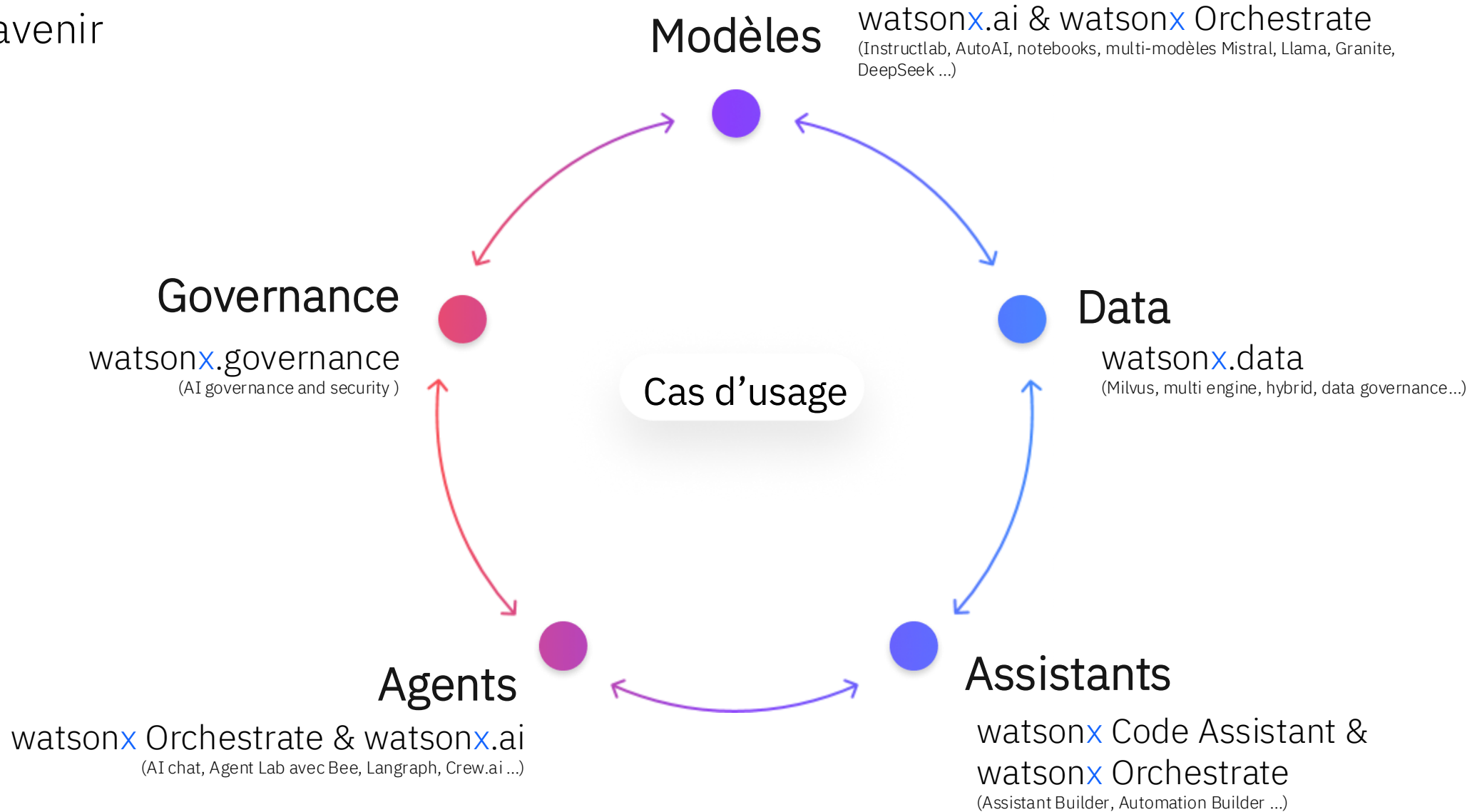


InstructLab :

