

Can We Make Model Alignment a Software Engineering Process?

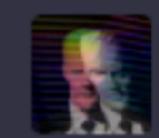
The AI Conference, San Francisco
September 2024

Dean Wampler, Ph.D.

The AI Alliance and IBM Research
thealliance.ai

deanwampler.com/talks



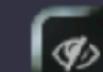


MikeMathia.com
@mikemathia@ioc.exchange

About the Images...

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

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



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



AI Alliance

AI and 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!



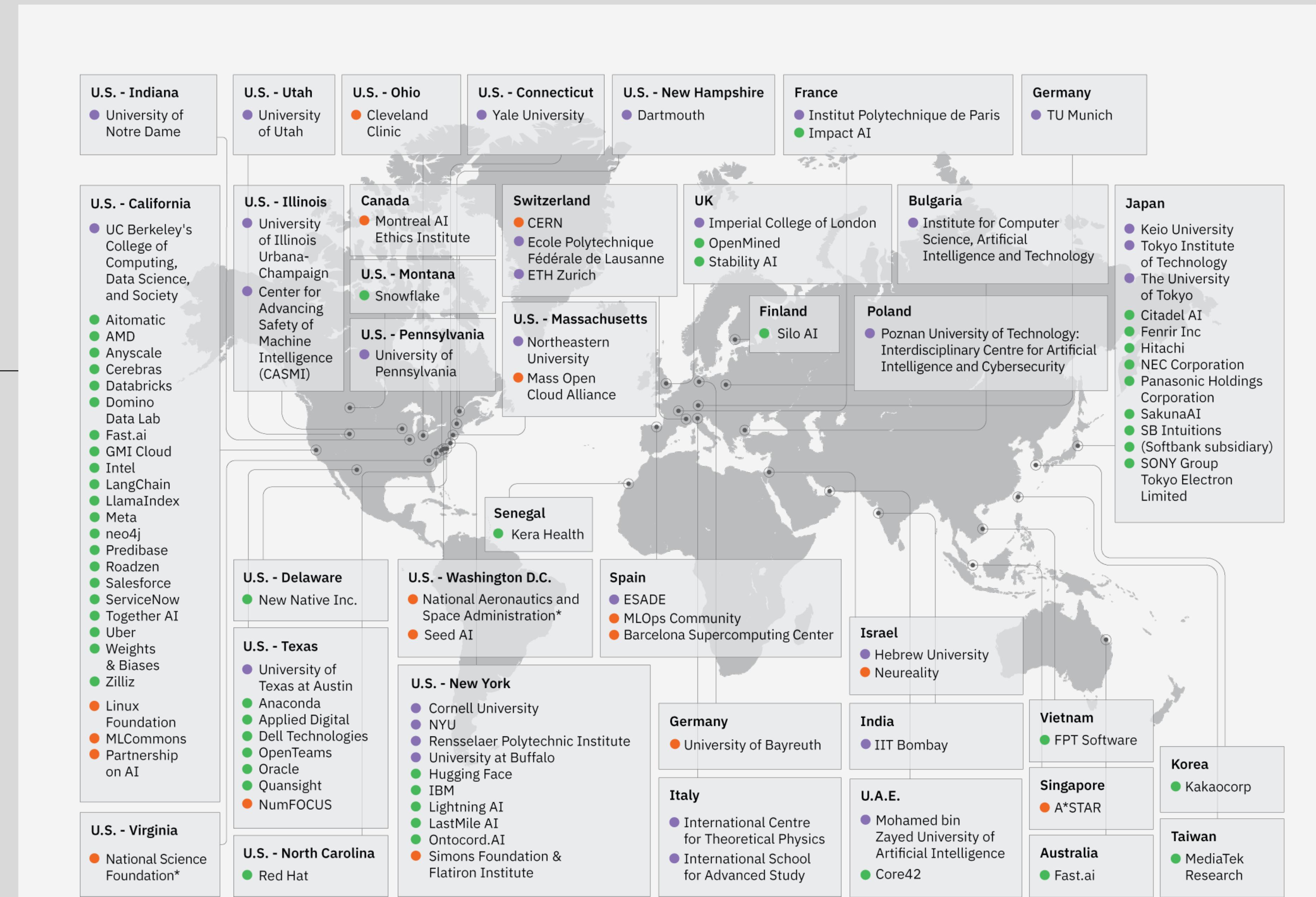
AI Alliance

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.

