Can We Make Model Alignment a Software Engineering Process?

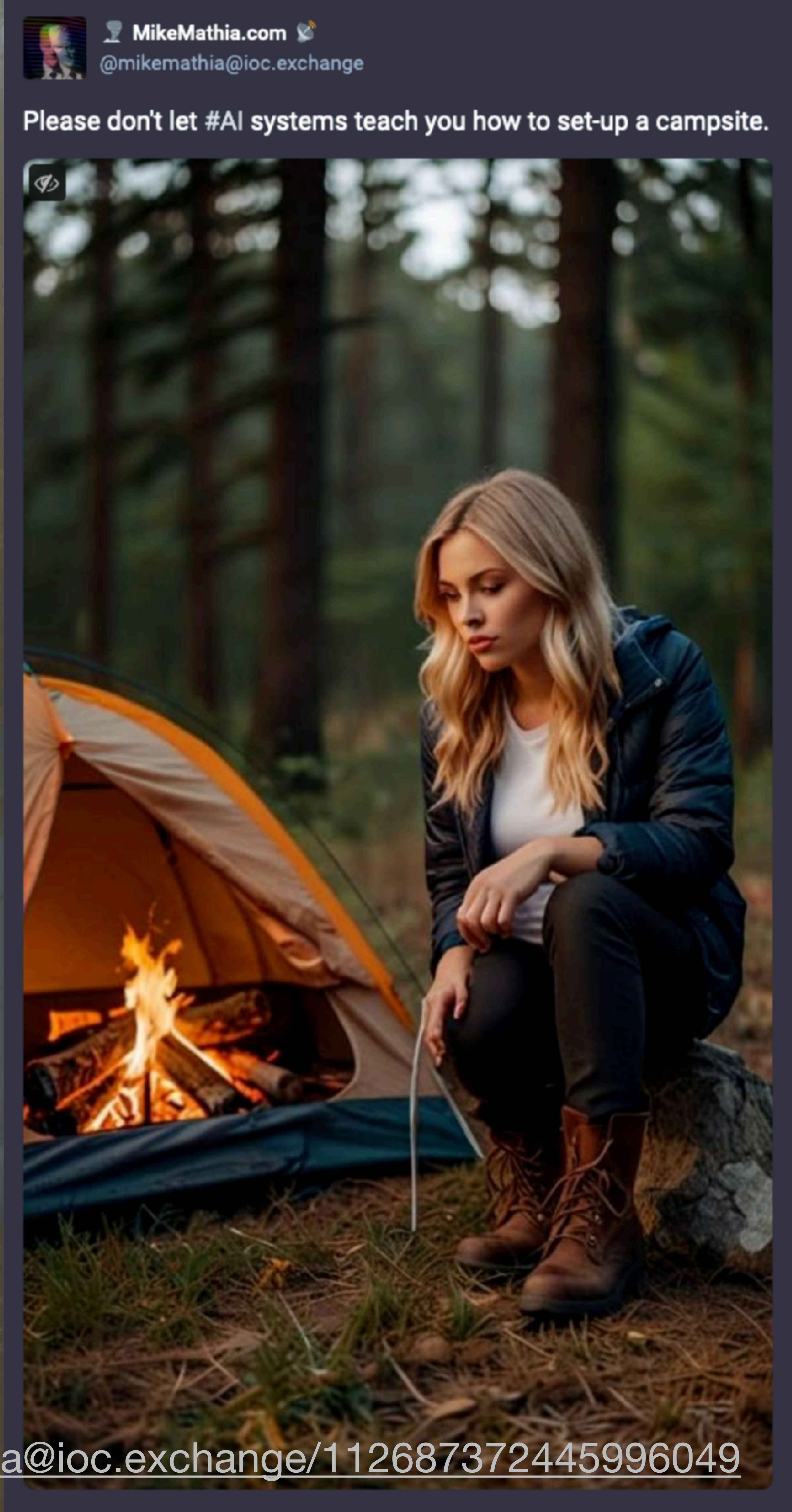
The AI Conference, San Francisco
September 2024
Dean Wampler, Ph.D.
The AI Alliance and IBM Research
thealliance.ai

deanwampler.com/talks

AI Alliance

About the Images…

I used Adobe Firefly to "enhance" my real photographs.



MikeMathia.com
@mikemathia@ioc.exchange

Please don't let #AI systems teach you how to set-up a campsite.

https://discuss.systems/@mikemathia@ioc.exchange/112687372445996049

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

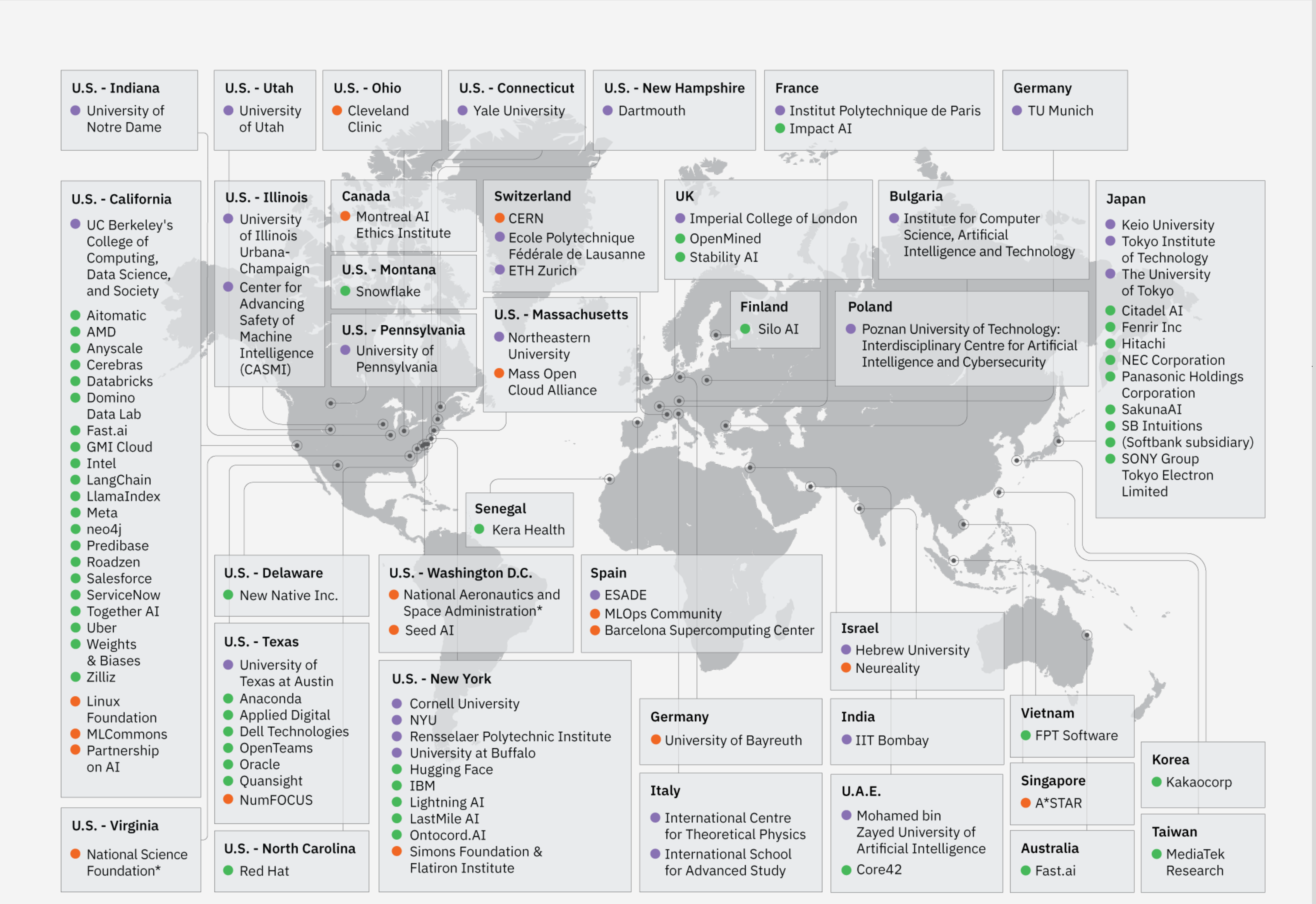
thealliance.ai

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



U.S. - Indiana
- University of Notre Dame

U.S. - Utah
- University of Utah

U.S. - Ohio
- Cleveland Clinic

U.S. - Connecticut
- Yale University

U.S. - New Hampshire
- Dartmouth

France
- Institut Polytechnique de Paris
- Impact AI

Germany
- TU Munich

U.S. - California
- UC Berkeley's College of Computing, Data Science, and Society
- Aitomatic
- AMD
- Anyscale
- Cerebras
- Databricks
- Domino Data Lab
- Fast.ai
- GMI Cloud
- Intel
- LangChain
- LlamaIndex
- Meta
- neo4j
- Predibase
- Roadzen
- Salesforce
- ServiceNow
- Together AI
- Uber
- Weights & Biases
- Zilliz
- Linux Foundation
- MLCommons
- Partnership on AI

U.S. - Illinois
- University of Illinois Urbana-Champaign
- Center for Advancing Safety of Machine Intelligence (CASMI)

Canada
- Montreal AI Ethics Institute

U.S. - Montana
- Snowflake

U.S. - Pennsylvania
- University of Pennsylvania

Switzerland
- CERN
- Ecole Polytechnique Fédérale de Lausanne
- ETH Zurich

U.S. - Massachusetts
- Northeastern University
- Mass Open Cloud Alliance

UK
- Imperial College of London
- OpenMined
- Stability AI

Finland
- Silo AI

Bulgaria
- Institute for Computer Science, Artificial Intelligence and Technology

Poland
- Poznan University of Technology: Interdisciplinary Centre for Artificial Intelligence and Cybersecurity

Japan
- Keio University
- Tokyo Institute of Technology
- The University of Tokyo
- Citadel AI
- Fenrir Inc
- Hitachi
- NEC Corporation
- Panasonic Holdings Corporation
- SakunaAI
- SB Intuitions
- (Softbank subsidiary)
- SONY Group
- Tokyo Electron Limited

Senegal
- Kera Health

U.S. - Delaware
- New Native Inc.

