

Konference NZIP & NZIS Open 2025

10. prosince 2025

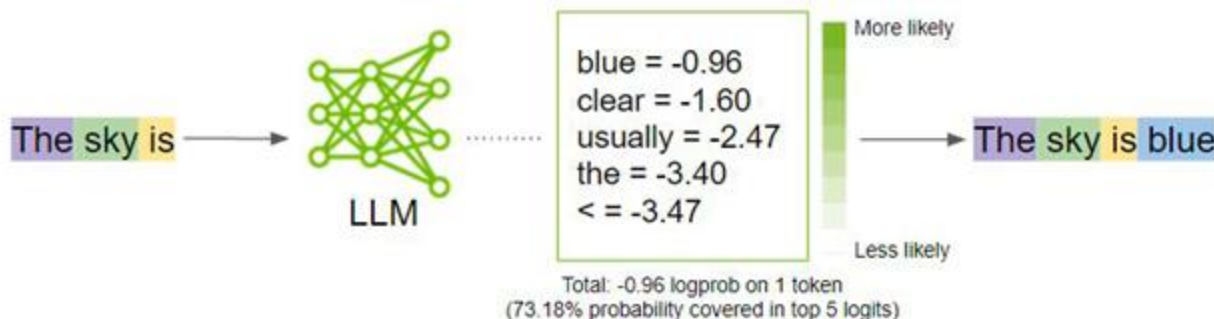
Název sekce: Zdravotnická data – technologie, využití, perspektivy I.

Ondřej Klempíř

České vysoké učení technické v Praze



Jak porozumět “jazyku” longitudinálních dat?



Projekt Stárnutí populace a související výzvy pro zdravotní a sociální systémy

AGEING - CZ

Reg.č. CZ.02.01.01/00/23_025/0008743, spolufinancován Evropskou unií



Spolufinancováno
Evropskou unií



datalab@fbmi.cvut.cz



FAKULTA
BIOMEDICÍNSKÉHO
INŽENÝRSTVÍ
ČVUT V PRAZE

MUNI
ECON



Attention Is All You Need			
Ashish Vaswani* Google Brain avaswani@google.com	Noam Shazeer* Google Brain noam@google.com	Niki Parmar* Google Research nikip@google.com	Jakob Uszkoreit* Google Research usz@google.com
Elliott J. Jones* Google Research elliotj@google.com	Aidan N. Gomez* University of Toronto aidan@cs.toronto.edu	Lukas Kaiser* Google Brain lukaskaiser@google.com	
Illia Polosukhin* illia.polosukhin@gmail.com			

- ❑ Role “AI” a velká data
- ❑ UK Biobank
- ❑ Our Future Health
- ❑ Úspěchy “AI” v medicíně
- ❑ AlphaFold 2
- ❑ Radiodiagnostika
- ❑ GPT
- ❑ **Fenotypování**
- ❑ Metody
- ❑ Příklady modelování

Research Letter

April 1, 2024

ONLINE FIRST

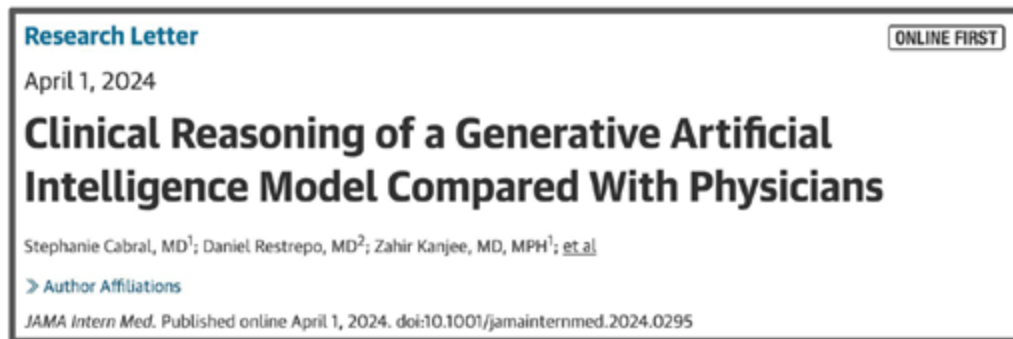
Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians

