

Can We Make Model Alignment a Software Engineering Process?

The AI Conference, San Francisco
September 2024

Dean Wampler, Ph.D.

The AI Alliance and IBM Research

thealliance.ai

deanwampler.com/talks



About the Images...

I used Adobe Firefly to “enhance” my real photographs.



AI as Software Engineering

- Two topics:
 1. Can we make model alignment (e.g., tuning) more iterative and incremental?
 2. Automated testing of probabilistic systems is dang hard!

AI Alliance

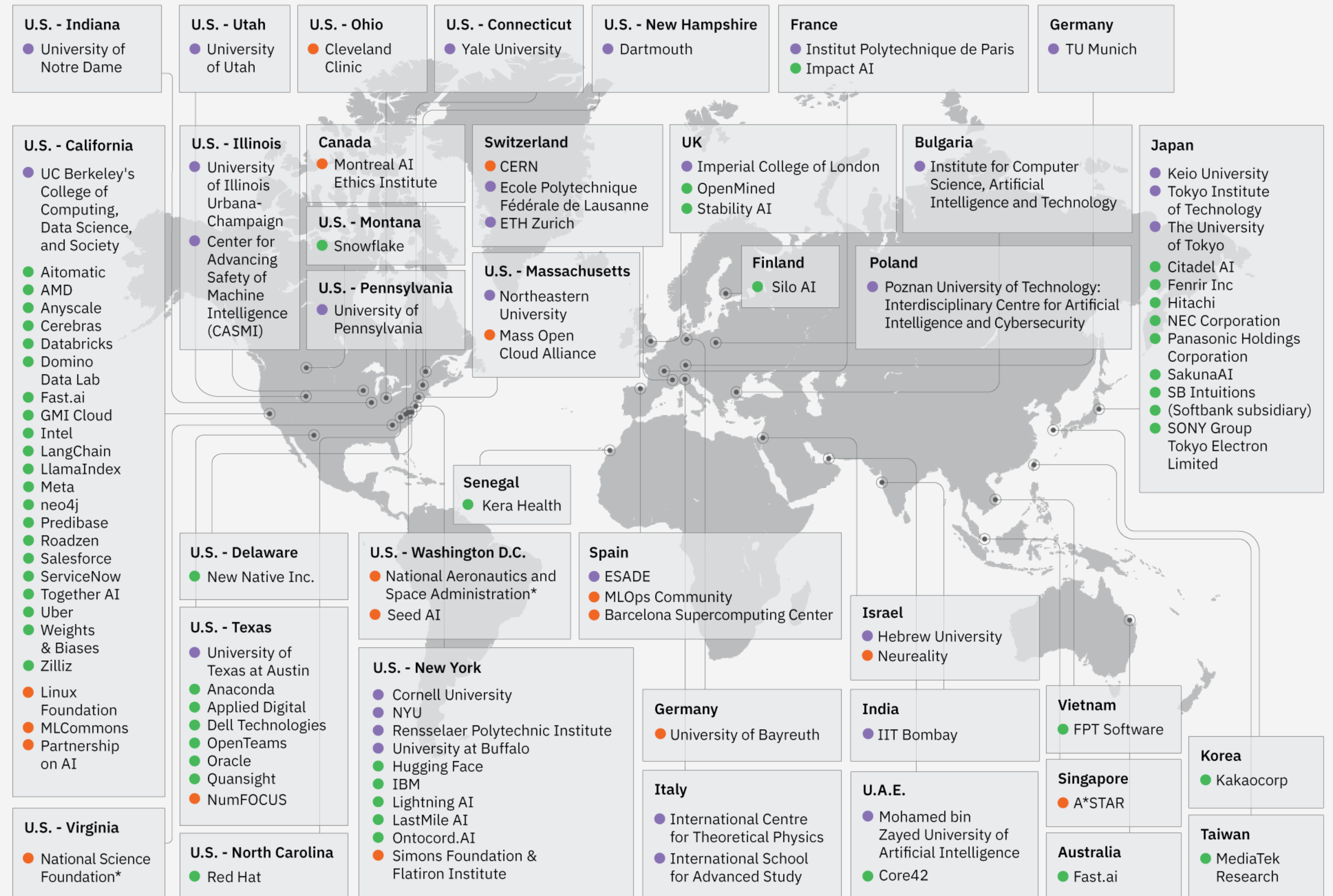
thealliance.ai

Our core beliefs in AI that is open is the tie that binds us, despite our differences.

Member organizations from academia, commercial, research and non-profits and span the globe.

Visit our booth, #129
(on the left as you enter the sponsor pavilion)

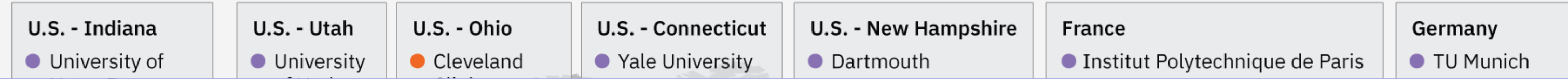
+100 organizations in +20 countries, and growing



AI Alliance

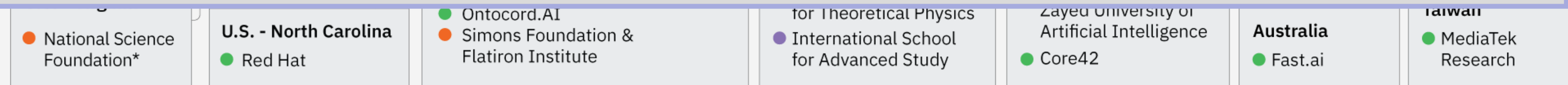
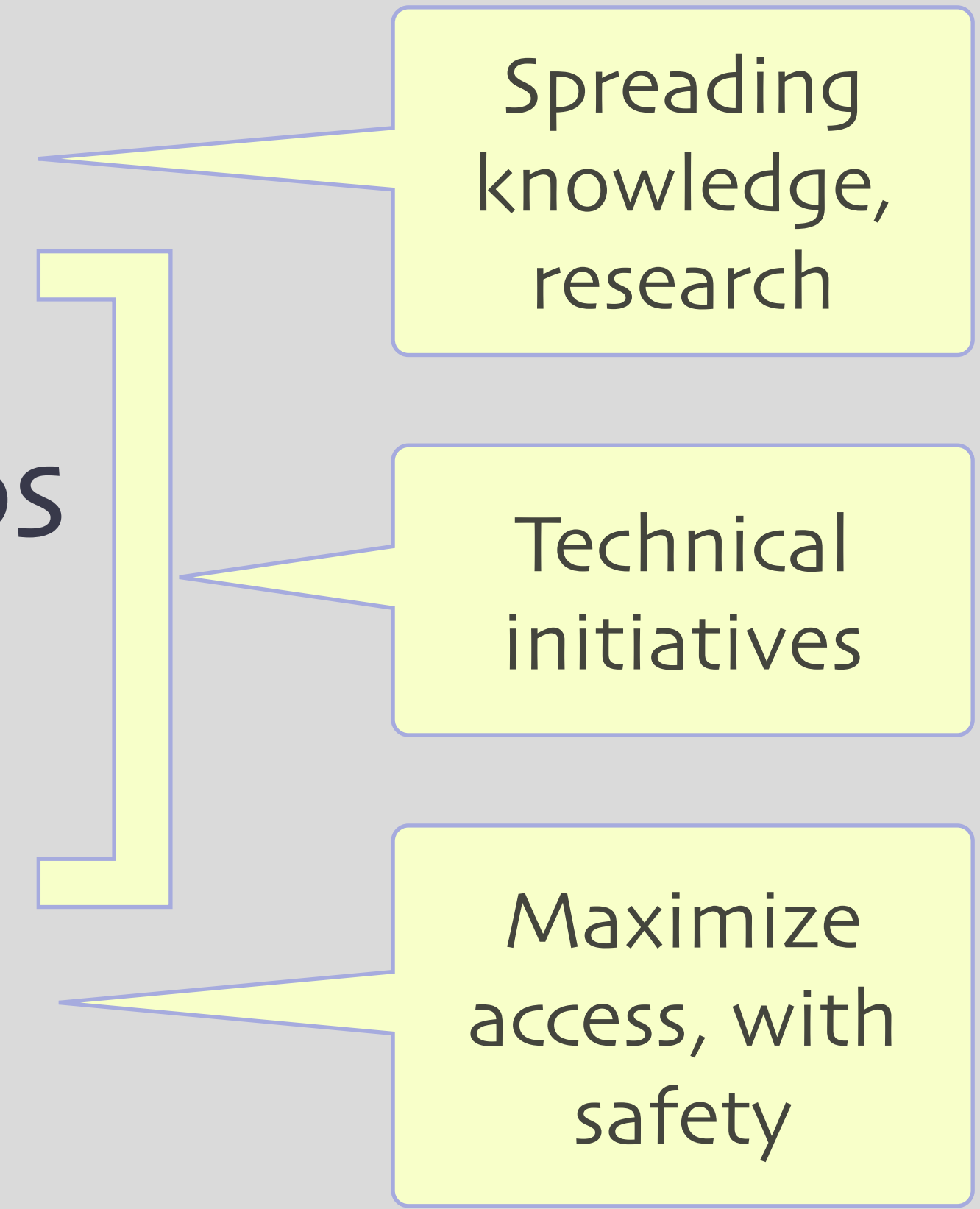
+100 organizations in +20 countries, and growing