U.S. - Washington D.C.
- National Aeronautics and Space Administration*
- Seed AI

Spain
- ESADE
- MLOps Community
- Barcelona Supercomputing Center

U.S. - Texas
- University of Texas at Austin
- Anaconda
- Applied Digital
- Dell Technologies
- OpenTeams
- Oracle
- Quansight
- NumFOCUS

U.S. - New York
- Cornell University
- NYU
- Rensselaer Polytechnic Institute
- University at Buffalo
- Hugging Face
- IBM
- Lightning AI
- LastMile AI
- Ontocord.AI
- Simons Foundation & Flatiron Institute

Israel
- Hebrew University
- Neureality

Germany
- University of Bayreuth

India
- IIT Bombay

Vietnam
- FPT Software

Korea
- Kakaocorp

Singapore
- A*STAR

Italy
- International Centre for Theoretical Physics
- International School for Advanced Study

U.A.E.
- Mohamed bin Zayed University of Artificial Intelligence
- Core42

Australia
- Fast.ai

Taiwan
- MediaTek Research

U.S. - Virginia
- National Science Foundation*

U.S. - North Carolina
- Red Hat

AI Alliance

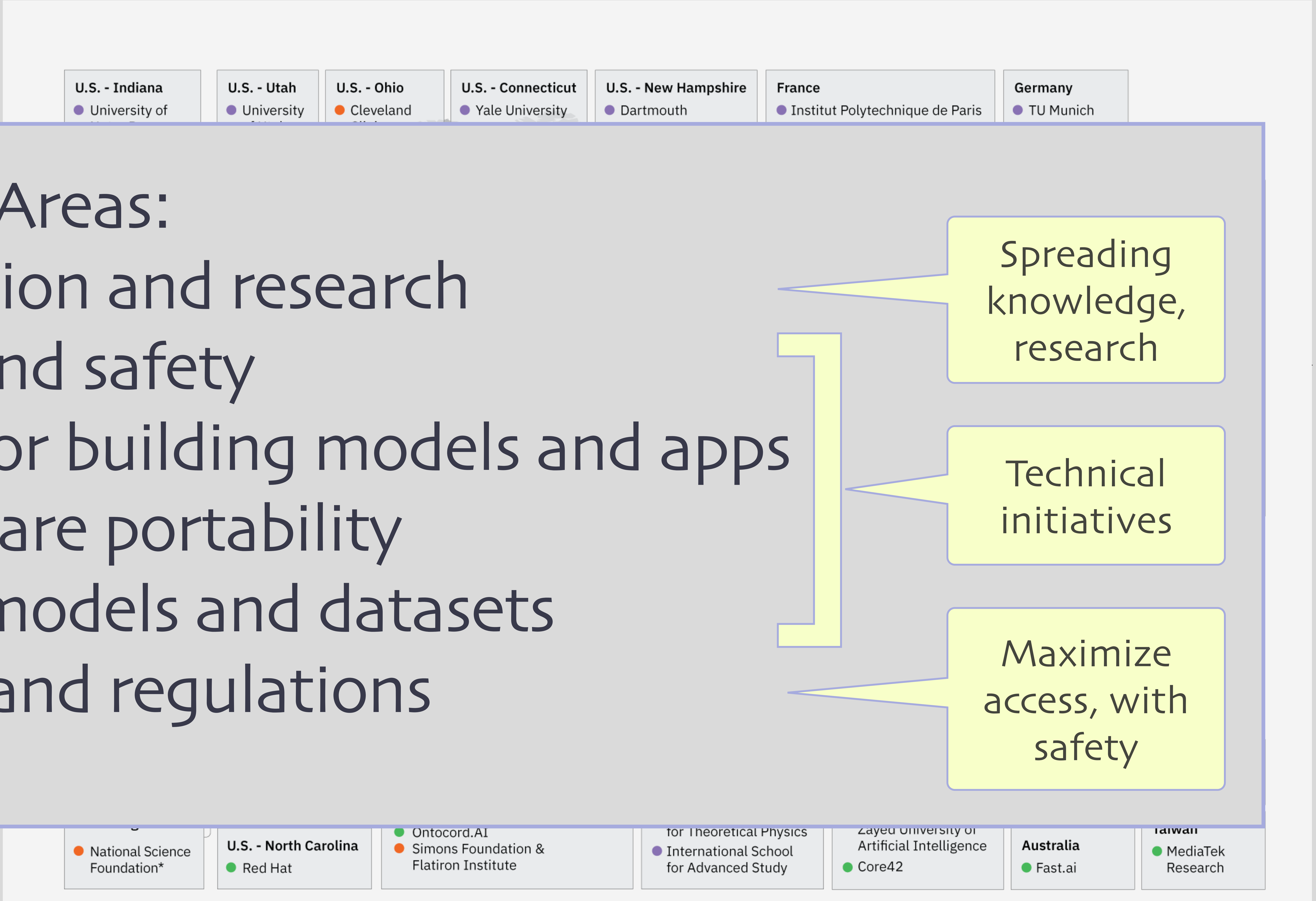
thealliance.ai

Our
ope
us,

Mer
aca
rese
anc

Vi

+100 organizations in +20 countries, and growing



- U.S. - Indiana
  - ● University of
- U.S. - Utah
  - ● University
- U.S. - Ohio
  - ● Cleveland
- U.S. - Connecticut
  - ● Yale University
- U.S. - New Hampshire
  - ● Dartmouth
- France
  - ● Institut Polytechnique de Paris
- Germany
  - ● TU Munich

- ● National Science Foundation*
- U.S. - North Carolina
  - ● Red Hat
- ● Ontocord.AI
- ● Simons Foundation & Flatiron Institute
- for Theoretical Physics
- ● International School for Advanced Study
- Zayed University of Artificial Intelligence
- ● Core42
- Australia
  - ● Fast.ai
- Taiwan
  - ● MediaTek Research

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

Spreading knowledge, research

Technical initiatives

Maximize access, with safety

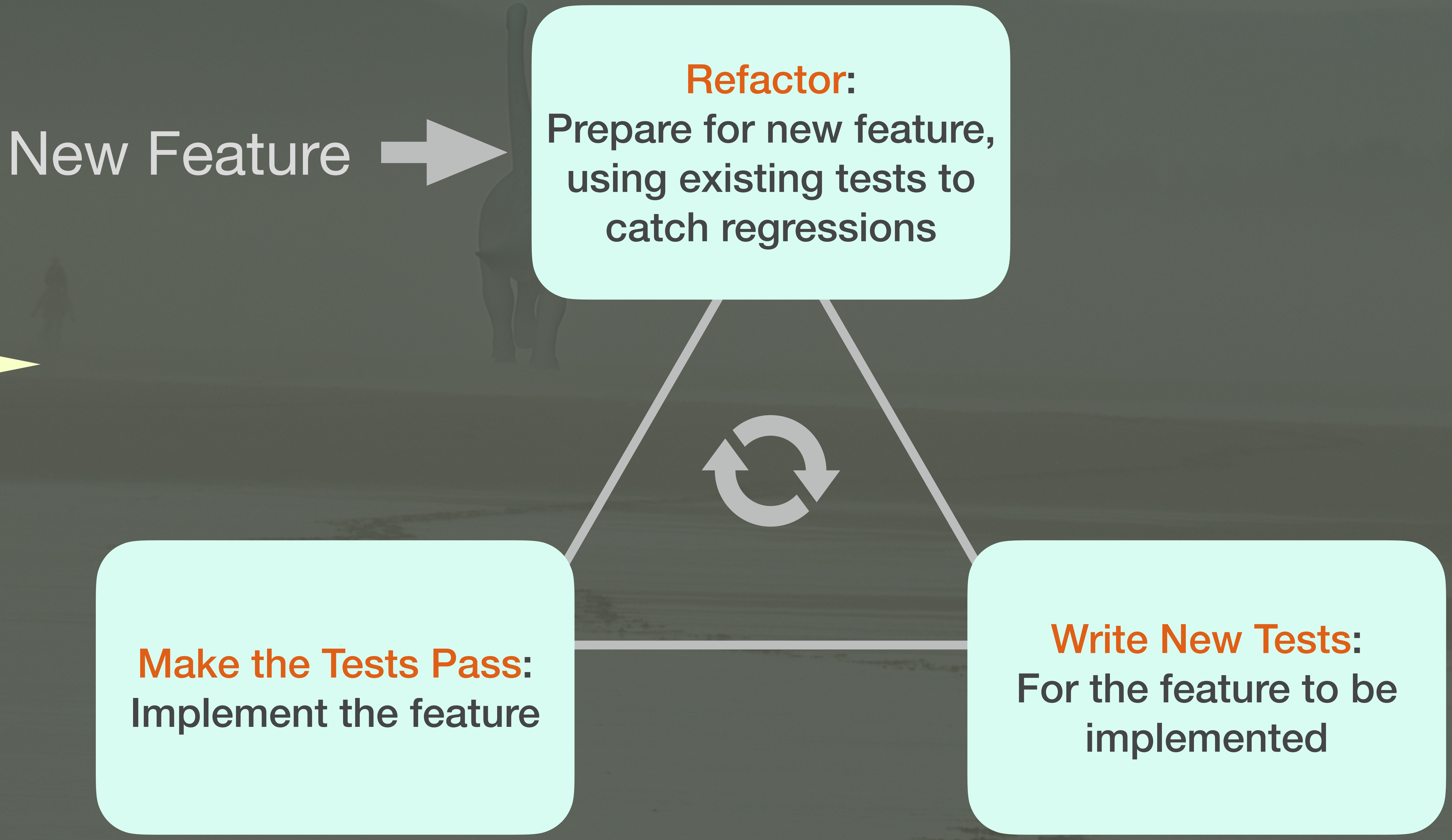
Iterative and Incremental Model Tuning

AI Alliance

What Software Developers Like

- Features are added incrementally.

