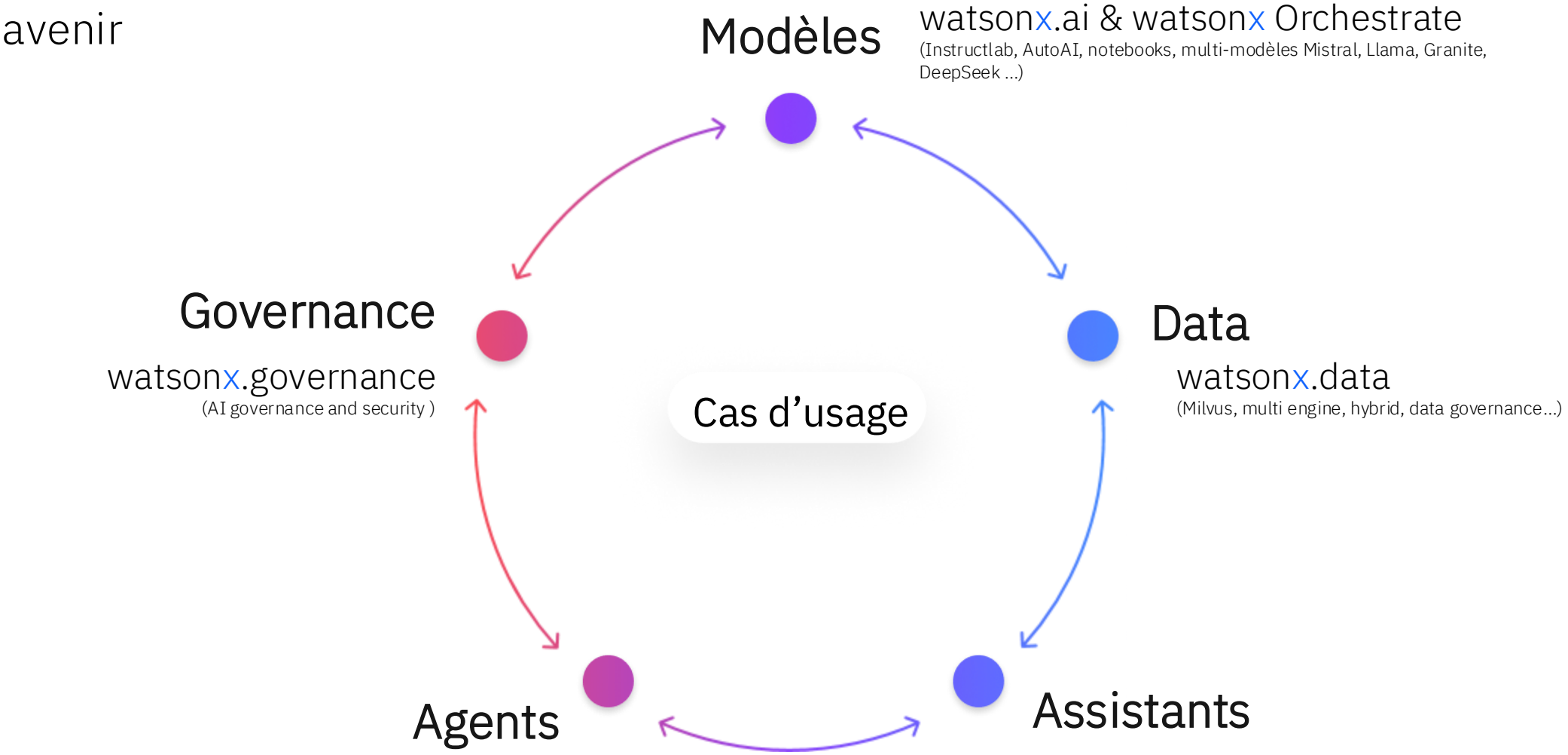
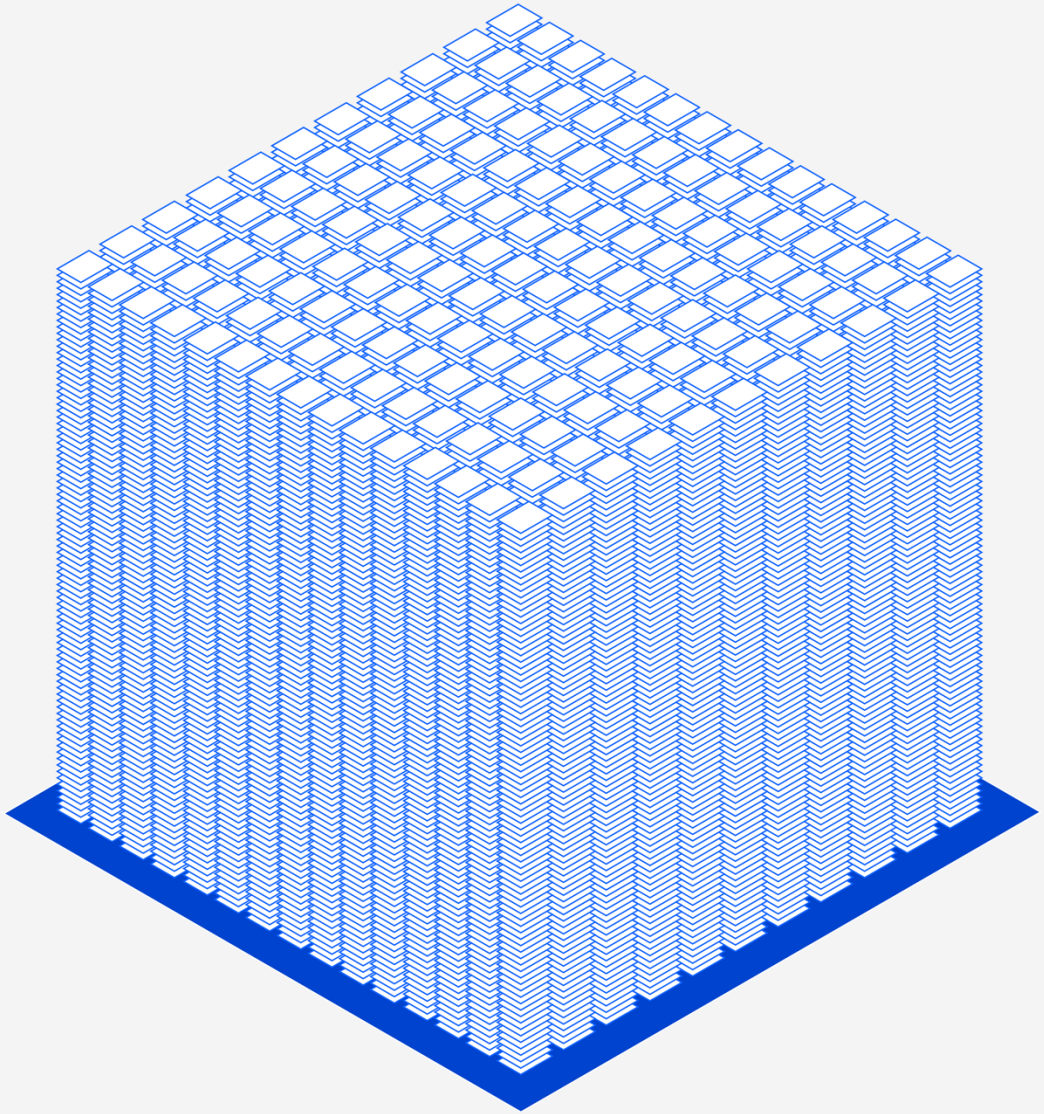
InstructLab : Finetuning collaborative des modèles de foundation

