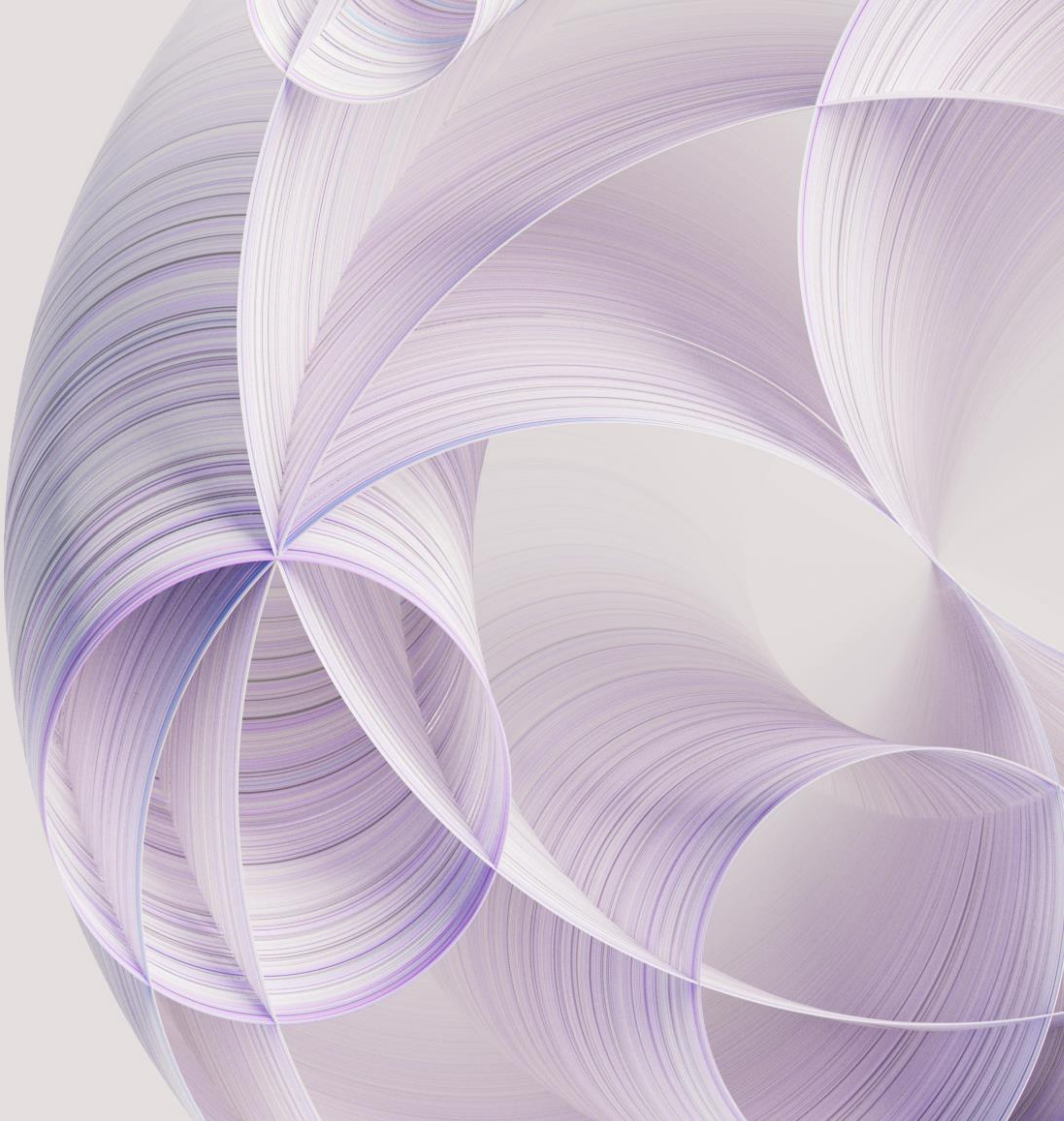
Watsonx.ai
Proof of experience (PoX) education

Prompt tuning

Felix Lee
felix@ca.ibm.com



Seller guidance and legal disclaimer

IBM and Business Partner
Internal Use Only

Slides in this presentation marked as **"IBM and Business Partner Internal Use Only"** are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by client.

