



**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
- CodeGen

**starcoder**  
15.5 billion params  
decoder only

Other data /  
modality

Natural  
Language

Code

# watsonx.ai – dostupné LLM models

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



Tuning

Code

InstructLAB  
support in  
watsonx.ai

ARICOMA



IBM

# 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

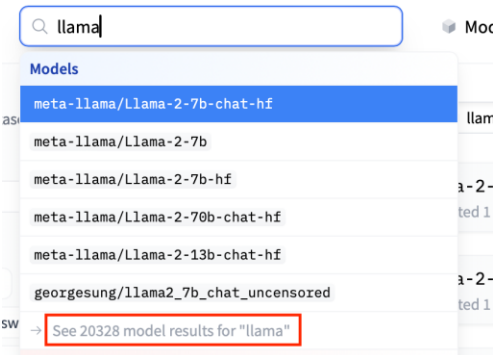
Žiadny spôsob zdieľania vylepšení LLM

# ARICOMA



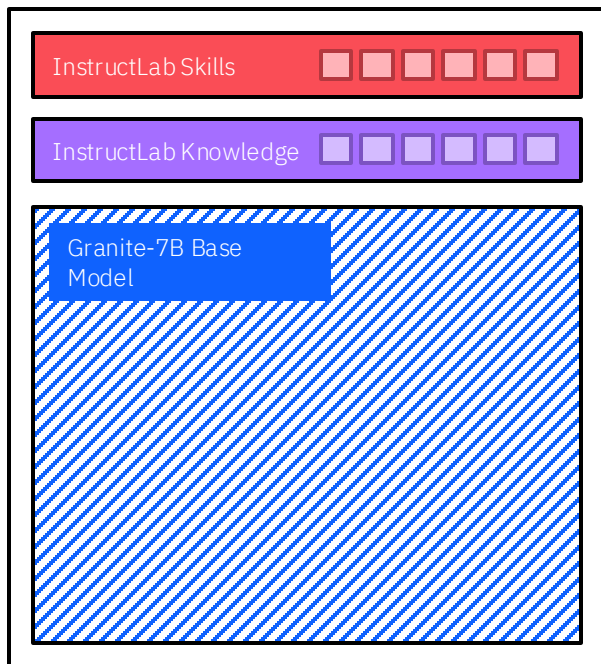
Viaceré vidlice modelov

20 000+ modelov Llama na HuggingFace

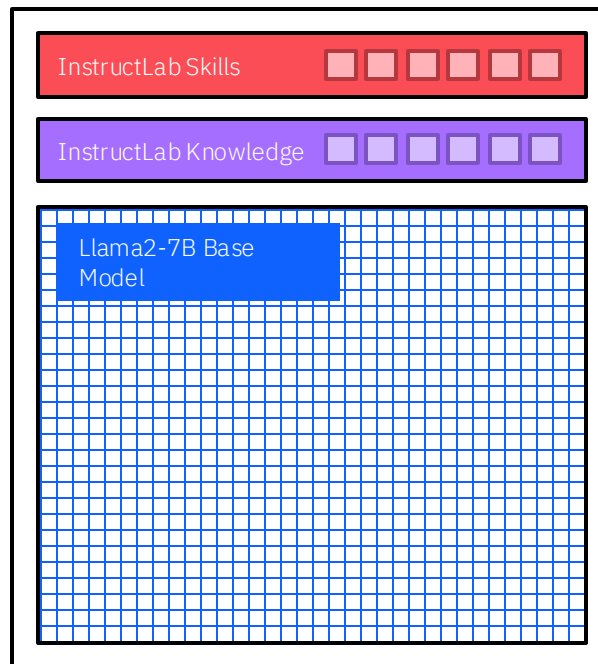


Všetky modely InstructLab budou mať konzistentné, jednotné skúsenosti bez ohľadu na základný model.

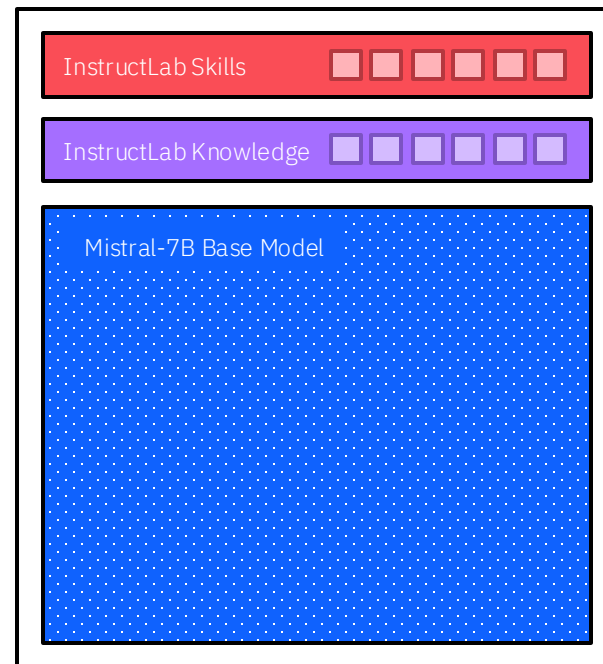
Granite-Community



Labradorite-7B



Merlinitite-7B



InstructLab bude zotrvačníkom rýchlych inovácií s otvoreným zdrojovým kódom.