M’hamed Benabid
Data & AI Technical Specialist

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

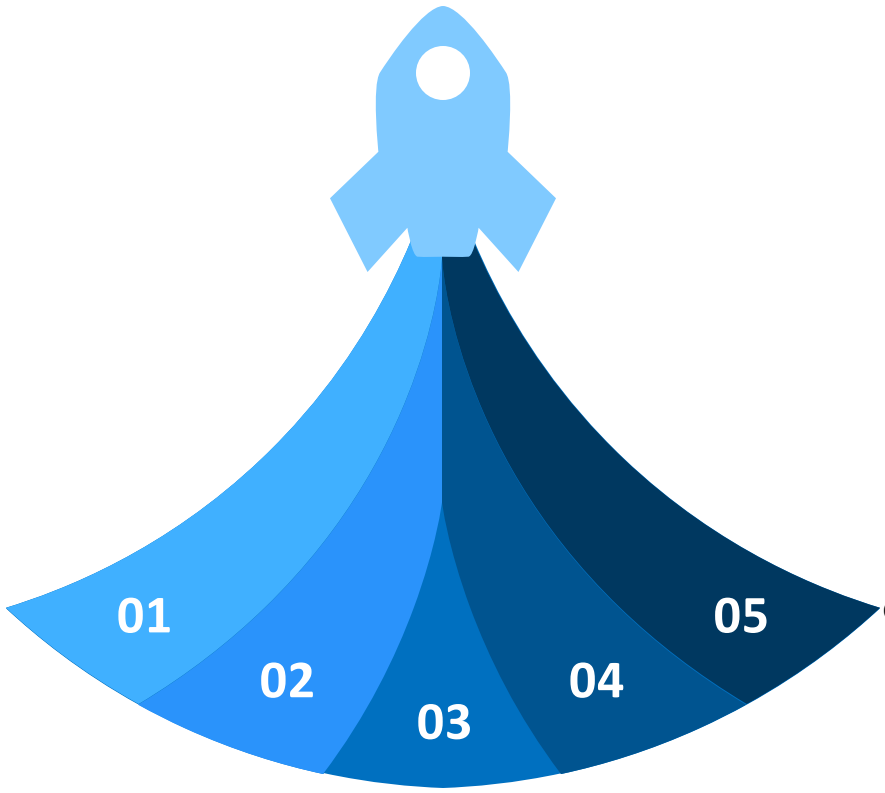
RAG

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



03

PEFT (LoRA, QLoRA)