thealliance.ai



Six Focus Areas:

1. Education and research
2. Trust and safety
3. Tools for building models and apps
4. Hardware portability
5. Open models and datasets
6. Policy and regulations



Our
ope
us,

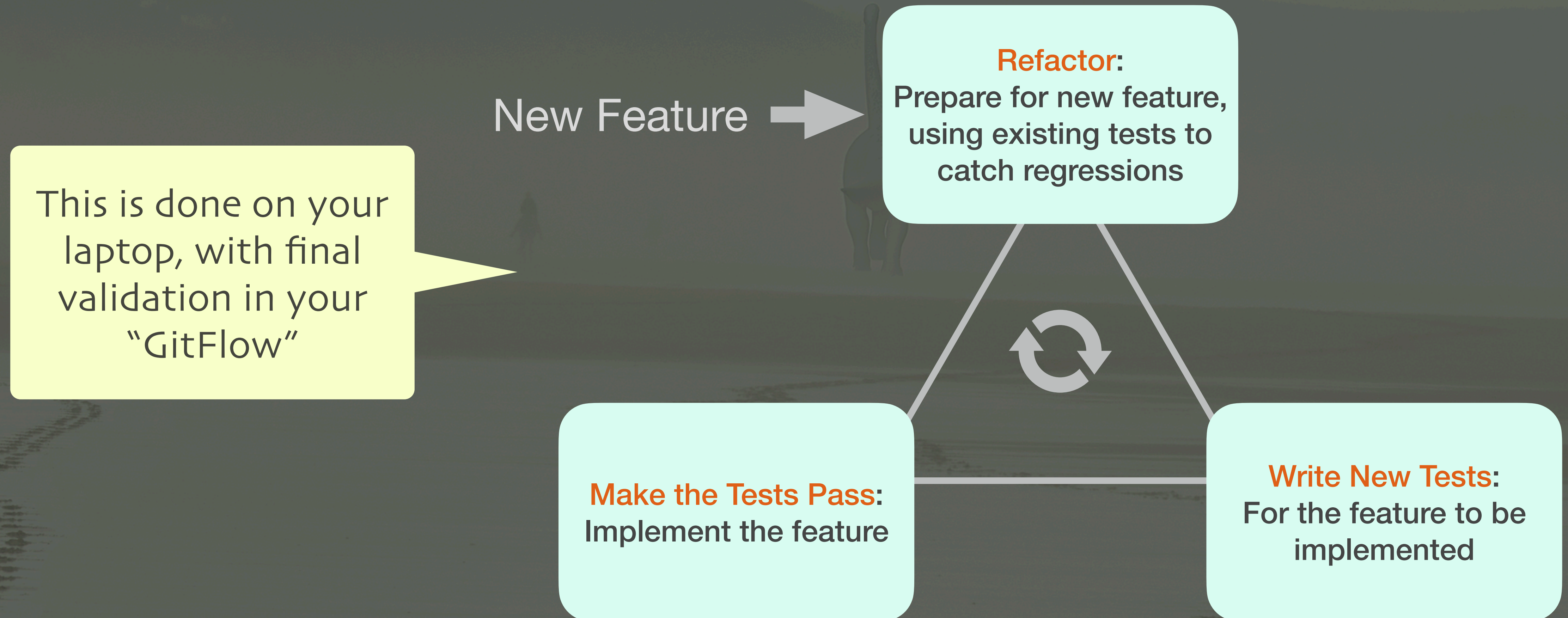
Me
aca
res
anc

v

Iterative and Incremental Model Tuning

What Software Developers Like

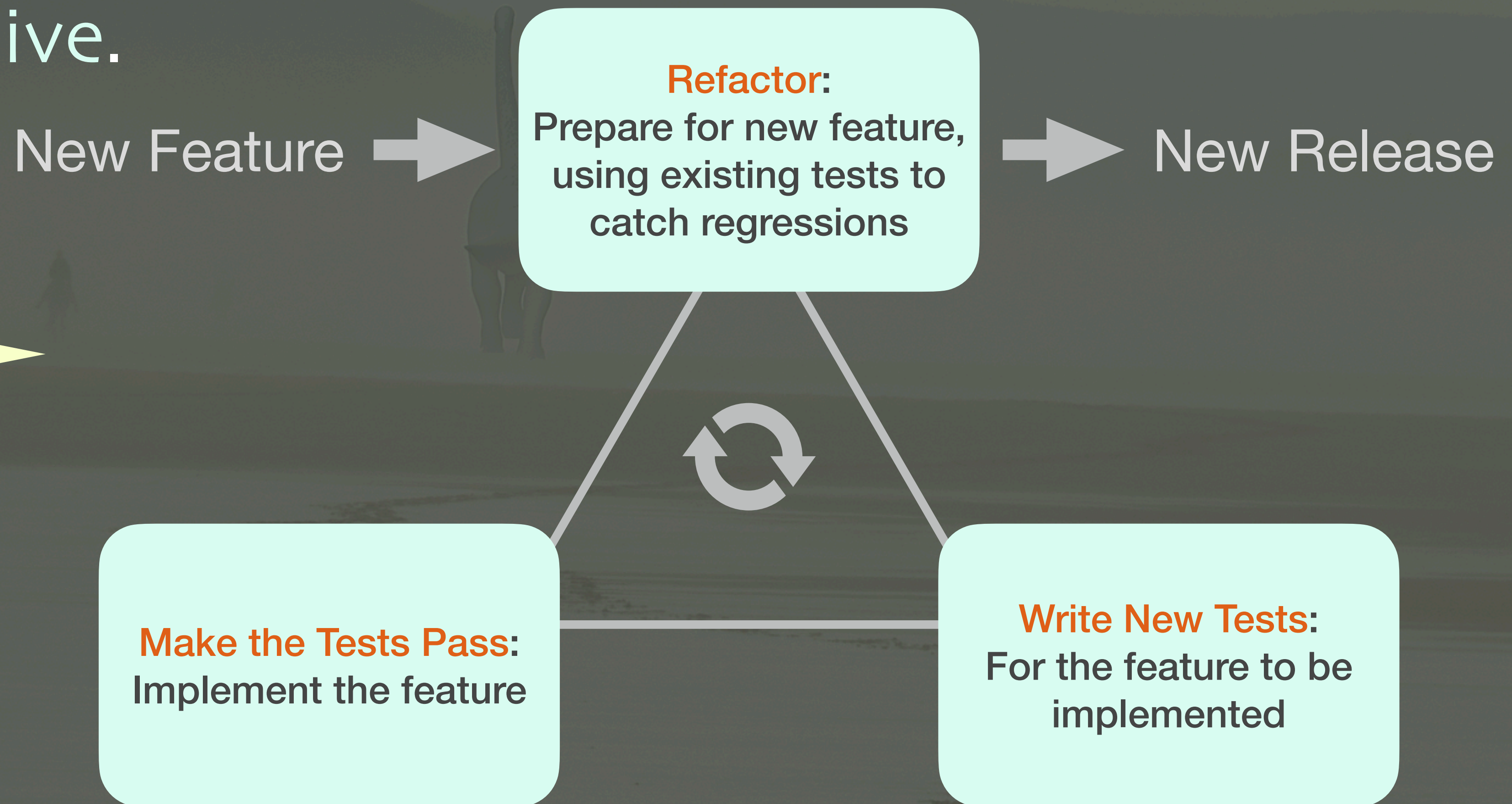
- Features are added incrementally.



