Can We Make Model Alignment a Software Engineering Process?

The AI Conference, San Francisco
September 2024

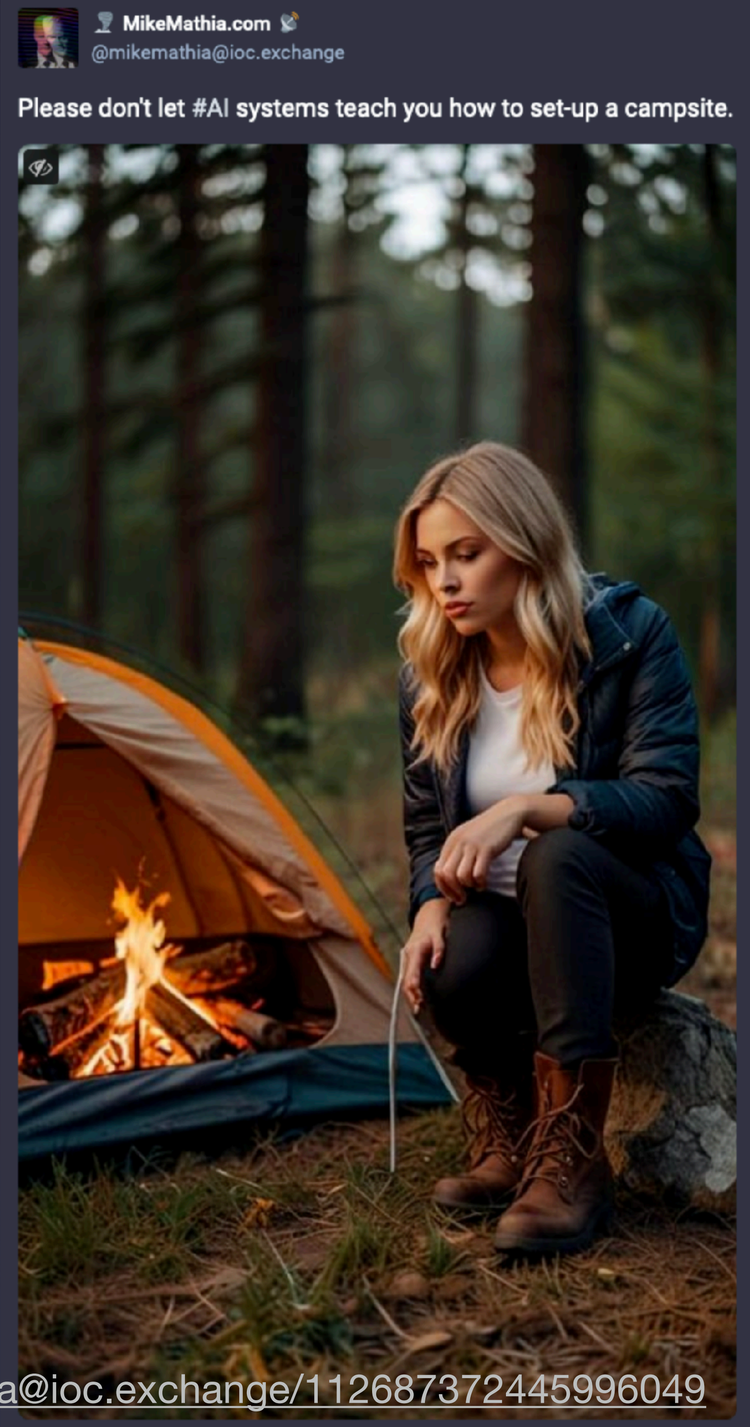
Dean Wampler, Ph.D.
The AI Alliance and IBM Research
thealliance.ai