This is done on your laptop, with final validation in your "GitFlow"



New Feature

Refactor: Prepare for new feature, using existing tests to catch regressions

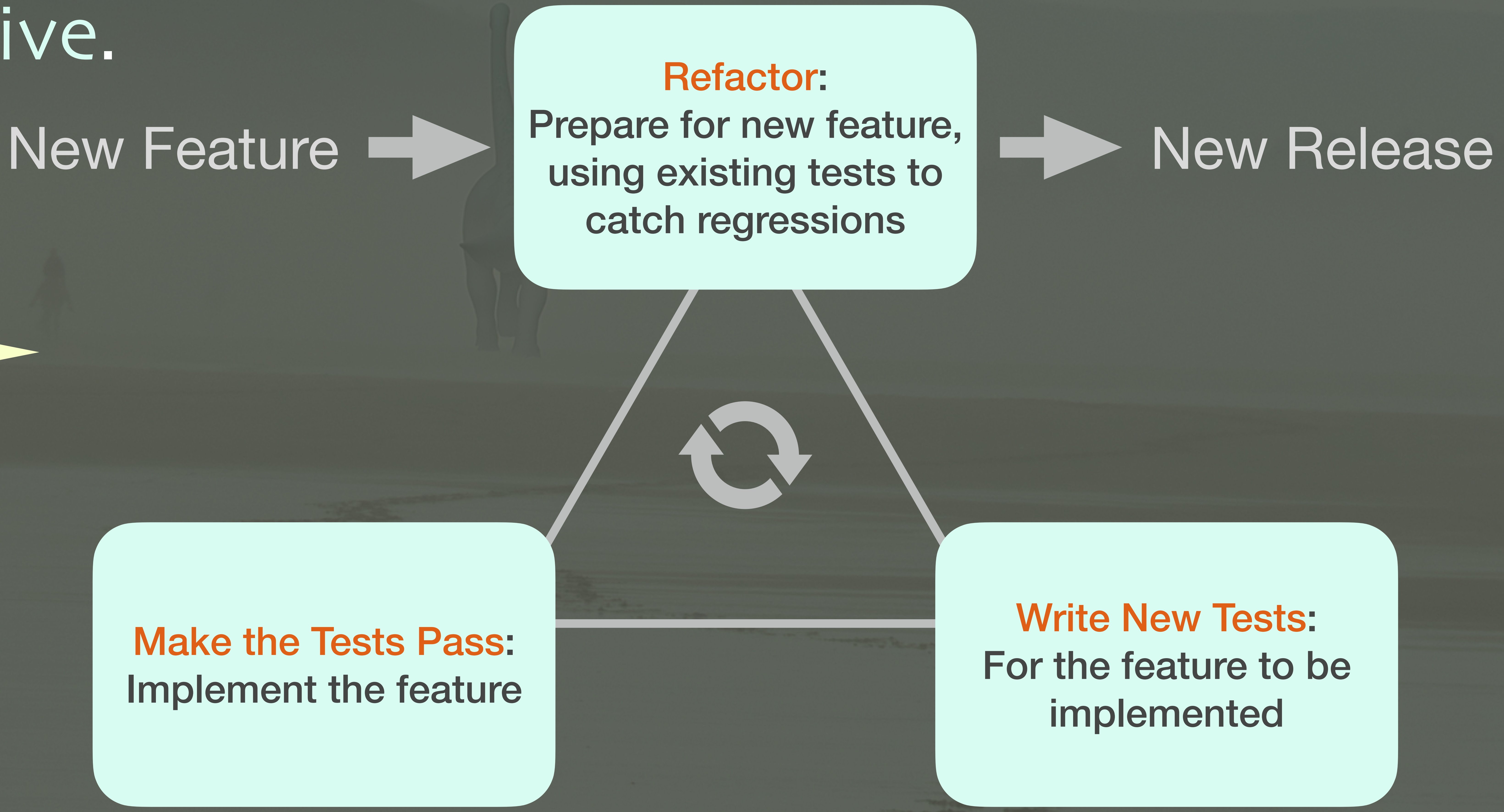
Write New Tests: For the feature to be implemented

Make the Tests Pass: Implement the feature

What Software Developers Like

- Features are added incrementally.
- Releases are iterative.

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



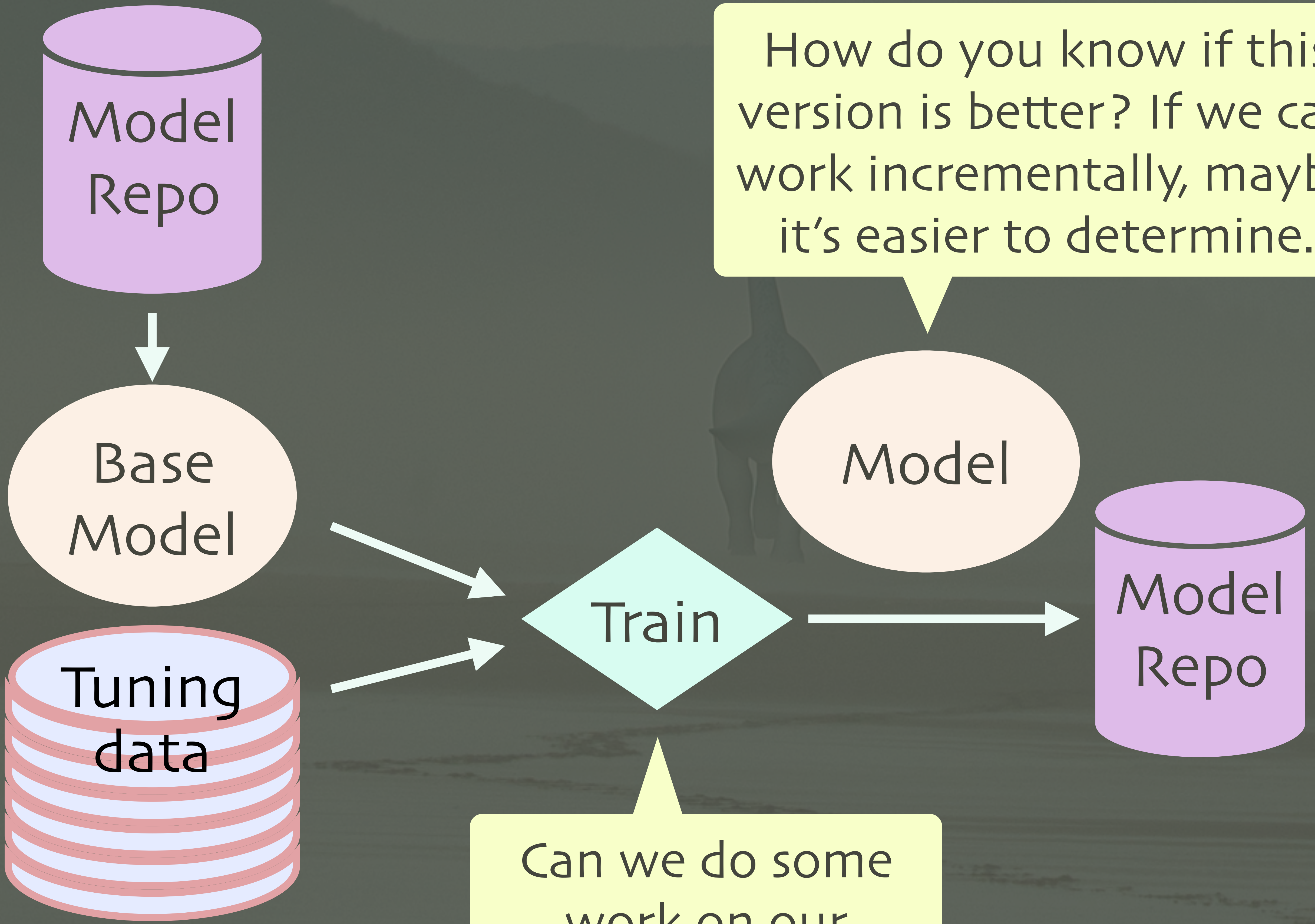
New Feature → **Refactor**: Prepare for new feature, using existing tests to catch regressions → New Release

Write New Tests: For the feature to be implemented

Make the Tests Pass: Implement the feature

What Model Tuning Is Often Like

Can we make this incremental, iterative, and local to a laptop with a final “GitFlow”-like verification and completion?



