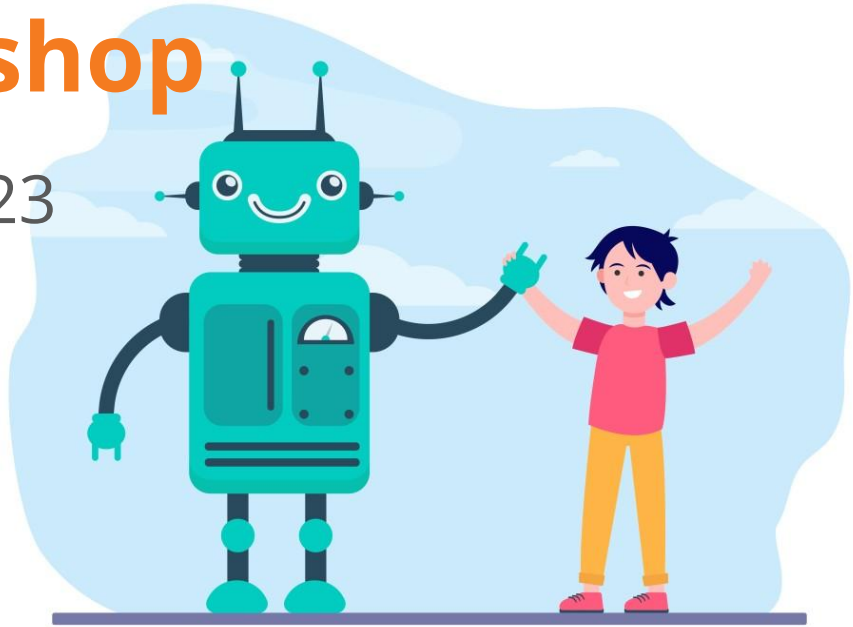
Rudolf Rosa, ÚFAL MFF UK

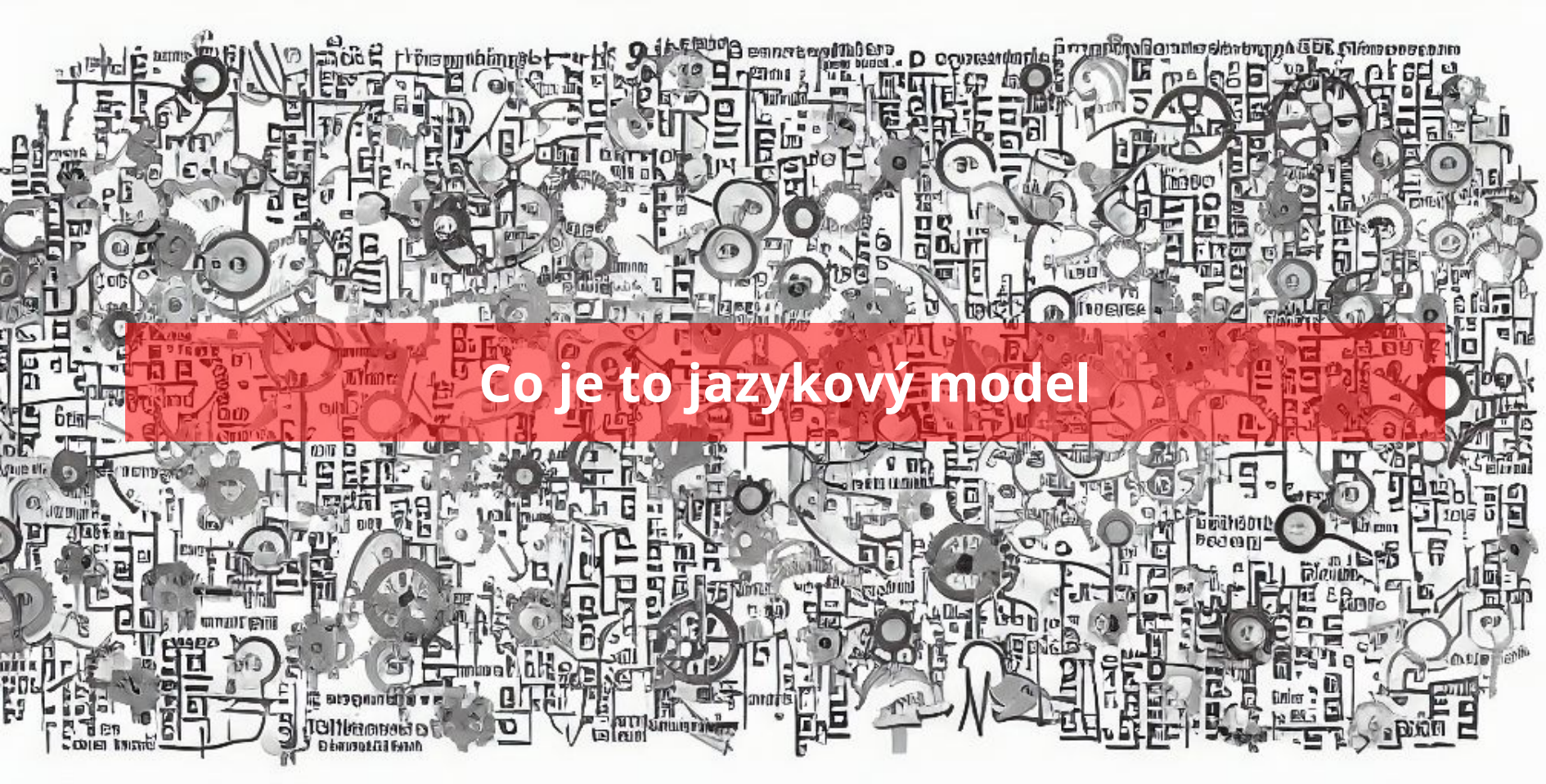
AI workshop

RUK 12.9.2023



bit.ly/ruk-ai-2023

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky


Co je to jazykový model

Jazykový model (např. GPT-3)

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do...
    - ???

Jazykový model

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do...
    - práce
    - koupelny
    - kina
    - koňské

Jazykový model

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do… práce/koupelny/kina/koňské
- Jazykový model
  - Potřebuje se naučit, jak vypadá jazyk
  - Obrovské množství textů: noviny, knihy, webové stránky, filmové titulky… (miliardy slov)

Jazykový model

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do… práce/koupelny/kina/koňské
- Jazykový model