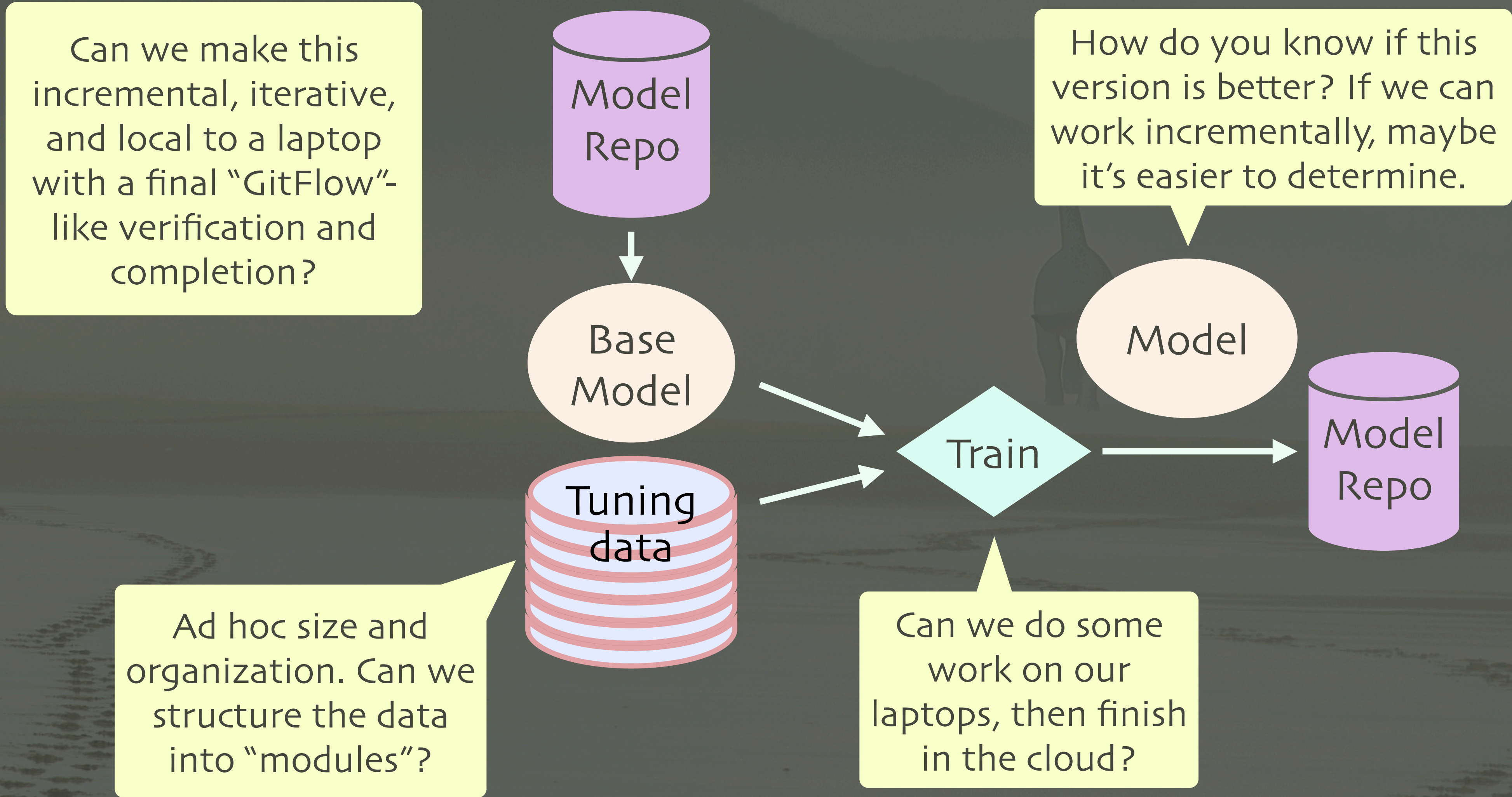
What Software Developers Like

- Features are added incrementally.
- Releases are iterative.

Releases are usually done with server-side automation, including final validation steps.



