IBM watsonx.ai

watsonx.ai – Dostupné modely

granite-instruct/chat-V2
13/20 billion params
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG
- Tunning

llama-2*
13/*70 billion params*
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG
- Tunning

mixtral instruct
8x *7 billion params*
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG
- CodeGen

flan-t5-xl-3b
3 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG
- Tunning

flan-t5-xxl-11b
11 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

flan-ul2-20b
20 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

mt0-xxl-13b
13 billion params
encoder/decoder

- Q&A
- Generate
- Summarize
- Classify

mpt-instruct2-7b
7 billion params
decoder only

- Q&A
- Generate

starcoder
15.5 billion params
decoder only

- CodeGen

Other data / modality

Natural Language

Code

watsonx.ai – dostupné LLM models

ARICOMA

granite-instruct/chat-V2 *13/20 billion params* decoder only Tunning	flan-t5-xl-3b *3 billion params* encoder/decoder Tunning	flan-t5-xxl-11b *11 billion params* encoder/decoder	flan-ul2-20b *20 billion params* encoder/decoder	mt0-xxl-13b *13 billion params* encoder/decoder	mpt-instruct2-7b *7 billion params* decoder only	starcoder *15.5 billion params* decoder only CodeGen
llama-3 *8/70 billion params* decoder only	mixtral-8x7b-instruct-v01 8x *7 billion params* decoder only CodeGen	codellama-34b-instruct-hf *3 billion params* encoder/decoder CodeGen	llama-2 *13/70 billion params* decoder only Tunning	merlinite-7b *7 billion params* decoder only		

Tunning Code

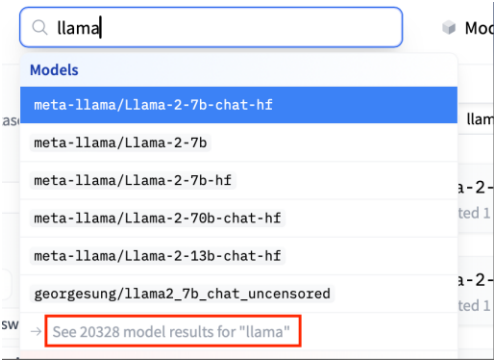
InstructLAB support in watsonx.ai

Čo je InstructLab?

InstructLab je open source projekt na vylepšenie veľkých jazykových modelov (LLM) používaných v aplikáciách generatívnej umelej inteligencie (gen AI). Komunitný projekt InstructLab, ktorý vytvorili IBM a Red Hat, poskytuje nákladovo efektívne riešenie na zlepšenie zosúladenia LLM a otvára dvere tým, ktorí majú minimálne skúsenosti so strojovým učením, aby prispeli.

Potreba trhu

Existuje jasná potreba, aby AI vývojári spolupracovali a aktualizovali existujúce jazykové modely bez vytvárania viacerých klonov