Stephanie Cabral, MD¹; Daniel Restrepo, MD²; Zahir Kanjee, MD, MPH¹; [et al](#)

> Author Affiliations

JAMA Intern Med. Published online April 1, 2024. doi:10.1001/jamainternmed.2024.0295

Computational (EHR-based) Phenotyping



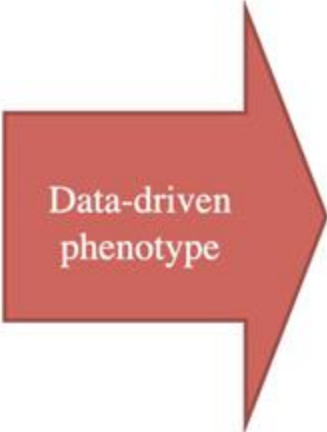
- ❑ Role "AI" a v...ta
- ❑ "I"...
- ❑ ...uture Health
- ❑ Úspěchy "AI" v medicíně
- ❑ AlphaFold 2
- ❑ Radiodiagnostika
- ❑ GPT
- ❑ *Fenotypování*
- ❑ Metody
- ❑ Příklady modelování

Computational (EHR-based) Phenotyping

Fenotyp
onemocnění

=

**POZOROVATELNÉ
CHARAKTERISTIKY ČI
ZNAKY ONEMOCNĚNÍ**



Data-driven
phenotype

**PROJEVUJE SE SEKUNDÁRNĚ V
DATECH**

Fenotypy → Symptomy → Břemeno → Čerpání
zdrav. péče → Charakteristický “podpis” v datech
o vykázané péči

(„datový fenotyp nemoci“)

Generátory velkých dat ve zdravotnictví

- EHR/EMR
- genomika
- PACS
- **administrativní data**
- laboratoře
- domácí monitoring

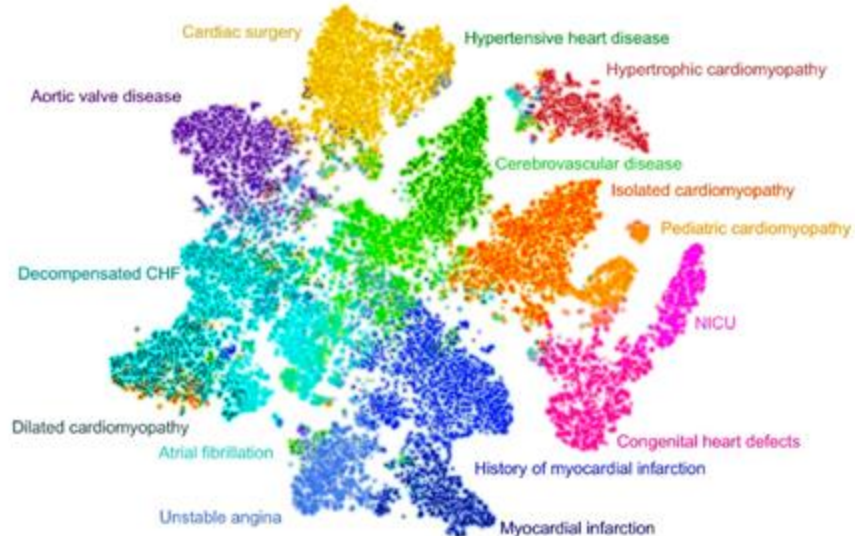
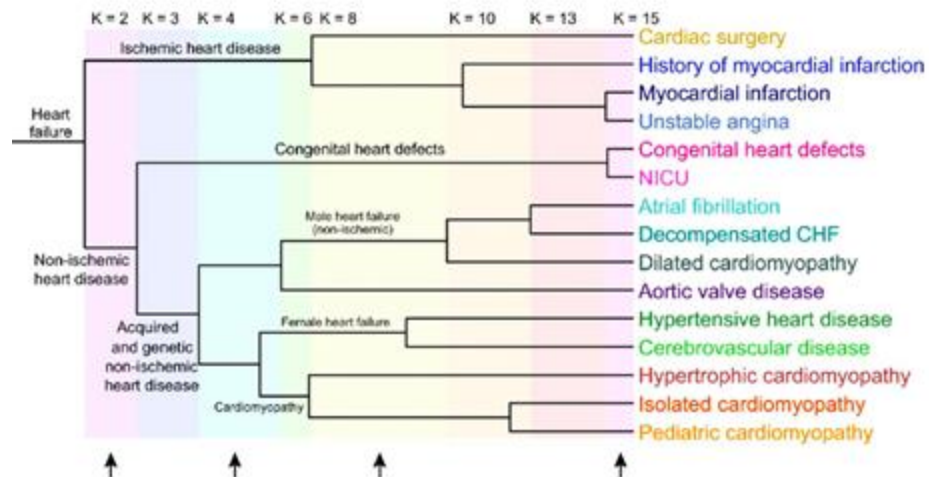