How do you know if this version is better? If we can work incrementally, maybe it’s easier to determine.

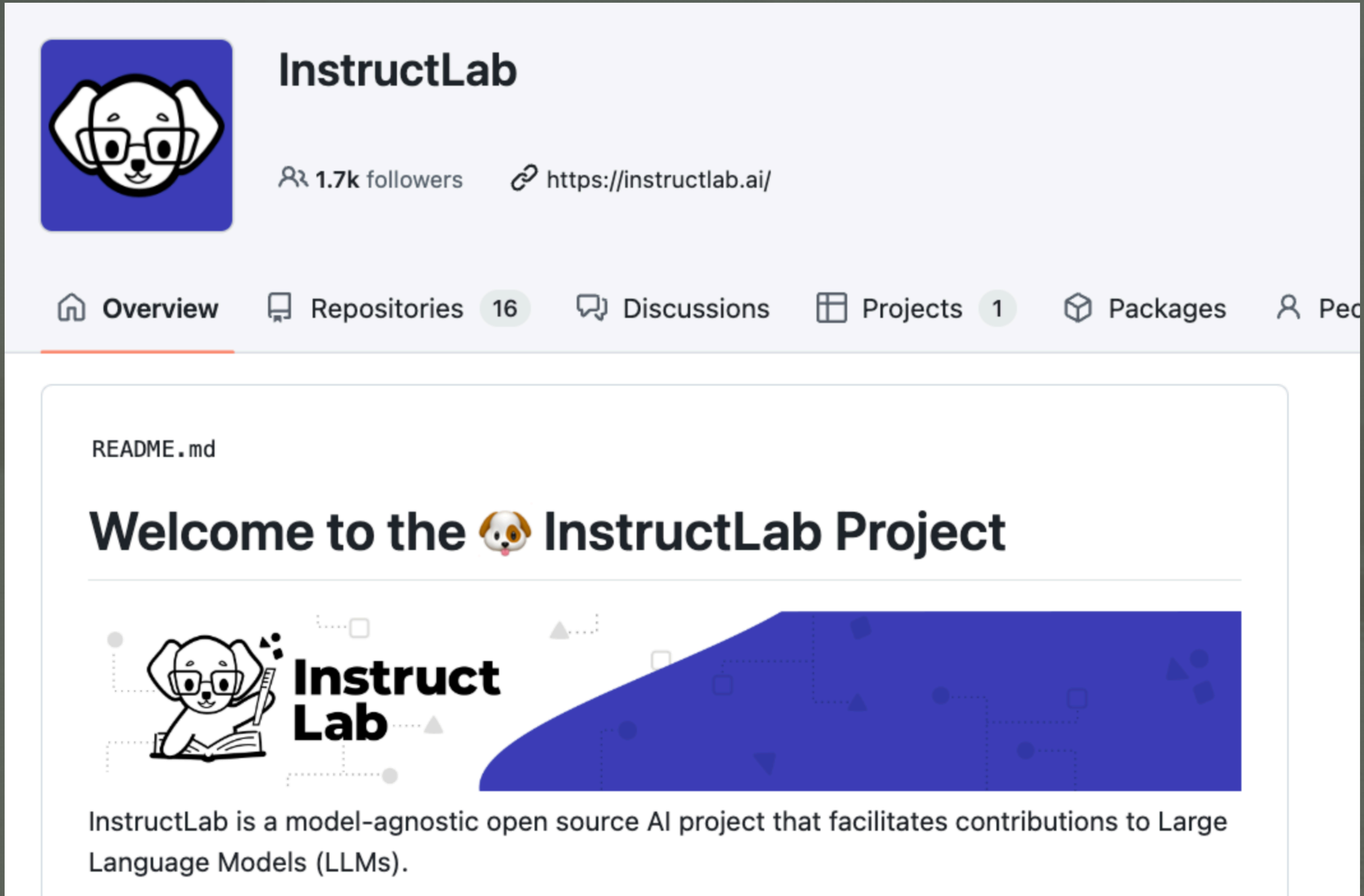
Ad hoc size and organization. Can we structure the data into “modules”?

Can we do some work on our laptops, then finish in the cloud?