Contents

Watsonx.ai Tuning Studio

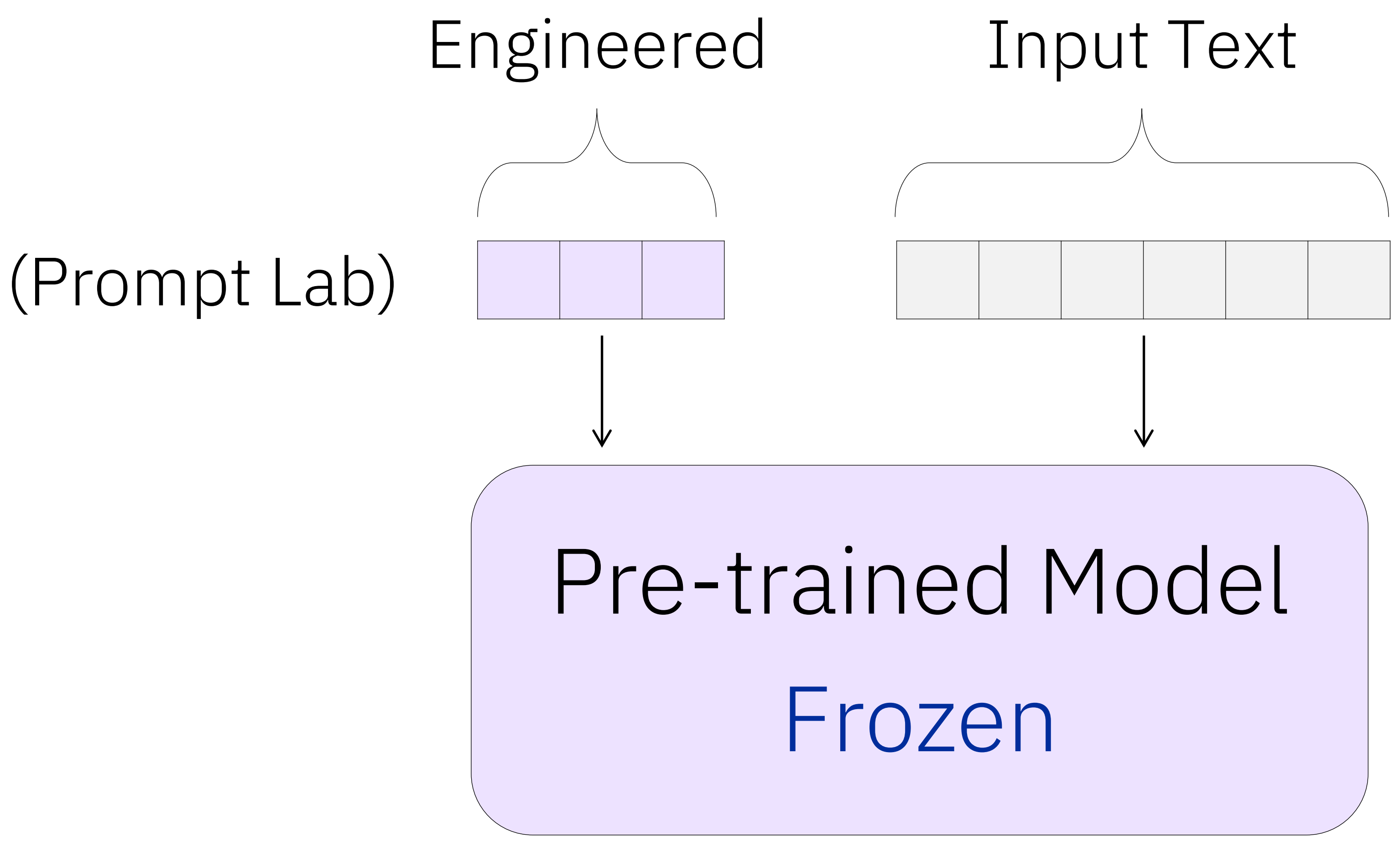
- What is prompt tuning?
- Prompt tuning vs. prompt engineering
- Prompt tuning vs. fine-tuning
- Prompt tuning and RAG
- Prompt tuning parameters
- Labeled data for prompt tuning
- Prompt tuning in a PoX

IBM watsonx.ai Tuning Studio

- The watsonx.ai Tuning Studio enables:
  - Prompt tuning (available now)
  - Fine-tuning (roadmap item)
- Prompt tuning is a powerful tool in any PoX
  - Base models do not have specific business knowledge to perform various tasks. These include:
    - Business processes and workflow
    - Industry-specific terminologies and categories
    - Operational details
  - Several business tasks that are not based on retrieving information from a corpus of truth (so RAG does not apply)
    - Providing output in a certain style, format, and tone
    - Determine the next action-taker based on defined workflows and current input
    - Handling edge cases in specific manners
  - Easier and much less costly than fine-tuning, and can be very effective.