+100 organizations in +20 countries, and growing



thealliance.ai

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

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 advocacy

Spreading knowledge, research

Technical initiatives

Effective, measured regulations

• 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

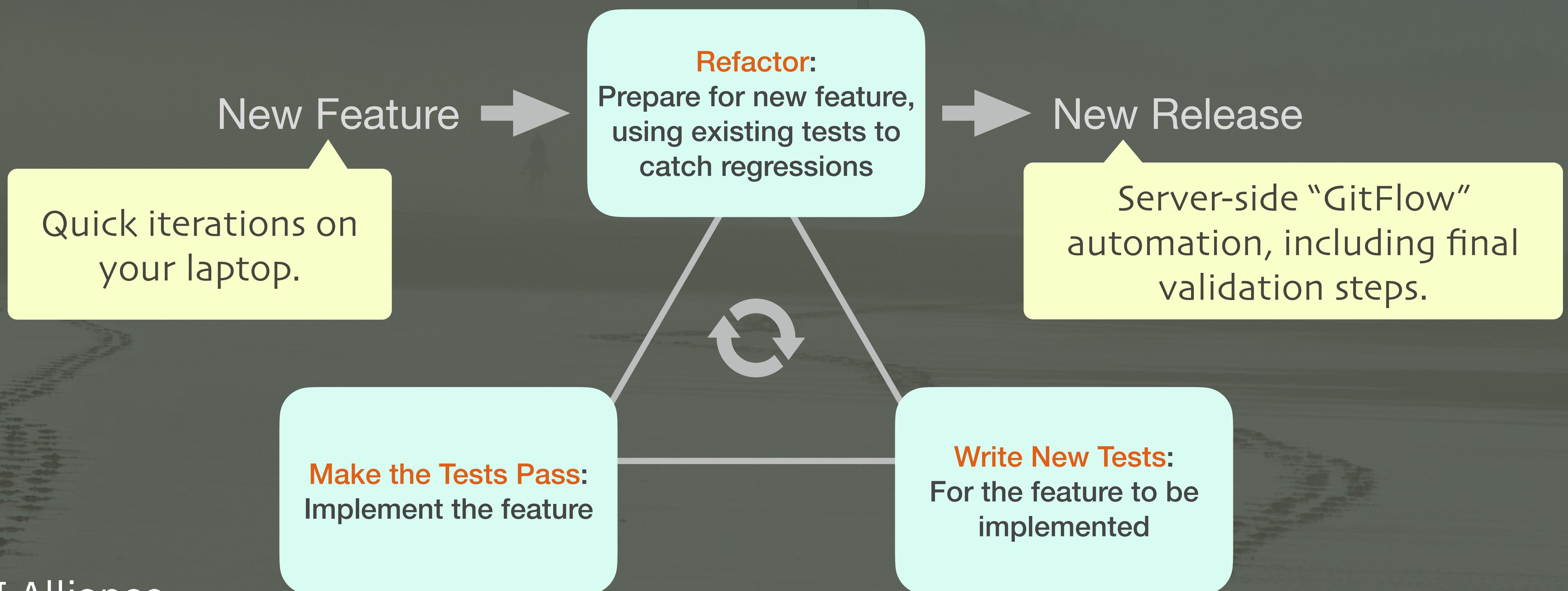
Taiwan
• MediaTek Research

Iterative and Incremental Model Tuning

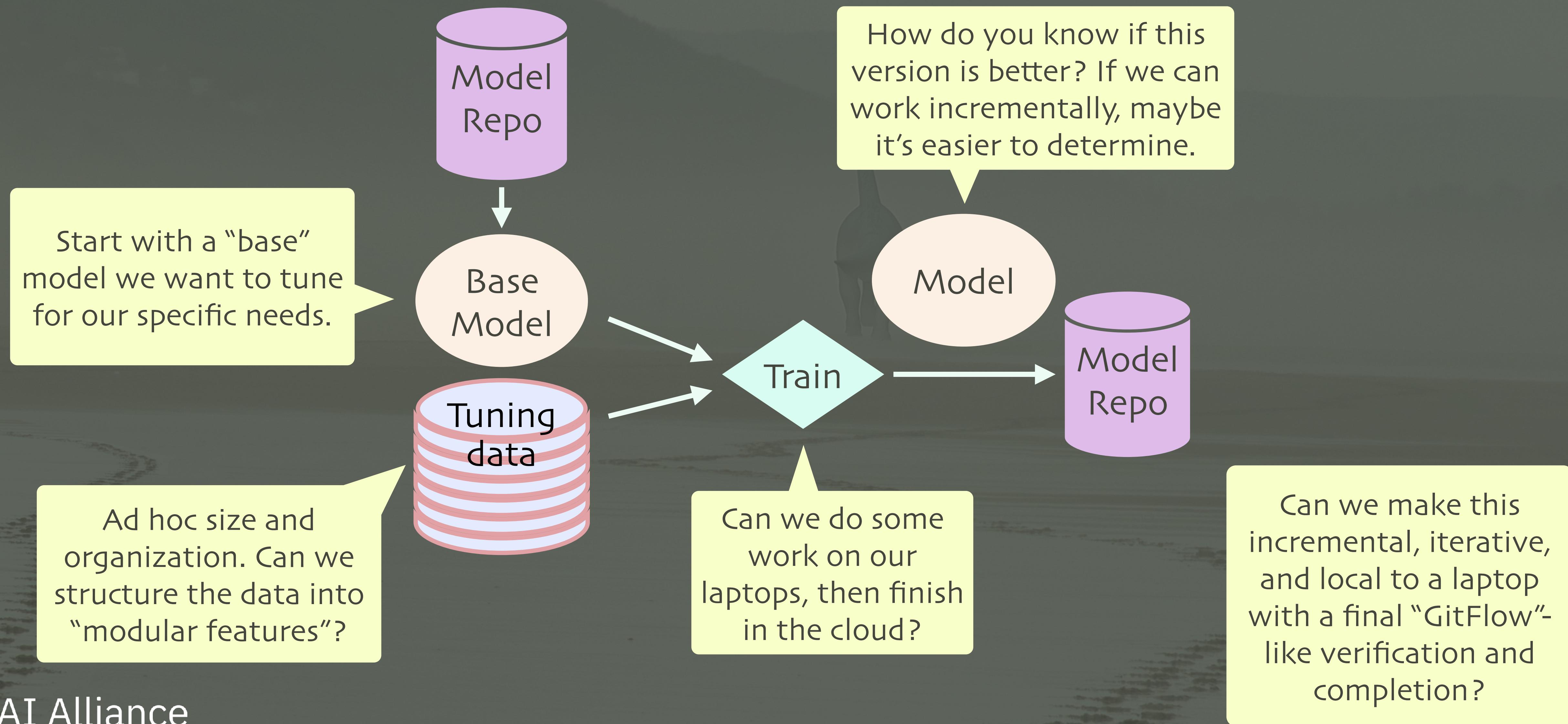


What Software Developers Like

- Features are implemented and released incrementally.
- The process is iterative.