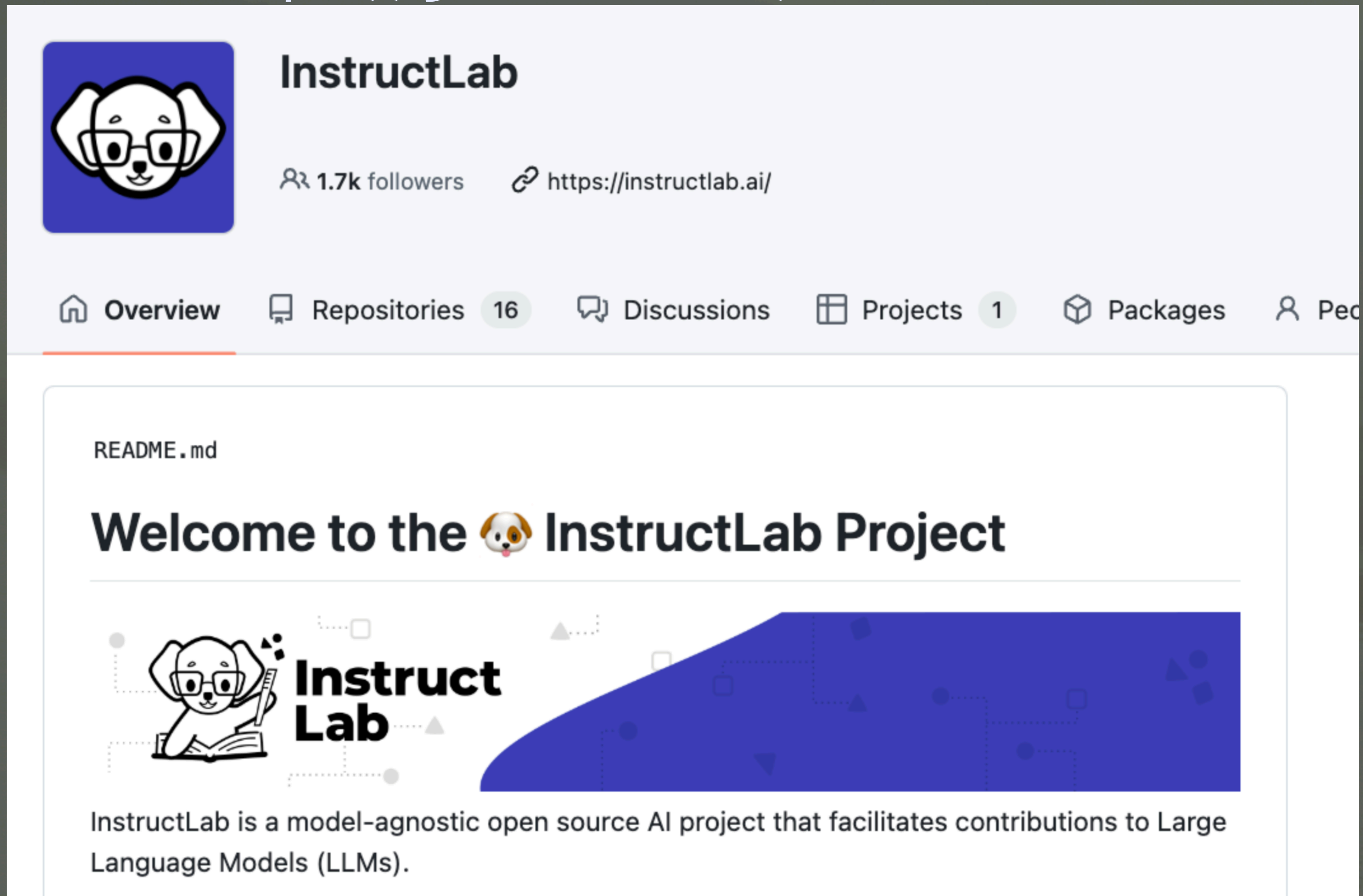
What Model Tuning Is Often Like



One Approach: InstructLab

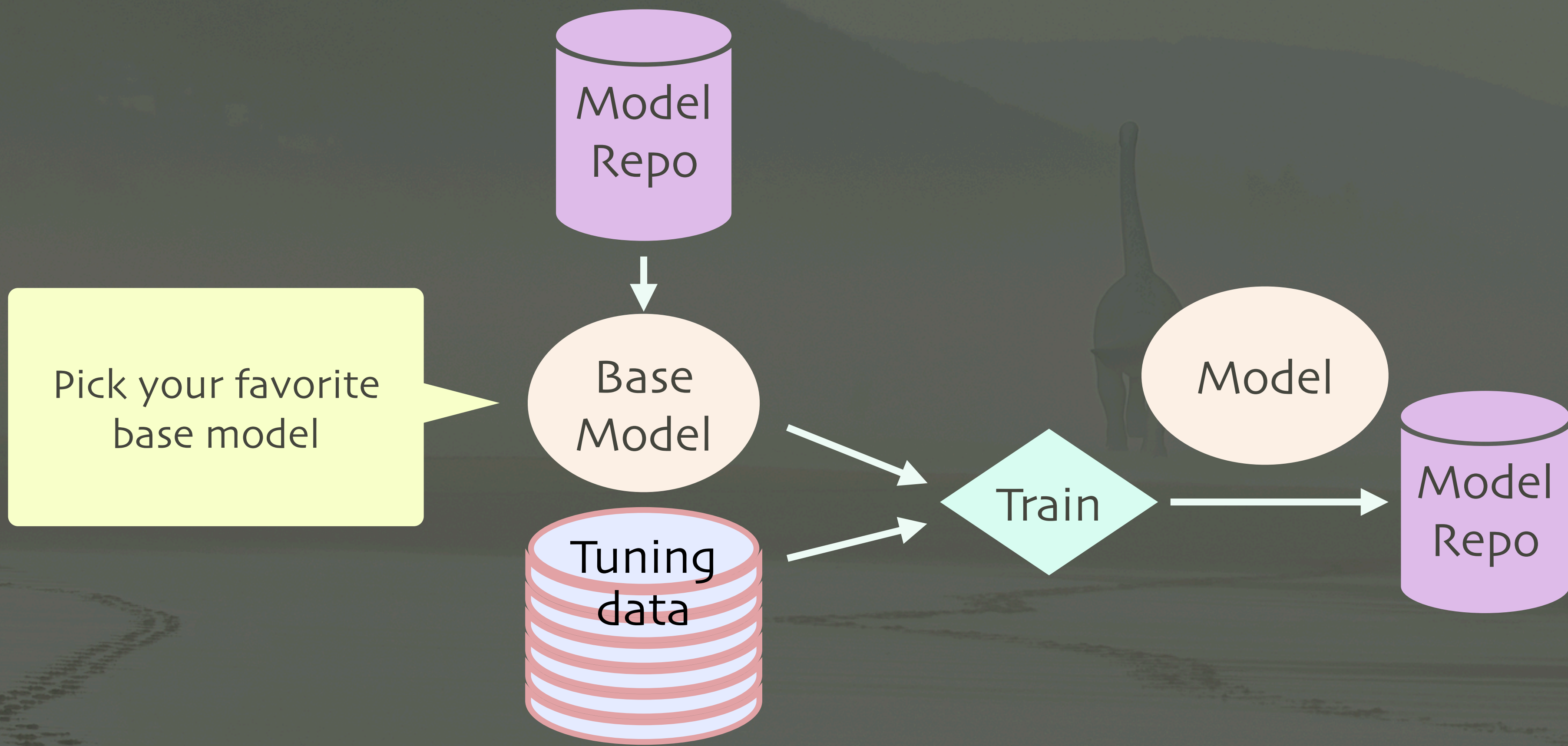
<https://github.com/instructlab>

Open sourced by
IBM and Red Hat

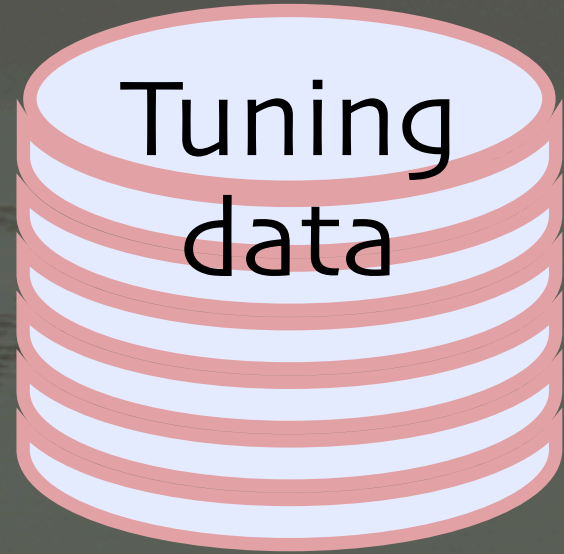
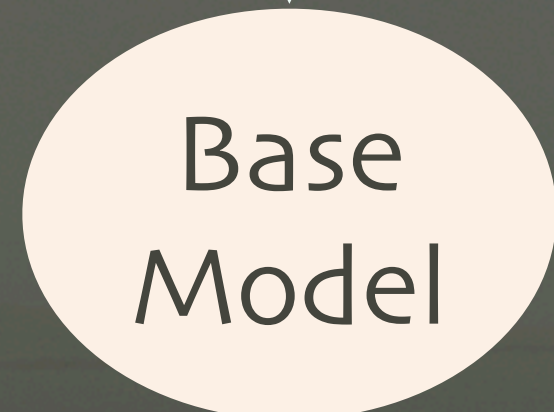
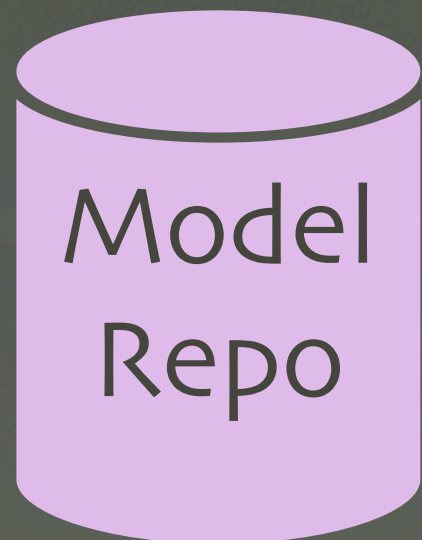


The screenshot shows the GitHub profile for InstructLab. The profile picture is a blue square with a white dog wearing glasses. The name "InstructLab" is displayed in bold black text. Below the name, it shows "1.7k followers" and a link to "https://instructlab.ai/". The navigation bar includes "Overview" (selected), "Repositories 16", "Discussions", "Projects 1", and "Packages". The main content area shows a "README.md" file with the heading "Welcome to the 🐶 InstructLab Project". Below the heading is a graphic featuring the dog logo and the text "Instruct Lab". The text below the graphic reads: "InstructLab is a model-agnostic open source AI project that facilitates contributions to Large Language Models (LLMs)."