Úhly pohledu na longitudinální cestu pacienta

- Co se vyskytuje?
- Co se vyskytuje a kolikrát?
- Jaká je sekvence?
- Jaká je sekvence a časová delta mezi událostmi?

Code:	57243	87433	87223	87223	96167
ICD10 Code:	J44.8	C45.2	C34.3	C34.3	J44.8
Day:	1	2	5	5	8


temporal dimension



Co jsou ty jednotlivé datové body ve shledáních?

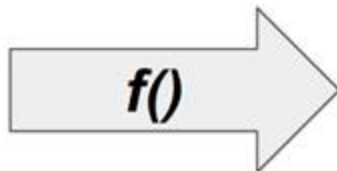
Nagamine, T., Gillette, B., Pakhomov, A. et al. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. *Sci Rep* 10, 21340 (2020).

Vektorová reprezentace sekvence

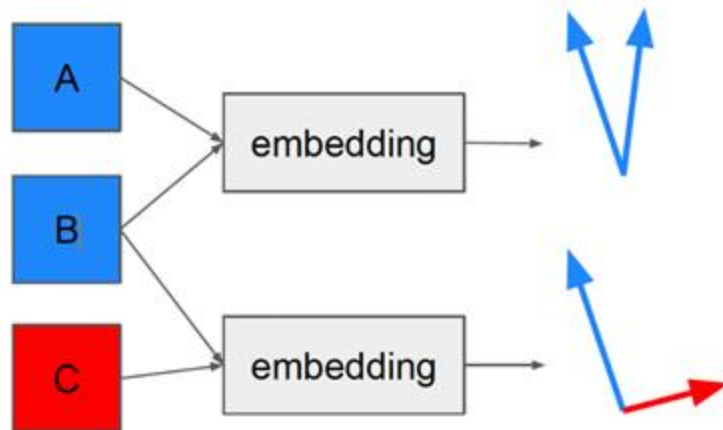
- S číselnými vektory se “dobře” počítá
 - Lineární algebra
 - Vhodný vstup pro neuronovou síť (hluboké učení)
 - Urychlení výpočtu na grafické kartě
 - Paralelizace
- Vektory umí zachovat různé vlastnosti

Code:	57243	87433	87223	87223	96167
ICD10 Code:	J44.8	C45.2	C34.3	C34.3	J44.8
Day:	1	2	5	5	8

temporal dimension →



- Tohle mohou být i objekty jiného typu
 - Medicínské obrazy
 - Audio
 - EKG
 - ...
 - multimodální...



[6.3, 2.5, 6.2, 4.4 ...]

- je kompaktní
- uchovává kontext
- dimenze bývá důležitá

Hledání fenotypů “s učitelem” vs. “bez učitele”

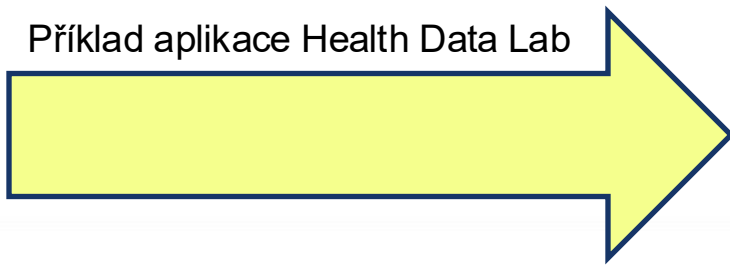
Známe rozdělení do skupin

- Překlasifikování chybně stanovených diagnóz
- Subtypování a kvantifikace podobnosti mezi diagnózami
- Stanovení pravděpodobnosti výskytu prodromální fáze
- Predikce (identifikace rizikových faktorů) pro určitou kondici, např. COVID