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 repository page for InstructLab. The repository has 1.7k followers and a link to https://instructlab.ai/. The main content area features a README.md file and a large banner image. The banner image includes the text "Welcome to the 🐕 InstructLab Project" and a logo of a dog wearing glasses reading a book. Below the banner, a description states: "InstructLab is a model-agnostic open source AI project that facilitates contributions to Large Language Models (LLMs)."

InstructLab

1.7k followers <https://instructlab.ai/>

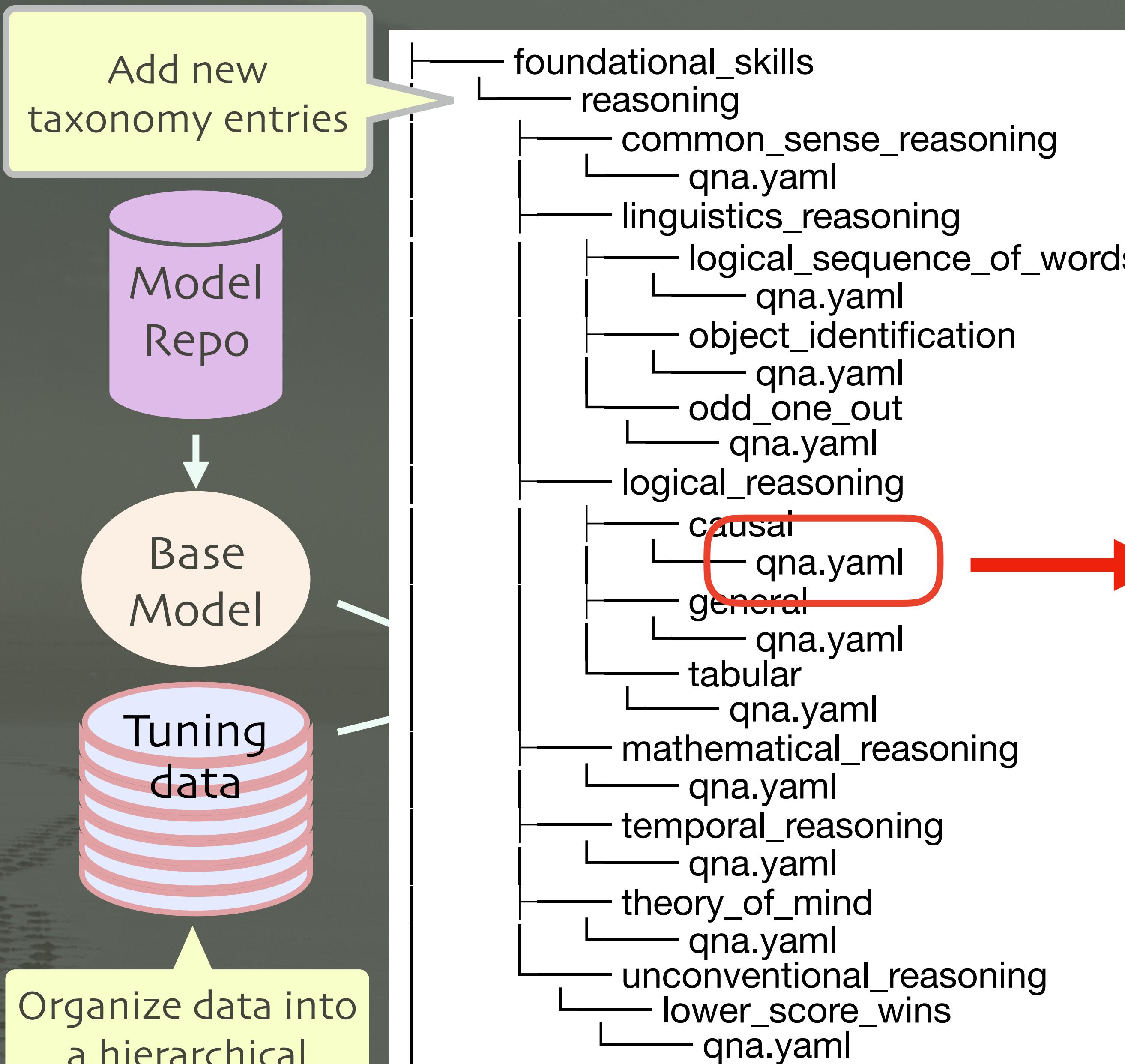
Overview Repositories 16 Discussions Projects 1 Packages People