  - Potřebuje se naučit, jak vypadá jazyk
  - Obrovské množství textů: noviny, knihy, webové stránky, filmové titulky… (miliardy slov)
- N-gramový jazykový model (např. 3-gramový)
  - Jak často po slovech A B následuje slovo **C**?
    - "šel do **práce**" > "šel do **koňské**"?
    - "šel do **koupelny**" > "šel do **kina**"?

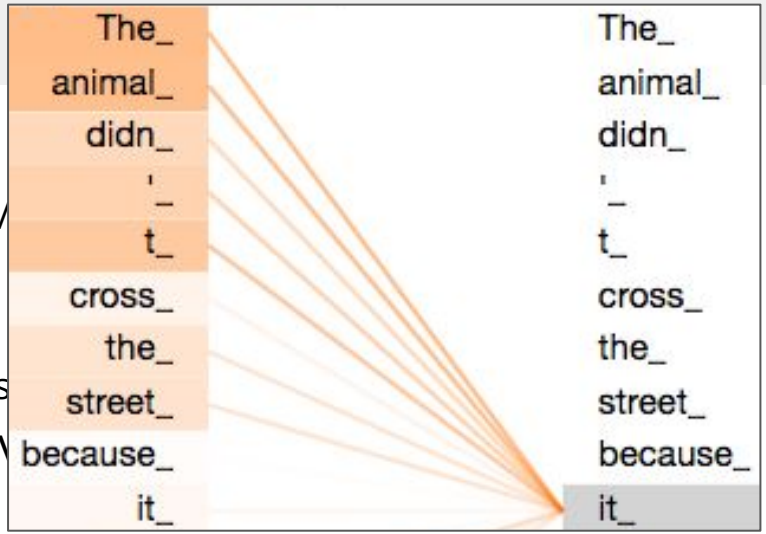
Jazykový model

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do… práce/koupelny/kina/koňské
- Jazykový model
  - Potřebuje se naučit, jak vypadá jazyk
  - Obrovské množství textů: noviny, knihy, webové stránky, filmové titulky… (miliardy slov)
- N-gramový jazykový model (např. 3-gramový)
  - Jak často po slovech A B následuje slovo **C**?
    - "šel do **práce**" > "šel do **koňské**"? 1640 > 8
    - "šel do **koupelny**" > "šel do **kina**"?