One Approach: InstructLab

https://github.com/instructlab

Open sourced by IBM and Red Hat



InstructLab

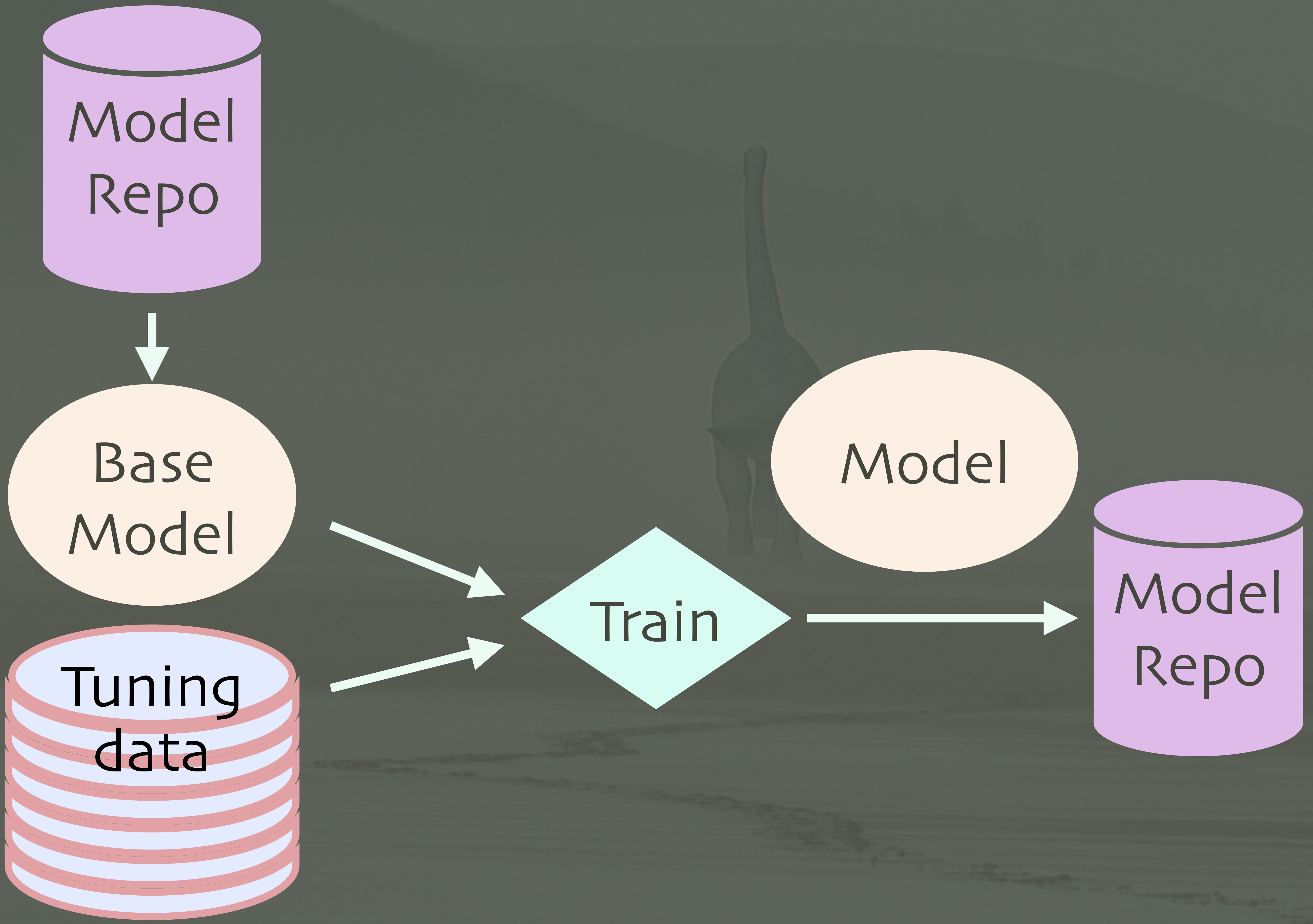
1.7k followers https://instructlab.ai/

README.md

Welcome to the 🐶 InstructLab Project

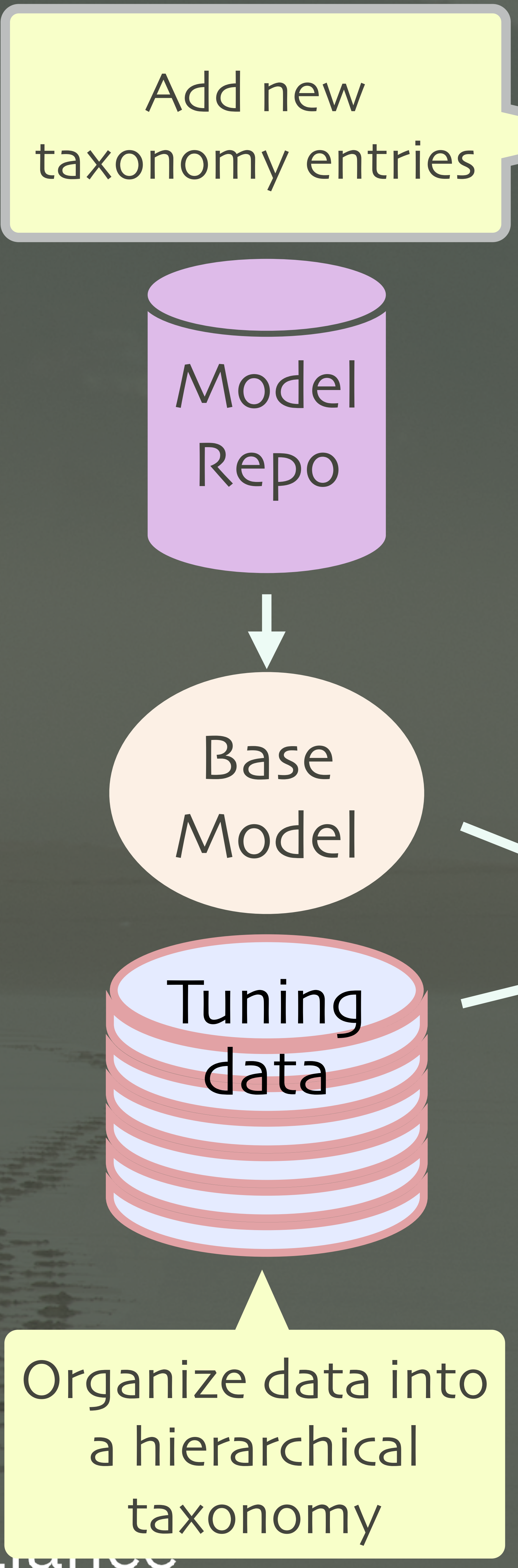
InstructLab is a model-agnostic open source AI project that facilitates contributions to Large Language Models (LLMs).