README.md

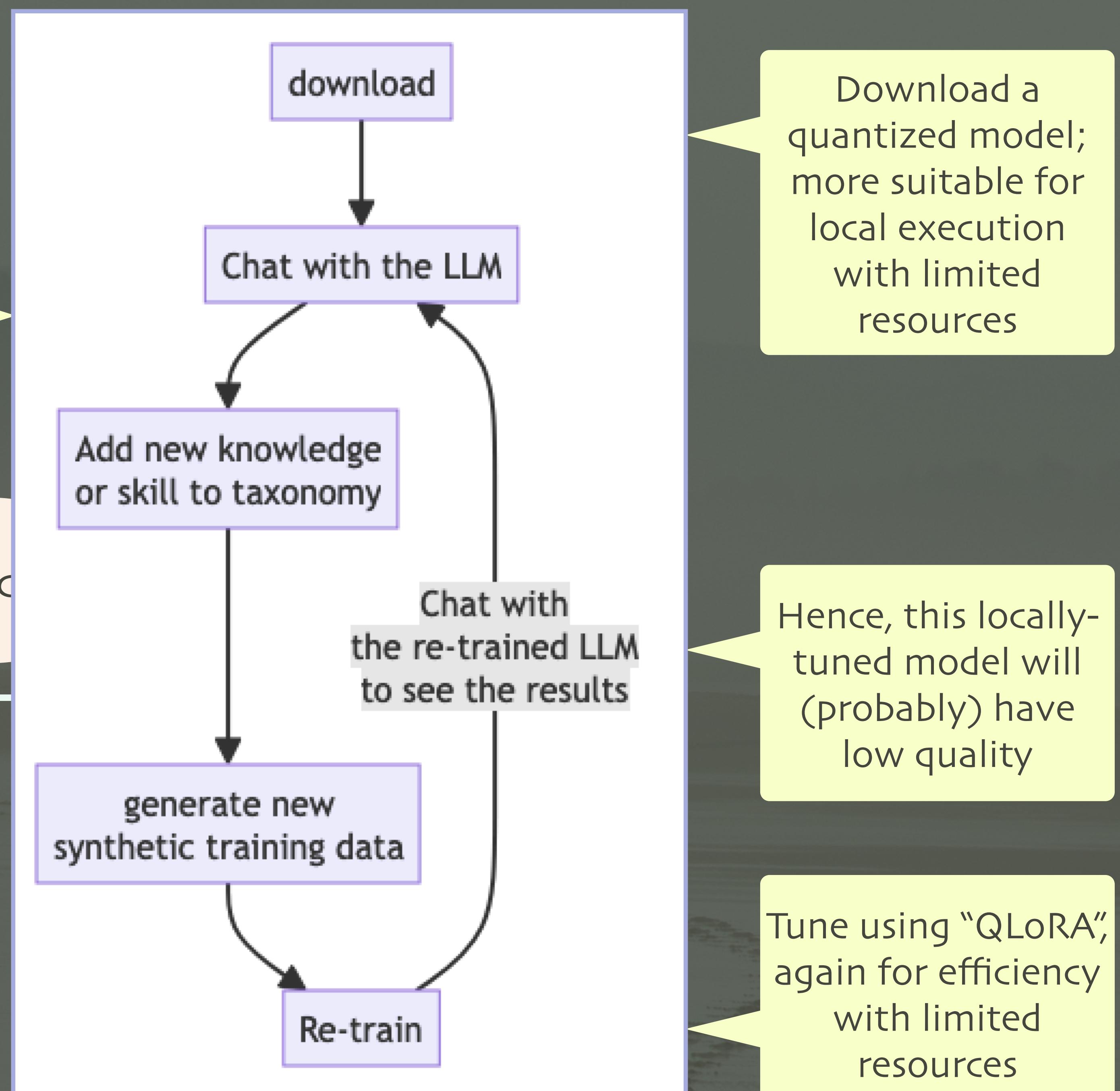