deanwampler.com/talks

AI Alliance

About the Images…

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



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

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

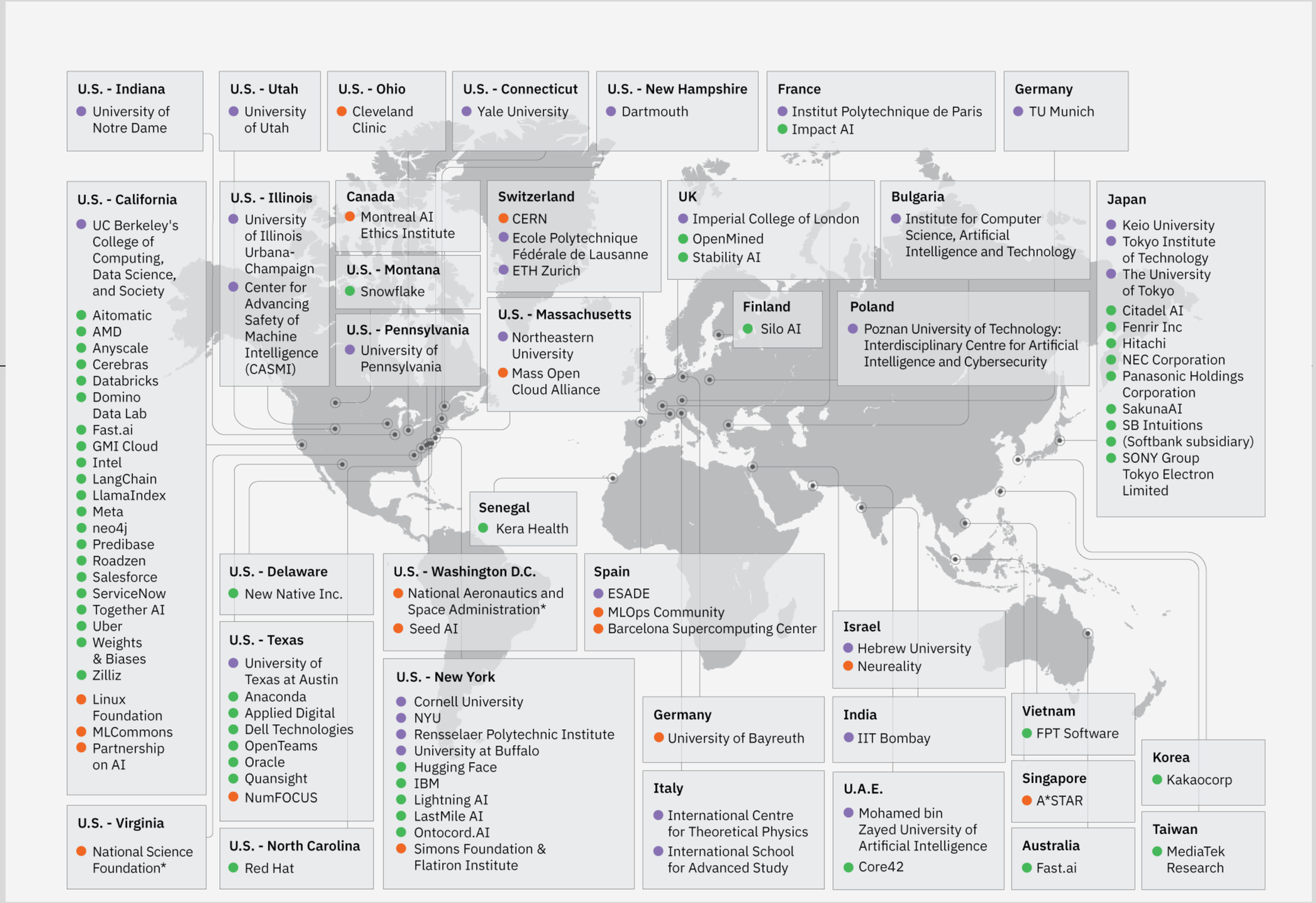
thealliance.ai

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.