Jazykový model

- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do… práce/koupelny/kina/koňské
- Jazykový model
  - Potřebuje se naučit, jak vypadá jazyk
  - Obrovské množství textů: noviny, knihy, webové stránky, filmové titulky… (miliardy slov)
- N-gramový jazykový model (např. 3-gramový)
  - Jak často po slovech A B následuje slovo **C**?
    - "šel do **práce**" > "šel do **koňské**"? 1640 > 8
    - "šel do **koupelny**" > "šel do **kina**"? 372 > 287

Jazykový model



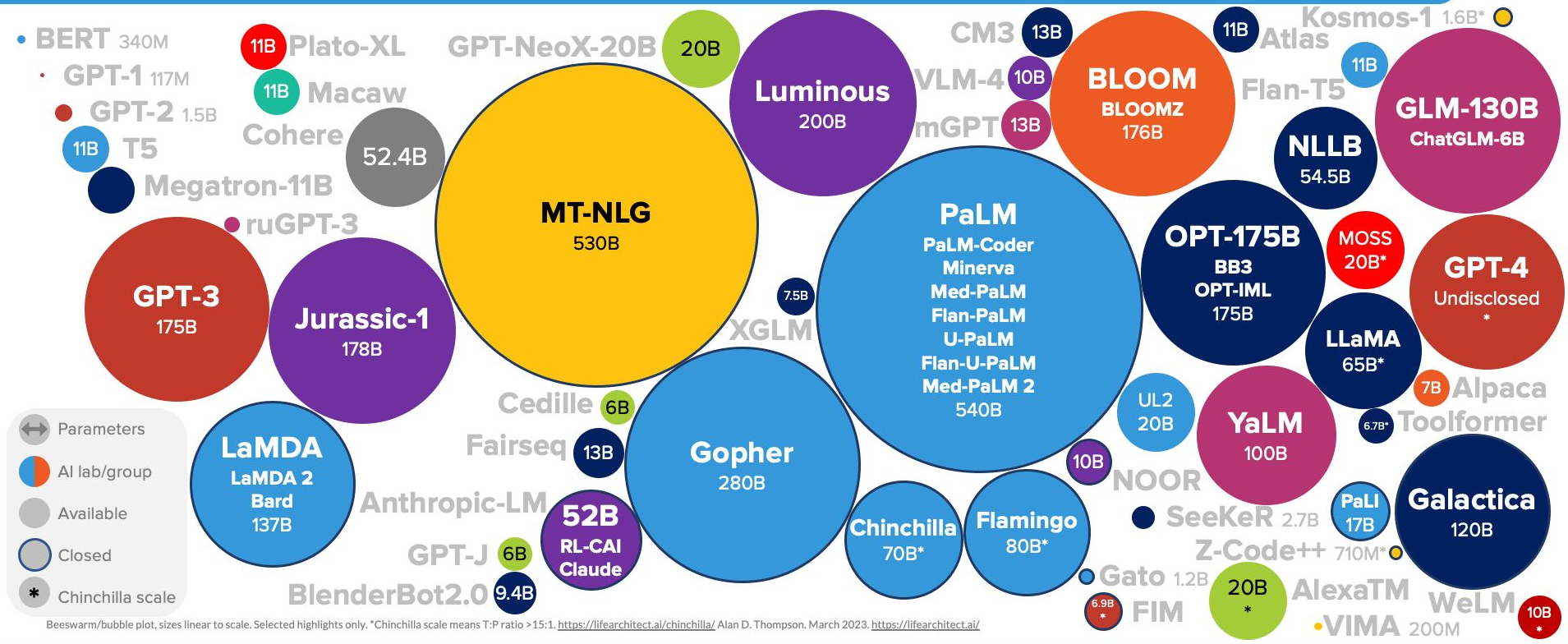
- Úloha: jaké slovo má následovat?
  - Tatínek ráno vstal a šel do… práce/koupelny/kina/
- Jazykový model
  - Potřebuje se naučit, jak vypadá jazyk
  - Obrovské množství textů: noviny, knihy, webové s
- N-gramový jazykový model (např. 3-gramov
  - Jak často po slovech A B následuje slovo **C**?
    - "šel do **práce**" > "šel do **koňské**"? 1640 > 8
    - "šel do **koupelny**" > "šel do **kina**"? 372 > 287
- Jazykový model založený na umělých neuronových sítích
  - Dívá se na větší počet předchozích slov (GPT-2: až 1023 předchozích slov)
  - Vybírá si, na která předchozí slova se bude dívat (attention)
  - Umí odhadnout podobnost slov (kosinová vzdálenost slovních embeddinků)
  - Složitější vnitřní reprezentace a operace místo prostého porovnávání četností
  - …