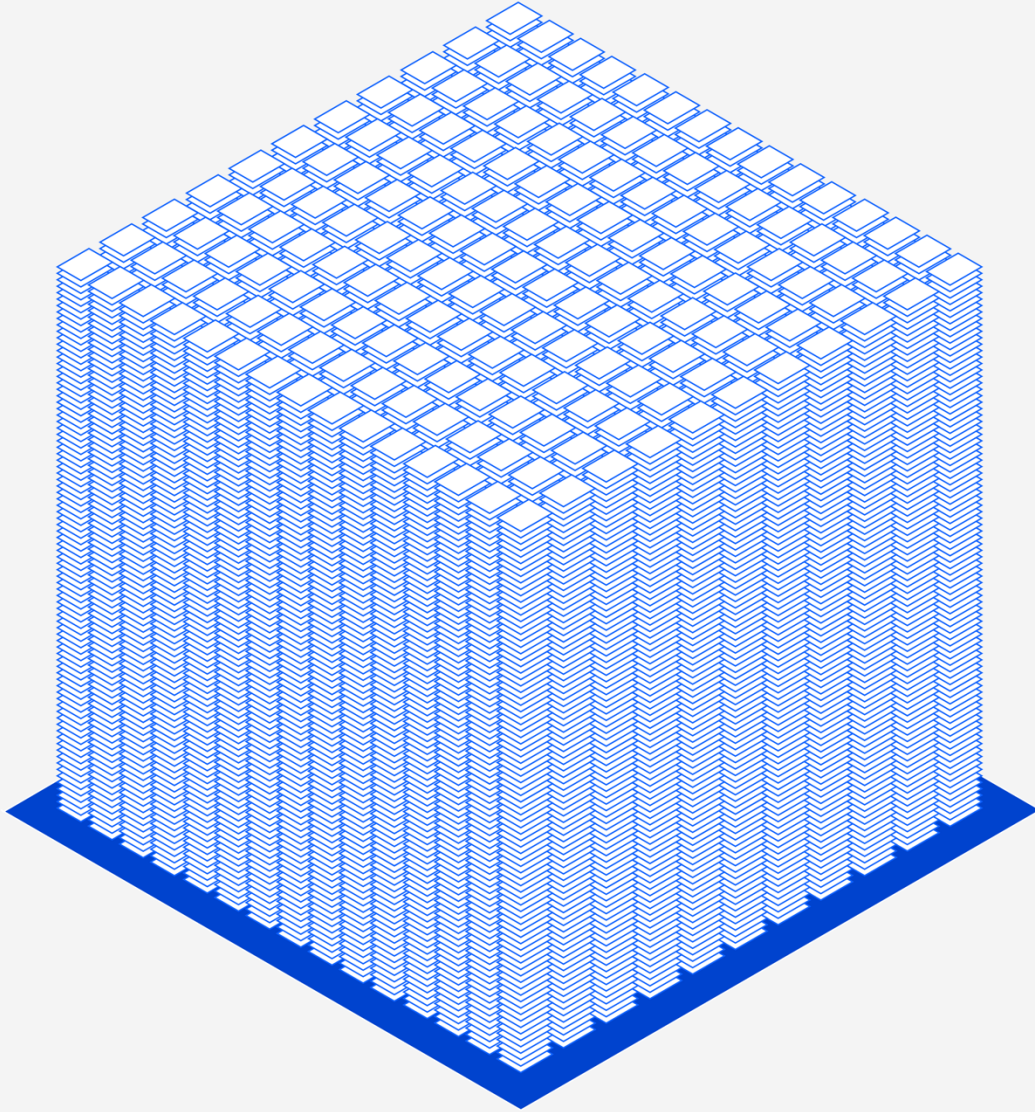
Finetuning collaborative des modèles de foundation



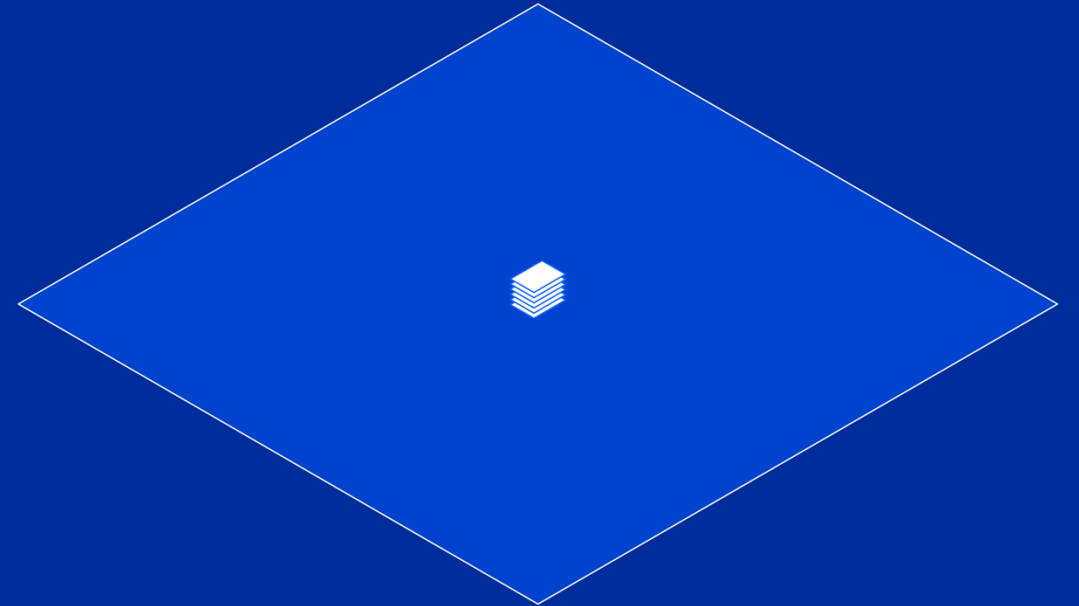
Les éléments constitutifs de l'IA pour l'avenir



Presque toutes les données publiques disponibles sont représentées dans des modèles de fondation



Une très petite quantité de toutes les données d'entreprise est représentée dans les modèles de base



Comment augmenter les modèles avec des données d'entreprise ?