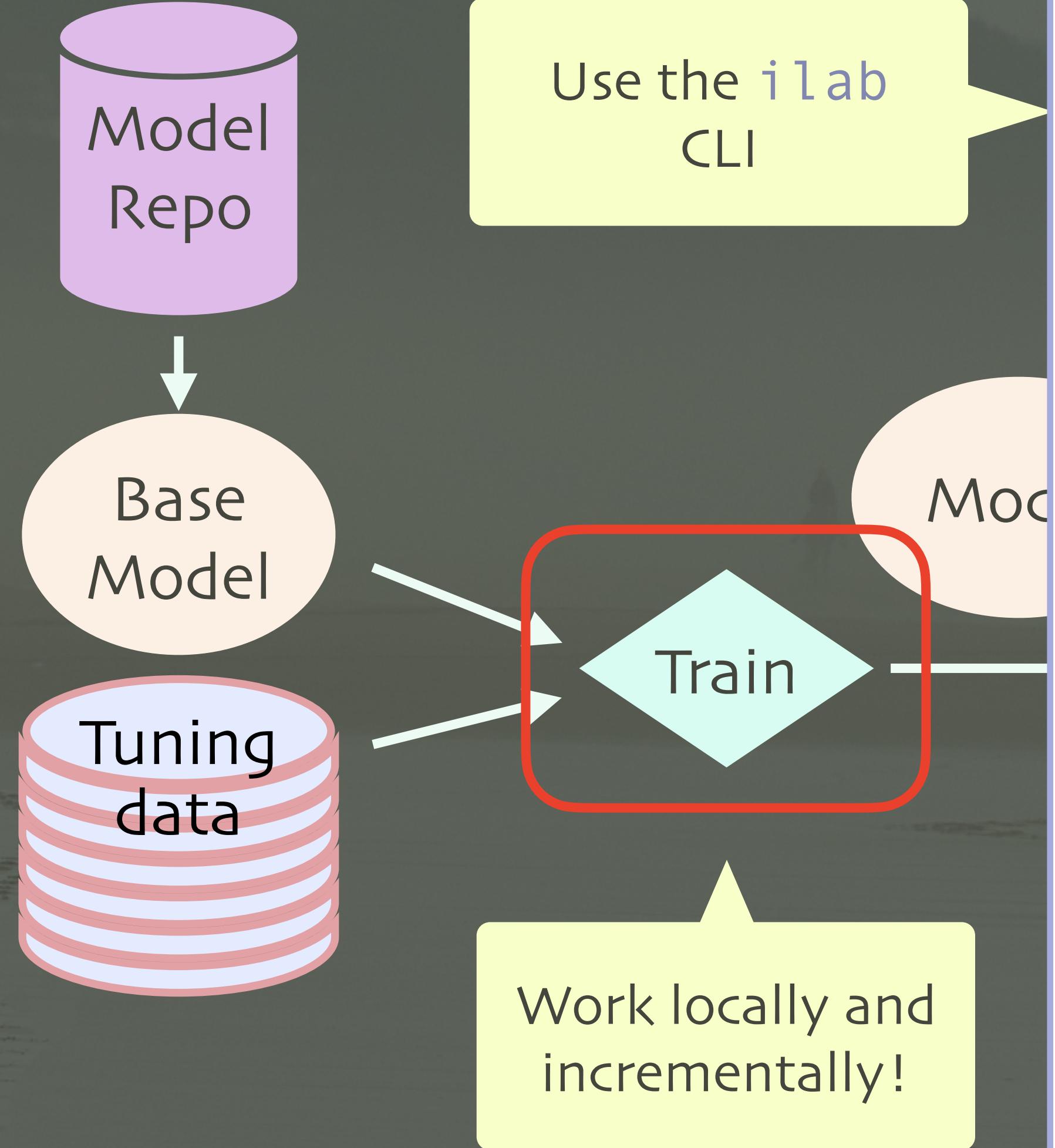
Welcome to the 🐕 InstructLab Project