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



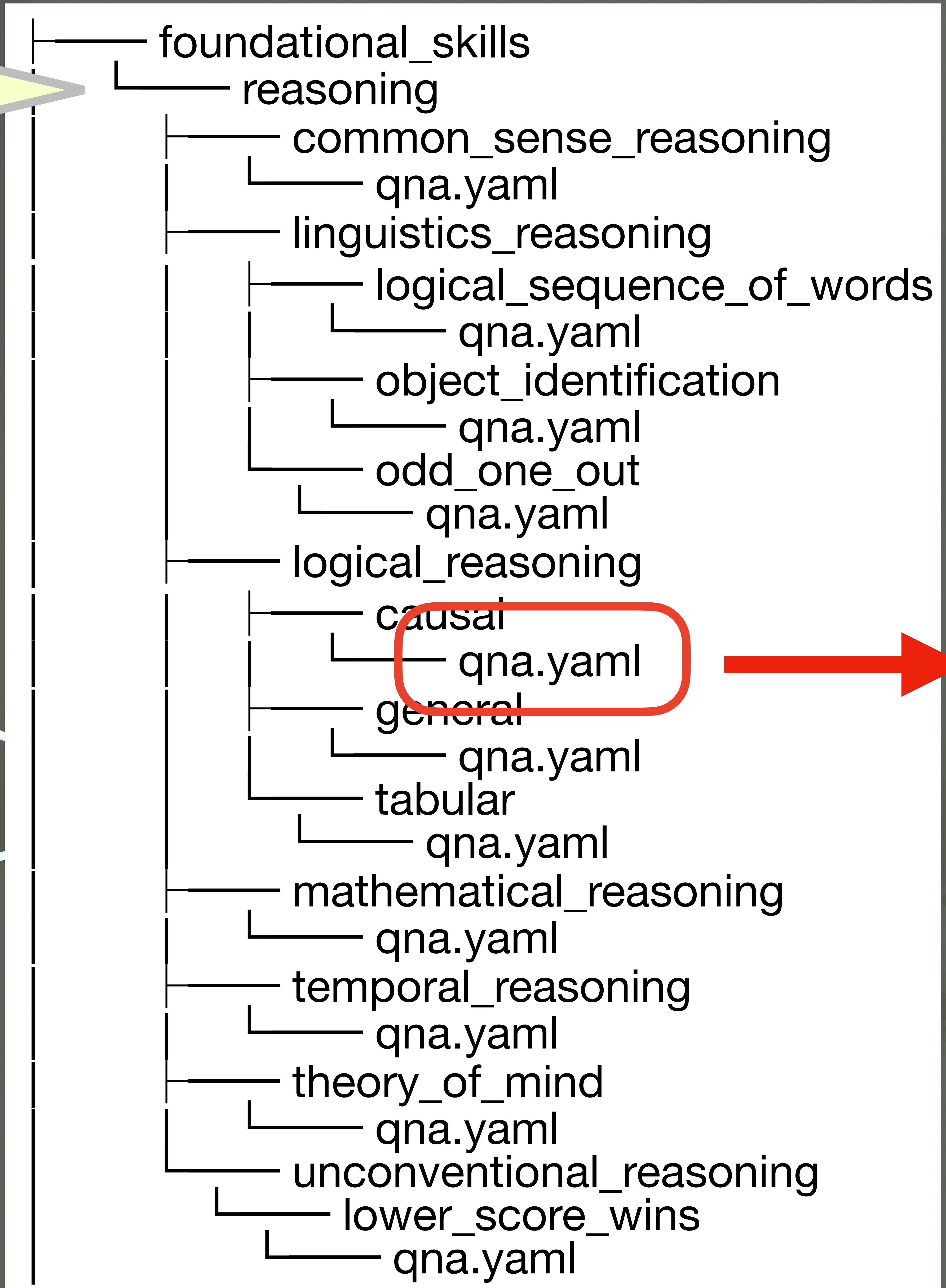
Pick your favorite base model

AI Alliance

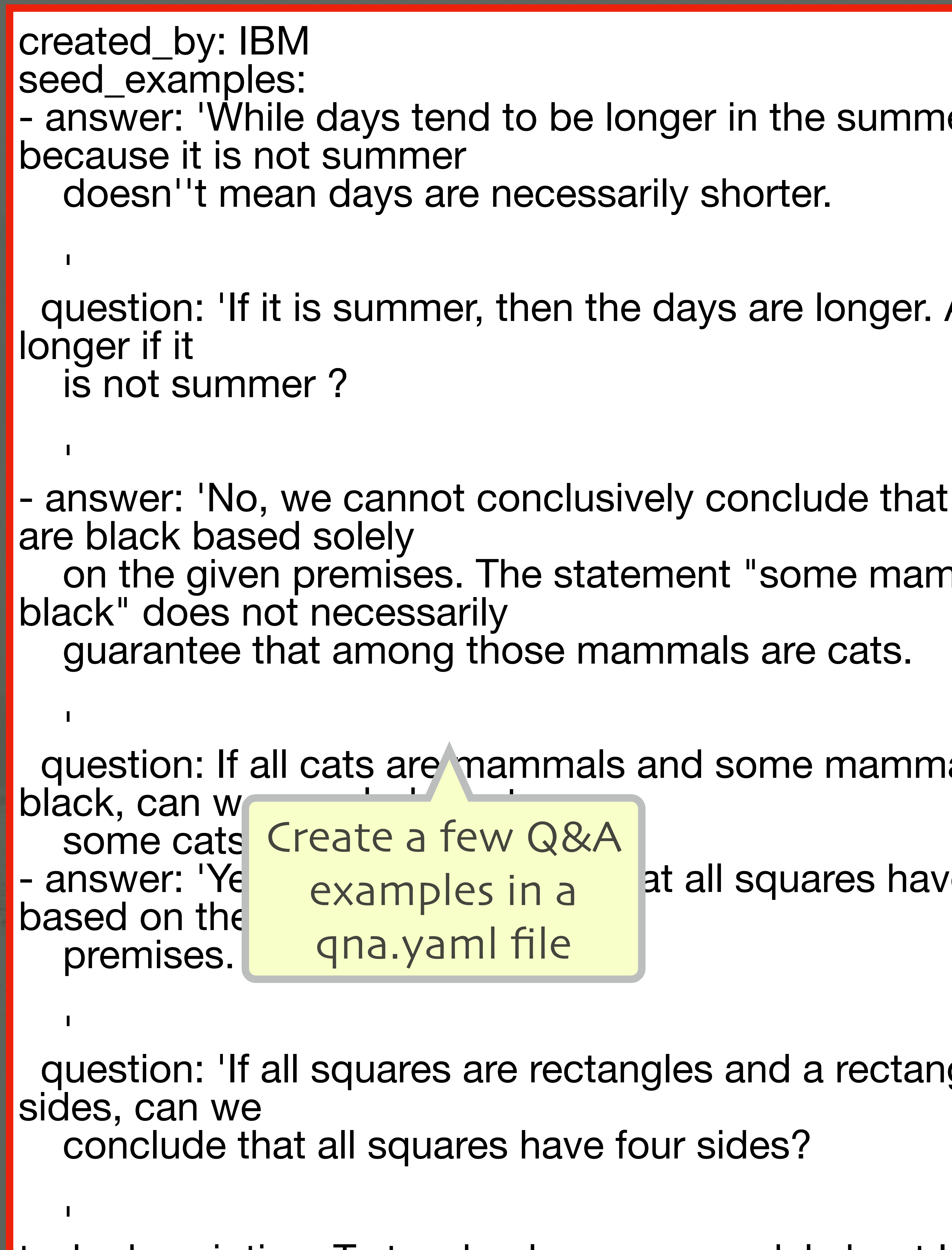


Add new taxonomy entries

Organize data into a hierarchical taxonomy

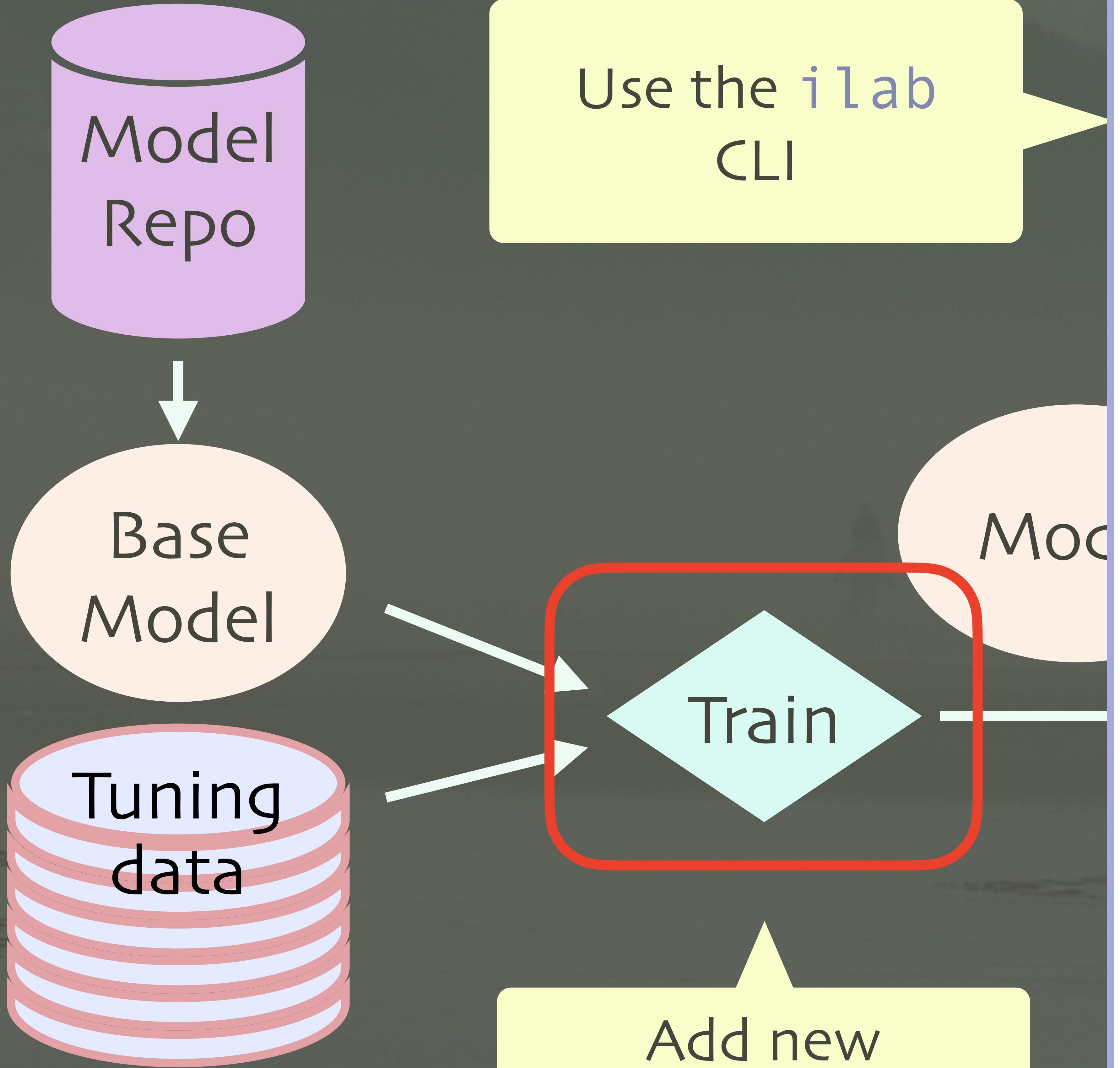


```
├──── foundational_skills
│    └──── reasoning
│          ├──── common_sense_reasoning
│          │    └──── qna.yaml
│          ├──── linguistics_reasoning
│          │    ├──── logical_sequence_of_words
│          │    │    └──── qna.yaml
│          │    ├──── object_identification
│          │    │    └──── qna.yaml
│          │    └──── odd_one_out
│          │         └──── qna.yaml
│          ├──── logical_reasoning
│          │    ├──── causal
│          │    │    └──── qna.yaml
│          │    ├──── general
│          │    │    └──── qna.yaml
│          │    └──── tabular
│          │         └──── qna.yaml
│          ├──── mathematical_reasoning
│          │    └──── qna.yaml
│          ├──── temporal_reasoning
│          │    └──── qna.yaml
│          ├──── theory_of_mind
│          │    └──── qna.yaml
│          └──── unconventional_reasoning
│               └──── lower_score_wins
│                    └──── qna.yaml
```