Žiadny spôsob, ako prispieť k jednému modelu

Nie je riadený komunitou

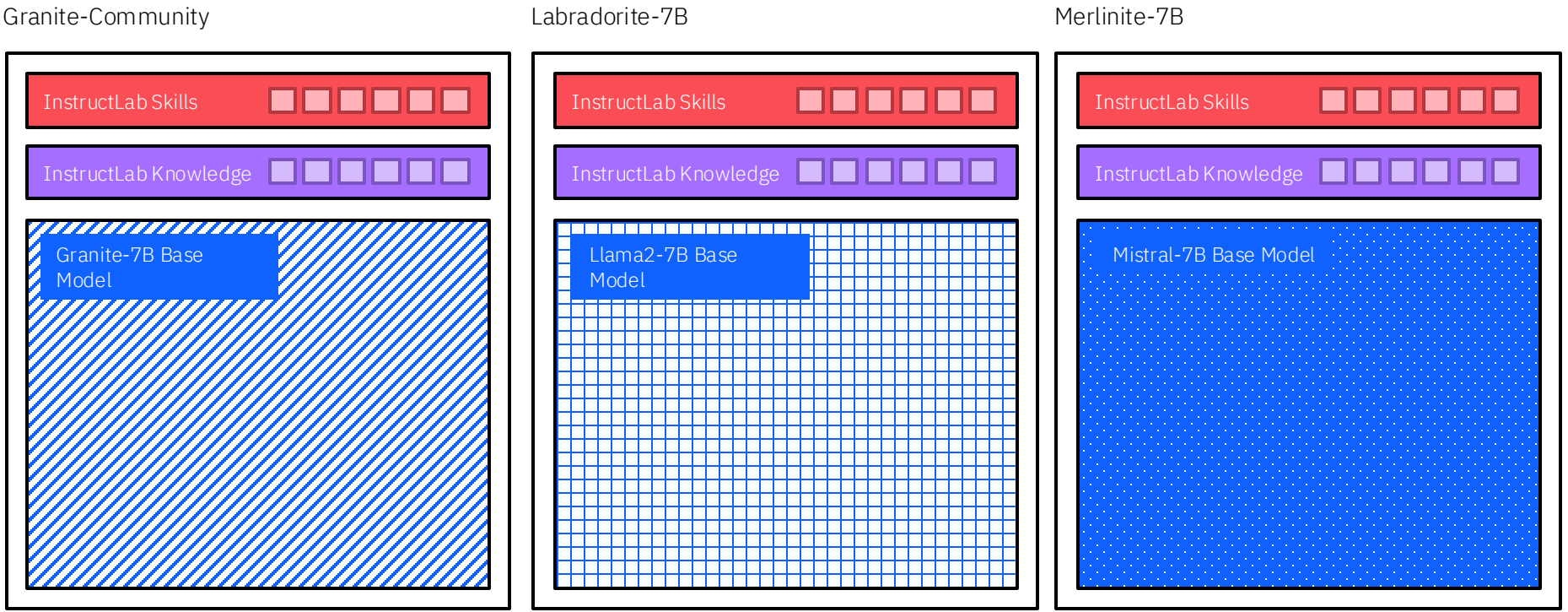
Žiadny spôsob zdieľania vylepšení LLM

Viaceré vidlice modelov

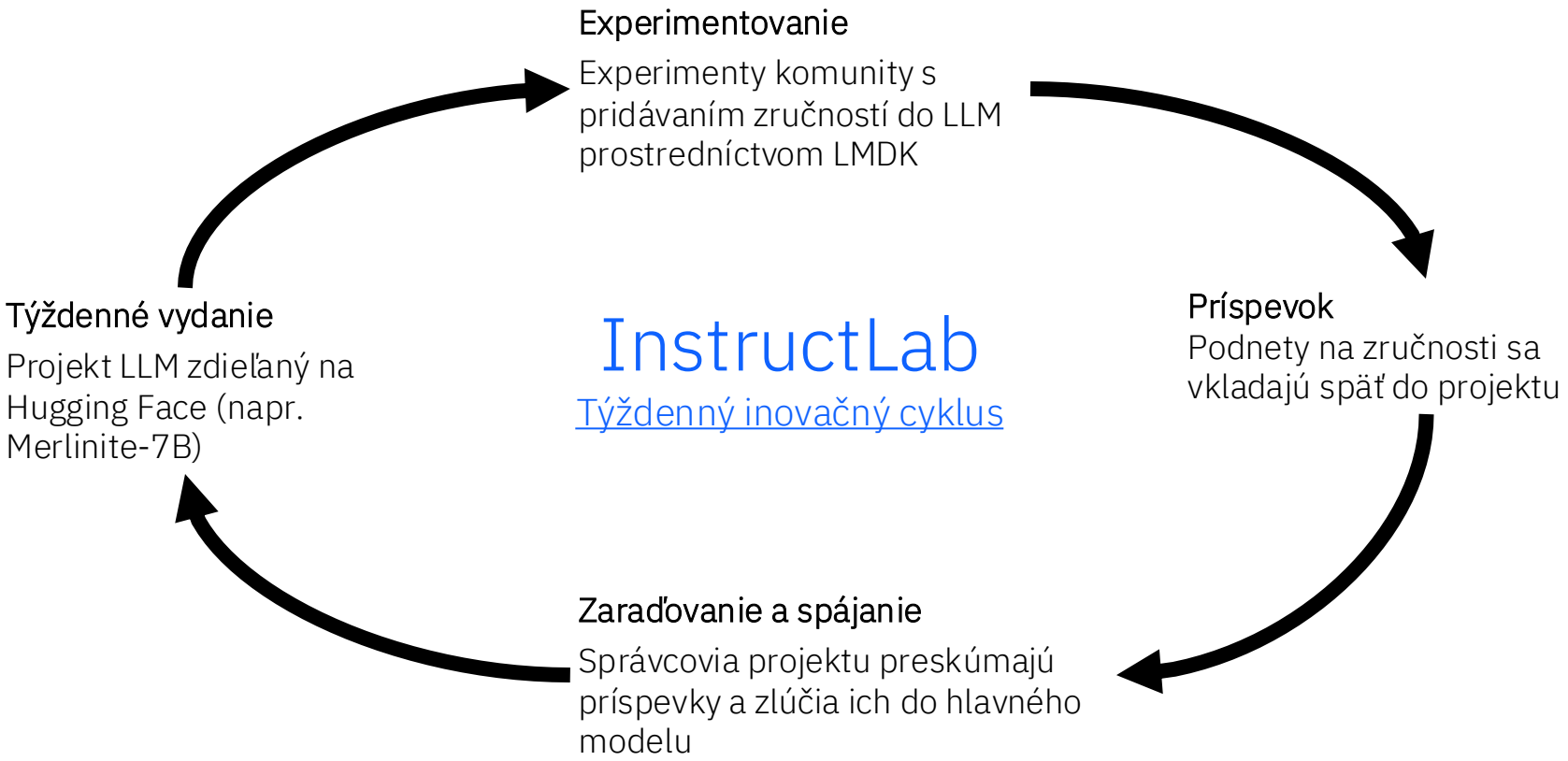
