Can We Make Model Alignment a Software Engineering Process?

The AI Conference, San Francisco
September 2024

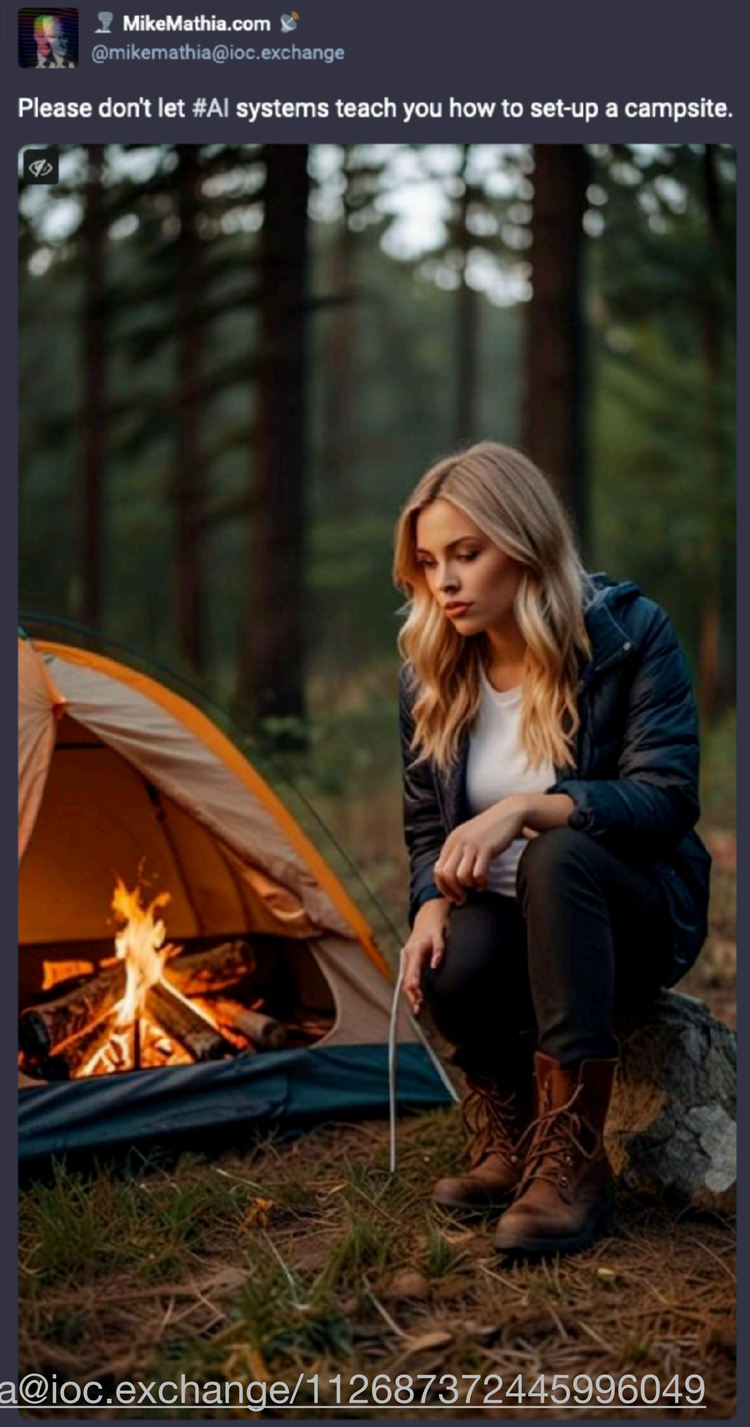
Dean Wampler, Ph.D.
The AI Alliance and IBM Research
thealliance.ai

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!