01

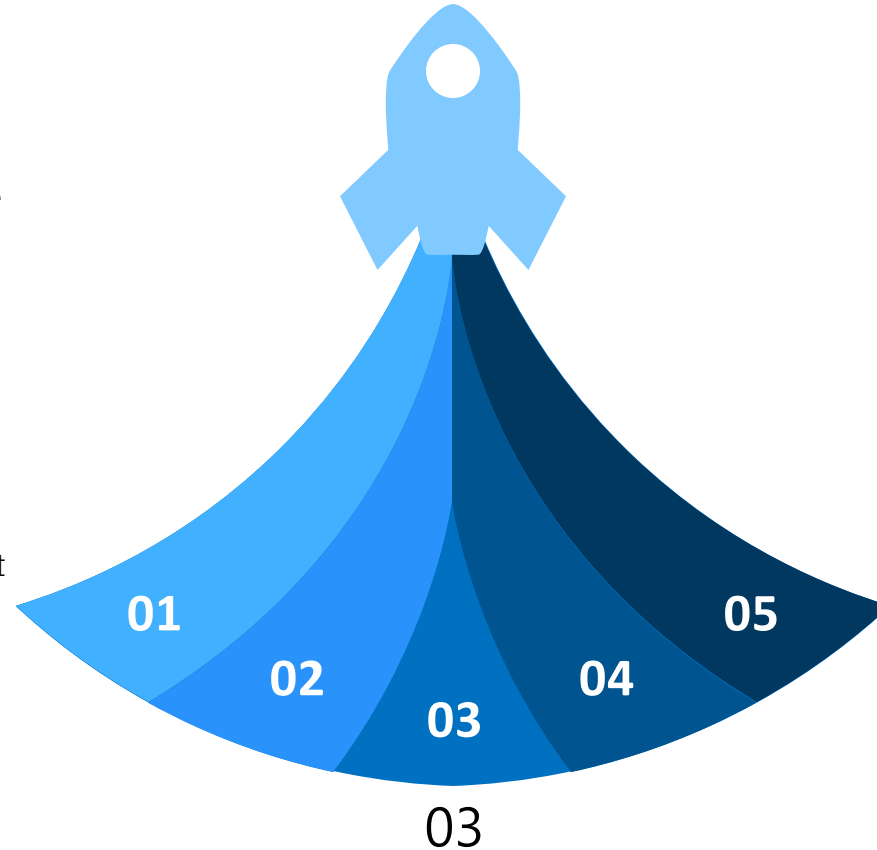
RAG

Adapté aux données modifiées de manière dynamique. Données d'entreprise non représentées dans le modèle. N'améliore pas le modèle

02

Prompt Tuning

Le plus adapté aux tâches d'adaptation rapide, en particulier lorsque les ressources de calcul sont limitées ou que la tâche est étroitement liée au modèle pré-entraîné



PEFT (LoRA, QLoRA)

Idéal pour les environnements aux ressources limitées ou lorsque vous avez besoin d'affiner efficacement plusieurs tâches.

05



InstructLab

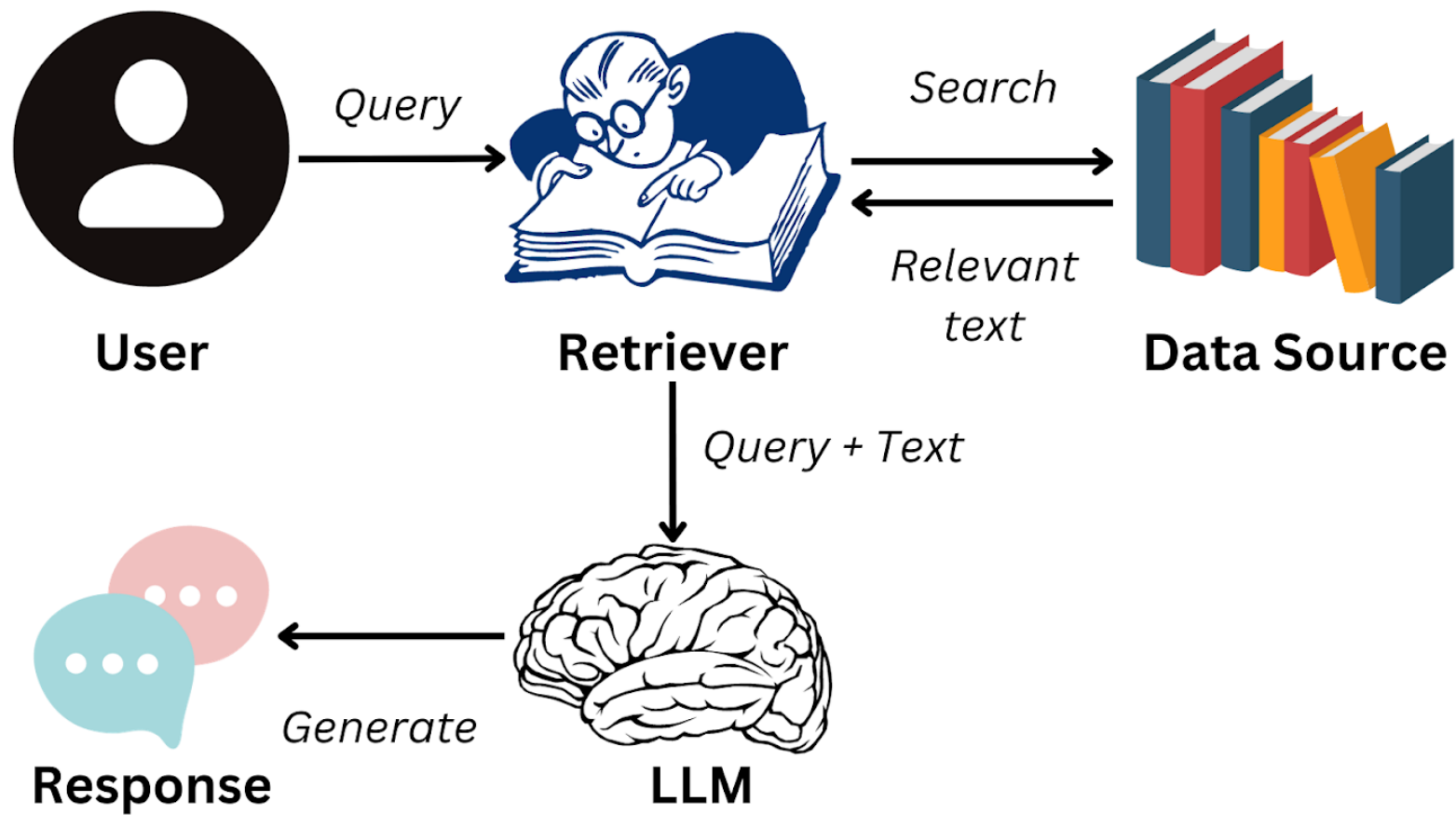
Ne nécessite pas d'expertise technique et de préparation des données. Fournit une plateforme collaborative, la génération de données synthétiques et une meilleure précision des modèles