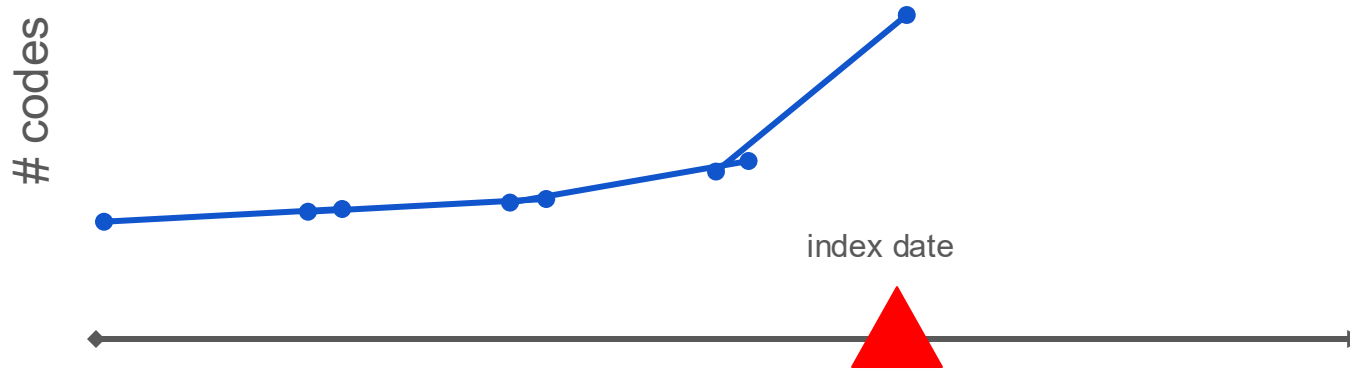
Neznáme přiřazení

- Detekce shluků, které vykazují podobné vlastnosti
- Detekce netypických cest pacienta

Příklad aplikace Health Data Lab



LONGITUDINÁLNÍ OBSERVAČNÍ DATA



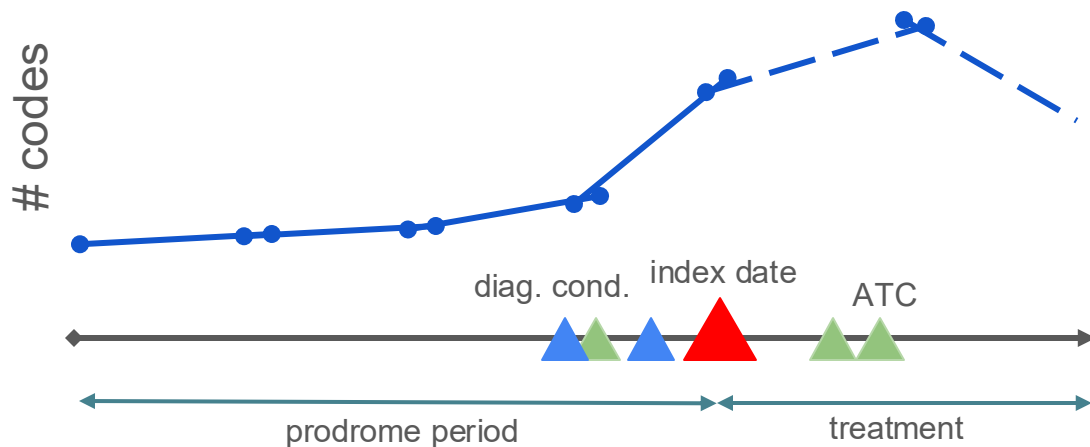
Health-care use before a first demyelinating event suggestive of a multiple sclerosis prodrome: a matched cohort study