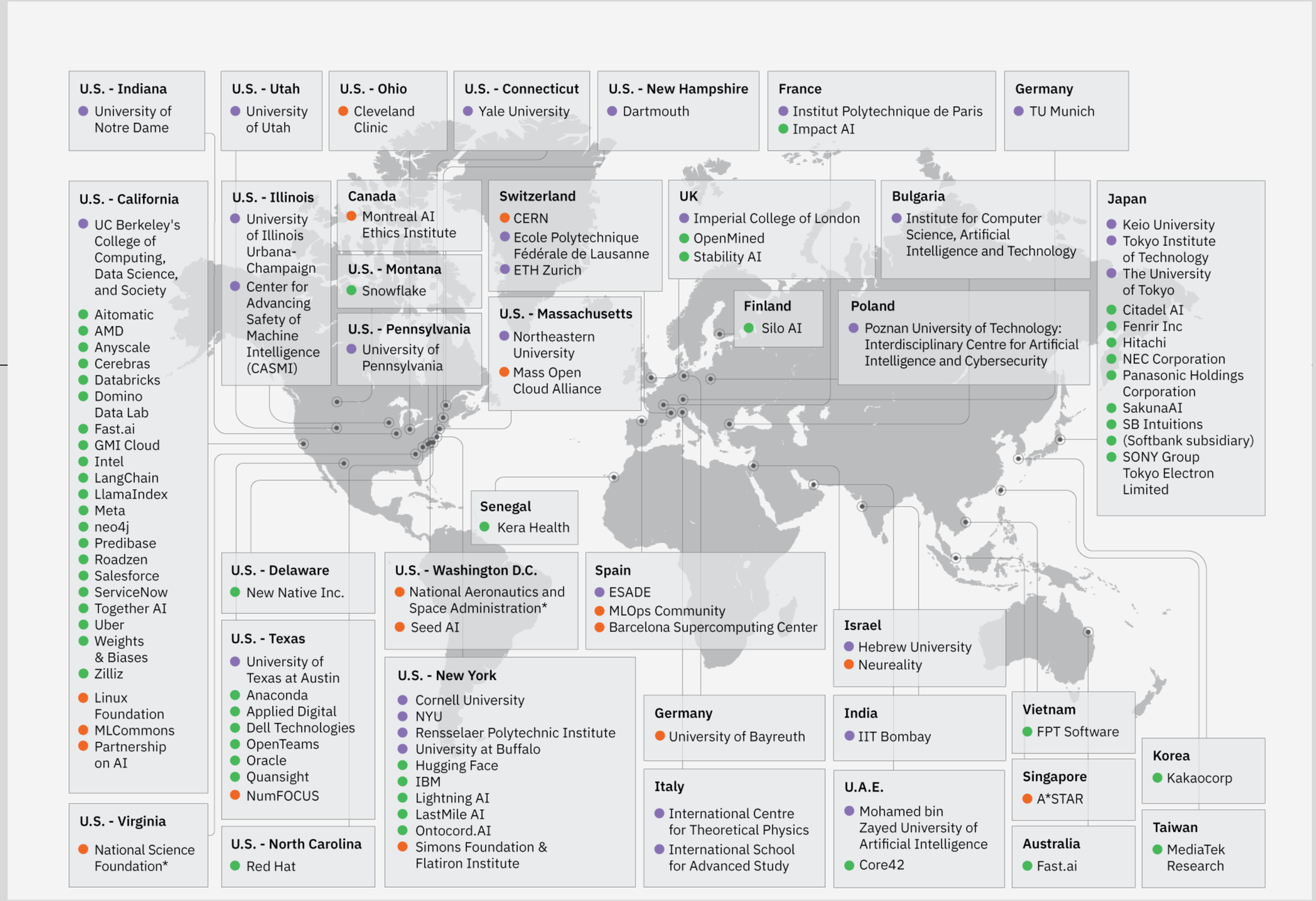
AI Alliance

+120 organizations in +20 countries, and growing

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.



AI Alliance

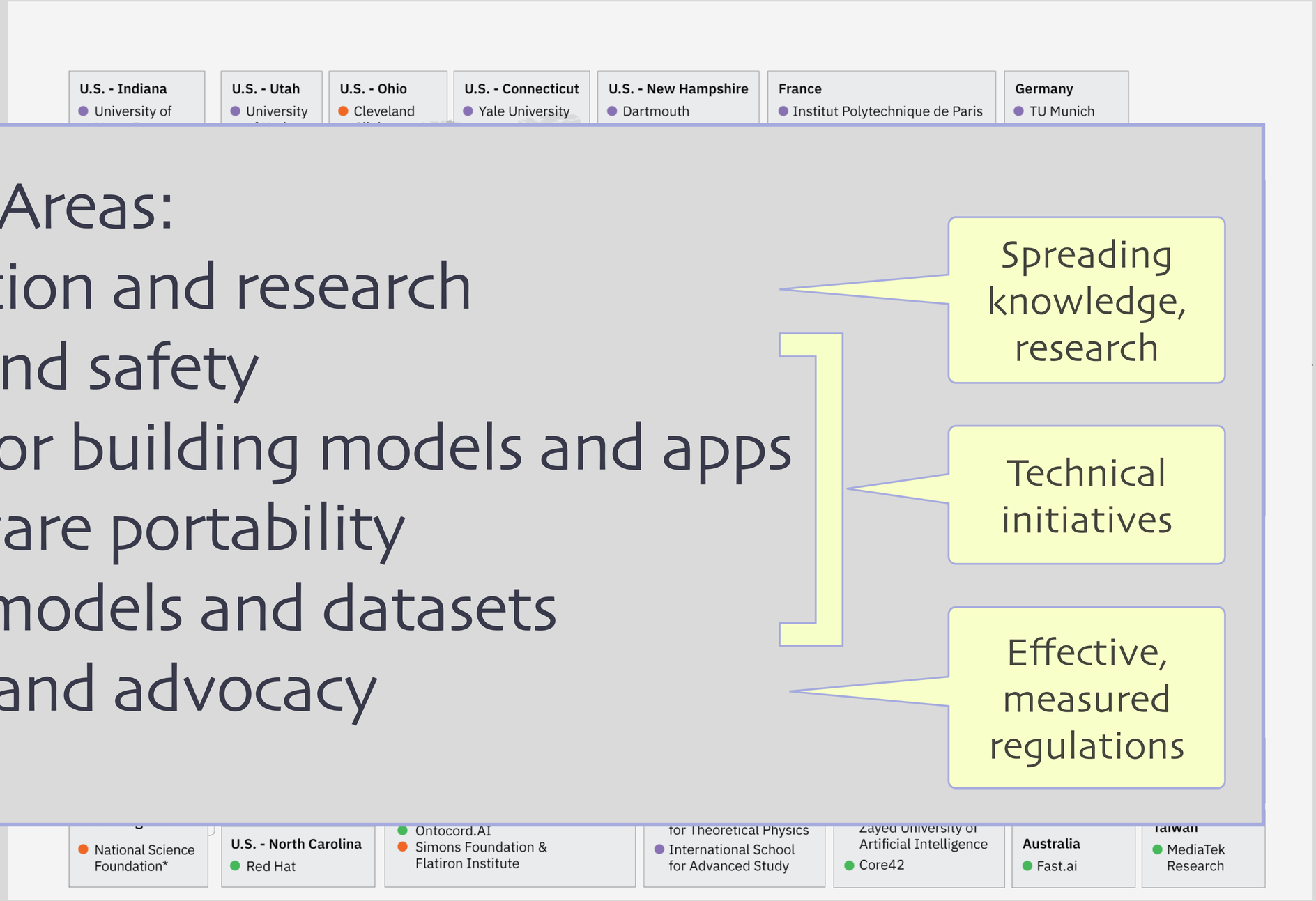
+100 organizations in +20 countries, and growing