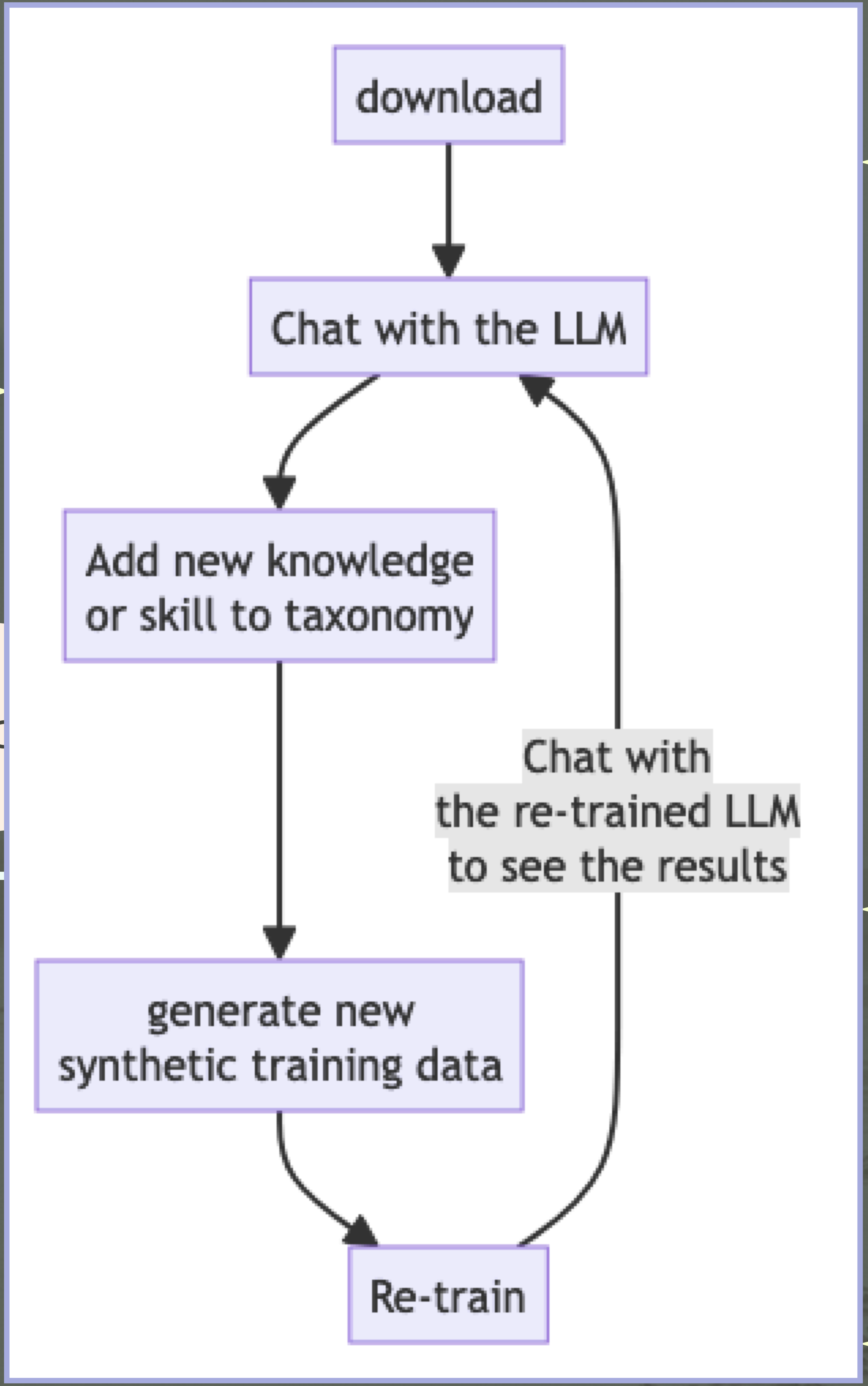
```
created_by: IBM
seed_examples:
- answer: 'While days tend to be longer in the summe
because it is not summer
    doesn''t mean days are necessarily shorter.

    '
  question: 'If it is summer, then the days are longer. A
longer if it
    is not summer ?

    '
- answer: 'No, we cannot conclusively conclude that
are black based solely
    on the given premises. The statement "some mam
black" does not necessarily
    guarantee that among those mammals are cats.

    '
  question: If all cats are mammals and some mamma
black, can w
    some cats
- answer: 'Ye                at all squares hav
based on the
    premises.

    '
  question: 'If all squares are rectangles and a rectan
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

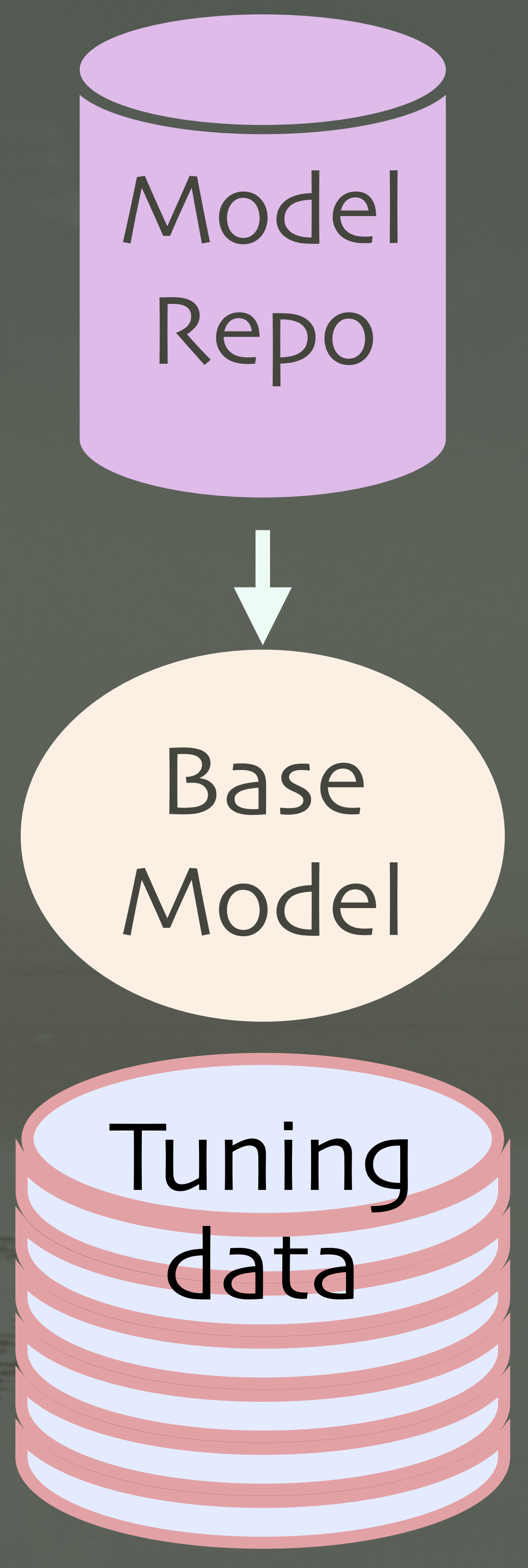

download
Chat with the LLM
Add new knowledge or skill to taxonomy
generate new synthetic training data
Re-train
Chat with the re-trained LLM to see the results

Download a quantized model version suitable for local execution

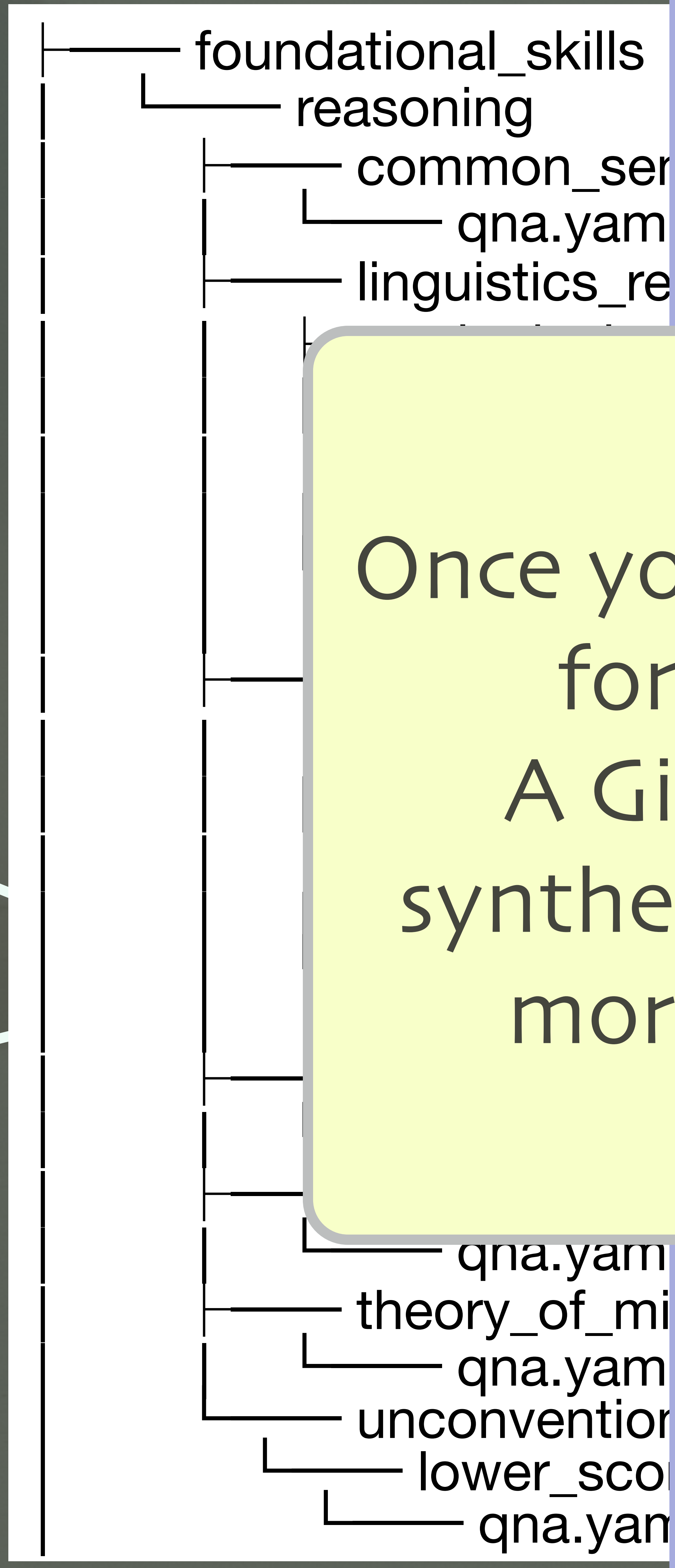
Results locally will be low fidelity

Uses QLoRA for efficiency

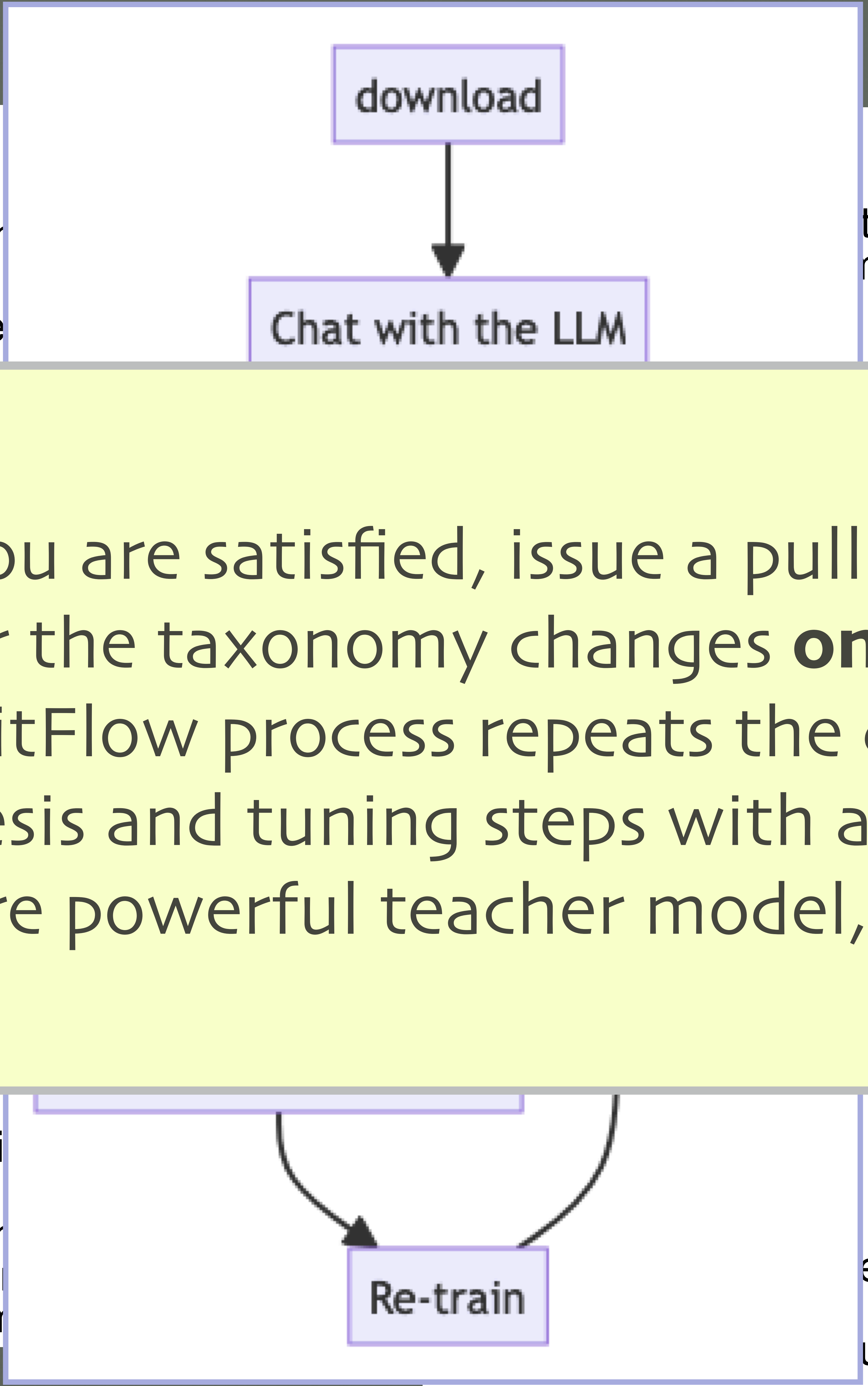
Image: https://github.com/instructlab/instructlab

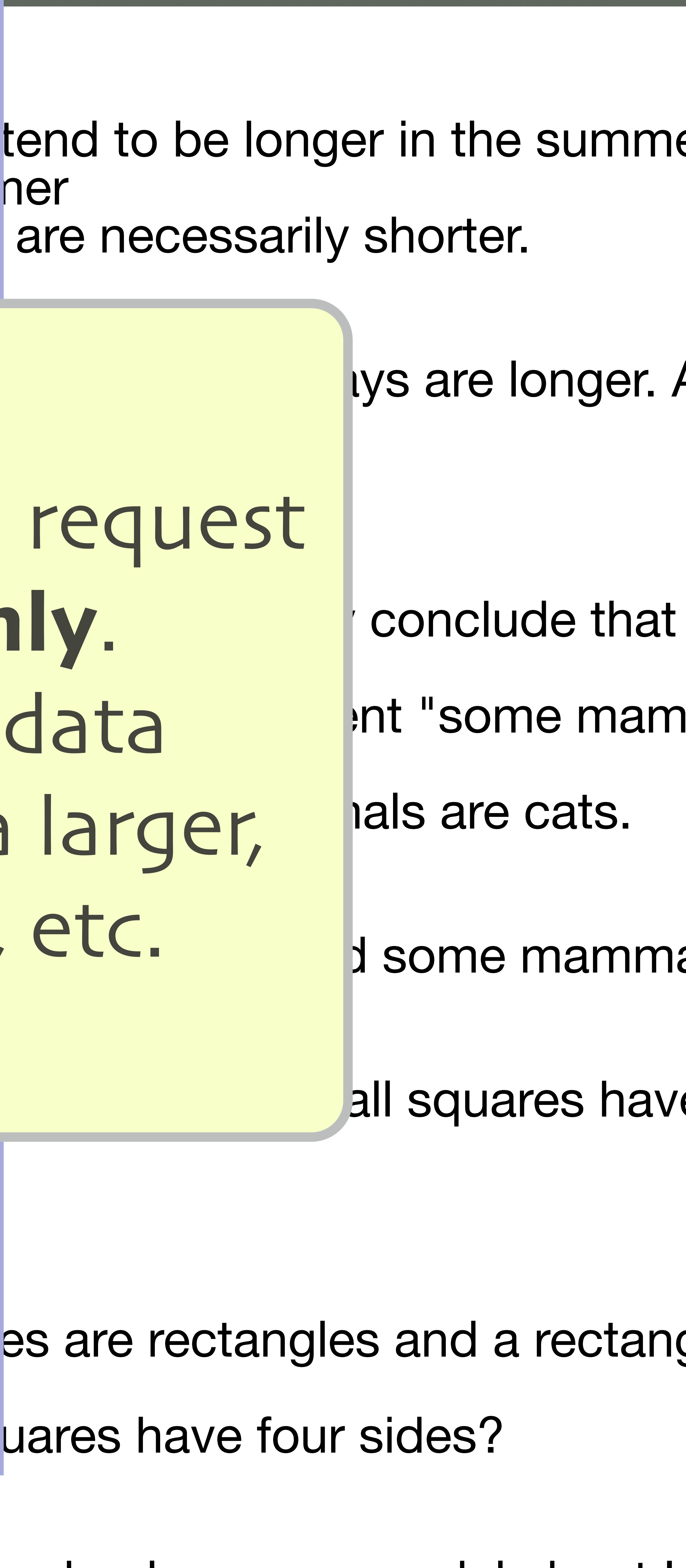


Organize data into a hierarchical taxonomy