Některá omezení většiny jazykových modelů

- Založené na dostupných datech
  - dobré pro běžné časté věci, potřeba velkých trénovacích dat
  - biasy: předsudky, rasismus, sexismus, agresivita, heteronormativita, klišé…
- Technická omezení
  - omezený kontext (ale: GPT-4 údajně až 32 000 tokenů ~ desítky stránek textu)
  - nevhodná reprezentace čísel a matematiky
  - špatné chápání časové souslednosti (“vidí texty ze všech časových okamžiků najednou”)
- Založené pouze na textu
  - nemají zkušenosti z reálného světa, nemají lidské vnímání světa a sebe
  - nemají jiné kanály (omezeně: multimodální modely)
  - protiřečení, nesmysly
  - divadelní hra: jako dramatik, který nikdy nebyl v divadle
- Chybí “knowledge of knowledge” či “confidence estimation”
  - neumí dobře odhadnout kvalitu vygenerovaného výstupu
  - míchání faktických znalostí a jazykových dovedností
- Skutečné porozumění jazyku, obecná/silná umělá inteligence…???

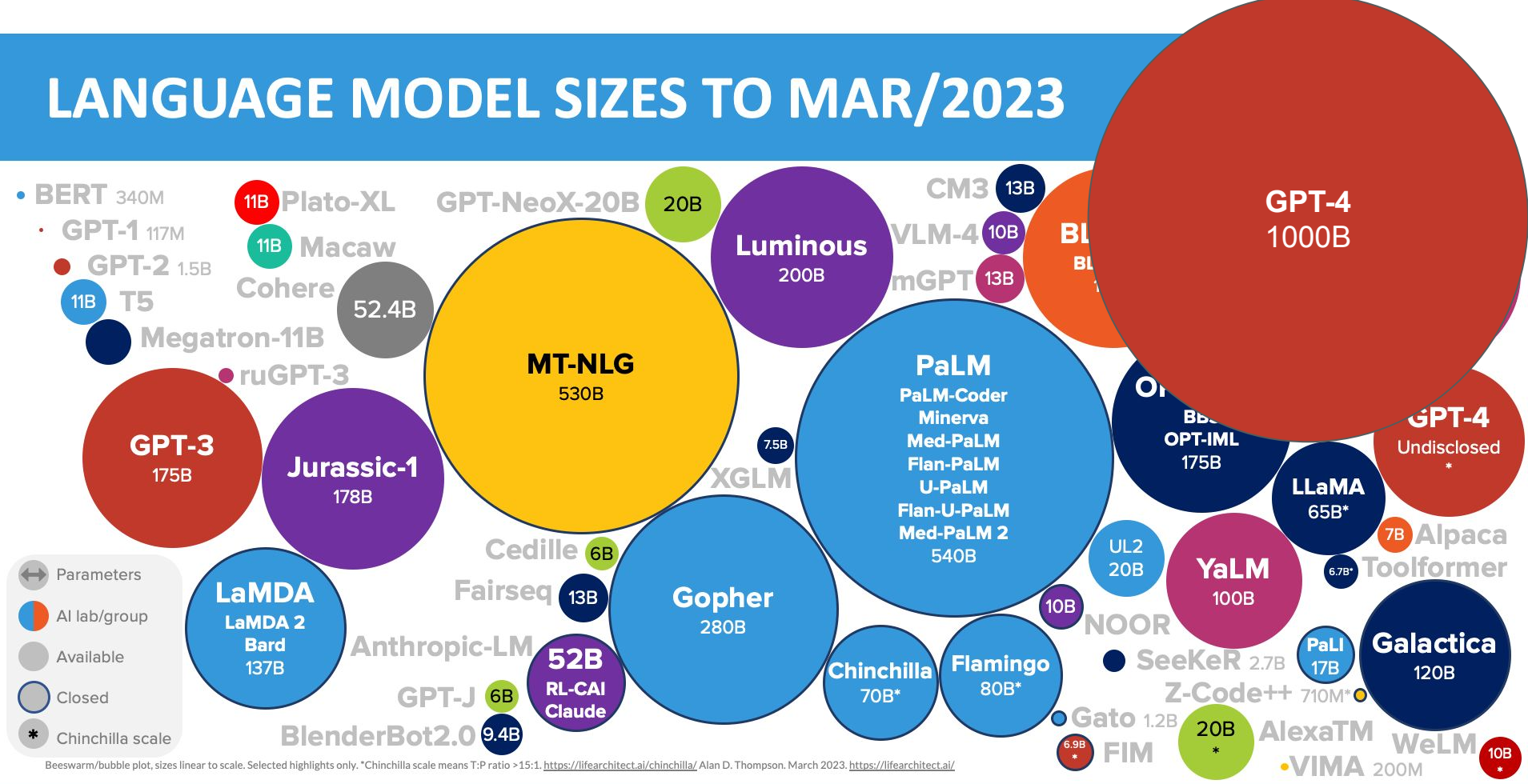
Jazykové modely nejsou jen GPT

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

Některé významnější jazykové modely a postupy

- generativní jazykové modely
  - OpenAI: GPT models (GPT-1 … GPT-4, ChatGPT)
  - Meta: OPT, LLaMA → Alpaca, Vikuna, gpt4all
  - akademická sféra: HPLT, OpenGPT-X, OpenAssistant (vše in progress)
- masked modely: dobré pro reprezentaci textu, špatné pro generování
  - BERT, Multilingual BERT, RoBERTa, XLM-RoBERTa, BART
- destilování modelů (e.g. DistillBERT)
  - teacher-student: vezmu velký model, učím malý model simulovat velký model
- nejen “přirozeně se vyskytující text” (~internet crawl)
  - InstructGPT: instrukce: “Summarize the following text: …”
  - Copilot: zdrojové kódy: “Write a Python script to sort an array…”
  - GPT3.5: reinforcement learning with human feedback (RLHF): 👍👎 “correct answer is: …”
- nejen text → multimodální (obrázky), externí tooly, embodiment (roboti)