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

05

★ InstructLab

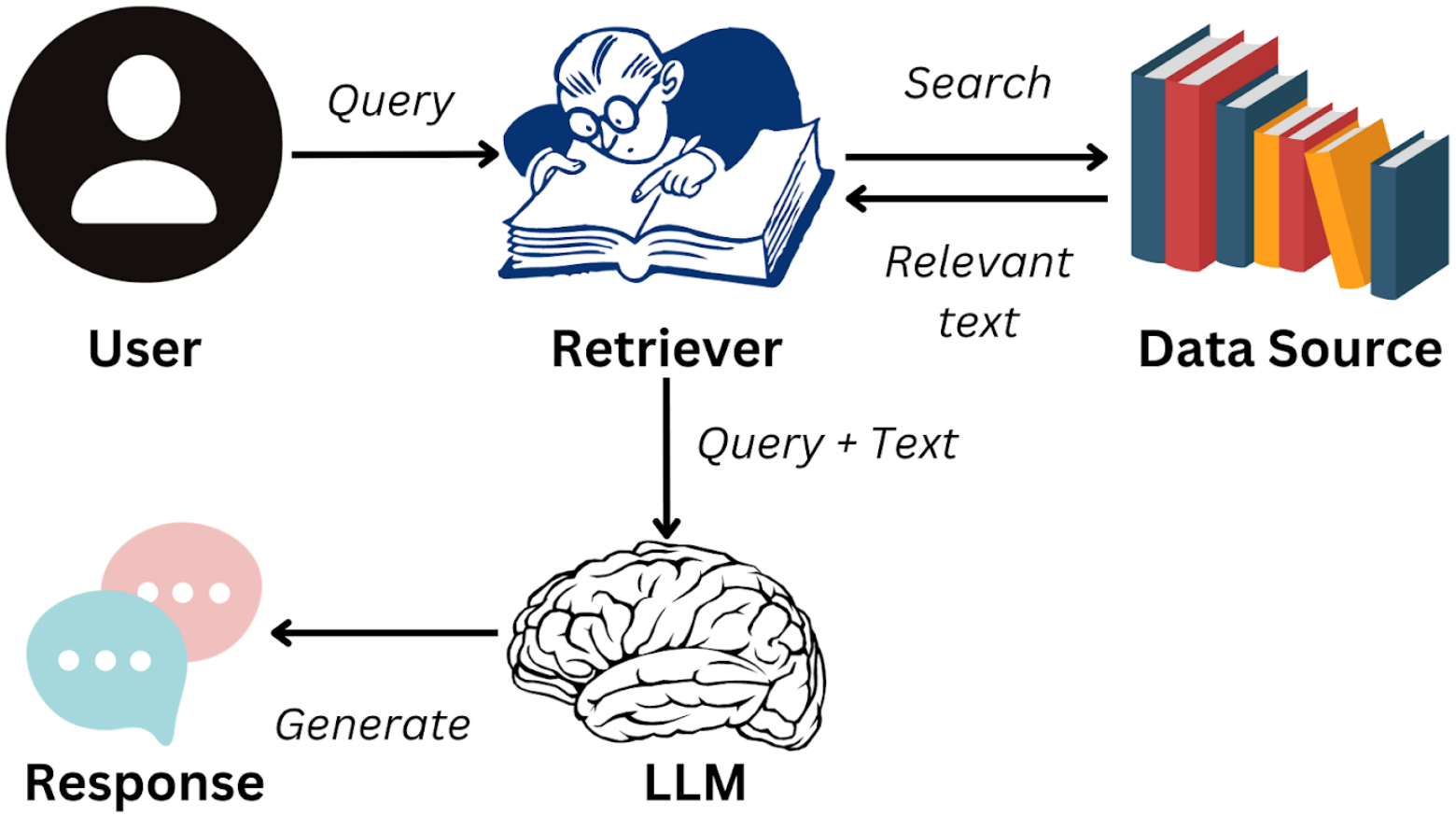
Ne nécessite pas d'expertise technique et de préparation des données. Fournit une plateforme collaborative, la génération de données synthétiques et une meilleure précision des modèles