```
├──── foundational_skills
│     └──── reasoning
│           ├──── common_ser
│           │     └──── qna.yam
│           ├──── linguistics_re
│           │
│           │
│           │     └──── qna.yam
│           ├──── theory_of_mi
│           │     └──── qna.yam
│           └──── unconvention
│                 └──── lower_sco
│                       └──── qna.yam
```





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

ys are longer. A

conclude that

nt "some mam

als are cats.

d some mamma

all squares have

es are rectangles and a rectang

uares have four sides?

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.

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

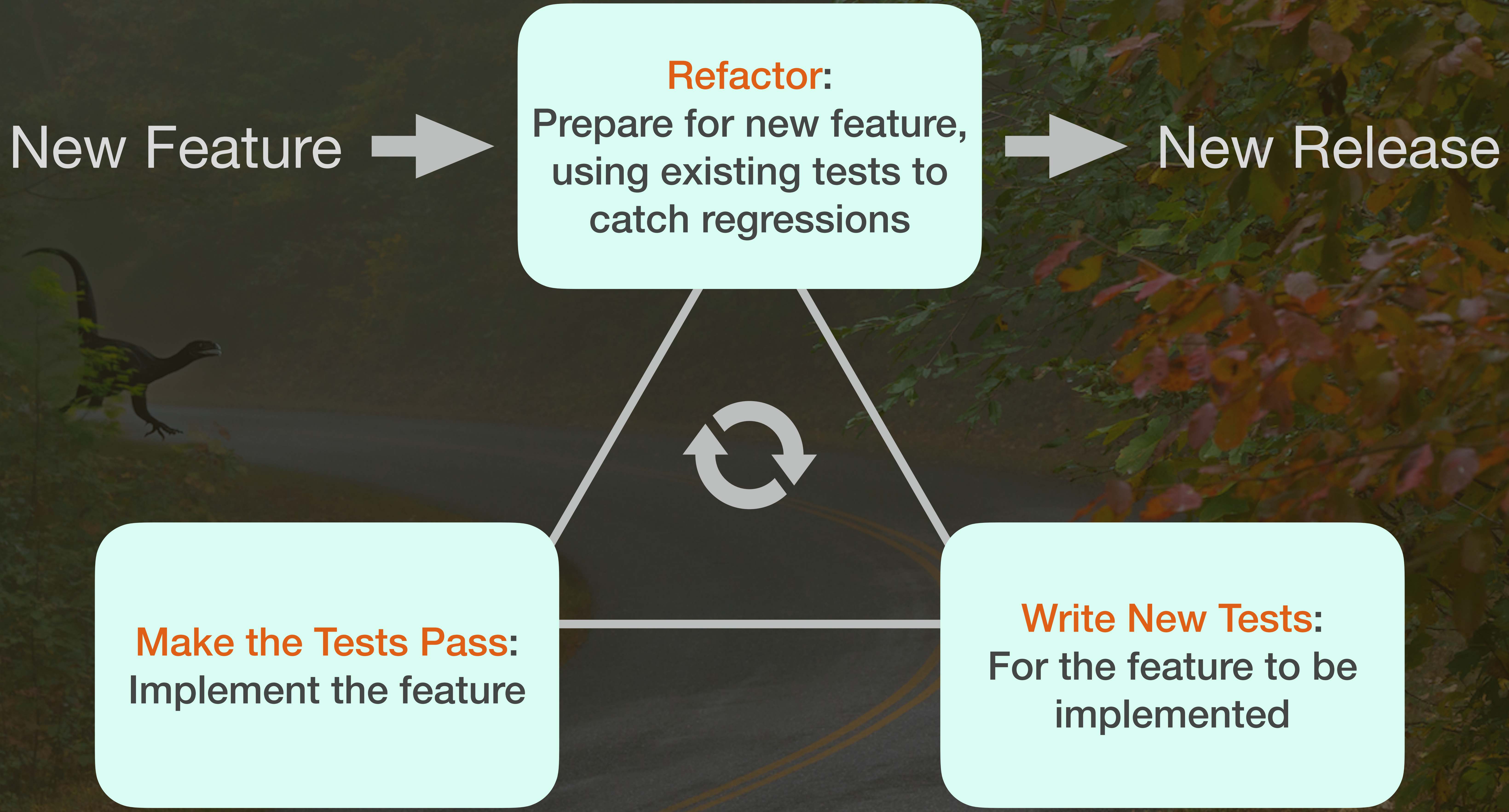
- Useful conventions for the taxonomy structure and Q&A examples for each taxonomy topic.
- `ilab` command hides and automates much of the grunt work for local, incremental steps.
- You can work locally and incrementally!

https://github.com/instructlab

Automated Tests of Probabilistic Gen. AI??

Automated Tests of Probabilistic Gen. AI??

Remember this?



New Feature

Refactor: Prepare for new feature, using existing tests to catch regressions

New Release

Make the Tests Pass: Implement the feature

Write New Tests: For the feature to be implemented

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(π) = -1
- The output is different? Something is broken!
- Developers rely on determinism to help ensure correctness and reproducibility.

[‡] Distributed systems break this clean picture.

What Do Developers Expect?

Developers expect software to be deterministic‡:

- The s
  - e.g.
- The c oken!
- Deve ensure corre

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

‡ Distributed systems break this clean picture.

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

‡ A tunable “temperature” controls how probabilistic.

What's new with Gen. AI?

Generative models are probabilistic[‡]:

- The
  - ch
- Wit

  repe
  - Is … rse

    th
  - Did any regressions in other behavior occur?

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

[‡] A tunable "temperature" controls how probabilistic.

AI Alliance

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

AI Alliance

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/), 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.