  - GPT-4, Retrieval-enriched LMs, Bing AI Bot, ToolFormer, ChatGPT plugins, PaLM-E

Praktické použití existujících nástrojů

Existující nástroje

- generování textu, práce s textem…
  - ChatGPT: chat.openai.com, Codebreaker Chat: codebreakeredu.com/chat, Google Bard
- odpovídání na otázky, vyhledávání odpovědí…
  - Bing Chat: bing.com, Perplexity AI: perplexity.ai
- odpovídání na otázky podle PDF dokumentu
  - ChatPDF: chatpdf.com, AskPDF: askpdf.xyz
- detekce vygenerovaného obsahu (není a nikdy nebude zcela spolehlivá)
  - Turnitin Originality, Copyleaks, Hive
- strojový překlad
  - Google, DeepL, Bing, LINDAT Translation, Charles Translator for Ukraine
- OCR (převod obrazu na text), i čeština, i psané rukou
  - nahrát obrázek na Google Drive → “otevřít v aplikaci Dokumenty Google”
  - PicWish, Online OCR, Tesseract OCR
- řada dalších AI nástrojů a technologií
  - převod zvuku na text (ASR) a textu na zvuk (TTS), generování a analýza obrázků a videí…

Základní tipy na používání velkých jazykových modelů

- klasický generativní jazykový model: dokončení textu
  - zadám začátek textu, model vygeneruje pokračování
  - *Vážení studující, rádi bychom vás pozvali na slavnostní zahájení akademického roku.*
- zadat instrukce, co má udělat
  - jasné, jednoznačné, detailní
  - *Napiš e-mail pro doktora Rosu, ve kterém jej požádáš o přednášku o AI a zeptáš se na detaily.*
- důležité věci v promptu zopakovat
  - hlavně na začátku a na konci; to jsou nejdůležitější pozice
  - *Napiš e-mail… (další pokyny)… Napiš to ve formě e-mailu.*
- vygenerovat víc možností
  - a vybrat si z nich ručně, anebo i požádat model, ať z nich vybere
  - buď rovnou požádat o několik možností, nebo několikrát nechat přegenerovat
- když je výstup moc krátký, skončí předčasně, má jakékoliv problémy…
  - zkusit přegenerovat výstup, zkusit požádat o úpravu, zkusit formulovat zadání jinak…
  - *Pokračuj. Napiš to podrobněji. Napiš to jednodušeji. Napiš to ve stylu obchodního dopisu.*