### Experimentovanie

Experimenty komunity s pridávaním zručností do LLM prostredníctvom LMDK

### Príspevok

Podnety na zručnosti sa vkladajú späť do projektu

### Zaradovanie a spájanie

Správcovia projektu preskúmajú príspevky a zlúčia ich do hlavného modelu

### Týždenné vydanie

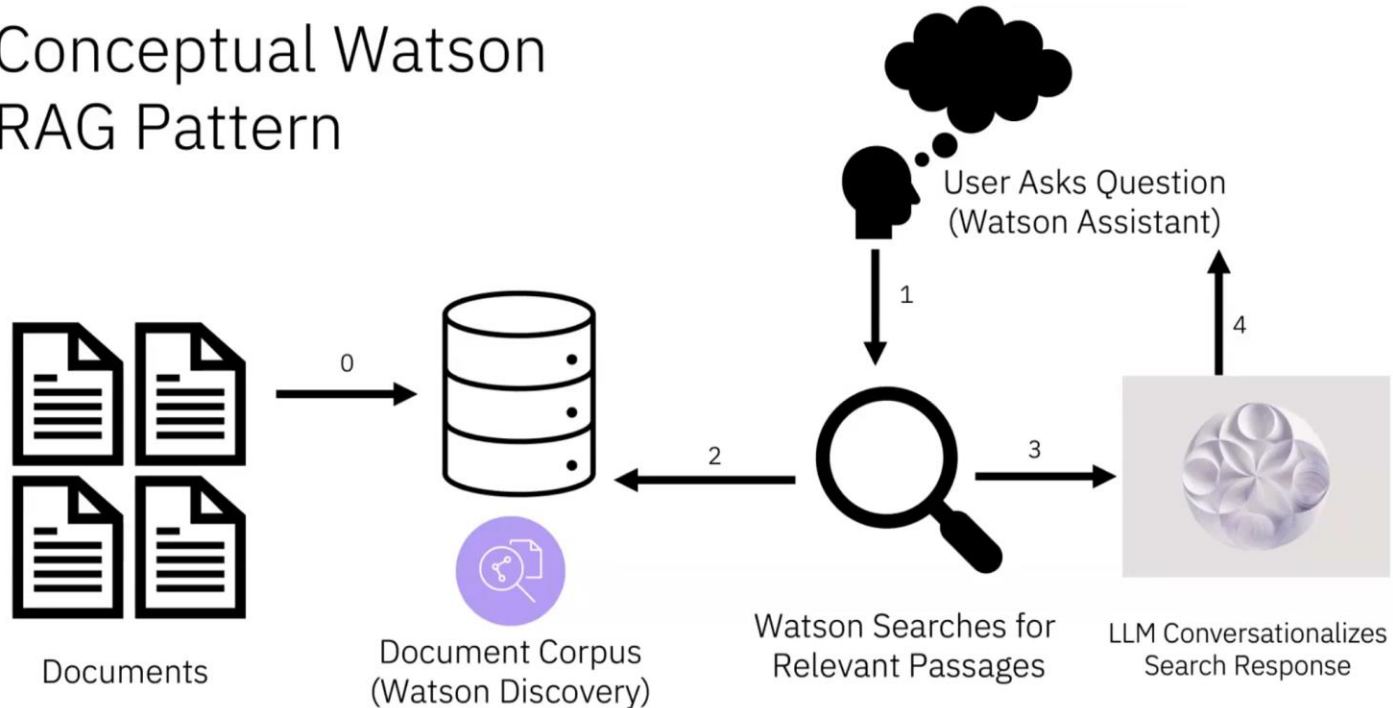
Projekt LLM zdieľaný na Hugging Face (napr. Merlinite-7B)

**InstructLab**  
[Týždenný inovačný cyklus](#)



# Konverzačné vyhľadávanie – rozšírený vzor generovania vyhľadávania

Conceptual Watson  
RAG Pattern



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

# ARICOMA

The screenshot shows the IBM Watsonx Demo Hub interface. At the top, there is a navigation bar with a home icon, 'IBM watsonx Demo Hub', a 'Submit a Demo!' button, and an 'FAQ' link. On the right, there are search and utility icons. The main content area is titled 'Featured Demos' and features a large video player for 'watsonx.governance Demonstration'. To the right of the video are two buttons: 'Try the Demo' and 'Learn More!'. Below the video, there are several category tags: 'Govern LLMs with watsonx.governance', 'Data, AI & Automation Demo Builder', 'RAGstar: Virtual Assistant with GenAI', 'watsonx', and 'Code Assistant'. Below this, there are filters for 'Use Case' and 'Patterns', and a 'Sort' dropdown set to 'Rating (High to Low)'. The main content area displays three demo cards:

- watsonx Cross Platform Demo (watsonx.data/Milvus + watsonx.ai)**  
5 (3 ratings)  
watsonx is IBM's platform committed to injecting generative AI into services that span across customer's data lifecycle. Each of the services offer a unique experience but when combined together, the business value is even stronger. Here, we have created a cross platform demo combining watsonx.ai
- RAGstar: Virtual Assistant with GenAI**  
3.875 (4 ratings)  
This demo showcases an HR virtual assistant that supports employees with compliance information at RAGstar (fictitious company).  
The demo will show how generative AI can be
- Scanned document Q&A**  
3.5 (3 ratings)  
The demo showcases how applying the Retrieval Augmented Generation (RAG) pattern enables the retrieval of answers from documents in a Q&A format. However, the quality of the answer depends on the quality of the text extracted from the documents, especially if the documents are scanned, have handwritten text and have quality issues. The demo shows how scanned PDF documents are converted to text using Watson Document Understanding technology from IBM Research to

Ďakujem!

ARICOMA