04

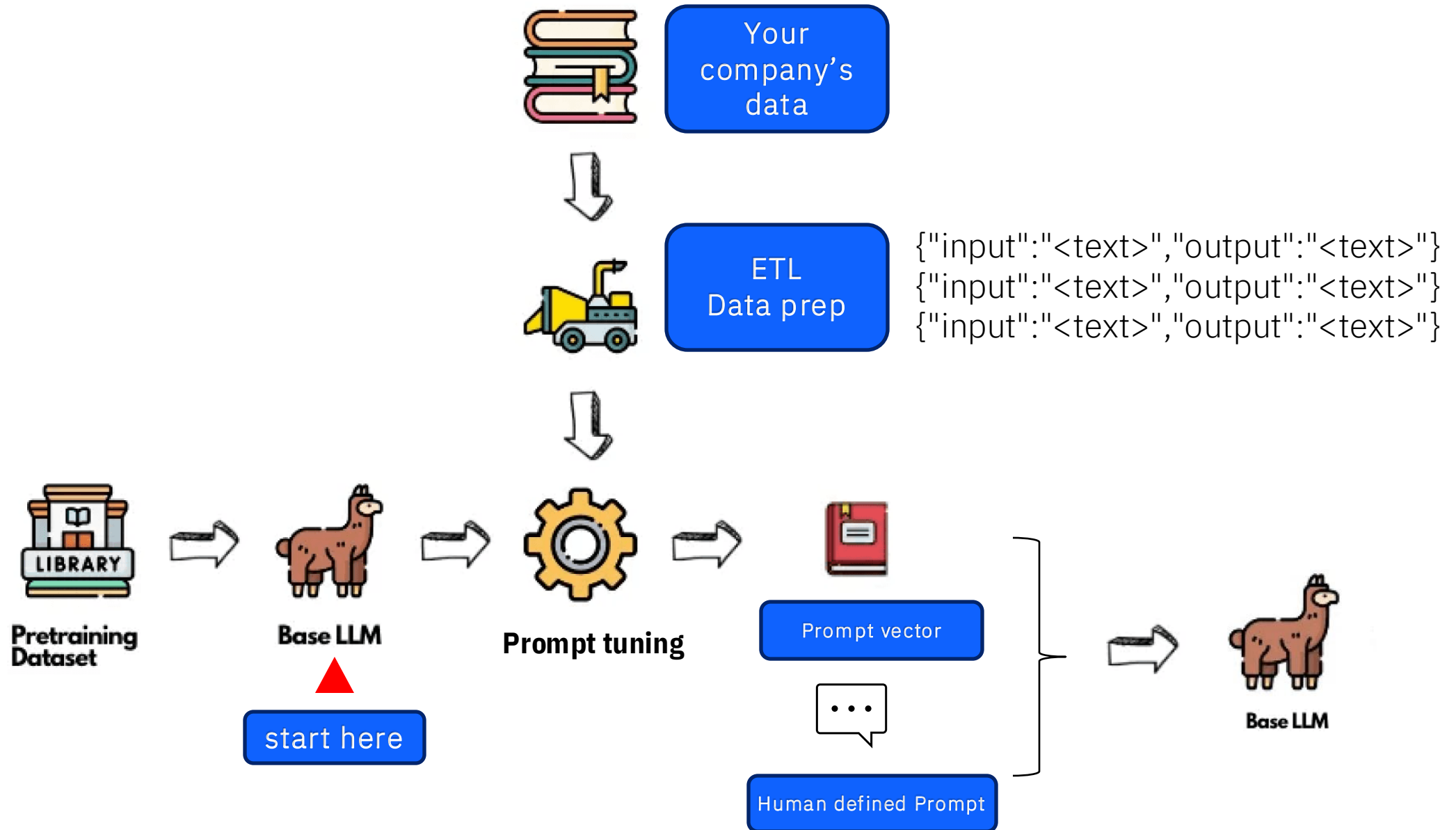
Full fine-tuning

Idéal pour les scénarios où une précision maximale et une adaptation spécifique à la tâche sont essentielles et où les ressources sont abondantes

RAG :



Prompt tuning :



watsonx.ai: Tuning Studio

Ajustez vos LLMs avec des données étiquetées

Résumé :

- Inclut des méthodes de finetuning qui permettent d'optimiser les performances des LLMs
- Le modèle finetuné peut être déployé et inféré via l'API ou Prompt Lab

Plusieurs méthodes disponibles :

- **Prompt Tuning :**
Une méthode de spécialisation sans modifications des poids des modèles.
- **InstructLab :**
Une approche communautaire qui permet à plusieurs personnes au sein de l'entreprise de contribuer des connaissances et des capacités à un modèle, qui sera ré-entraîné après une étape de génération de données synthétique.
- **Full Fine Tuning :**
Mettre à jour tous les poids du modèle pour répondre aux tâches les plus complexes.