04

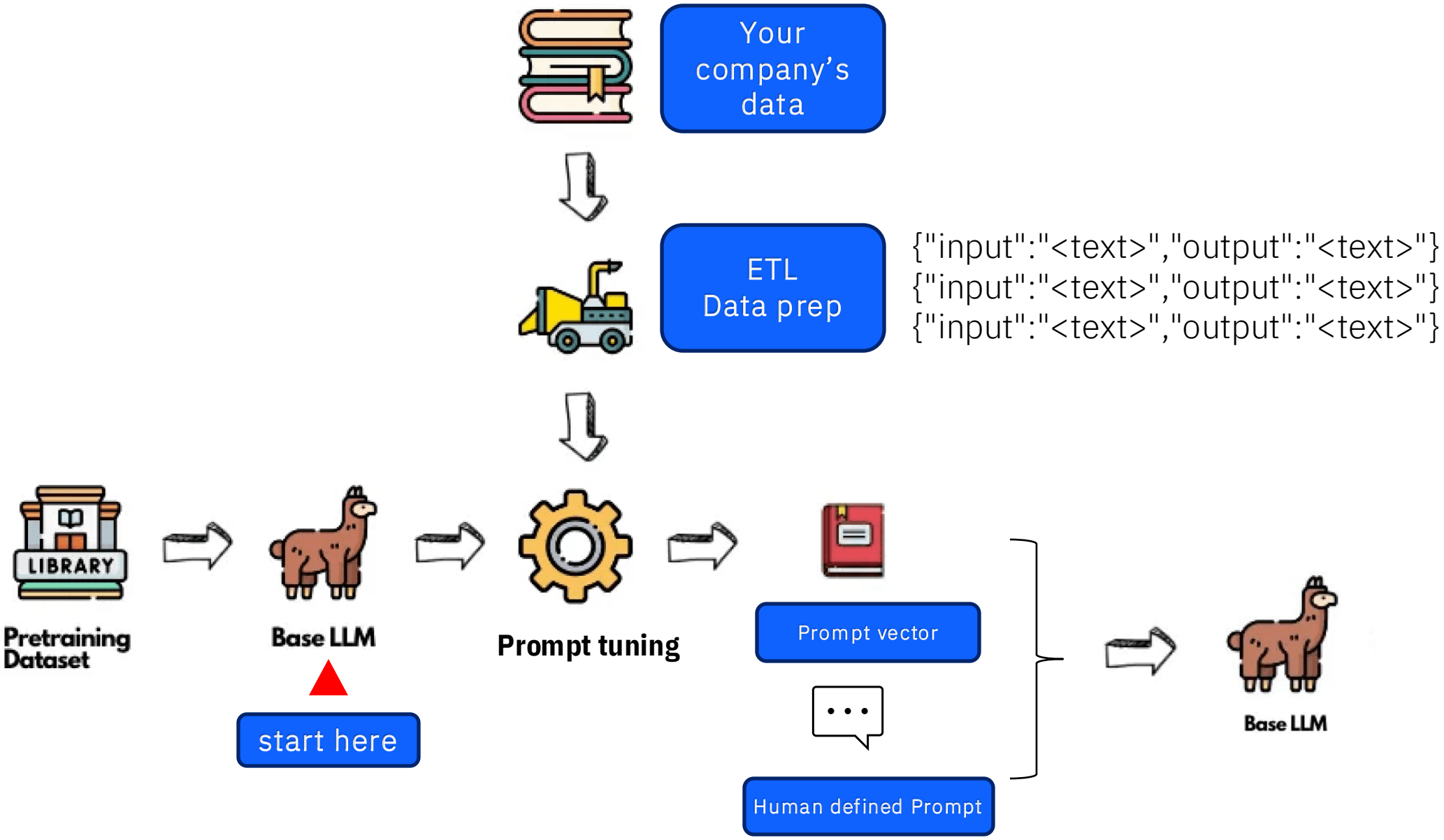
Full fine-tuning

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

RAG :


Query
Search
Relevant text
User
Retriever
Data Source
Query + Text
Generate
Response
LLM

Prompt tuning :



watsonx.ai: Tuning Studio

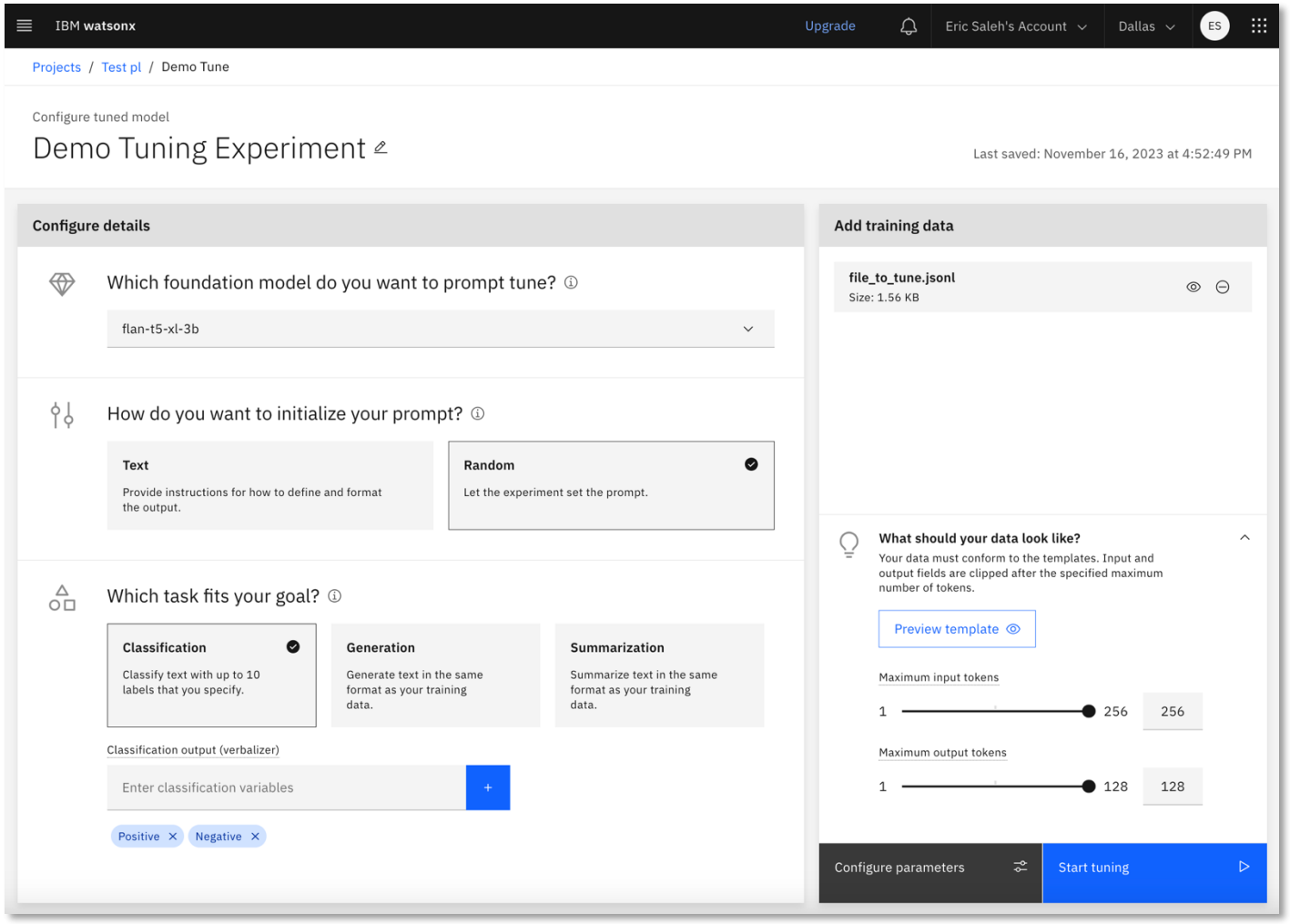
Ajustez vos LLMs avec des données étiquetées

<u>Résumé :</u>

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

<u>Plusieurs méthodes disponibles :</u>

- Prompt Tuning :
  Une méthode de specialisation sans modifications des poids des modèles.
- InstructLab :
  Une approche communautaire qui permet à plusieurs personnes au seins de l'entreprise de contribuer des connaissances et des capacités à un modèle, qui sera ré-entrainé 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 complexe.