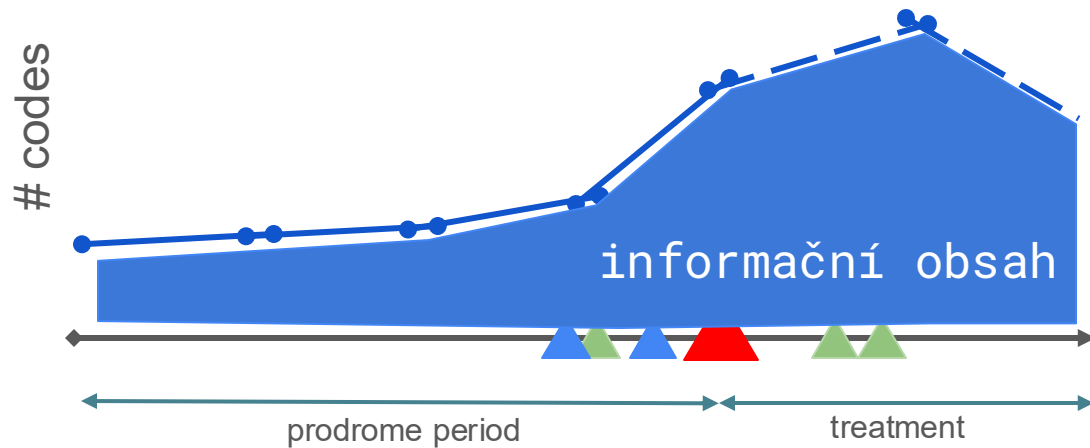
[José M A Wijnands, PhD](#) • [Elaine Kingwell, PhD](#) • [Feng Zhu, MSc](#) • [Yinshan Zhao, PhD](#) • [Tanja Högg, MSc](#) •

[Karen Stadnyk, MSc](#) • et al. [Show all authors](#)

10.1016/S1474-4422(17)30076-5

PRO ROZPOZNÁNÍ PRODROMÁLNÍ FÁZE VYUŽÍVÁME STROJOVÉ UČENÍ





Využití AI a Perspektivy pro NZIP & NZIS

Porovnání výkonu velkých jazykových modelů v komplexních úlohách na datech NZIP.cz

Motivace a záměr projektu

- Dokáže AI generovat strukturovaný a spustitelný kód?
 - Pokud ano, mohlo by to významně změnit přístupy v práci analytiků pracujících se zdrav. daty
- NZIP - veřejně dostupné datové sady
- Co jsou typické analytické otázky?
 - age-sex distribuce diagnóz
 - year-over-year porovnání incidence
 - a mnoho dalších...
- Takové dotazy jsou vesměs jednoduché na implementaci, avšak analytici potřebují přesto znát práce s datovou strukturou a mít programovací schopnosti v SQL nebo Python.

nzip.cz

Výše uvedené motivuje otázku:

Do jaké míry dokáží současné velké jaz. modely automaticky připravit správný a spustitelný kód pro takové úlohy?

Setup

- 11 dostupných LLMs v úloze generovat Python kód
 - Dokáží korektně odpovídat na zadané otázky v přirozeném jazyce s min. nebo žádnou lidskou opravou?
- 3 úrovně complexity
 - “Ground truth” - zreplikuj analytické otázky tak jak byly popsány v dokumentaci
 - “Go beyond” - teoreticky, model mohl být trénován na dokumentaci -> modifikace
 - “Information retrieval” - co když model neví nic o konkrétním datasetu -> “end-to-end helper”
- Evaluate
 - Obecná “Data Science” kritéria
 - Kritéria specifická pro medicínská data
- Prompty byly navrženy a formulovány v CZ
- Temperature setting je 0 (co to znamená?)

Evaluating large language models for natural-language-to-code generation on aggregate Czech public health data analysis

Ondrej Klempir¹, Ladislav Dusek², Radim Krupicka¹, Gleb Donin¹, Jan Zigmond¹, Radka Storchova¹, Ales Tichopad¹✉

¹ *Faculty of Biomedical Engineering, Czech Technical University in Prague, Sitna Square 3105, Kladno, Czech Republic, datalab@fbmi.cvut.cz*

² *Institute of Health Information and Statistics, Palackého náměstí 4, Prague, Czech Republic*

✉ *Corresponding author: ales.tichopad@gmail.com*

<https://gitlab.fel.cvut.cz/klempond/nzip-llms-eval/>

NELZE VŽDY ODPOVĚDĚT!