The screenshot displays the IBM watsonx Tuning Studio interface for configuring a tuned model. The top navigation bar includes the IBM watsonx logo, an 'Upgrade' button, a notification bell, the user's account 'Eric Saleh's Account', the location 'Dallas', and a profile icon 'ES'. The breadcrumb trail shows 'Projects / Test pl / Demo Tune'.

The main heading is 'Configure tuned model' with the subtitle 'Demo Tuning Experiment'. A timestamp indicates 'Last saved: November 16, 2023 at 4:52:49 PM'.

The interface is divided into two main sections: 'Configure details' and 'Add training data'.

Configure details:

- Which foundation model do you want to prompt tune?** A dropdown menu shows 'flan-t5-xl-3b'.
- How do you want to initialize your prompt?** Two options are shown: 'Text' (with a description: 'Provide instructions for how to define and format the output.') and 'Random' (selected, with a description: 'Let the experiment set the prompt.').
- Which task fits your goal?** Three task cards are displayed: 'Classification' (selected, with a description: 'Classify text with up to 10 labels that you specify.'), 'Generation' (with a description: 'Generate text in the same format as your training data.'), and 'Summarization' (with a description: 'Summarize text in the same format as your training data.'). Below these, there is a section for 'Classification output (verbalizer)' with a text input field 'Enter classification variables' and a blue '+' button. At the bottom, there are two buttons: 'Positive' and 'Negative', each with an 'x' icon.

Add training data:

- A file named 'file_to_tune.jsonl' is listed with a size of '1.56 KB'.

What should your data look like?

- A lightbulb icon and a note: 'Your data must conform to the templates. Input and output fields are clipped after the specified maximum number of tokens.'
- A 'Preview template' button.
- Two sliders for token limits: 'Maximum input tokens' (set to 256) and 'Maximum output tokens' (set to 128).

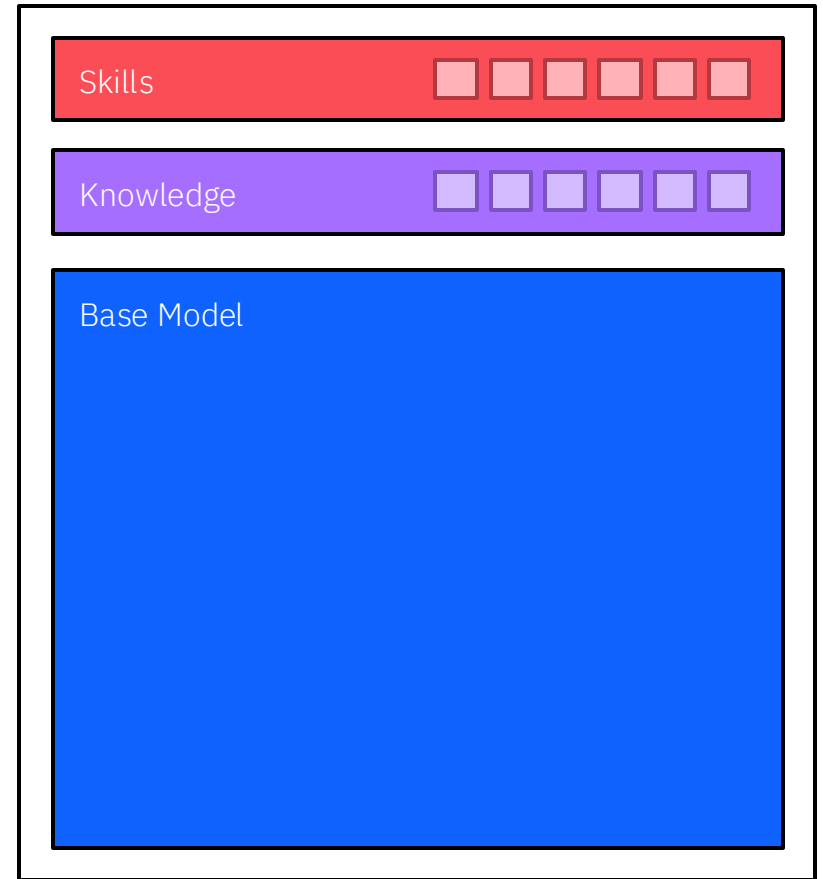
The bottom of the interface features a 'Configure parameters' button and a large blue 'Start tuning' button.

InstructLab : Augmenter les connaissances et améliorer les capacités des modèles

InstructLab: Moderniser la personnalisation des modèles

Faire en sorte que les LLM apprennent
comme nous :

avec des connaissances et des
compétences



InstructLab utilise une nouvelle technique d'alignement d'IBM Research appelée LAB (Large-scale Alignment for chatBots)