20 000+ modelov Llama na HuggingFace

ARICOMA

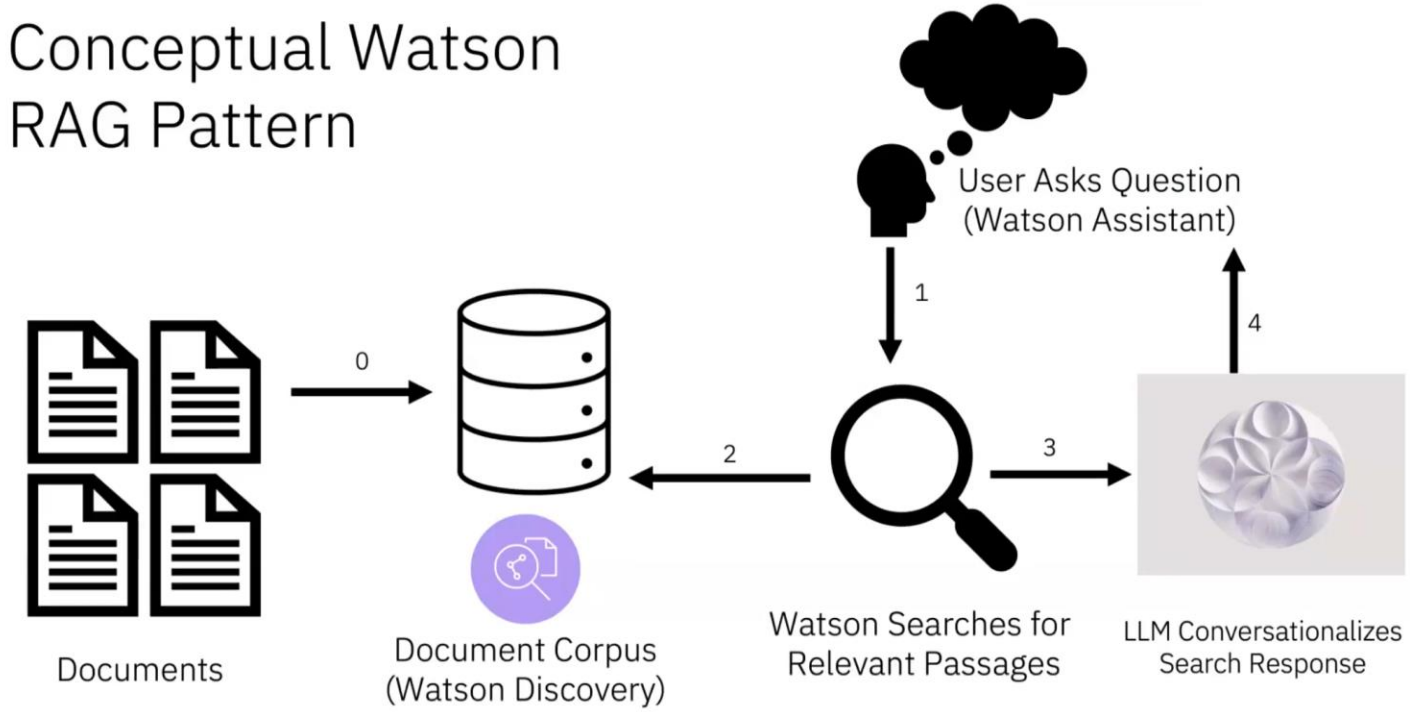
Všetky modely InstructLab budú mať konzistentné, jednotné skúsenosti bez ohľadu na základný model.