Jaký je celkový trend počtu pacientů s rakovinou plic v celé ČR v letech 2010-2012?

Atribut	**Popis**
url	`Otevrena-data-NR-04-84-asthma-bronchiale.csv`
dc:title	*Vývoj počtu pacientů s bronchiálním astmatem*
dc:description	Počty unikátních ošetřených pacientů (UOP) s bronchiálním astmatem v ČR a jednotlivých krajích od 2010.
Sloupce	1. `rok_pece` - kalendářní rok péče (int) 2. `kraj_kod` - alfanumerický kód kraje (str) 3. `pocet` - počet pacientů (int)
Zdroj	ÚZIS ČR (Národní registr hrazených zdravotnických služeb)
Licence	Volný přístup (`https://data.gov.cz/podminky-užiti/volny-pristup/`)
Poslední úprava	2025-07-17

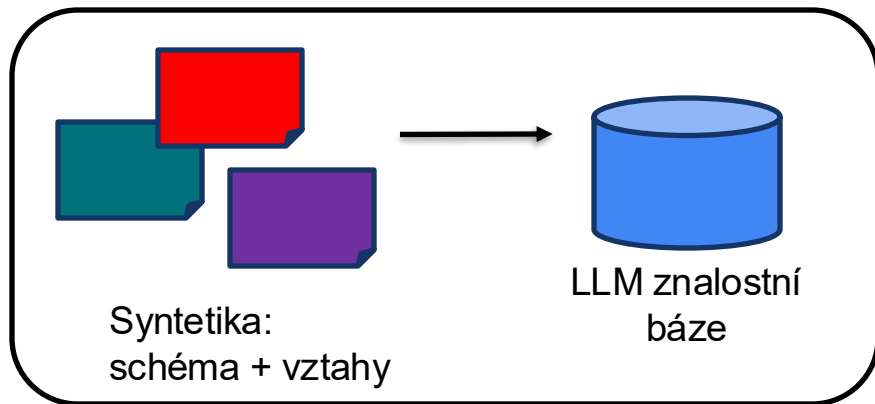
Nemohu připravit kód pro analýzu trendu počtu pacientů s rakovinou plic, protože poskytnutý dataset obsahuje údaje pouze o bronchiálním astmatu, nikoli o rakovině plic.

AI pro Syntetiku - příležitosti

1. Může nám AI pomoci s návrhem SQL pro syntetiku?

- Self-hosted LLM

- gpt-oss-120b - "ohromný" kontext dokáže zpracovat
- Metacentrum API



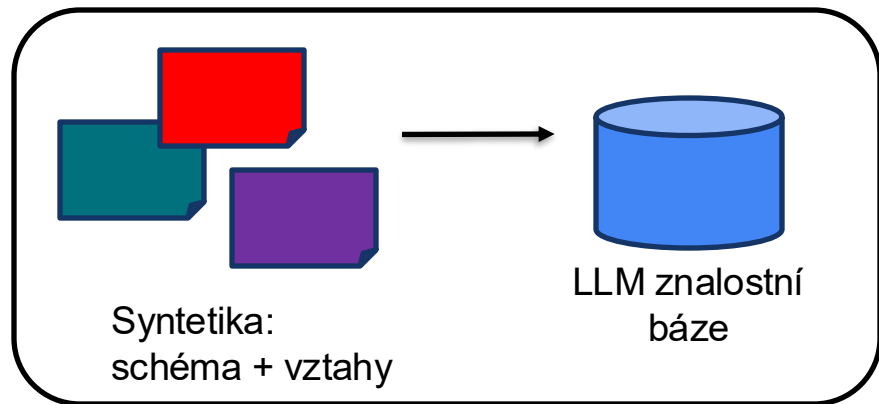
```
messages=[      {"role": "system", "content": "Chovej se  
jako seniorní biostatistik. Vrať pouze SQL kód, nic víc.  
SQL kód musí být kompatibilní s Vertica. Pokud  
nedokážeš SQL kód připravit, přiznej to."},      {  
"role": "user",      "content": context,      },    ]
```

?