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

+100 organizations in +20 countries, and growing



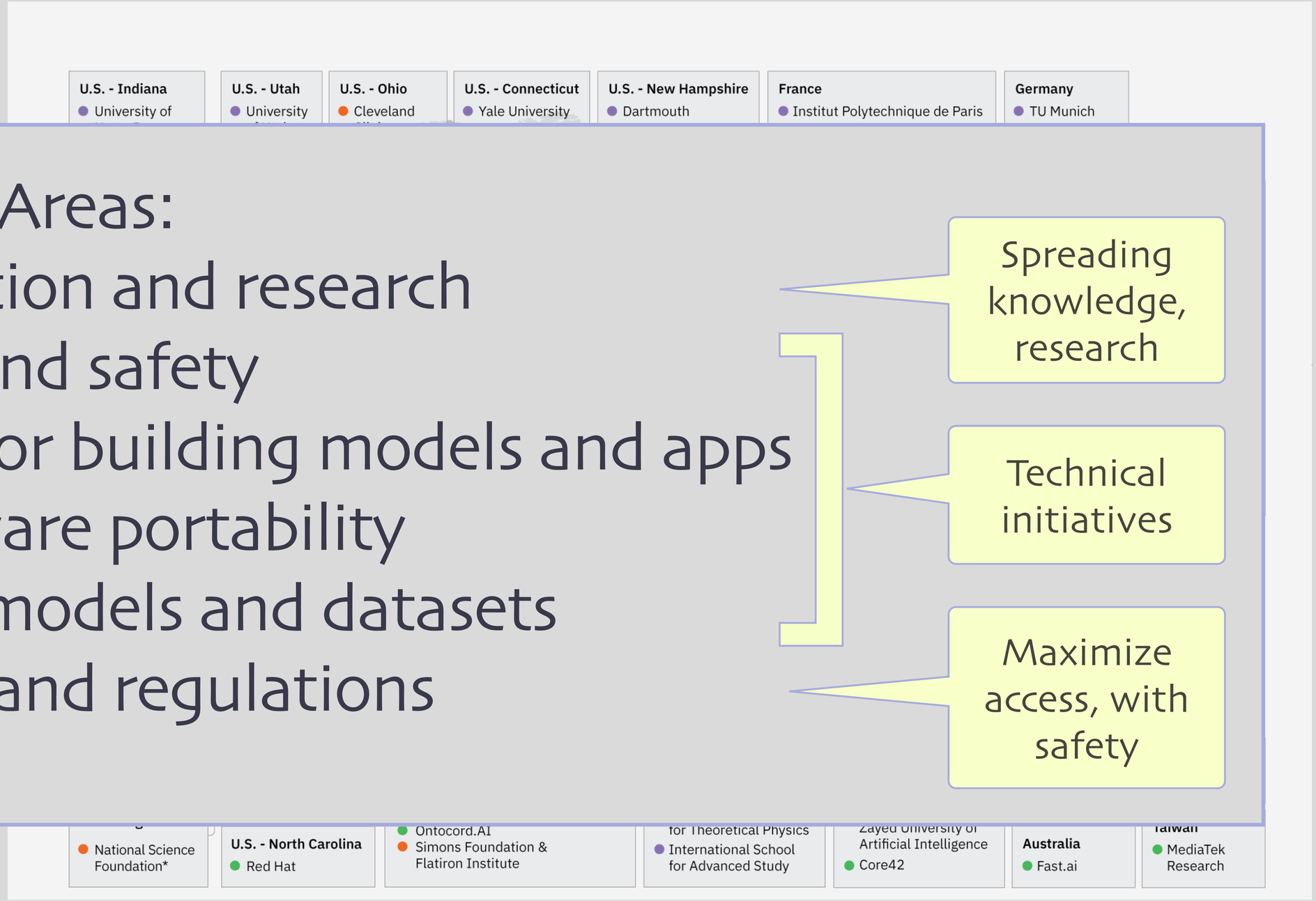
AI Alliance

+100 organizations in +20 countries, and growing

thealliance.ai

Our
ope
us,

Mer
aca
rese
and



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

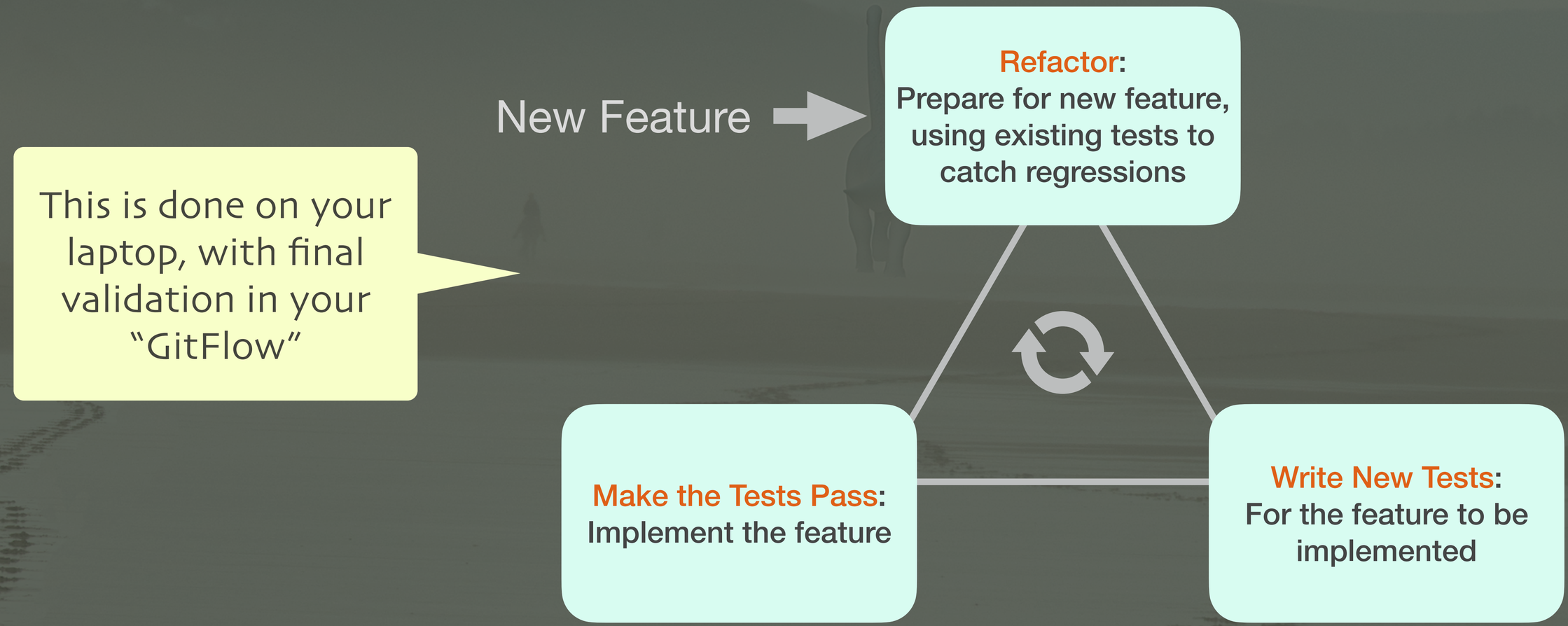
Maximize access, with safety

Iterative and Incremental Model Tuning

AI Alliance

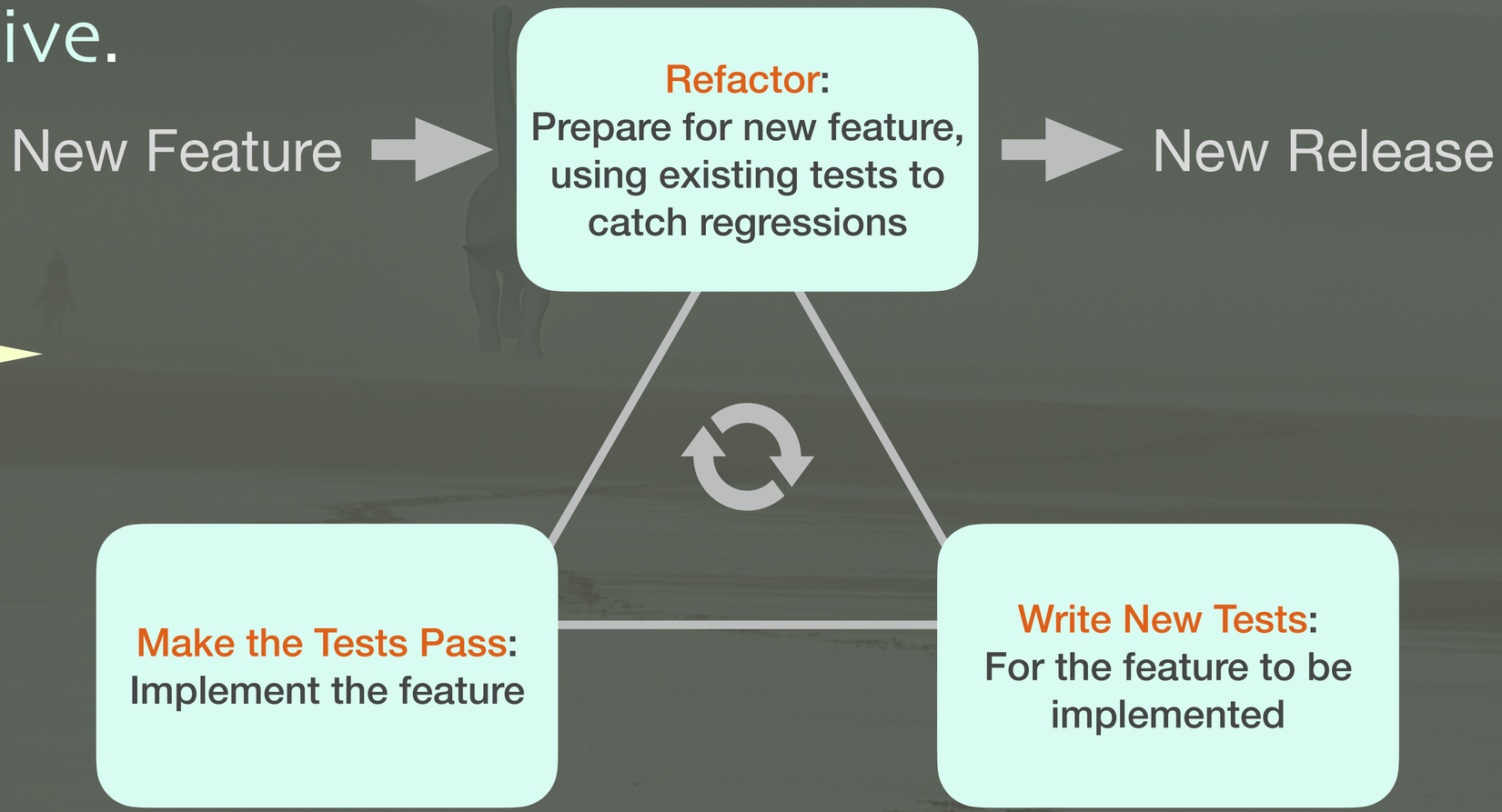
What Software Developers Like

- Features are added incrementally.