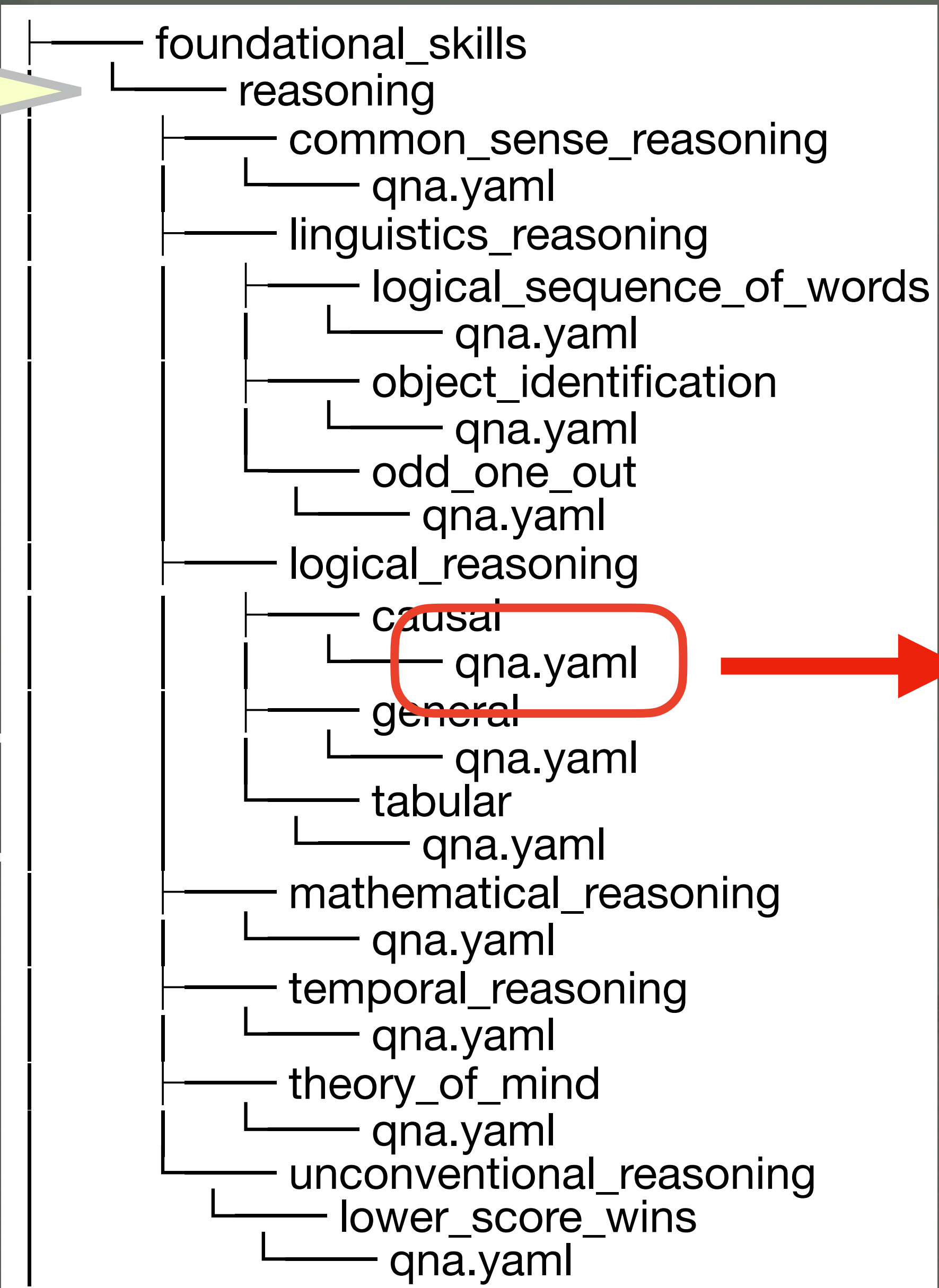
See also AgentInstruct:
<https://arxiv.org/abs/2407.03502>