What is prompt tuning?

Prompt Engineering



Engineered

Input Text

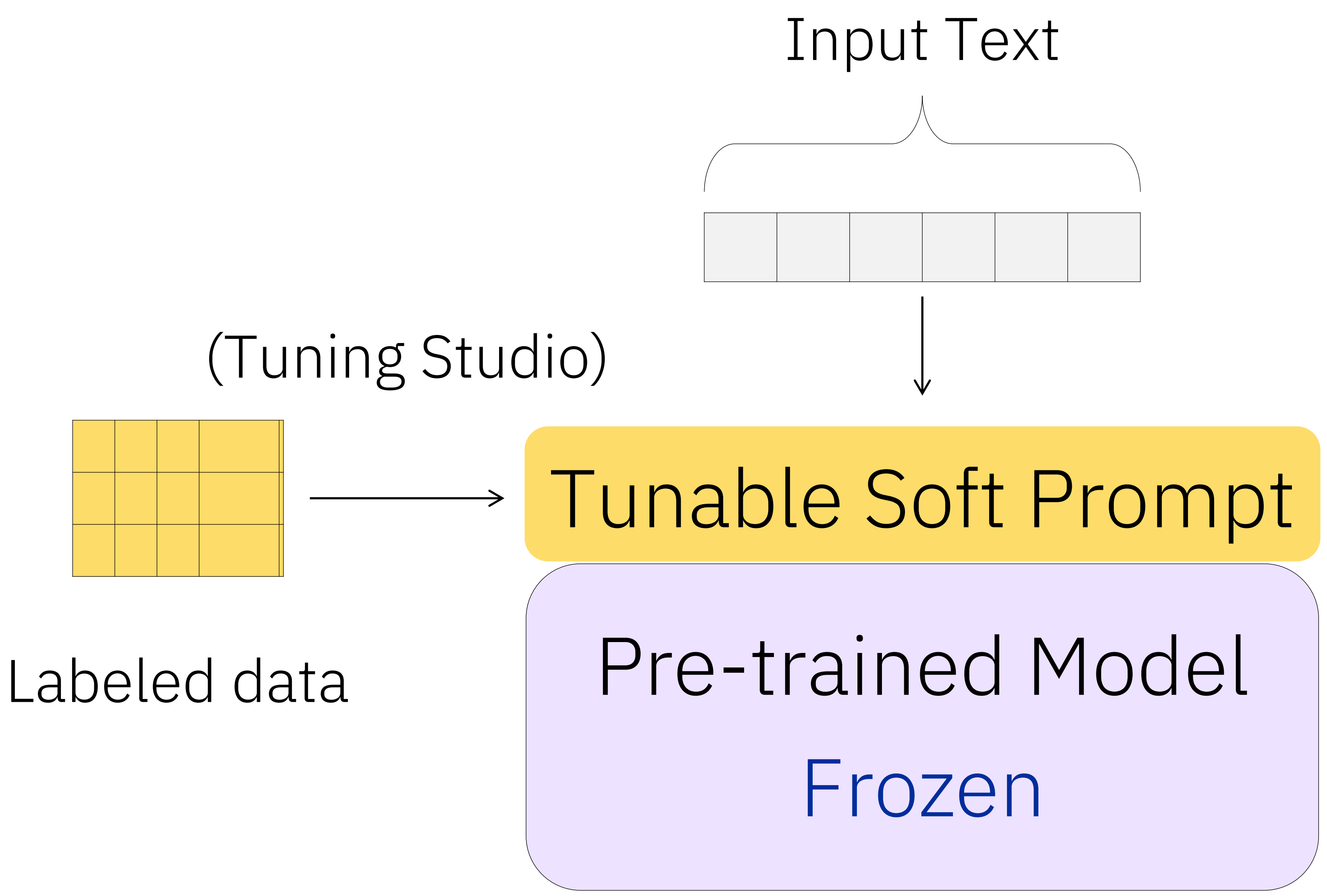
(Prompt Lab)

Pre-trained Model

Frozen

The input prompt, plus any engineered input (such as multi-shot etc.) are passed into the LLM (which remains frozen).