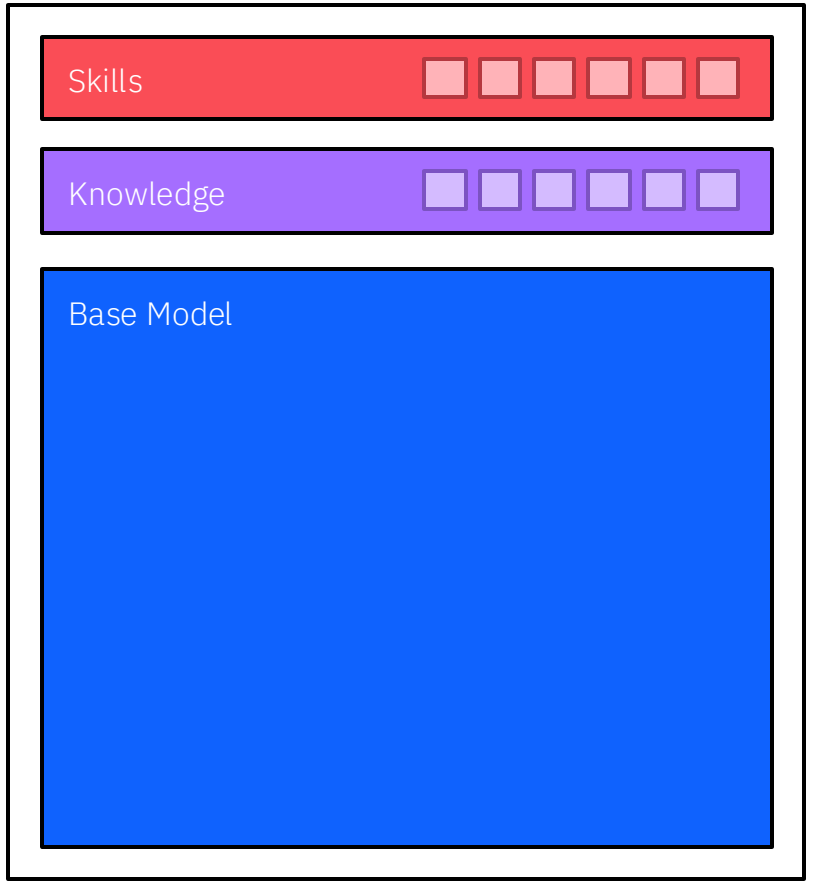


InstructLab : Augmenter les connaissances et améliorer les capacités des modèles

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

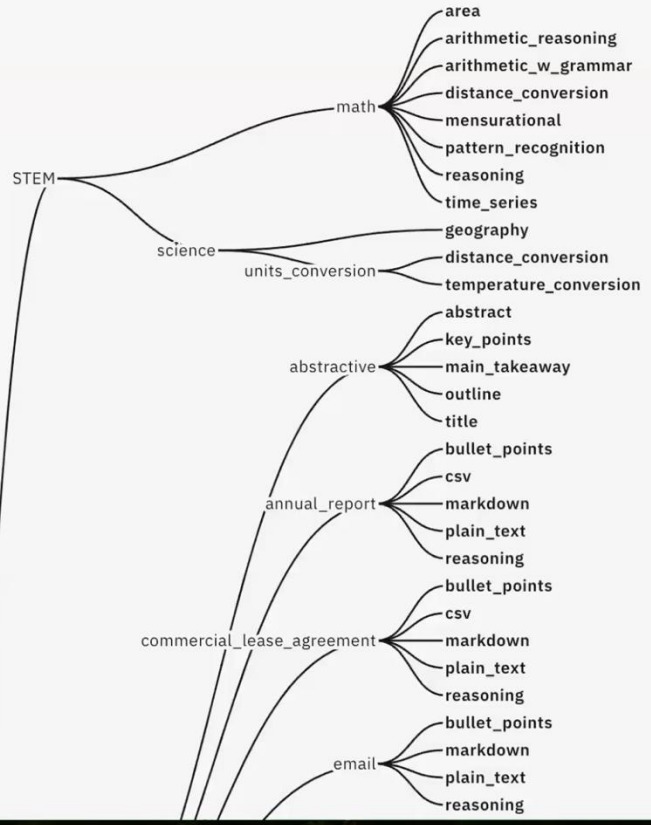
Alignement du modèle

granite-20b-multilingual

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

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

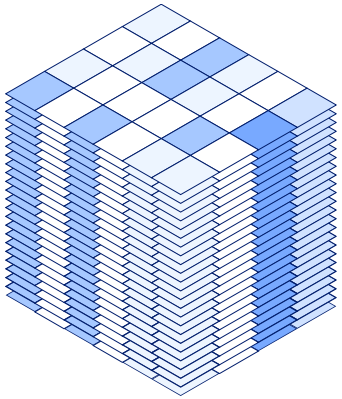
Details | **Training taxonomy**

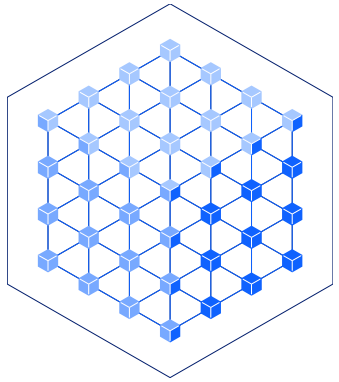


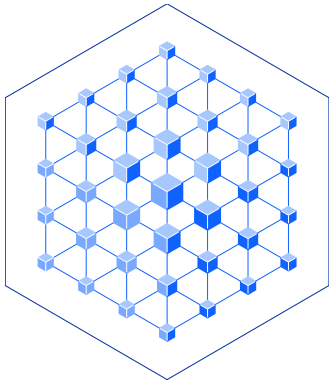
Explore the taxonomy for this model

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 donné.