1

Curation des données basée sur la taxonomie

2

Génération d'une large base de donnée

3

Alignement du modèle

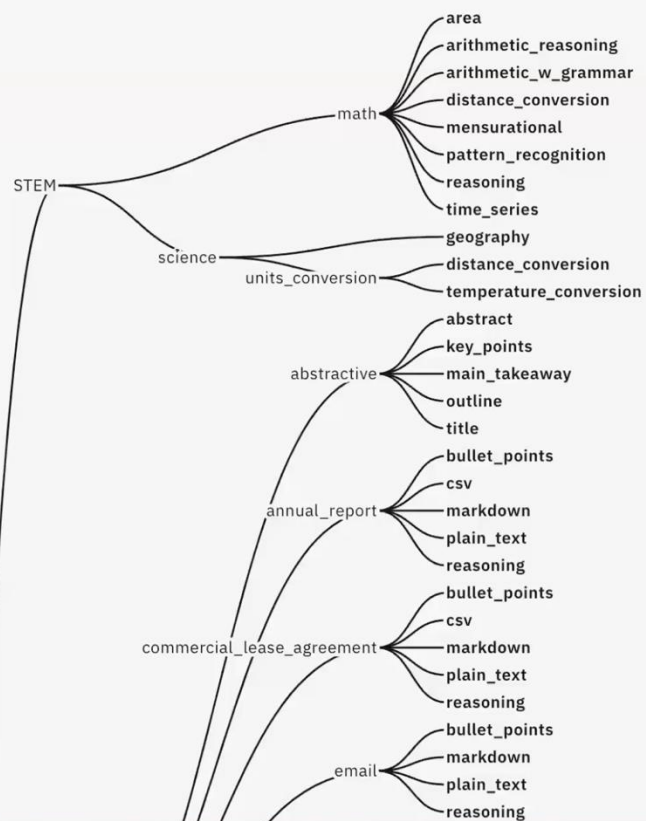
granite-20b-multilingual

Provider: IBM | Version: 1.1 | Type: Provided model

Question answering Summarization Retrieval-Augmented Generation(...) Classification Generation Extraction Translation

Details

Training taxonomy



Explore the taxonomy for this model

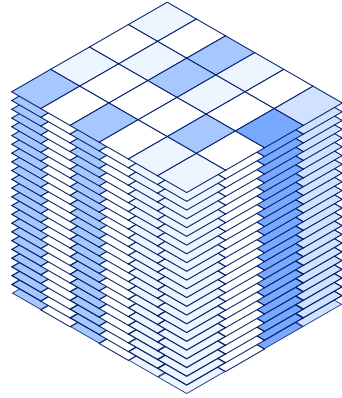
The training taxonomy shows the public view of how the model was trained with InstructLab. Click an end node to view details about knowledge or a skill added to the model. [Learn more](#)

InstructLab (Large-scale Alignment for chatBots)



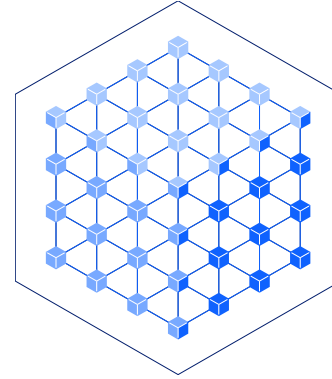
Taxonomie

Des sources de connaissances de haute qualité, organisées à la main, ainsi qu'une taxonomie des tâches avec des exemples générés par l'homme pour chacune.



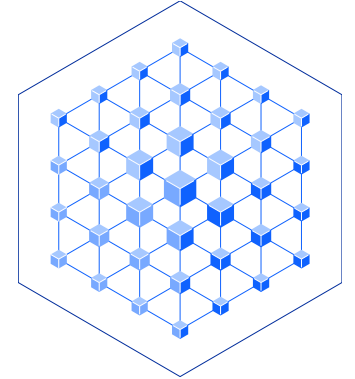
Teacher model(s)

Un modèle enseignant génère une base de millions de questions et de réponses à partir de la taxonomie donnée.



Critic model(s)