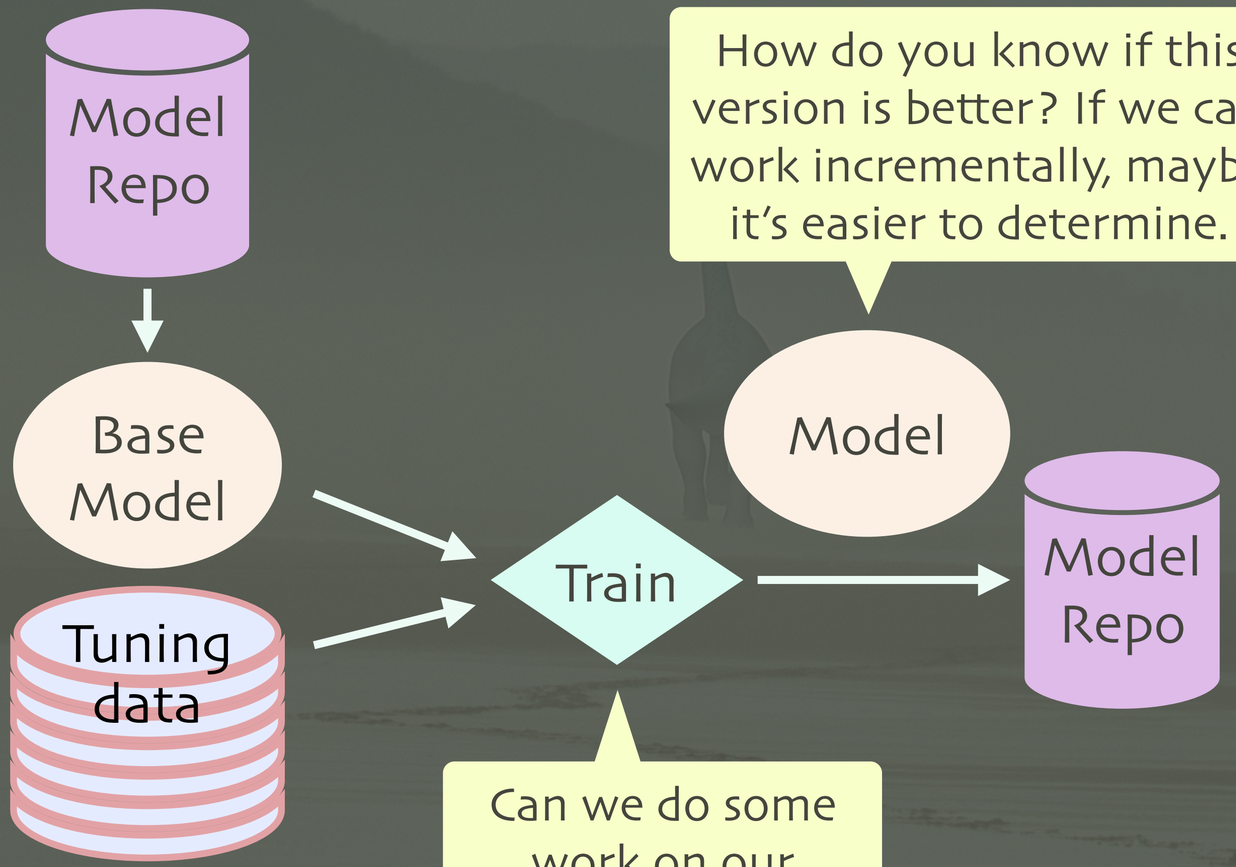
What Software Developers Like

- Features are added incrementally.
- Releases are iterative.



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 "modular features"?