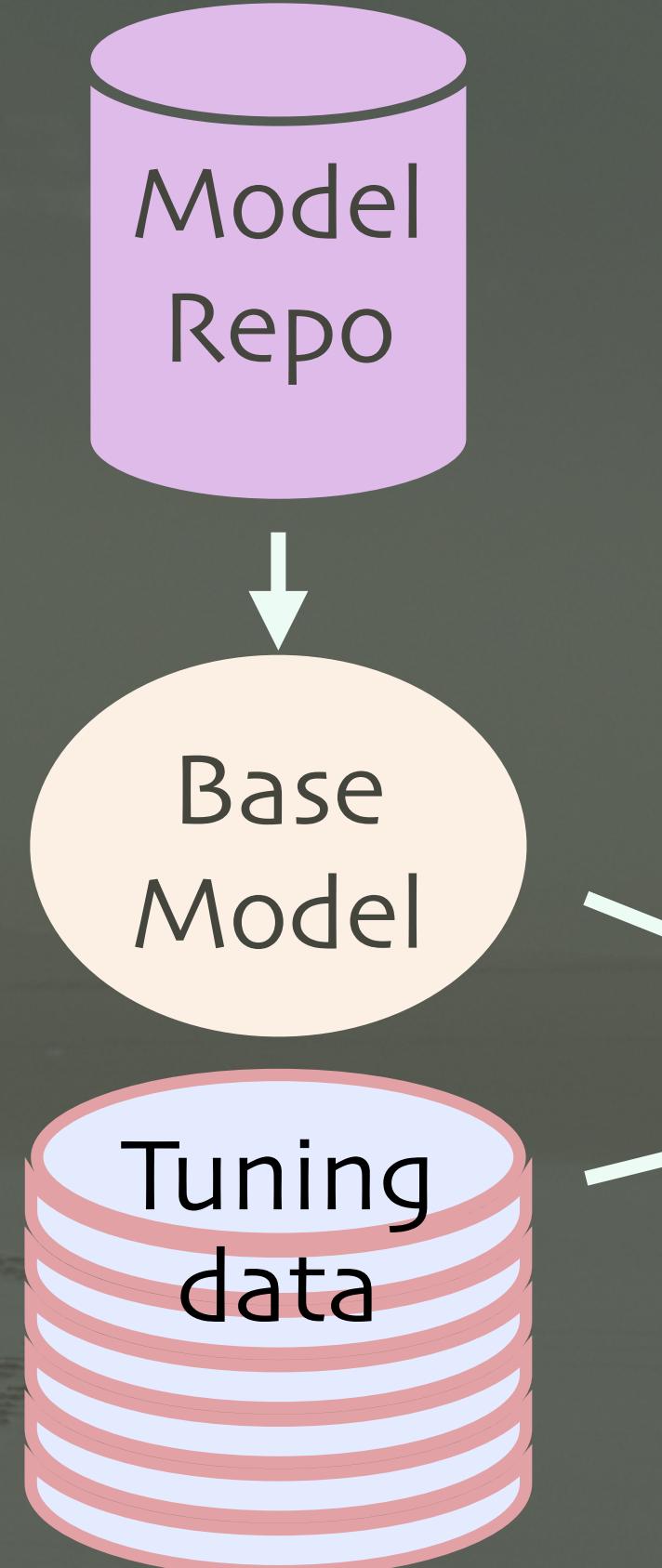
InstructLab is a model-agnostic open source AI project that facilitates contributions to Large Language Models (LLMs).

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



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. Are days longer if it is not summer ?'
'
- answer: 'No, we cannot conclusively conclude that all mammals are black based solely on the given premises. The statement "some mammals are black" does not necessarily guarantee that among those mammals are cats.'
'
question: If all cats are mammals and some mammals are black, can we conclude that some cats are black?
Create a few Q&A examples in a qna.yaml file
- answer: 'Yes, we can conclude that some cats are black based on the given premises.'
'
question: 'If all squares are rectangles and a rectangle has four sides, can we conclude that all squares have four sides?'
'





```
foundational_skills
  reasoning
    common_sense_reasoning
      qna.yaml
    linguistics_reasoning
```

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.'

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 and unquantized base model, etc., etc.

```
qna.yaml
theory_of_mind
  qna.yaml
  unconventional_reasoning
    lower_score_wins
      qna.yaml
```

based on the given premises.

question: 'If all squares are rectangles and a rectangle has four sides, can we conclude that all squares 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 older, unchanged taxonomy areas get worse!
- We’ll return to this topic...