Add new taxonomy entries

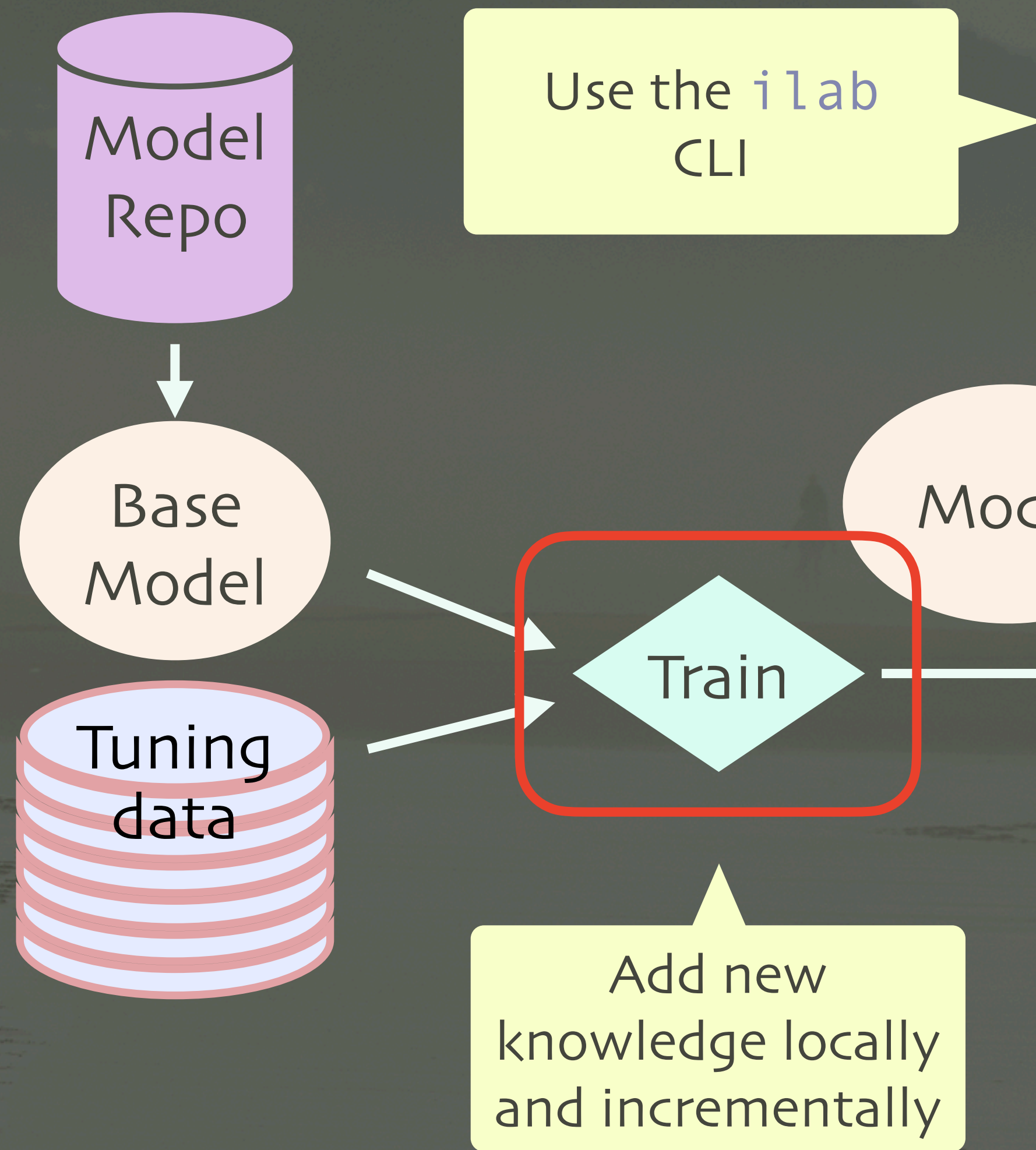


Organize data into a hierarchical taxonomy



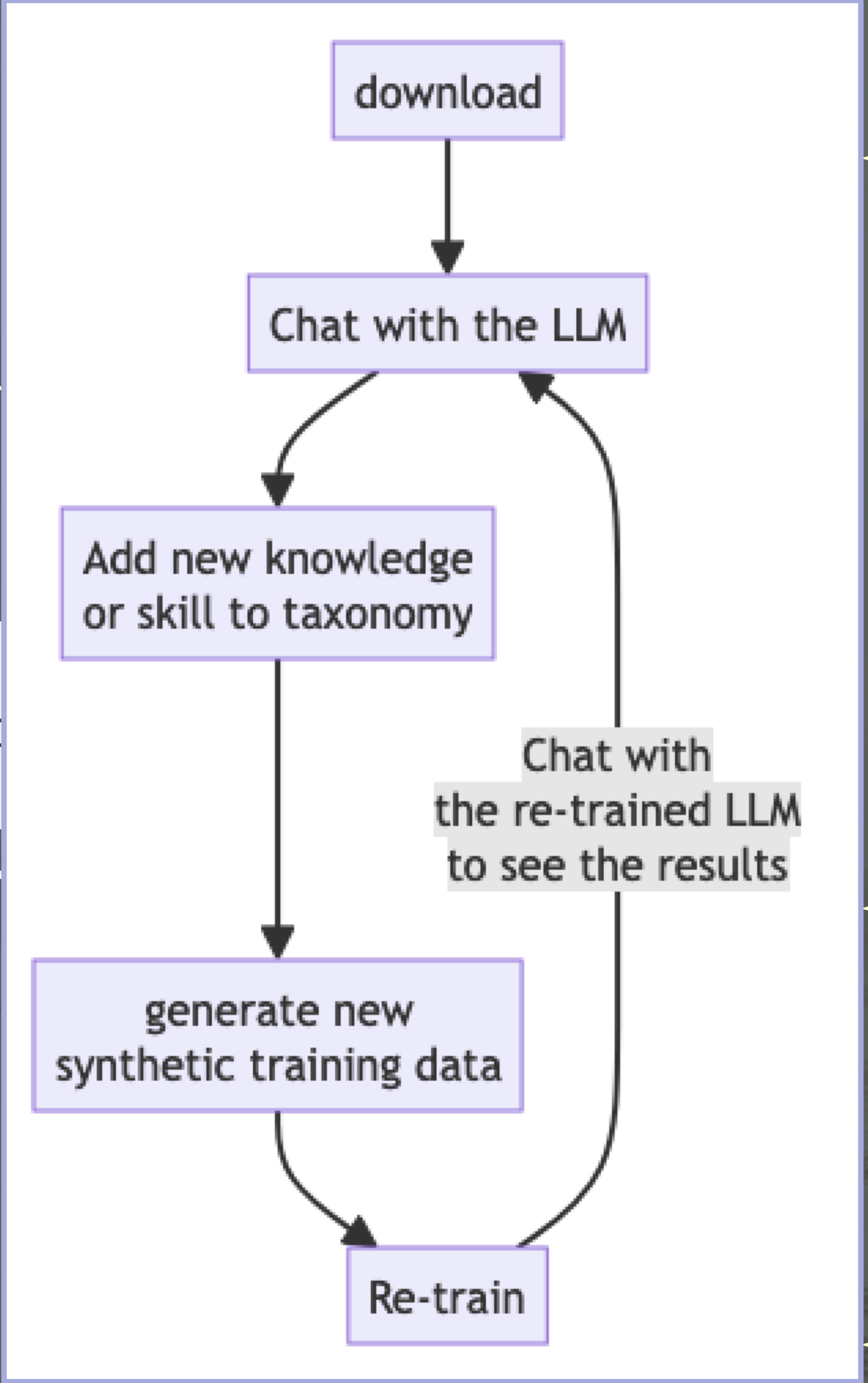
created_by: IBM
seed_examples:
- answer: 'While days tend to be longer in the summer because it is not summer doesn't mean days are necessarily shorter.'
,
question: 'If it is summer, then the days are longer. / longer if it is not summer ?'
,
- answer: 'No, we cannot conclusively conclude that are black based solely on the given premises. The statement "some mammals black" does not necessarily guarantee that among those mammals are cats.'
,
question: 'If all cats are mammals and some mammals black, can we conclude that some cats are black?'
- answer: 'Yes, we can conclude that some cats are black based on the premises.'
,
question: 'If all squares are rectangles and a rectangle has four sides, can we conclude that all squares have four sides?'