Pokročilé tipy na používání velkých jazykových modelů

- chain of thought (řetězec myšlenek)
  - *Popiš postup, jak jsi k odpovědi došel.*
  - *Postupuj krok za krokem.*
  - *Vysvětli svou odpověď.*
- “few shot learning”: ukázat, co chci
  - jedna či několik ukázek vstupu a výstupu, pak vlastní vstup
  - *Slovo banán má 5 písmen. Slovo lokomotiva má 10 písmen. Slovo rarášek má*
- výběr z možností
  - říct, že chci odpověď A nebo B
  - *Je následující text adresován rektorce? Odpověz “ano” nebo “ne”. Text: Vážená paní rektorko…*
- rozdělit na víc podúloh
  - generovat po částech, nechat rozvést, nechat shrnout, rozdělovat, spojovat…
  - *Navrhni skupiny hostů na udělení čestného doktorátu…*
  - *Napiš pozvánku pro první skupinu hostů…*

Další tipy na používání velkých jazykových modelů

- angličtina
  - modely obvykle v angličtině fungují lépe než v jiných jazycích
- říct, co chci
  - model se mi snaží dát to co chci
  - neví že chci “dobrý” text, v datech má i “špatné” texty, co když chci špatný?
  - říct si, že to má být kvalitní, pravdivé…
  - ale ani tak není zaručen dobrý výsledek
- říct, jako kdo má psát
  - pokud používám model přes programovací rozhraní, můžu mu přes system message popsat, jak mi má odpovídat
  - výchozí pro ChatGPT je “You are a helpful assistant.” (“Jsi užitečný pomocník.”)
  - ale můžu zadat například “You are an administrative employee at the university.”
  - celkem dobře lze zadat i přímo do promptu (stojí zato zkusit nejen česky ale i anglicky)

Jak nepoužívat ChatGPT a na co si dát pozor

- není to Google vyhledávač
  - na otázky odpoví, ale klidně si odpověď vymyslí
  - nerozlišuje pravdivé a pravděpodobné…
  - necituje zdroje, vymýšlí si zdroje…
  - lepší: Bing AI, Perplexity AI (skutečné zdroje, byť ne nutně důvěryhodné)
- má omezenou velikost vstupu
  - typicky několik tisíc slov
  - na práci s delším dokumentem lze použít např. ChatPDF či AskPDF
- odpovědi mohou být problematické
  - může vydat toxické odpovědi, nezákonné, neetické…
  - existují nějaké kontroly a filtry, ale nejsou dokonalé (tím spíš mimo angličtinu)

Existující nástroje

- generování textu, práce s textem…
  - ChatGPT: chat.openai.com, Codebreaker Chat: codebreakeredu.com/chat, Google Bard
- odpovídání na otázky, vyhledávání odpovědí…
  - Bing Chat: bing.com, Perplexity AI: perplexity.ai
- odpovídání na otázky podle PDF dokumentu
  - ChatPDF: chatpdf.com, AskPDF: askpdf.xyz
- detekce vygenerovaného obsahu (není a nikdy nebude zcela spolehlivá)
  - Turnitin Originality, Copyleaks, Hive
- strojový překlad
  - Google, DeepL, Bing, LINDAT Translation, Charles Translator for Ukraine
- OCR (převod obrazu na text), i čeština, i psané rukou
  - nahrát obrázek na Google Drive → “otevřít v aplikaci Dokumenty Google”
  - PicWish, Online OCR, Tesseract OCR
- řada dalších AI nástrojů a technologií
  - převod zvuku na text (ASR) a textu na zvuk (TTS), generování a analýza obrázků a videí…

