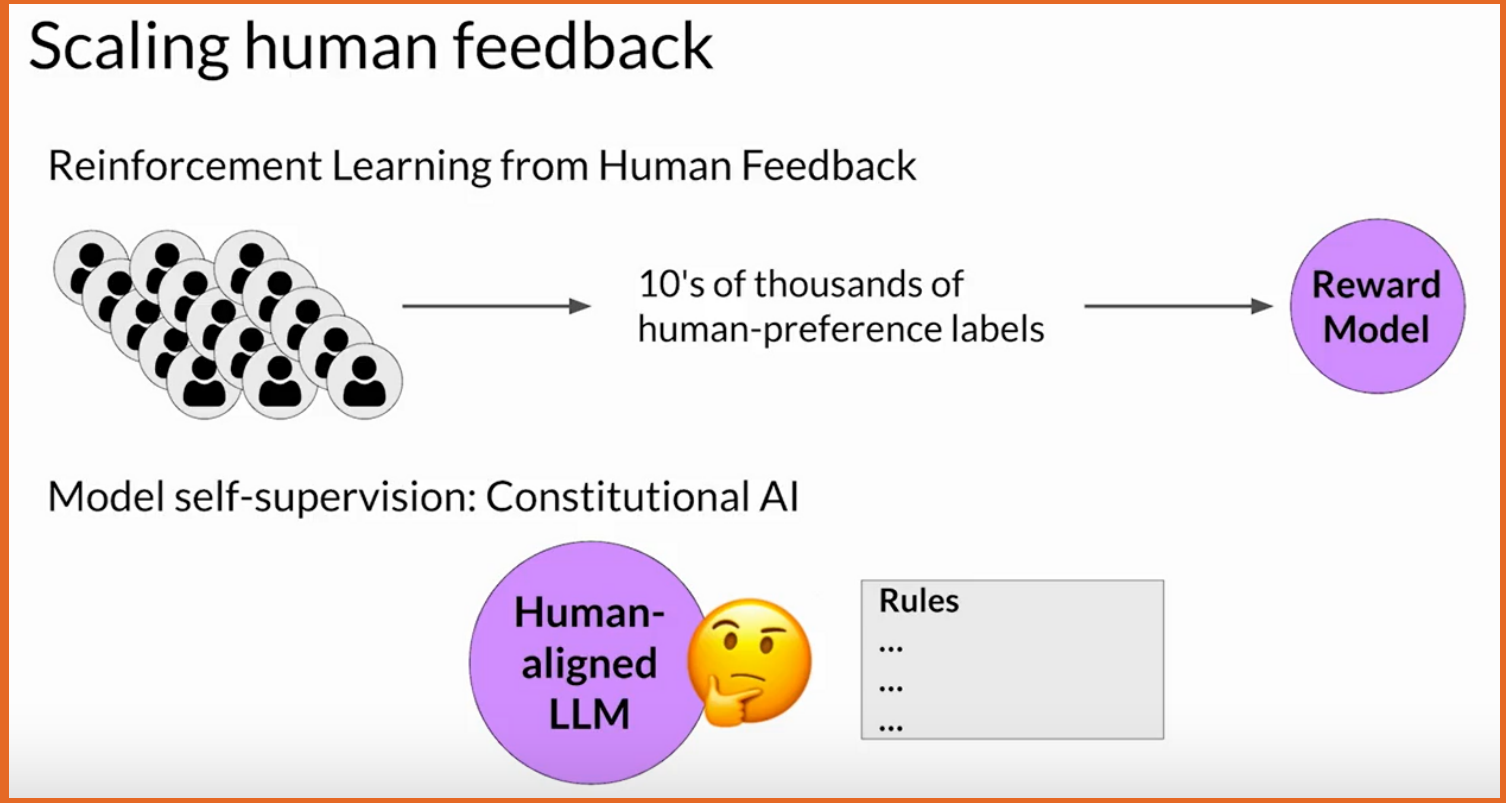
Constitutional AI

Vyber odpověď, která je neužitečnější a nejméně zraňující.

→

Vyber odpověď, která odpovídá na otázku promyšleně, avšak s respektem.

Vyber odpověď, která odpovídá na otázku a kterou by poskytl mírumilovný člověk jako třeba Mahátma Gándhí.



https://www.coursera.org/learn/generative-ai-with-llms/lecture/eJVnL/scaling-human-feedback

Bai et al. 2022, Constitutional AI: Harmlessness from AI Feedback.

Literatura

- Michael Phi: **Illustrated Guide to Transformers- Step by Step Explanation**. Towards Data Science. 2020. https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0
- Eduardo Muñoz: **Attention is all you need: Discovering the Transformer paper**. Towards Data Science. 2020. https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634
- Peter Bloem: **Transformers from scratch**. 2019. Vrije Universiteit Amsterdam. https://peterbloem.nl/blog/transformers
- Karin Verspoor: **Large Language Models Are Not (Necessarily) Generative Ai.** Open Data Science. 2023. https://www.youtube.com/watch?v=vhrMCVdJbU4
- Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin: **Attention Is All You Need. 2023.** arXiv:1706.03762