Prompt tuning



Input Text

(Tuning Studio)

Tunable Soft Prompt

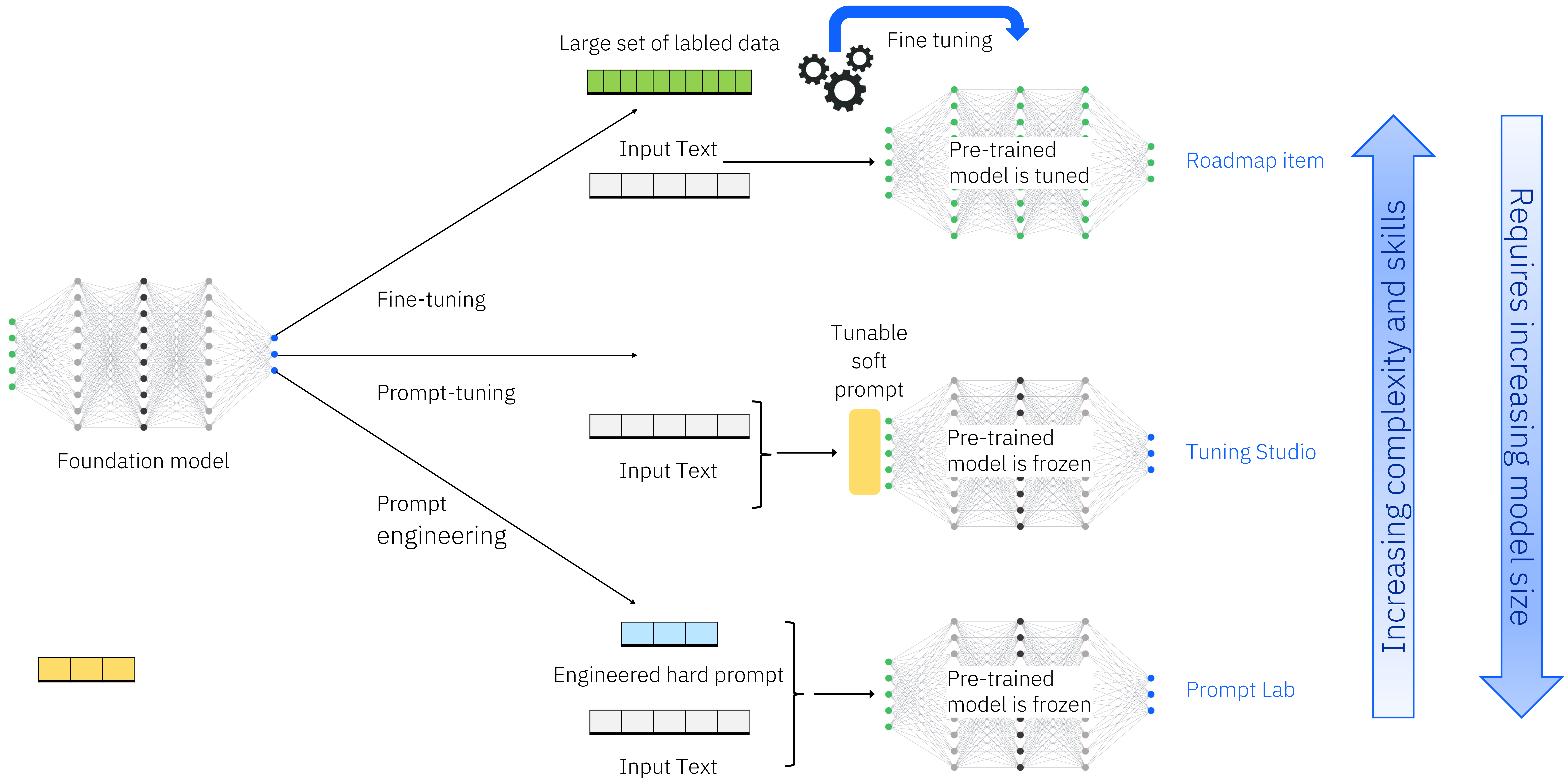
Labeled data

Pre-trained Model

Frozen

A tunable soft prompt is created using the input labeled data and sits on top of the LLM (which remains frozen). This soft prompt is always there for the new model, and can be re-tuned.