Právní aspekty AI

Varování

- Nemám právní vzdělání
  - leccos už jsme konzultovali s různými právníky
  - ale v závažných otázkách se obraťte na skutečné právníky
    - např. Jan Zibner, Jan Hořeňovský, Jáchym Stolička, Jan Barták

Právní situace dosti nejasná

- Právní podchycení zatím dost omezené
  - většinu otázek právní řád zatím uspokojivě neřeší
- Nástroje mají licenční podmínky
  - legálnost a právní vymahatelnost v lecčems sporná
  - je vhodné je dodržovat
- Legislativa se začíná tvořit
  - AI Act, autorskoprávní spory…

Práva k vygenerovaným dílům

- Typicky má uživatel veškerá práva
  - viz licenční podmínky nástroje
  - obvykle komerční i nekomerční využití za libovolným legálním účelem
  - obvykle není nutné uvádět použitý nástroj
    - ale typicky vhodné (citování zdrojů)
  - někdy mají některá práva i poskytovatelé nástroje
- Vygenerované dílo *spíše není* autorským dílem
  - “Předmětem práva autorského je dílo (…), které je jedinečným výsledkem tvůrčí činnosti autora (…)”
  - “Autorem je fyzická osoba, která dílo vytvořila.“

Práva k zadávaným vstupům (promptům)

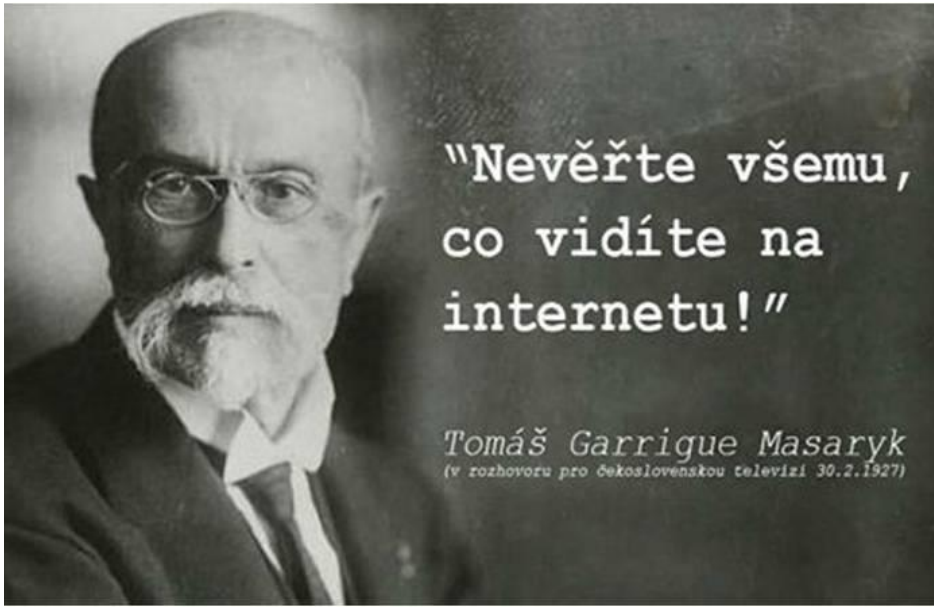
- Uživatel je jistě autorem promptu
  - ale je to autorské dílo?
  - typicky užitím nástroje dáváte k promptu některá práva poskytovatelům nástroje
  - typicky prompty mohou být využity pro trénování dalších verzí nástroje
- K zadávanému vstupu byste pravděpodobně měli mít vhodná práva
  - např. úryvek knihy
  - např. práce napsaná studentem
    - poskytnutí díla provozovatelům AI nástroje zjevně nespadá mezi povolená užití školního díla → raději si vyžádat souhlas
- Rozhodně nezadávejte cokoliv obsahující osobní či utajované údaje
  - např. osobní údaje studentů, číslo kreditní karty, hesla, platy zaměstnanců…
- Některé nástroje mohou garantovat důvěrnost dat
  - např. GPT využívané přes Microsoft Azure OpenAI Service (placené)
  - např. některé nástroje, které si instalujete a provozujete lokálně