Create a few Q&A examples in a qna.yaml file



Use the `ilab` CLI

Add new knowledge locally and incrementally

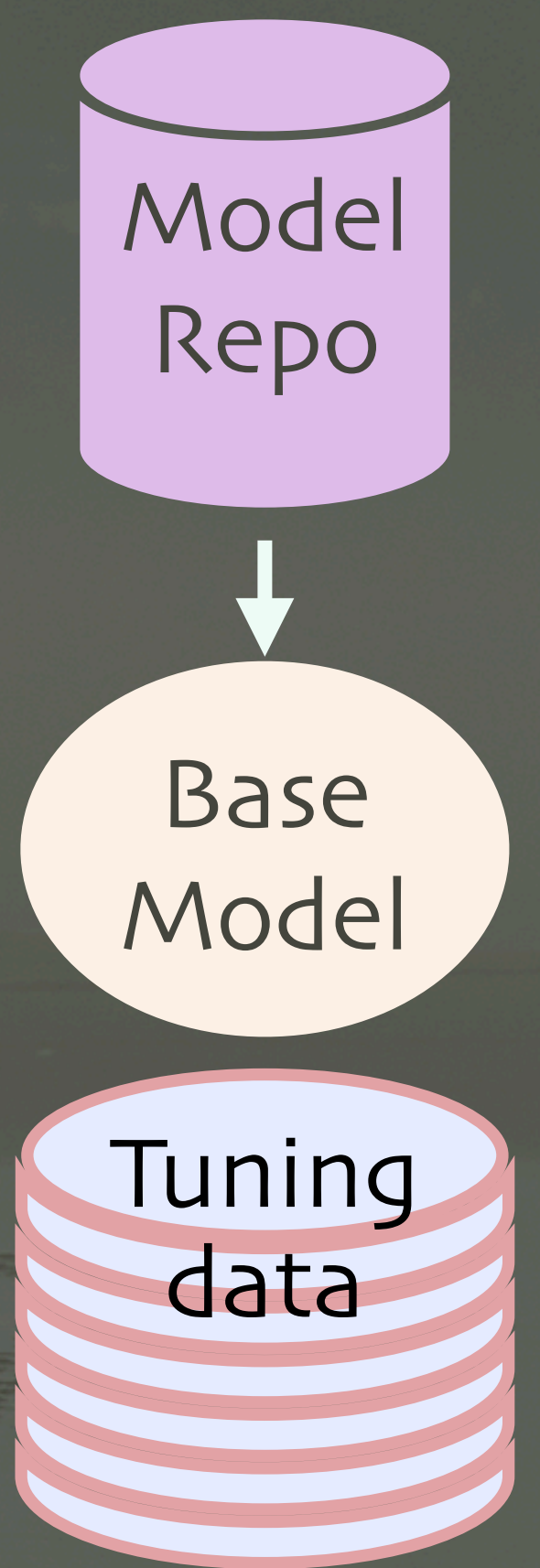


Download a quantized model version suitable for local execution

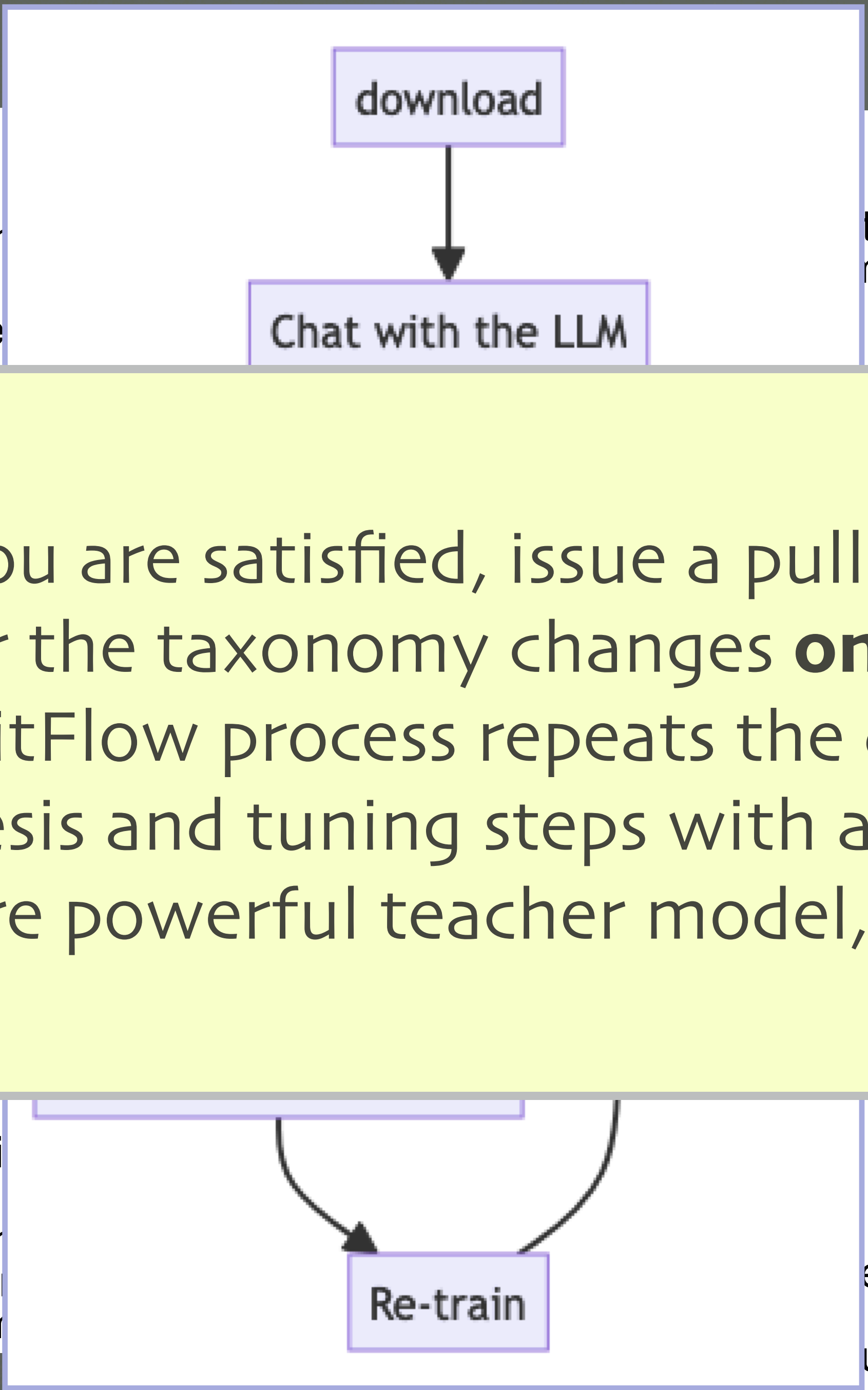
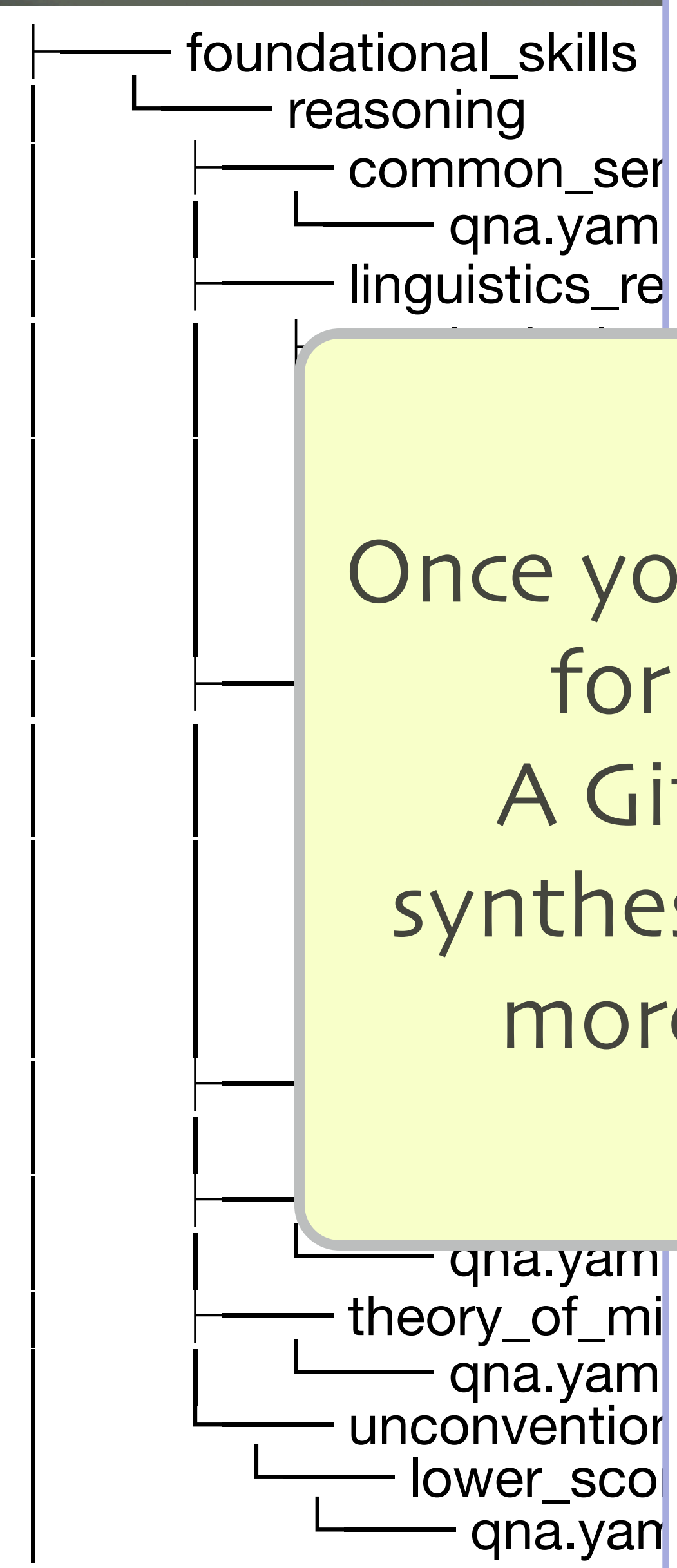
Results locally will be low fidelity

Uses QLoRA for efficiency

Image: <https://github.com/instructlab/instructlab>



Organize data into a hierarchical taxonomy



Once you are satisfied, issue a pull request for the taxonomy changes **only**. A GitFlow process repeats the data synthesis and tuning steps with a larger, more powerful teacher model, etc.

tend to be longer in the summer
ner
are necessarily shorter.

ays are longer. A

conclude that

ent "some mam

als are cats.

d some mamma

all squares have

es are rectangles and a rectang

uares have four sides?