Prompt tuning, prompt engineering and fine-tuning



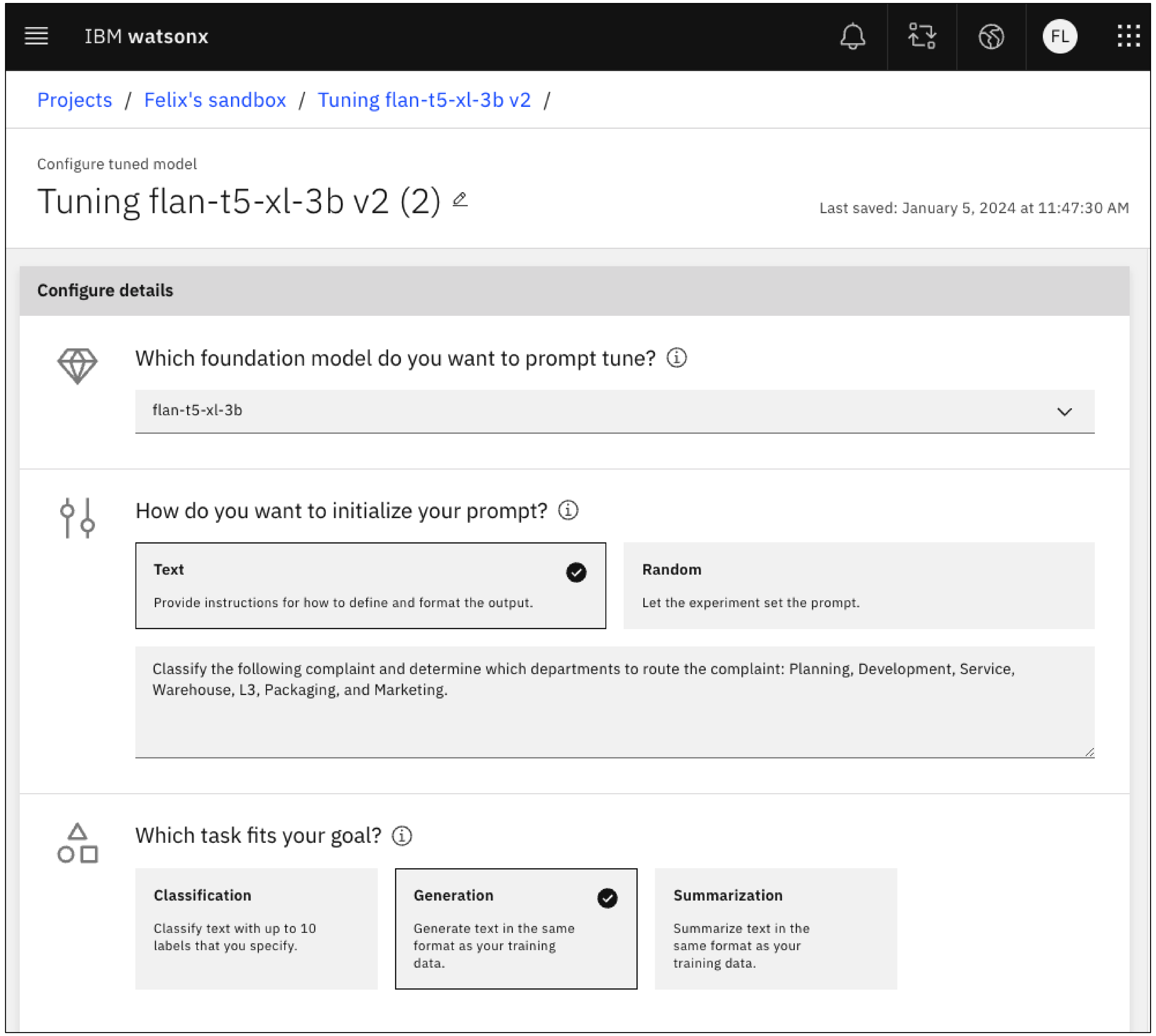
Watsonx.ai Tuning Studio – Prompt tune LLM with labeled data

Prompt tuning

- Tune the prompts with no changes to the underlying base model or weights
- Provides an efficient and low-cost way of adapting an LLM to new downstream tasks
- Allows clients to further train the model with focused, business data,
- Teach the LLM knowledge specific to, but not available outside of, the business.