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

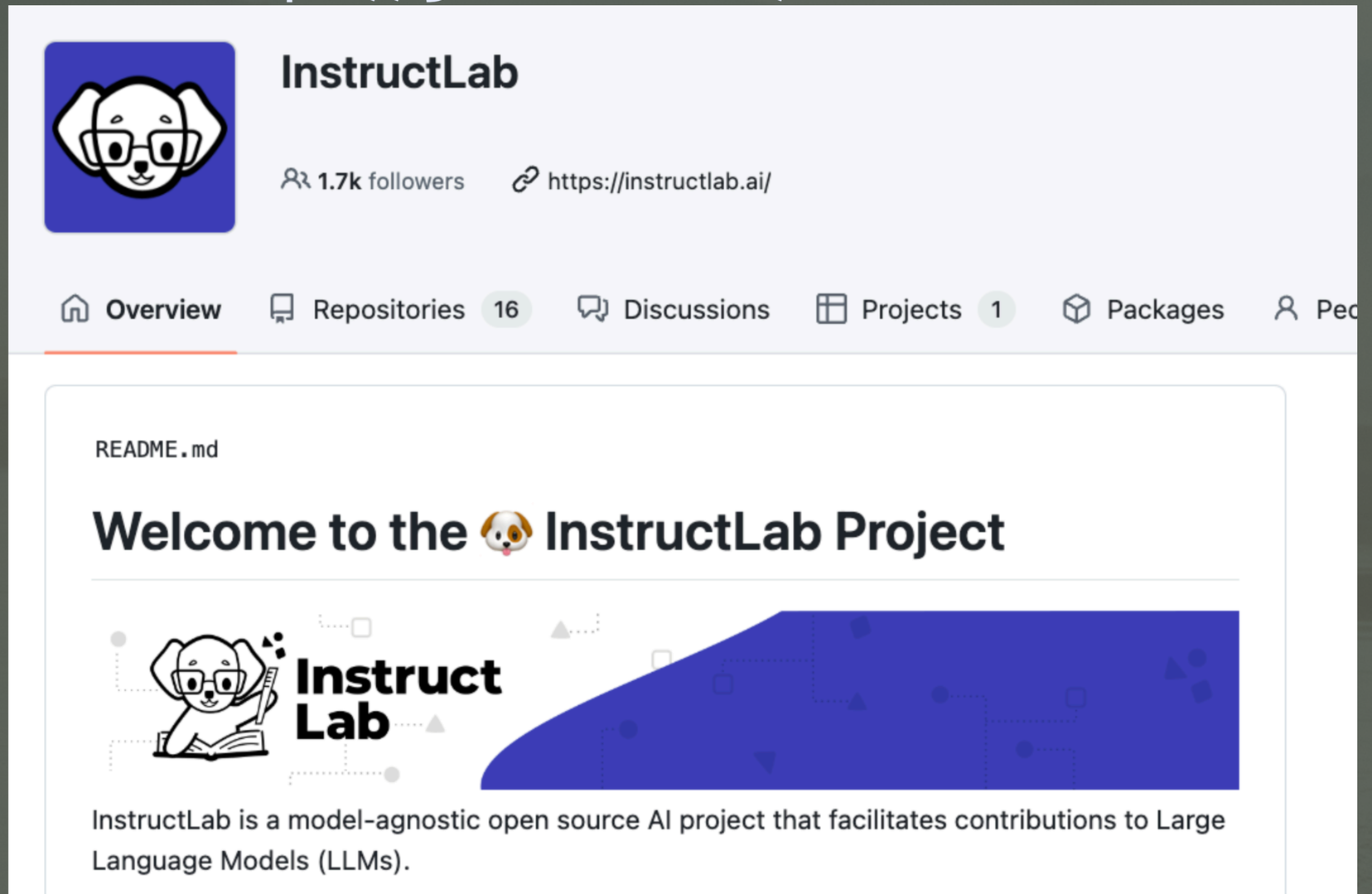
One Approach: InstructLab

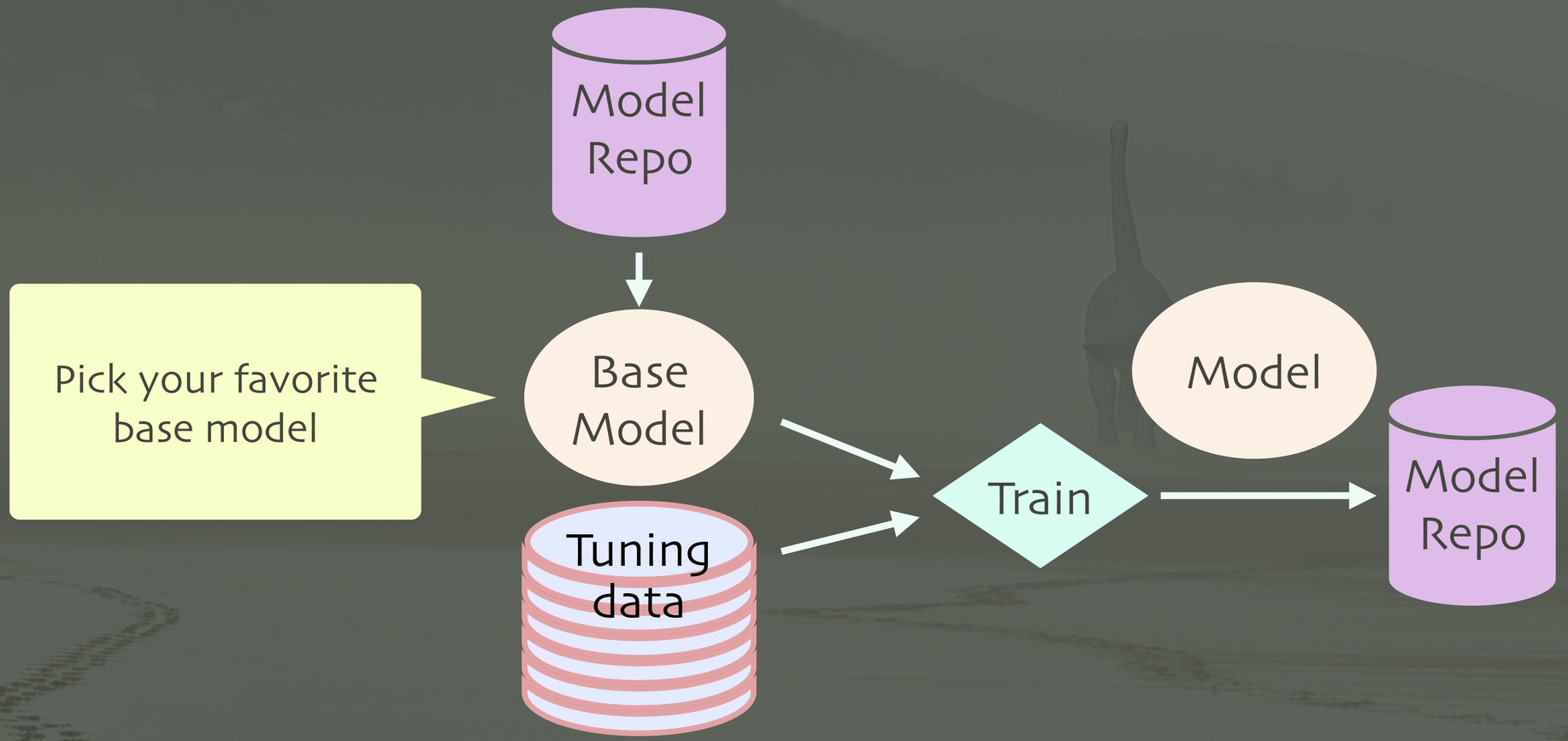
https://github.com/instructlab

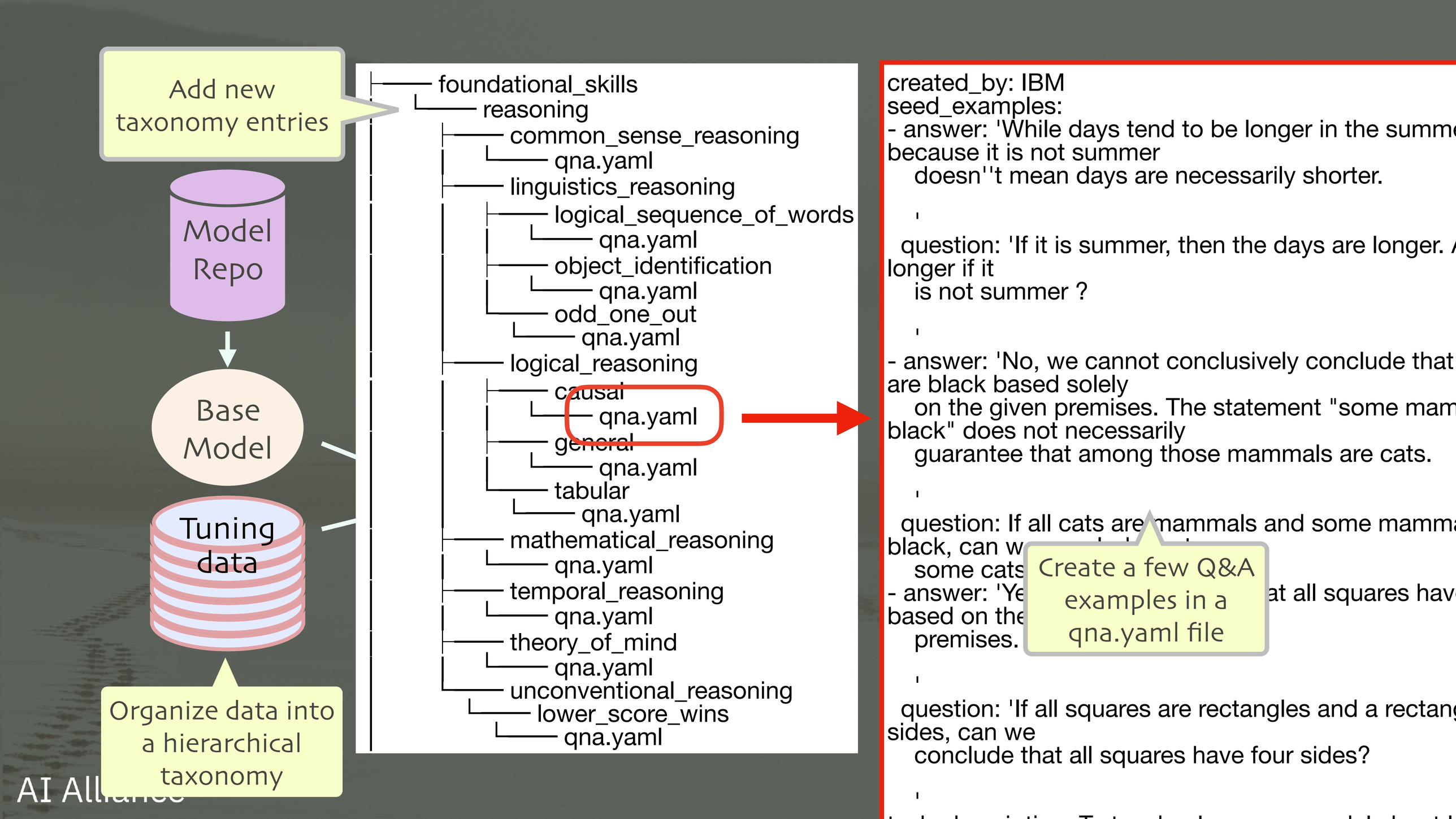
Open sourced by
IBM and Red Hat

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

AI Alliance




Model Repo
Pick your favorite base model
Base Model
Tuning data
Train
Model
Model Repo


Add new taxonomy entries
Model Repo
Base Model
Tuning data
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
based on the
premises.
'
question: 'If all squares are rectangles and a rectang
sides, can we
conclude that all squares have four sides?
'
Create a few Q&A examples in a qna.yaml file