InstructLab

Cons (1/2)

- Testing!
 - Supports a combination of standard benchmarks and “try it out”, but...
 - Still need “real” test-driven development.
 - It’s still easy to miss regressions, like in older, unchanged taxonomy areas!
 - (We’ll come back to this...)

InstructLab

Cons (2/2)

- Still need server-side infrastructure for final tuning stage.
- While the InstructLab project is setting up the ability for community collaboration on models, for your private needs, you still need to tune yourself.
- Might be too expensive for tuning on each PR.

InstructLab

Pros

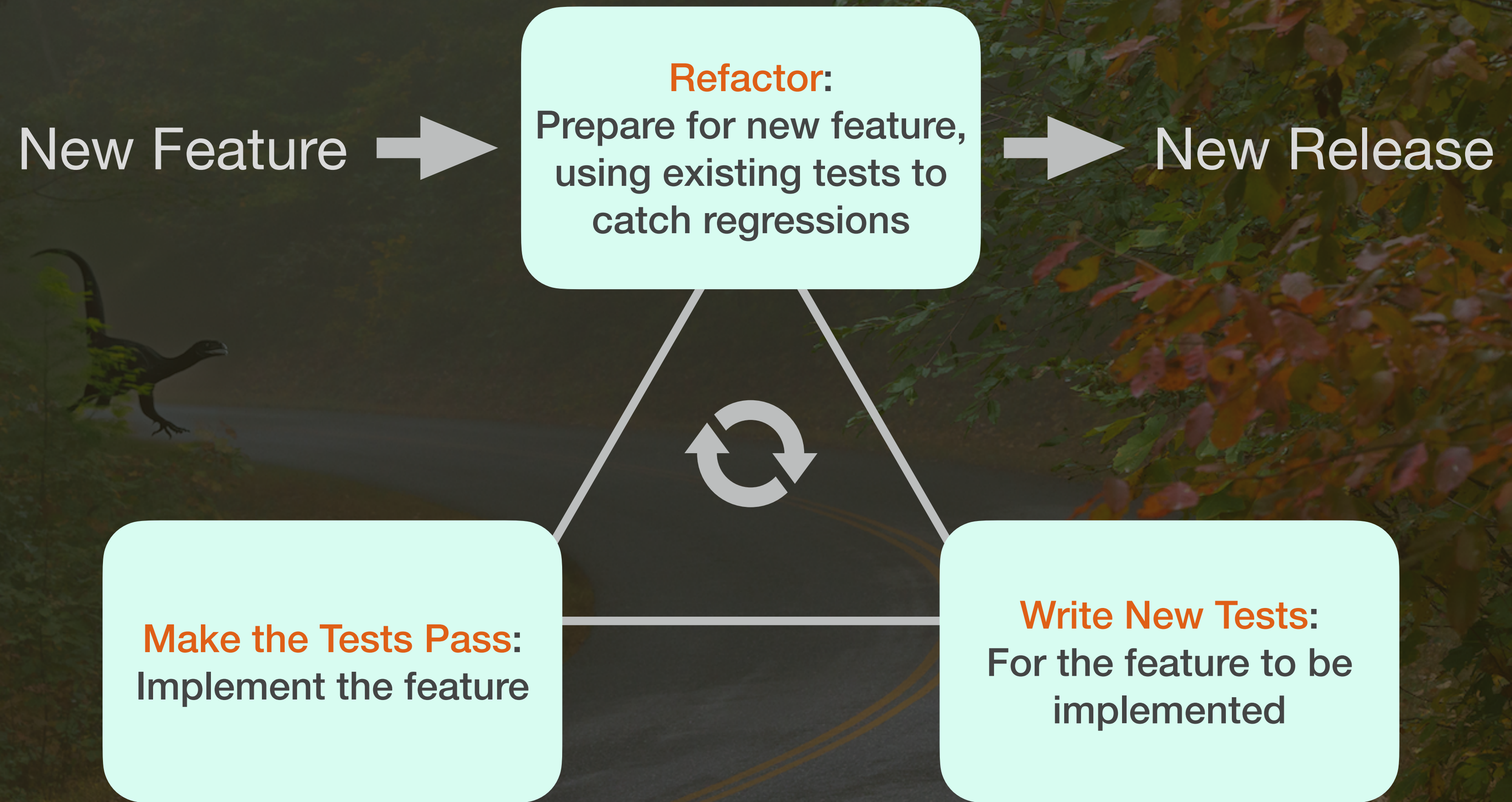
- Useful conventions for the taxonomy structure and Q&A examples for each taxonomy topic.
- `i lab` command hides and automates much of the grunt work for local, incremental steps.
- You can work locally and incrementally!

Automated Tests of Probabilistic Gen. AI??



Automated Tests of Probabilistic Gen. AI??

Remember this?



Testing is integral to this process.

What Do Developers Expect?

Developers expect software to be deterministic[‡]:

- The same input → the same output.
- e.g., $\sin(\pi) = -1$
- The output is different? Something is broken!
- Developers rely on determinism to help ensure correctness and reproducibility.

What Do Developers Expect?

Developers expect software to be deterministic[‡]:

- The s
- e.g.
- The o
- Deve
- corre

Put another way, the determinism makes it easier to specify the **system invariants**, what should remain true from one iteration to the next.

oken!
ensure

What's new with Gen. AI?

Generative models are probabilistic[‡]:

- The same prompt → **different** output.
- chatgpt("Write a poem") → **insanity**
- Without determinism, how do you write repeatable, reliable tests? Specifically,
- Is that new model actually better or worse than the old model?
- Did any regressions in other behavior occur?

"Insanity is doing the same thing over and over again and expecting different results."
— not Einstein

What's new with Gen. AI?

Generative models are probabilistic[‡]:

- The
- ch
- Wit
- rep
- Is
- th

Put another way, the system invariants are much less clear and therefore much less enforceable.

- Did any regressions in other behavior occur?

Are Automated Tests Possible with Gen. AI??

- Existing benchmarks alone aren't sufficient.
 - Would more specific, use case focused benchmarks help?
- We developers need help from you data scientists to build statistically-appropriate testing techniques.

Thank you!

Visit thealliance.ai at booth #129

I'll talk about InstructLab tomorrow: 10:15-11:15

I'm signing books in the O'Reilly booth at 3:30 today!

dwampler@thealliance.ai

Mastodon and Bluesky: @deanwampler

deanwampler.com/talks



Notes

© Text 2023-2024, Dean Wampler, © Images 2004-2024, Dean Wampler, except where noted. Most of the images are based on my photographs ([flickr.com/photos/deanwampler/](https://www.flickr.com/photos/deanwampler/)), but they are manipulated by AI to add “new content”.

1. The title image is adapted from from [this image](#) taken on a foggy day on the Blue Ridge Parkway.
2. The “Automated testing” image is from the same foggy day on the Blue Ridge Parkway (not on Flickr).
3. The “Iterative and Incremental Model Tuning” image is based on [this image](#) from the Oregon coast of a real horse and rider in the fog.
4. The “Thank you” slide uses [this Chicago Park image](#).