Tuning Studio

- Supports a variety of tasks: Classification, Generation, and Summarization* (more in the roadmap
- Supports flan-t5-xl-3b and llama-2 -13b-chat* more to come)
- Requires a small set of labeled data
- Can achieve close to fine-tuning results without model modification and run at a much lower cost



Configure tuned model

Tuning flan-t5-xl-3b v2 (2)

Last saved: January 5, 2024 at 11:47:30 AM

Configure details

Which foundation model do you want to prompt tune?

flan-t5-xl-3b

How do you want to initialize your prompt?

Text ✓ — Provide instructions for how to define and format the output.

Random — Let the experiment set the prompt.

Classify the following complaint and determine which departments to route the complaint: Planning, Development, Service, Warehouse, L3, Packaging, and Marketing.

Which task fits your goal?

Classification — Classify text with up to 10 labels that you specify.

Generation ✓ — Generate text in the same format as your training data.

Summarization — Summarize text in the same format as your training data.

* as of March 2024

Prompt tuning versus Promt Engineering in a PoX

Prompt engineering

- No change to the underlying LLM’s weights
- Easier to do, no need for any labeled data input
- Relies more on multi-shot prompting
- Multi-shot is a hard prompt – counts against available tokens
- LLM leverage RAG and a validated corpus of truth to minimize hallucination
- Better for learning, less for PoX

Prompt tuning

- No change to the underlying LLM’s weights
- Need more resources and time to do
- Can use multi-shot but shouldn’t need it
- Soft prompt – does not count against available tokens
- Tuned LLM can benefit from RAG, but provides benefits beyond RAG
- **Can be a very powerful tool for PoX**
  - “Teach” the model business-specific knowledge
  - Can be customized to different downstream tasks

Prompt tuning versus fine-tuning in a PoX

Prompt tuning

- Requires a smaller set of labeled data – usually from 200 to 1000 entries
- Relatively low-cost to create/update
- Underlying LLM is not changed
- With the right set of labeled data, can have a performance close to that of a fine-tuned model
- Great option in a PoX when the LLM needs to learn business knowledge to perform specific tasks.

Fine-tuning

- Has the best performance for the specific task(s) it's fine-tuned for.
- Requires a very large set of labeled data in the order of thousands of entries
- Expensive to tune
- Underlying LLM will be modified
- Risk of catastrophic forgetting – losing characteristics that made the LLM attractive in the first place
- Better for production

Prompt tuning and RAG

Prompt tuning

- **LLM learns business knowledge** lacking in its original training.
- Augmented (base model not changed) model used to serve downstream tasks such as summarization, classification, generation, etc.
- Not required, but can work with a RAG pattern to solve business use cases.