InstructLab

Cons (2/2)

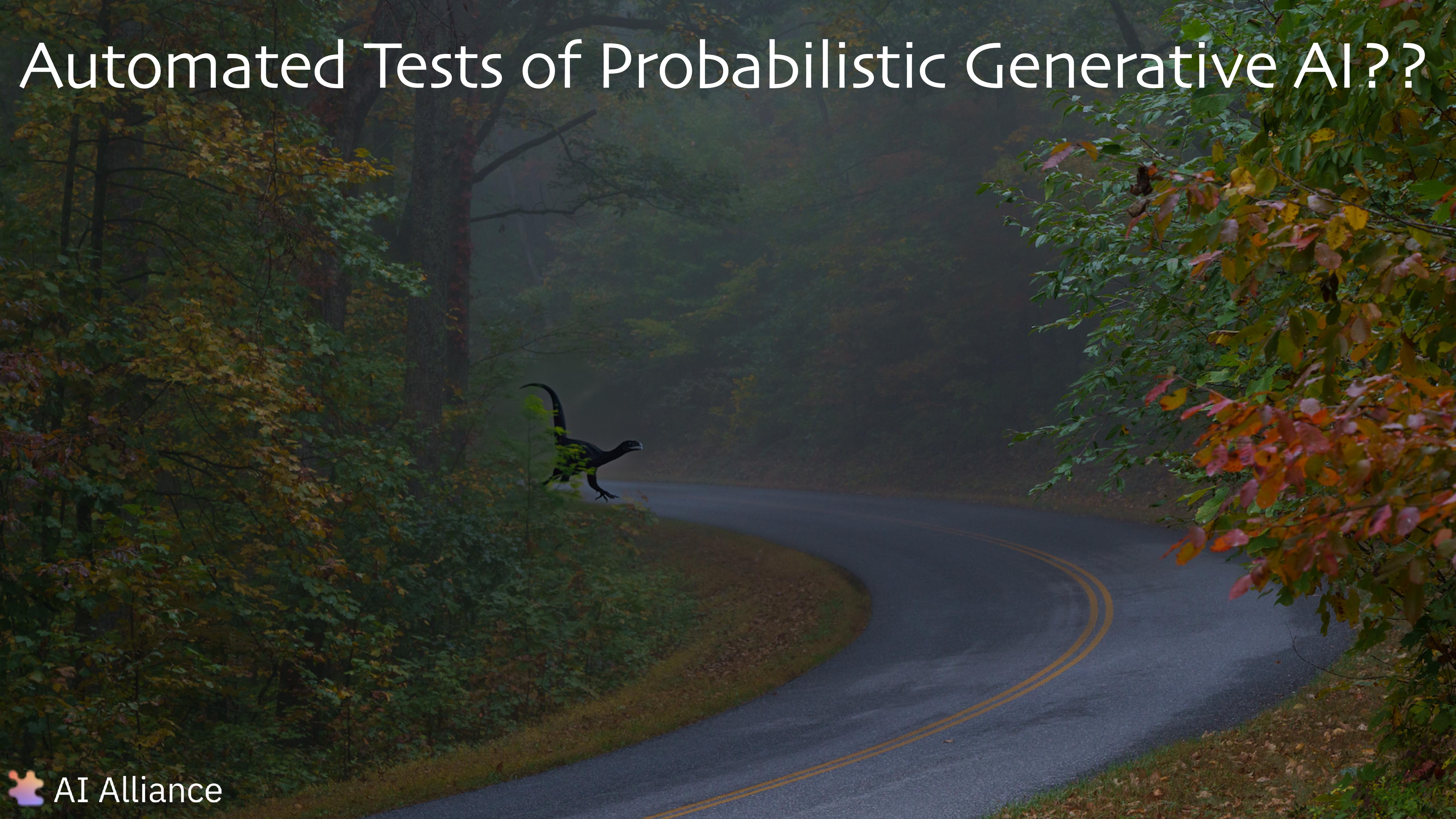
- You still need a server-side final tuning stage.
- The InstructLab project is setting up a community collaboration on public models where they provide the “GitFlow” process for PRs.
- But for your private needs, you still need the ability to do this server-side tuning yourself.
- It might be too expensive to tune on **each** PR.