InstructLab bude zotrvačníkom rýchlych inovácií s otvoreným zdrojovým kódom.



Konverzačné vyhľadávanie – rozšírený vzor generovania vyhľadávania



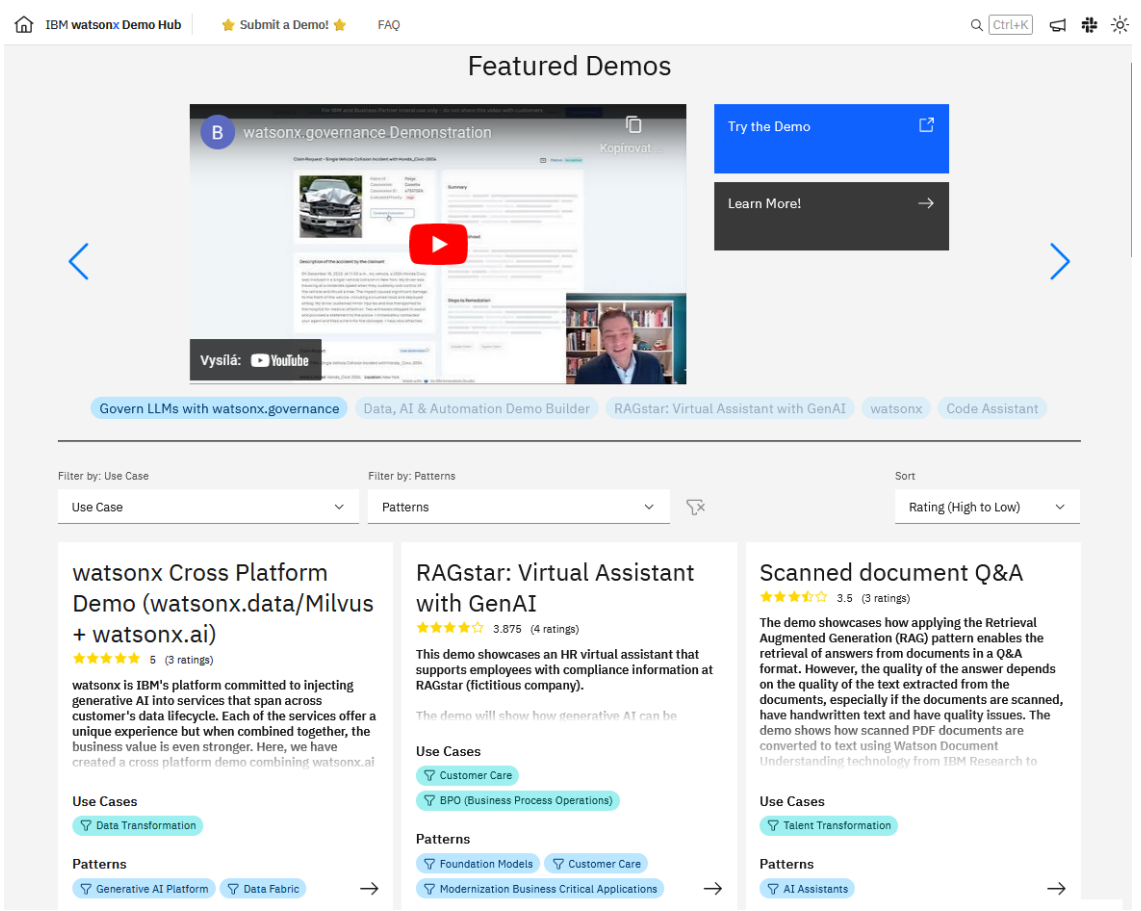
V súčasnosti máme k dispozícii **nadmerné množstvo informácií**, ktoré je pre jednotlivca ťažké spracovať a pochopiť.

Zákazníci a zamestnanci môžu **sami vykonávať rozsiahle informačné a transakčné úlohy** prostredníctvom interakcie s dôveryhodným a personalizovaným asistentom Gen AI.

watsonx Demo Hub

https://watsonx.demohub.techzone.ibm.com/

- obsahuje množstvo profesionálnych a klientských ukážok pokrývajúcich portfólio watsonx.
- v súlade s prípadmi použitia a technologickými vzormi, watsonx Demo Hub obsahuje všetko, čo potrebujete, aby ste boli informovaní



Ďakujem!