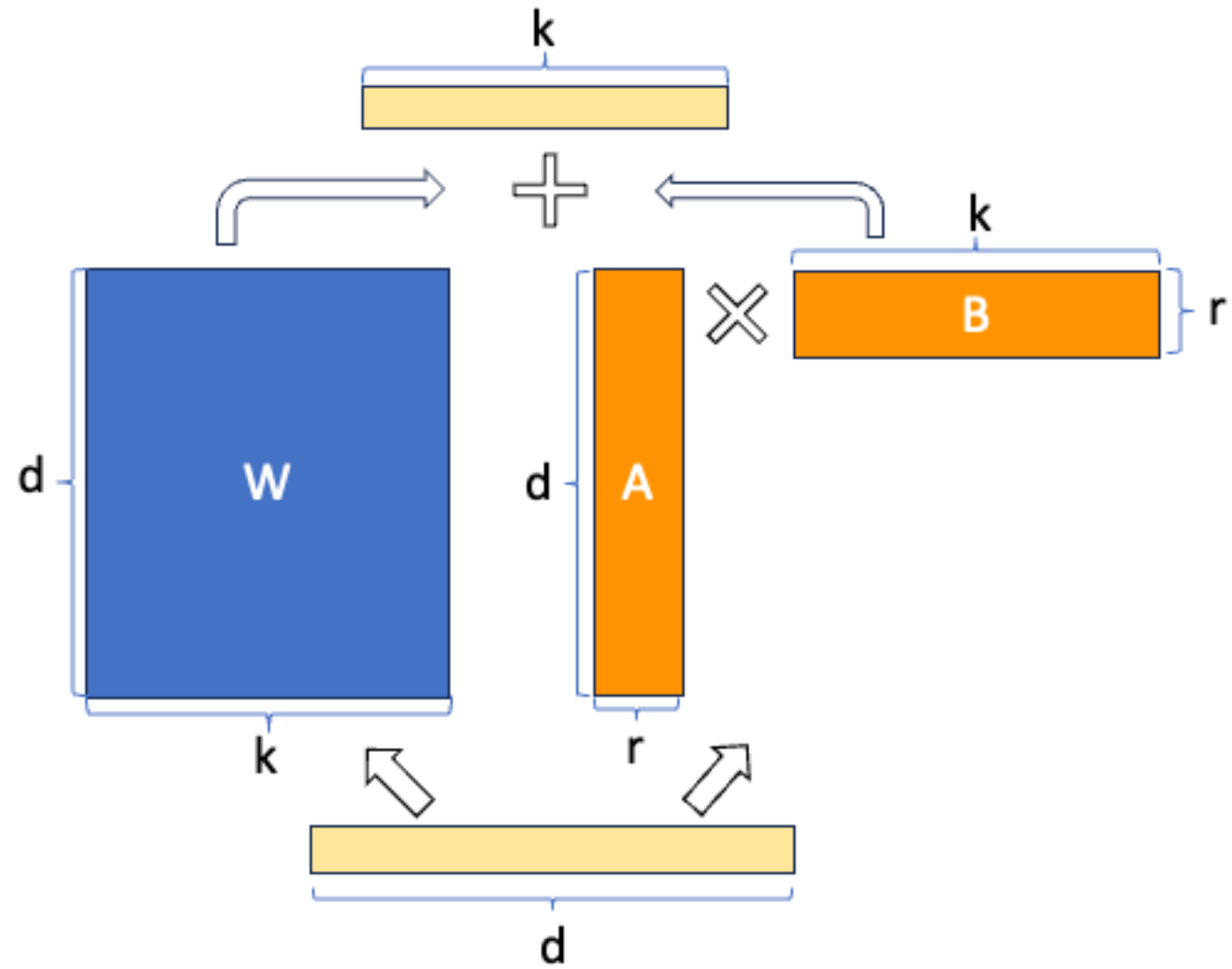
Un deuxième modèle va filtrer les questions pour en vérifier l'exactitude et la qualité. Les données synthétiques sont analysées à la recherche de matériel interdit.



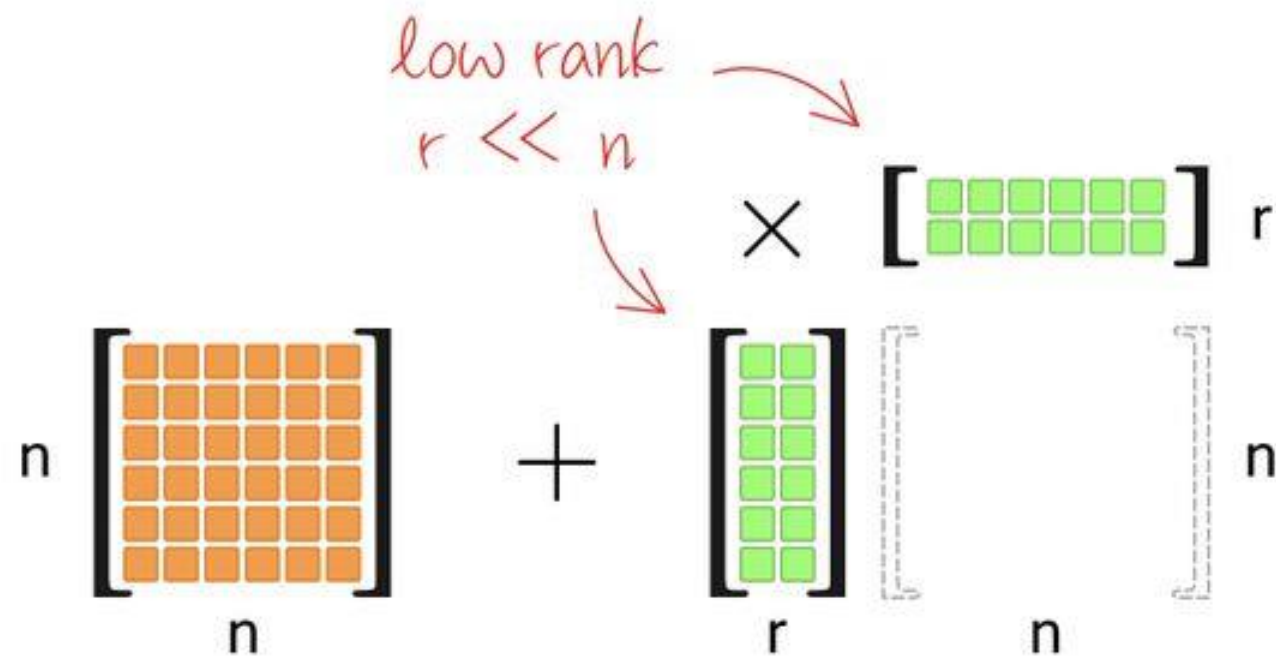
Student model(s)

The student model is trained with the curriculum using a novel training approach.