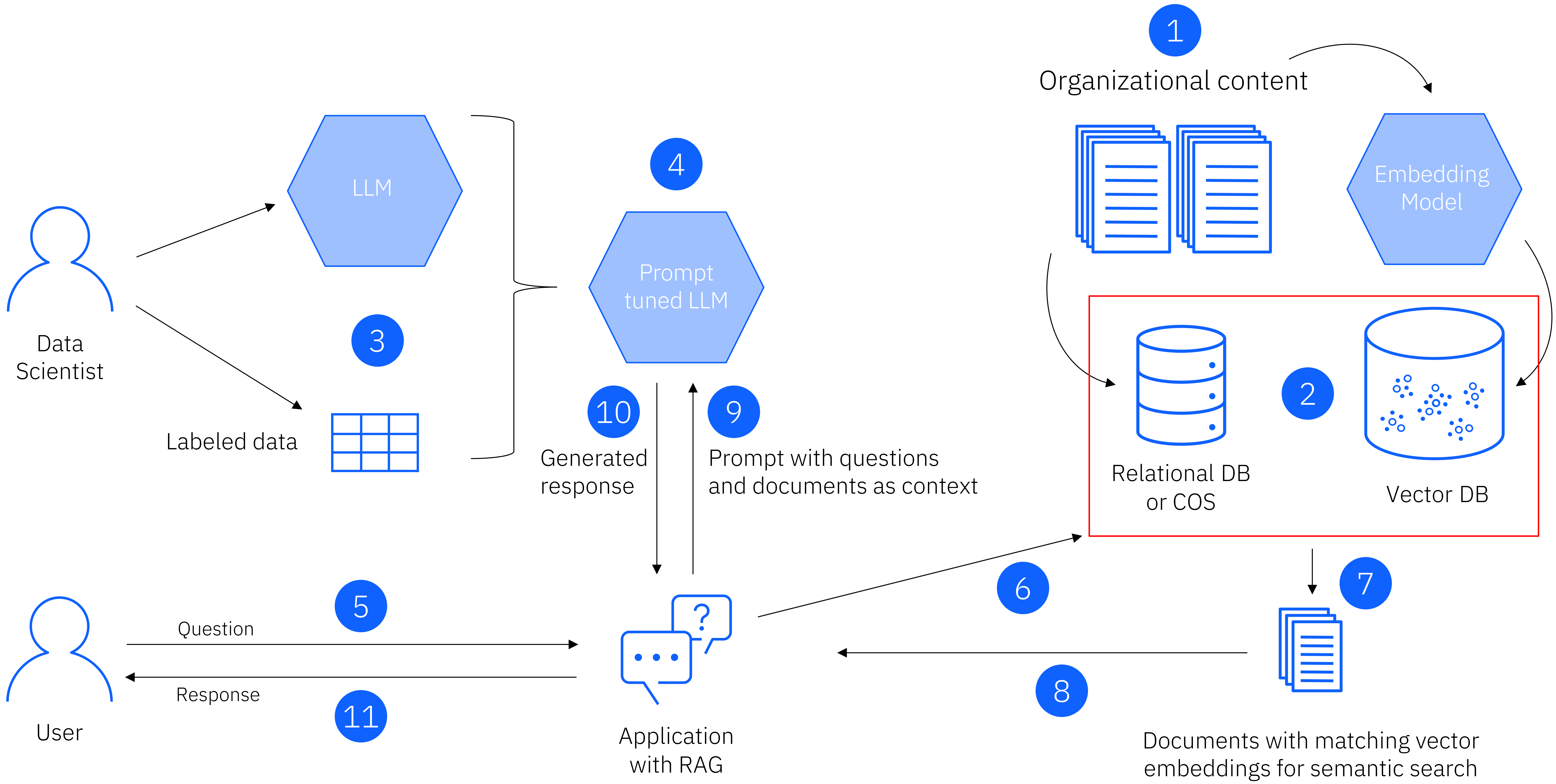
RAG

- The RAG pattern is useful when **completions should be derived from a corpus of truth**.
- The LLM generates completions based on a combination of the input prompt as well as extracted info from the corpus of truth.
- Its strength is in minimizing hallucination.

In a PoX:

- Can use prompt tuning or the RAG pattern to address specific needs
- Can use both to achieve even better results – but be aware of the resources require for doing each/both.