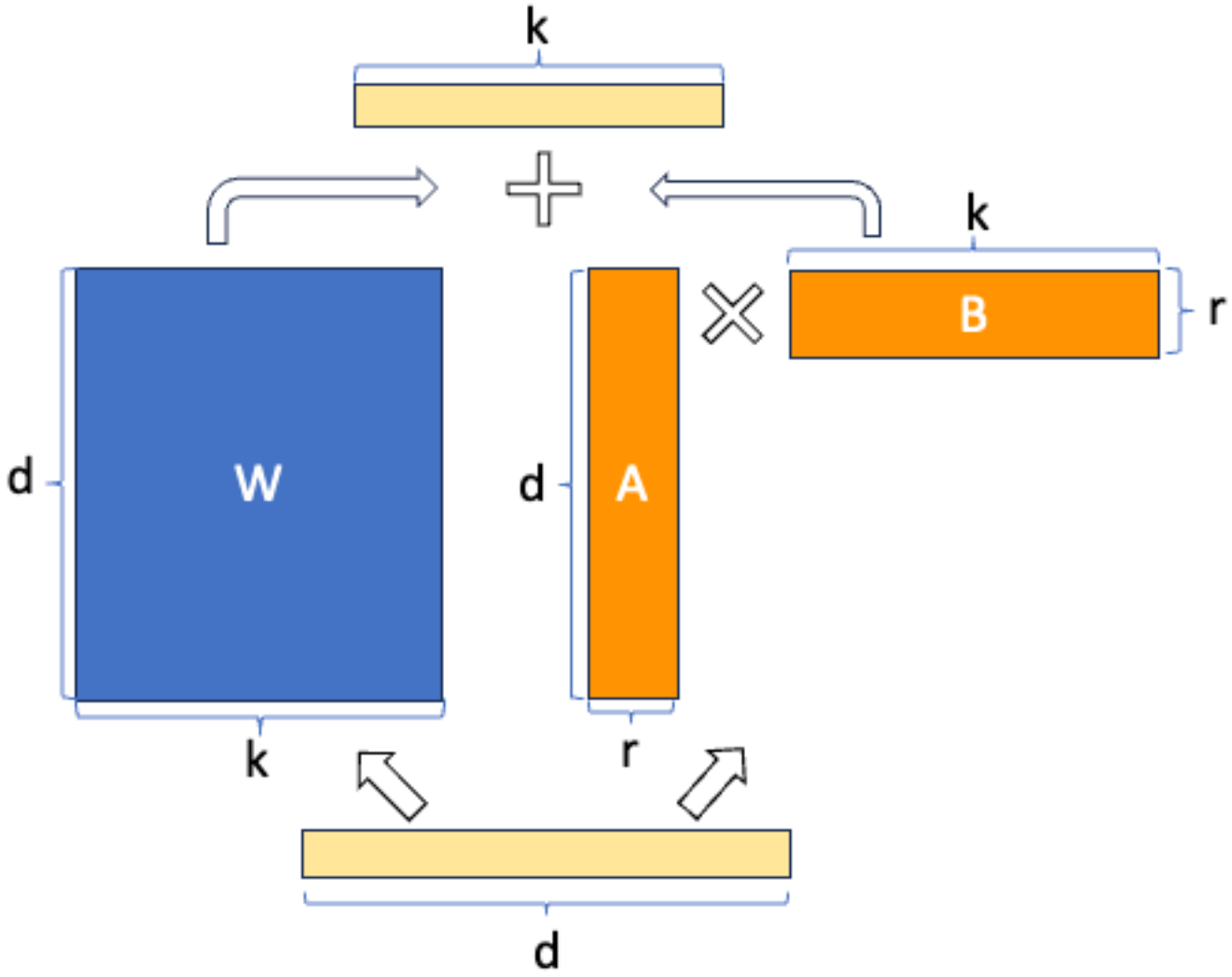
Critic model(s)

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.

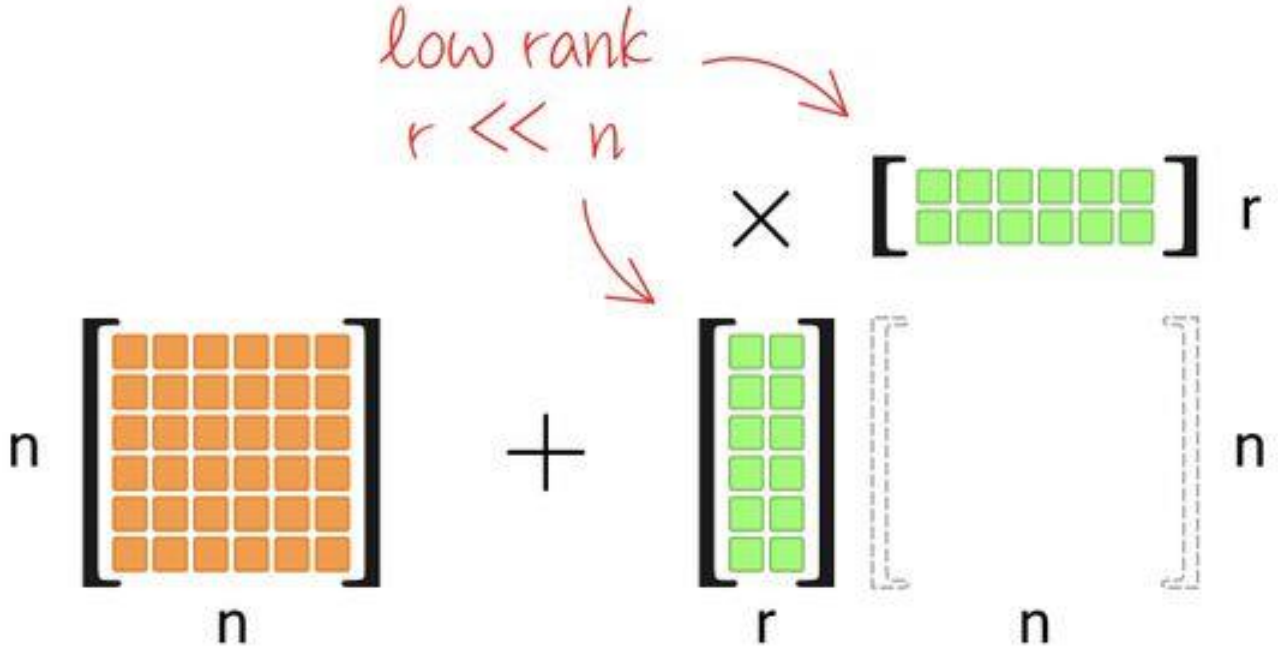
Student model(s)