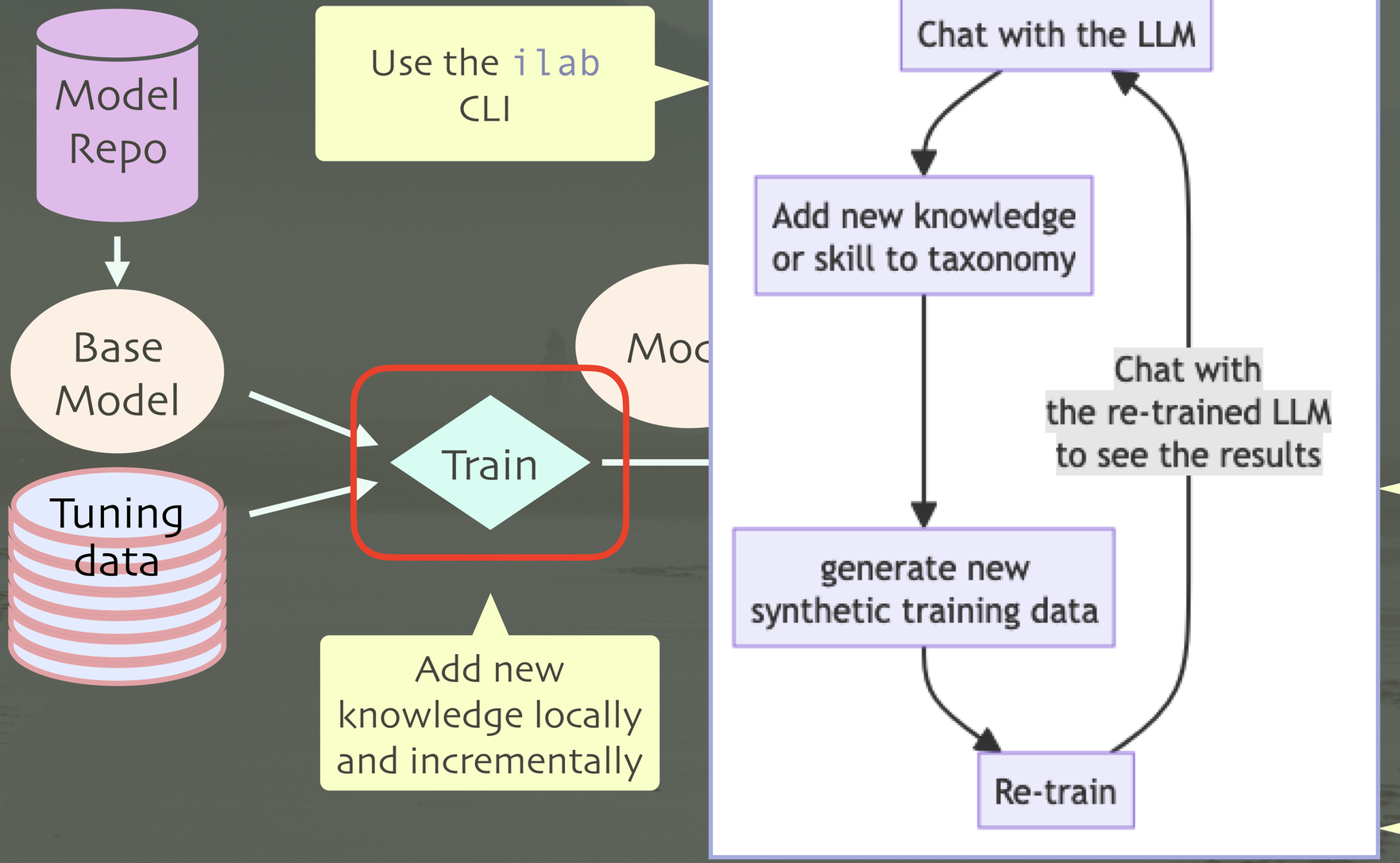
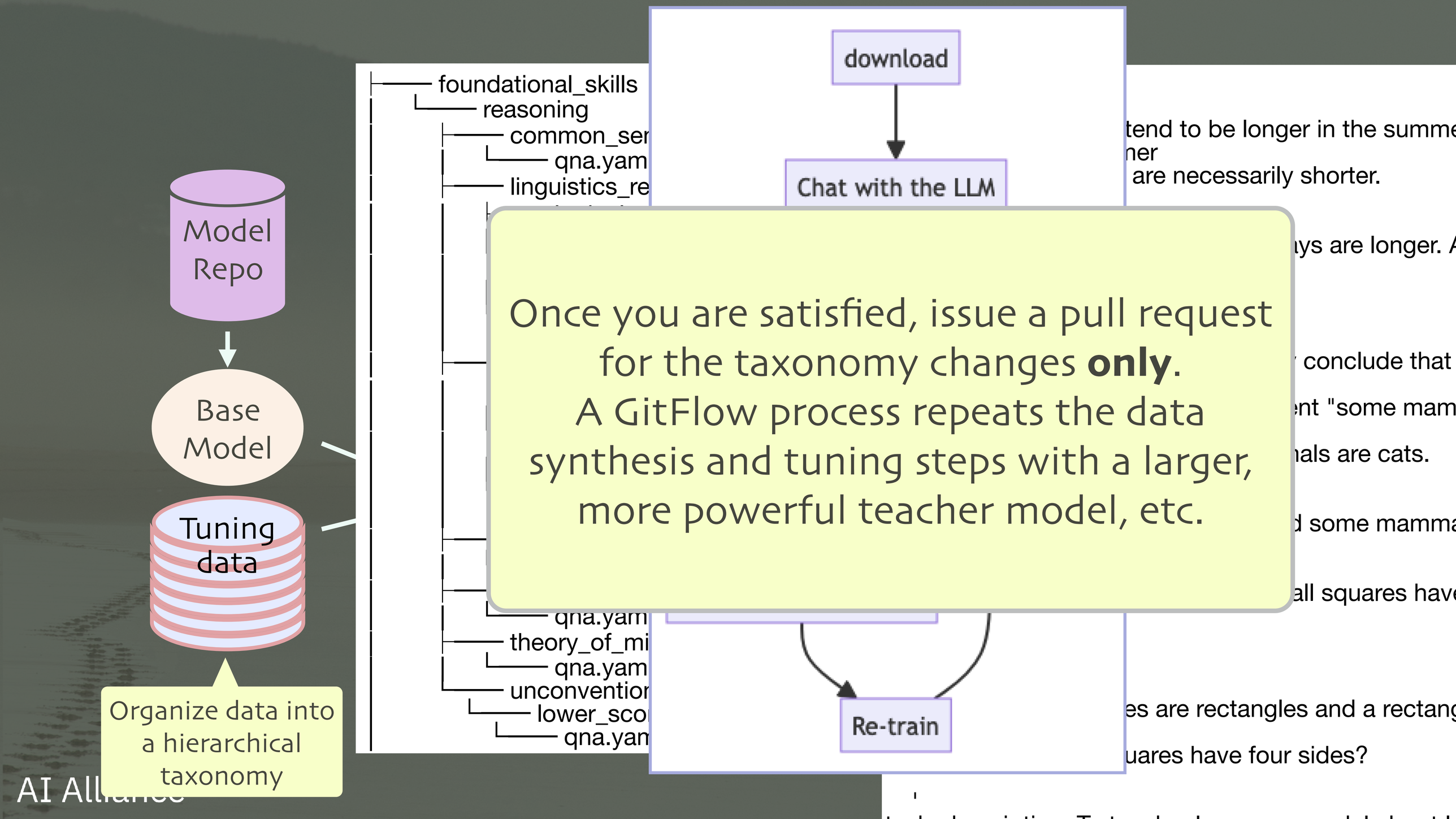


Image: https://github.com/instructlab/instructlab


Model Repo
Base Model
Tuning data
Organize data into a hierarchical taxonomy
foundational_skills
reasoning
common_ser
qna.yam
linguistics_re
qna.yam
theory_of_mi
qna.yam
unconventior
lower_sco
qna.yam
download
Chat with the LLM
Re-train
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 summe
ner
are necessarily shorter.
ys are longer. A
conclude that
nt "some mam
als are cats.
some mamma
all squares have
es are rectangles and a rectang
uares have four sides?
AI Alliance

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…)

https://github.com/instructlab

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.