Prompt: Jaká byla incidence pacientů s C35 v celé ČR v letech 2010-2012?

pseudonym_nrhzs neexistuje --> AI poukázala na nesoulad mezi DB schema a dokumentací

2. Definuj fenotypy



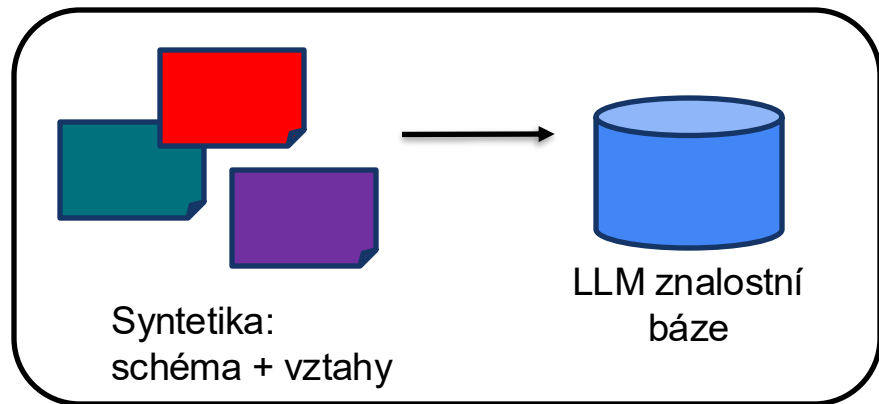
```
messages=[  
    {"role": "system", "content": "Chovej se jako  
seniorní biostatistik. Vrať analytický Python kód, nic víc.  
Pokud nedokážeš SQL kód připravit, přiznej to."},  
]
```

?

Prompt: Definuj fenotyp (cohort definition) pro diagnózu RS.

3. Může nám AI pomoci s omopizací syntetiky?

- OMOP CDM



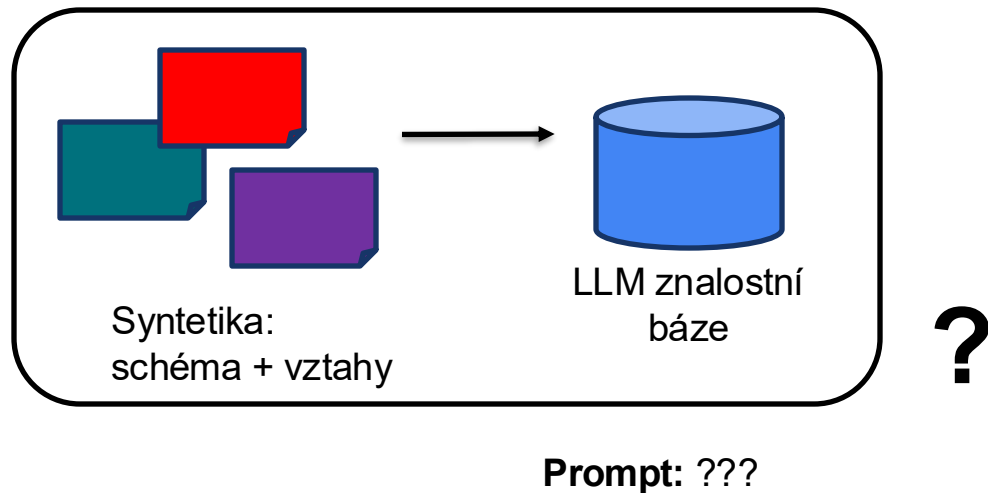
```
model="gpt-oss-120b", messages=[ {"role": "system",  
"content": "Chovej se jako seniorní databázový  
specialista se zaměřením na OMOP CDM."}, {"role":  
"user", "content": context, }, ]
```

?

Prompt: Na základě znalosti databáze (schéma, číselníky, atributy, vztahy), připrav Python skript pro konverzi dané databáze do OMOP CDM.

4. Vylepšuj syntetiku...

Data musí splňovat smysluplné distribuce a být funkční v multidimenzionálním prostoru!



Závěrem

- ☐ Příležitosti
- ☐ Výzvy
- ☐ Mnoho vývoje před námi



+ Gleb, Radka, Honza, Juliana a mnoho dalších!

Konference NZIP & NZIS Open 2025

10.–11. prosince 2025

Grandium Hotel Prague



nzip.cz



Spolufinancováno
Evropskou unií



nzis.open