Prompt tuning with RAG



Prompt tuning parameters

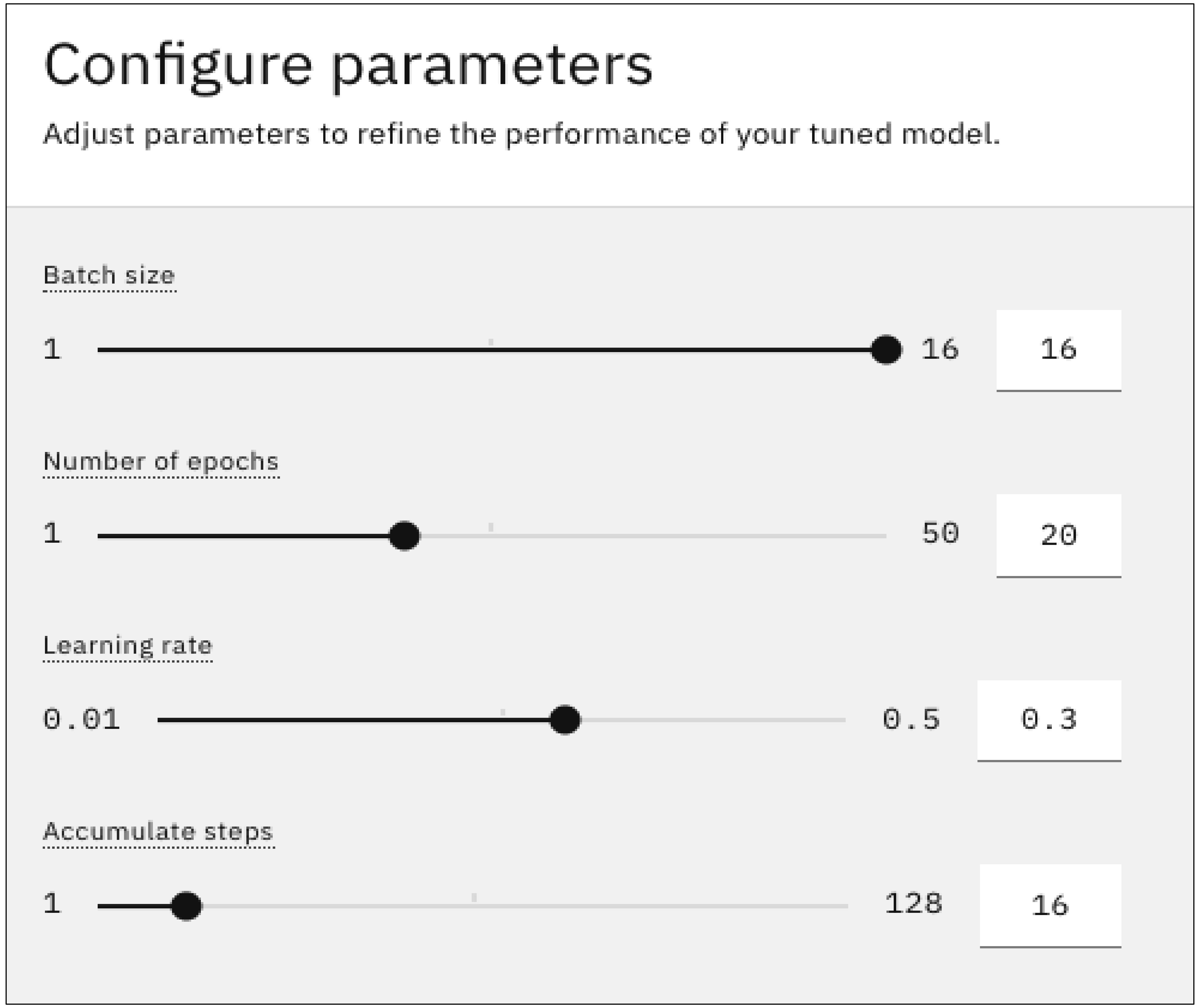
Can tune 4 parameters:

1. Batch size
   The number of samples to work through at one time

2. Number of epochs
   Number of times to cycle through the training data set

3. Learning rate
   How fast the neural network will progress towards the optimal "learn" state.

4. Accumulation steps
   The combined effect of the number of training steps to accumulate before updating the internal parameters



Configure parameters

Adjust parameters to refine the performance of your tuned model.

Batch size: 1 – 16, value 16

Number of epochs: 1 – 50, value 20

Learning rate: 0.01 – 0.5, value 0.3

Accumulate steps: 1 – 128, value 16

In a PoX where the number of samples is not too big (200 – 500), there is probably no need to change these parameters. There can be cost and optimization advantages in modifying some parameters in production when the sample size is much larger.

Prompt tuning: smaller model vs. larger model

Smaller model (such as flan-t5-3b-xxl)

- Likely perform less well as-is
- Takes less time to tune
- Less costly (CUH requirement is lower) to tune
- Easier to re-tune
- Might converge quicker to an optimal solution (for example: a lower value for its loss function given the number of epochs, and the same training data).

larger model (such as llama-2-13b-chat)

- Likely perform better as-is
- Take longer to tune
- More costly (higher CUH requirement) to tune
- More difficult to re-tune
- Tend to exhibit more peaks and valleys in its loss function, might take longer

Loss function: flan-t5-xl-3b vs llama-2-13b-chat

Loss function ⓘ
Loss
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Training run (epoch)

llama-2-13b-chat
tuning time 21 minutes

flan-t5-xl-3b
tuning time 5 minutes

Prompt tuning in a PoX

- Recognize the use cases:
  - The expected completions cannot be achieved via Prompt engineering with multi-shot prompting
  - Where an LLM needs to learn specific business logic and knowledge
  - Note that you are solving a class of problems/queries, not just a single problem/query.
  - Prompt-tuned models can be used in an RAG use case
  - A single LLM can be prompt-tuned for different downstream tasks

- Work with the client to identify
  - What client use cases can benefit from getting information from a “corpus of truth”?
  - What knowledge gaps and logic are missing from the LLM
  - The best set of data to address the gaps – it is better to spend more time and care to find the right set of data than randomly try with different sets of labeled data (even though they may be valid).

IBM