InstrctLab : LoRA



InstrctLab : LoRA



$$W = W_0 + \underline{A} \underline{B}$$

frozen \uparrow \uparrow trainable



The InstructLab Process

Enrich your training dataset

Thousands of new training examples get created

InstructLab

Recurring innovation cycle

Retraining the same model
Your model get retrained on the new training dataset

Deploy and inference

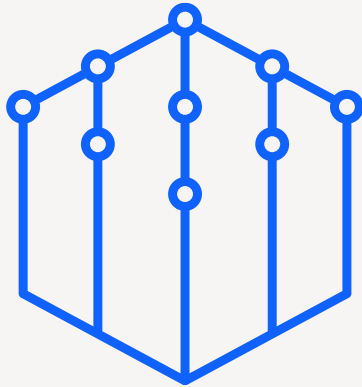
Deploy your newly trained model

Add new Knowledge and Skills

Update your taxonomy with knowledge and skills for your use cases

The IBM approach: fit-for-purpose models

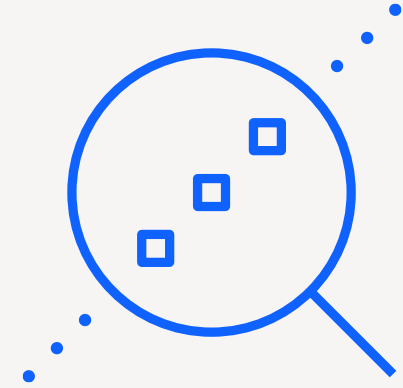
98% cost saving
with IBM Granite vs GPT-4



Vos données



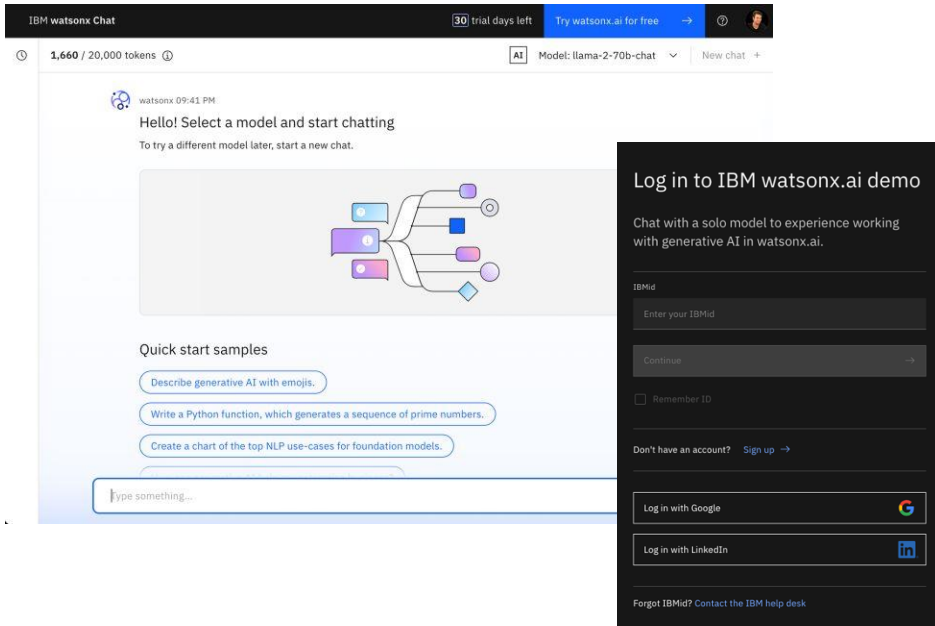
Le bon modèle



Cas d'usage ciblé
Fine tuning

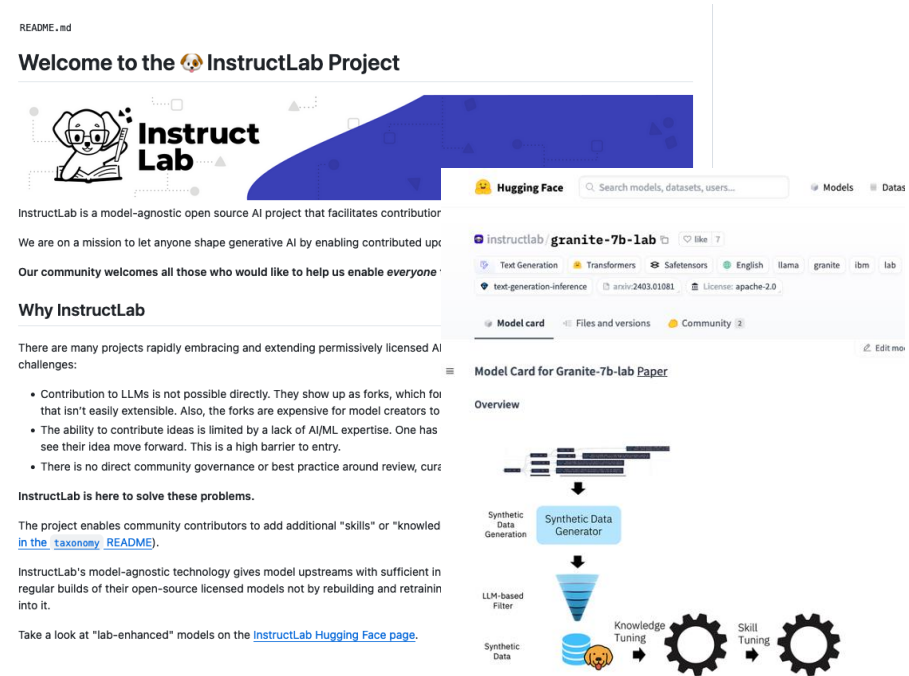
Essayez les modèles par vous-même

watsonx.ai interactive demo



<https://dataplatform.cloud.ibm.com/chat/>

InstructLab project



<https://github.com/instructlab>