InstructLab

Pros

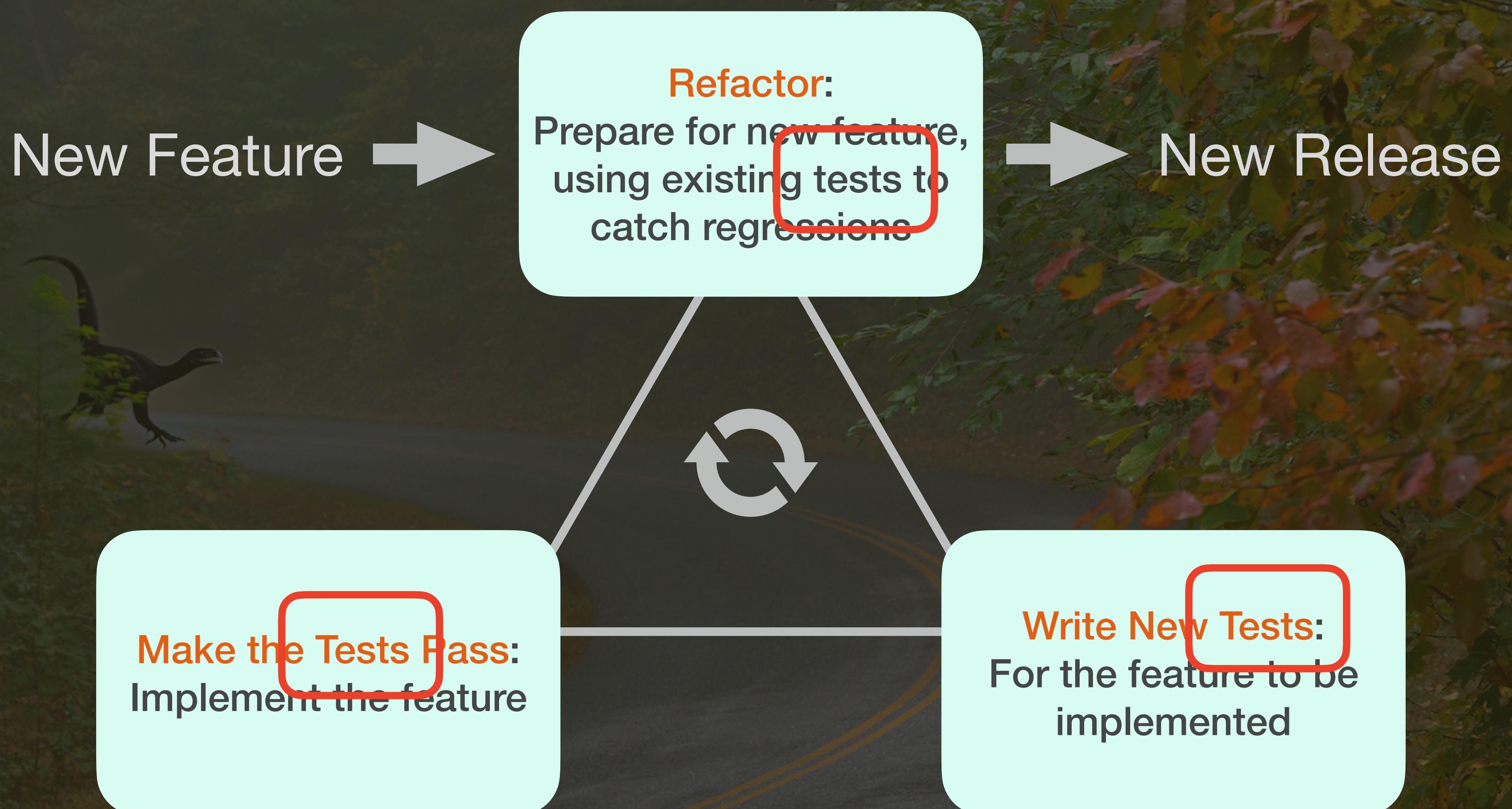
- Defines useful conventions for the taxonomy structure and Q&A examples for each taxonomy topic.
- i lab hides and automates much of the complexity for all the steps.
- You can work locally, incrementally, and iteratively!

Automated Tests of Probabilistic Generative AI??



Automated Tests of Probabilistic Generative 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, to catch regressions!

What Do Developers Expect?

Developers expect software to be deterministic[‡]:

- The system always produces the same output given the same input.
- e.g. $2 + 2 = 4$
- The code is predictable.
- Developers can reason about the correctness of their code.

Put another way, the determinism makes it easier to specify the *system* invariants. What should remain true before and after some step?

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 generated samples are not deterministic.
- They are generated from a distribution.
- With a different seed, the generated samples will be different.
- Is there a way to control the distribution? (Temperature)
- Put another way, the **invariants** are much less **clear** and therefore much less **enforceable**.
- Did any regressions in other behavior occur?

[‡] A tunable “temperature” controls how probabilistic.

Automated Tests of Probabilistic Generative AI??

- What about benchmarks?
 - Most existing benchmarks aren't sufficient specific, like good tests should be.
 - However, more specific, use case focused benchmarks might help.

Automated Tests of Probabilistic Generative AI??

- What about using a “critic” model or a “voting panel” of models as experts?
- The largest LLMs are being used this way.
 - For example, InstructLab uses a powerful teacher model to check synthetic data and reject bad Q&A examples.

Automated Tests of Probabilistic Generative AI??

- Data Science approach?
 - We developers could use help from you data scientists to build statistically-appropriate testing techniques.

What other ideas to you?
What techniques have you tried?

Thank you!

thealliance.ai

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/), 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](#).