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

InstrctLab : LoRA



InstrctLab : LoRA

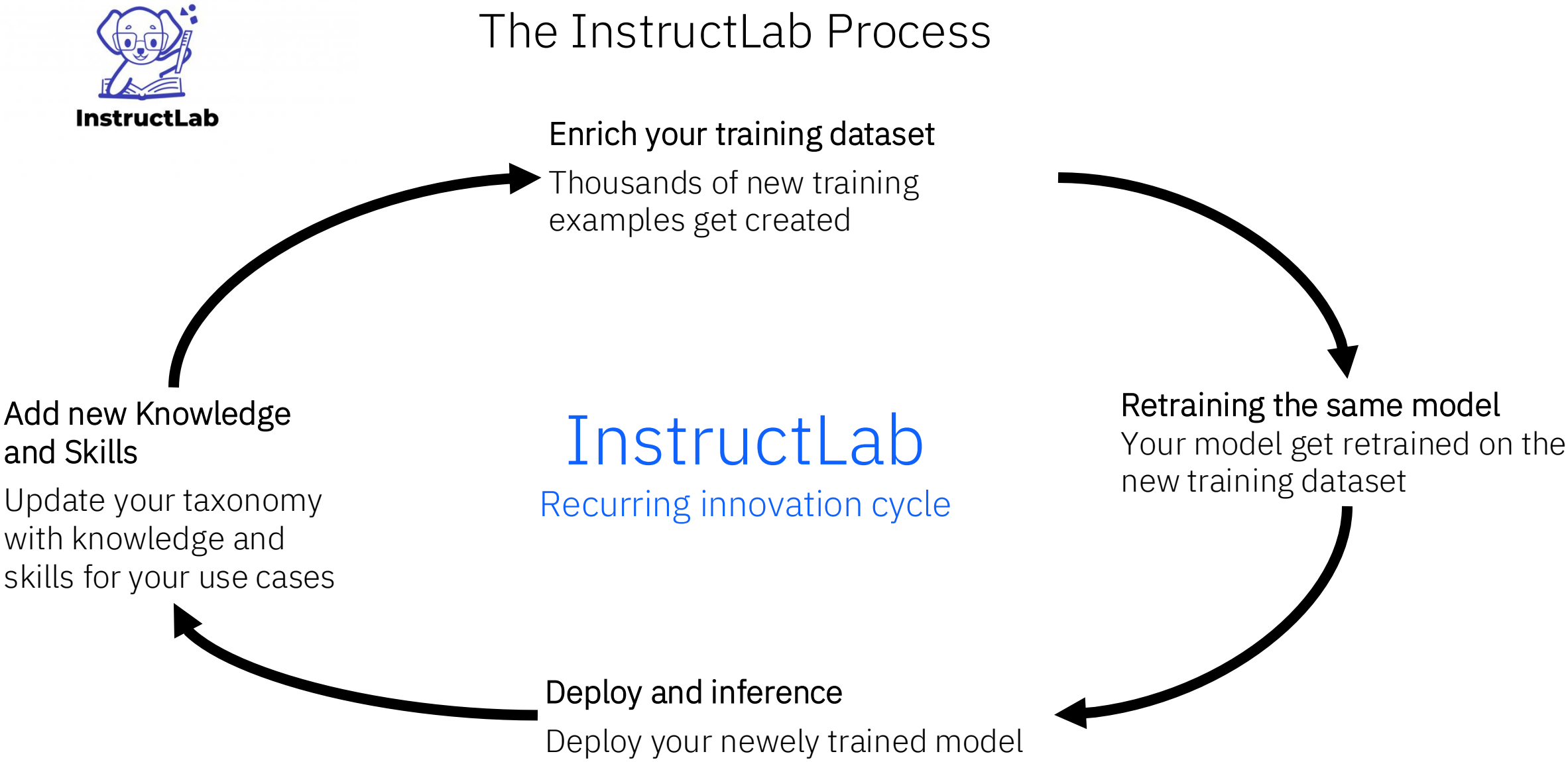


$$W = W_0 + A\,B$$

frozen

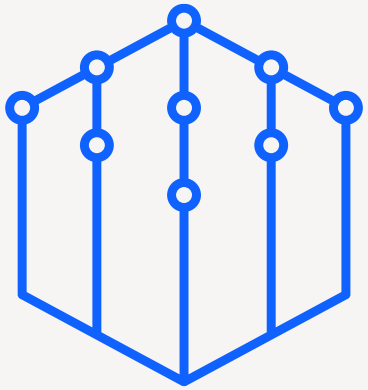
trainable

The InstructLab Process

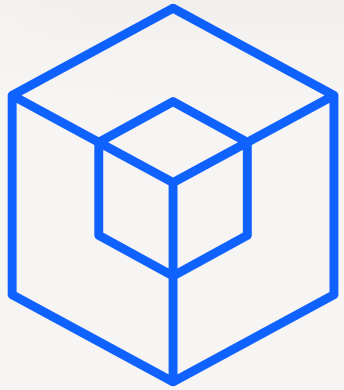


The IBM approach: fit-for-purpose models

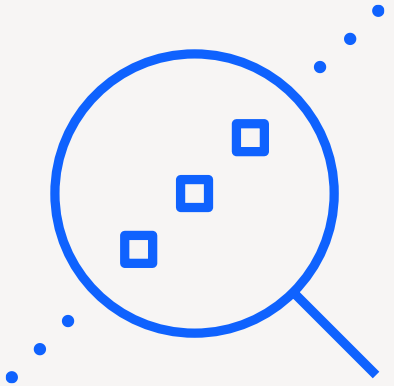