https://github.com/instructlab

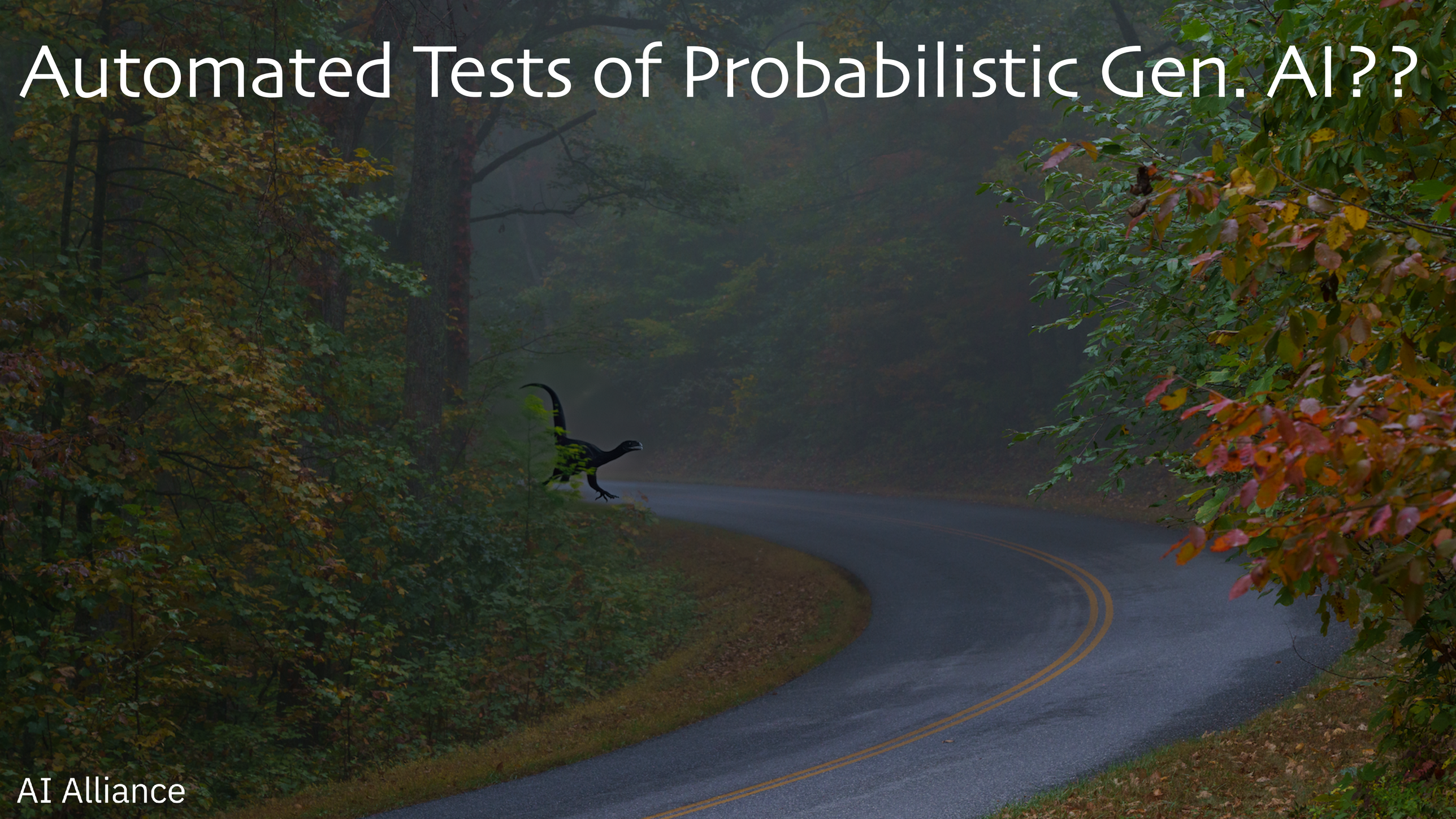
InstructLab

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

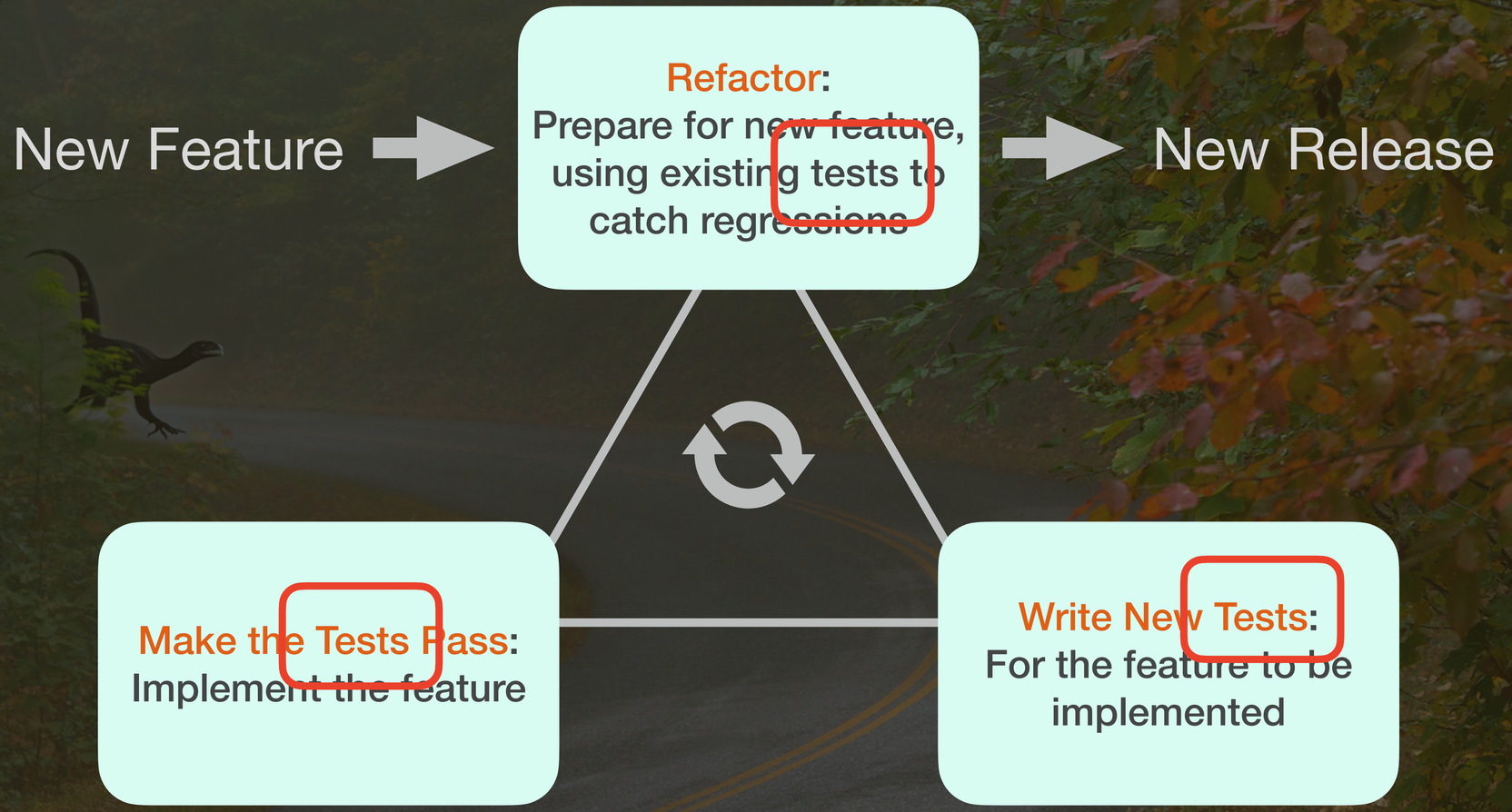
Automated Tests of Probabilistic Gen. AI??



AI Alliance

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.

‡ Distributed systems break this clean picture.

What Do Developers Expect?

Developers expect software to be deterministic‡:

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

oken!

ensure

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
  rep
  - Is
    th
  - Did any regressions in other behavior occur?

rse

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

‡ A tunable "temperature" controls how probabilistic.

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

 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.