Omezení daná podmínkami užití nástroje

- Často zákonná, etická a další omezení (viz licenční podmínky nástroje)
  - nevyužívat k ničemu nelegálnímu
  - neporušovat autorská práva
  - negenerovat urážlivý obsah, erotický obsah apod.
  - nástroj někdy upozorní na nevhodné užití, může vést i k zablokování uživatele
- Neostrá hranice
  - *vygeneruj mi obraz plačící ženy od Picassa*
    - asi není OK (obraz *Femme en pleurs (1937)*, chráněný autorským zákonem)
    - ale pokud se výsledek moc nepodobá originálu…?
  - *vygeneruj mi obraz plačící ženy*
    - asi OK
    - ale pokud se výsledek hodně podobá existujícímu obrazu od Picassa…?
  - konkrétní díla jsou chráněna autorským zákonem
    - i některé jejich části, které jsou samy autorským dílem (např. názvy, jména)
  - autorský styl jako takový chráněn není

Odpovědnost

- Odpovědnost za výsledky má vždy autor (nikoliv AI)
- Typicky se tvůrci a poskytovatelé nástroje zříkají veškeré odpovědnosti
- Veškerou odpovědnost má tedy uživatel
  - za zadávané vstupy
  - za použití vygenerovaných výstupů
  - za chyby nástroje
  - za autonomně jednající systémy (!)
  - ...
- Důležité uživatelské dovednosti
  - nevěřit hned výstupům AI nástrojů
    - často jsou správné, ale ne vždy
    - ChatGPT není Google!
  - kritické zhodnocení výstupů
  - dohledání a ověření zdrojů



Využitá trénovací data

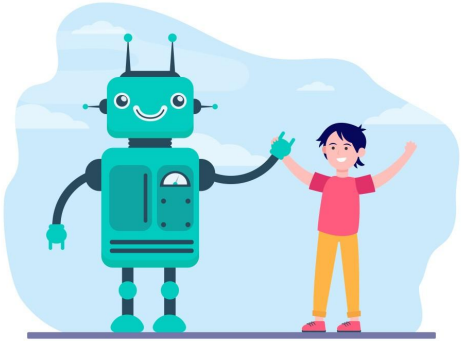
- Legálnost využití dat pro trénování nástrojů velmi sporná
  - málokterá data jsou zcela čistá pro využití, málokdy dává autor s tímto výslovný souhlas
  - trénování nástrojů autorský zákon pravděpodobně přímo neporušuje
    - nejde přímo o šíření chráněného díla
    - ačkoliv generátor může na výstup vydat autorsky chráněný obsah
    - rozmělnění pojmu “zveřejnění”, “šíření”…
  - využití je zjevně v rozporu s duchem autorského zákona
    - autoři a uživatelé nástroje mají nepřímo prospěch (i finanční) z užití díla
    - přičemž autor původního díla nic nezískává
    - autor ani zatím nemůže efektivně vyjádřit nesouhlas s takovým využitím svého díla
- Paralela s lidskými autory
  - člověk také vnímá existující díla jiných autorů a pak díky tomu tvoří vlastní díla
  - otázka vlastního tvůrčího vkladu *autora* (AI? uživatel? tvůrce nástroje?)
  - AI přináší vysokou efektivitu, veliký rozsah, užití je velmi snadné
  - i u lidských autorů je toto složité (v Česku např. “Upeč… třeba zeď”)

Současný a očekávatelný budoucí vývoj

- AI Act, kolektivní žaloby…
  - některé využití dat se legalizuje
  - některé využití dat se zakáže
  - možnost aktivně požádat o nezahrnutí/vyjmutí konkrétních dat
- Povinnost informovat o využití AI
- Budoucí nástroje z právních důvodů omezenější
  - omezená trénovací data → omezené schopnosti nástroje
- Tvorba a svěr vlastních dat legální cestou
  - např. ChatGPT legálně sbírá data od uživatelů
  - (ale jsou ta data poskytnutá legálně…?)

AI workshop

- Další odkazy
  - aireaktor.ujep.cz - návody a rady pro různé AI nástroje (UJEP)
  - aignos.cz - vzdělávací AI workshopy a semináře (spolek)
  - ainautes.com - komerční konzultace nasazení AI (tým)
  - ai.cuni.cz - doporučení pro AI ve vzdělávání (UK)
  - aivk.cz - interdisciplinární skupina “AI v kontextu” (UK)
- Otázky?



bit.ly/ruk-ai-2023

rosa@ufal.mff.cuni.cz