98% cost saving
with IBM Granite vs GPT-4



Vos données



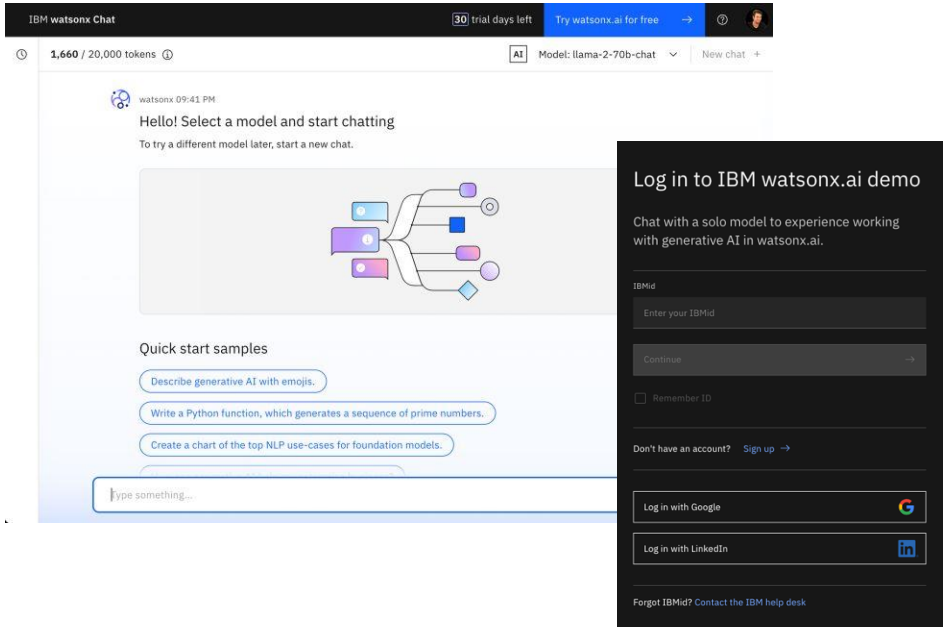
Le bon modèle



Cas d’usage ciblé
Fine tuning

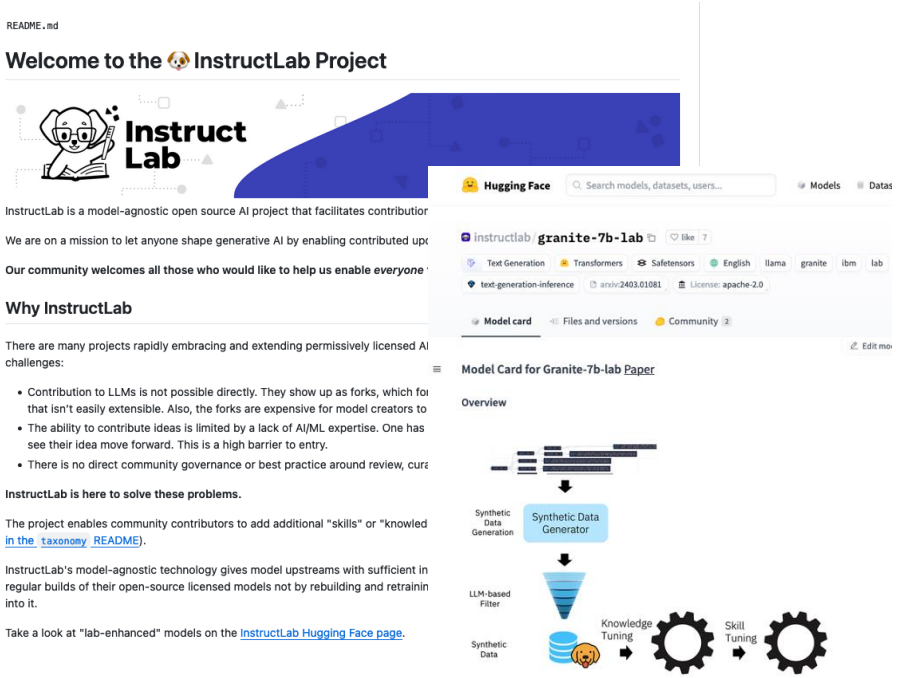
Essayez les modèles par vous-même

watsonx.ai interactive demo



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

InstructLab project



https://github.com/instructlab