thealliance.ai

Our
ope
us,

Mer
aca
rese
and

Vi



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

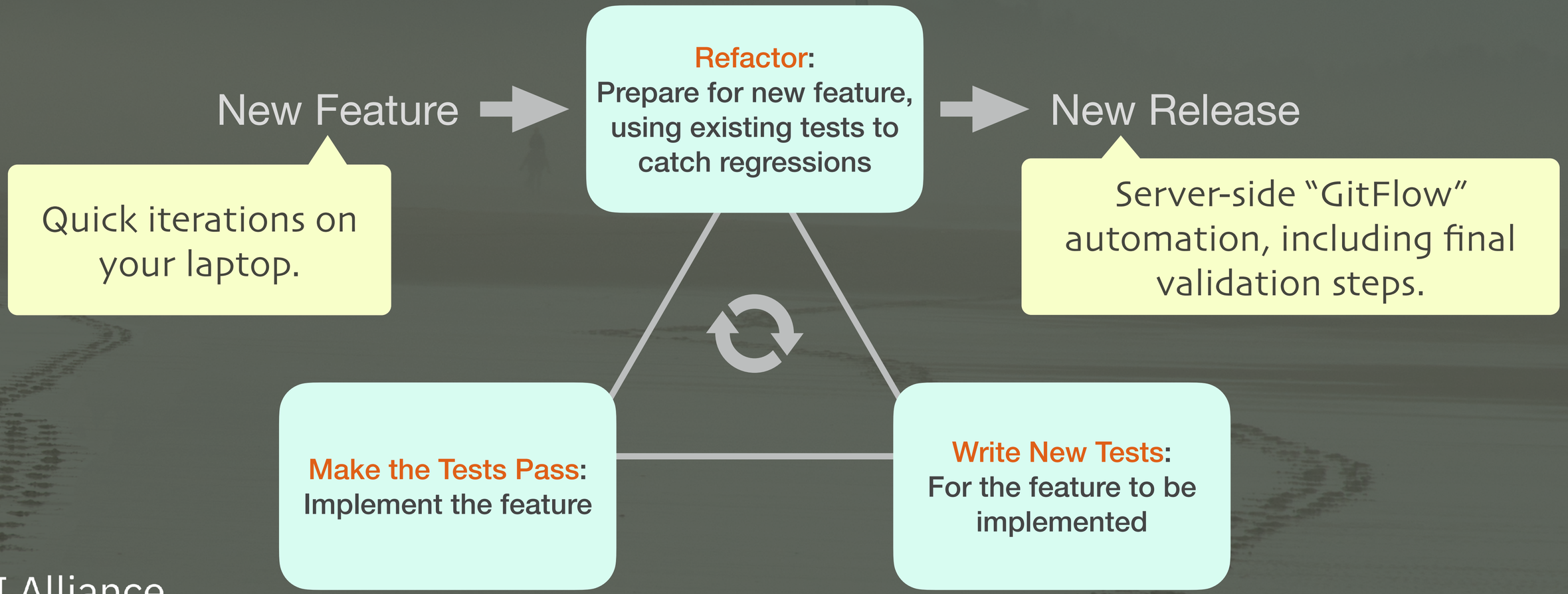
Effective, measured regulations

Iterative and Incremental Model Tuning

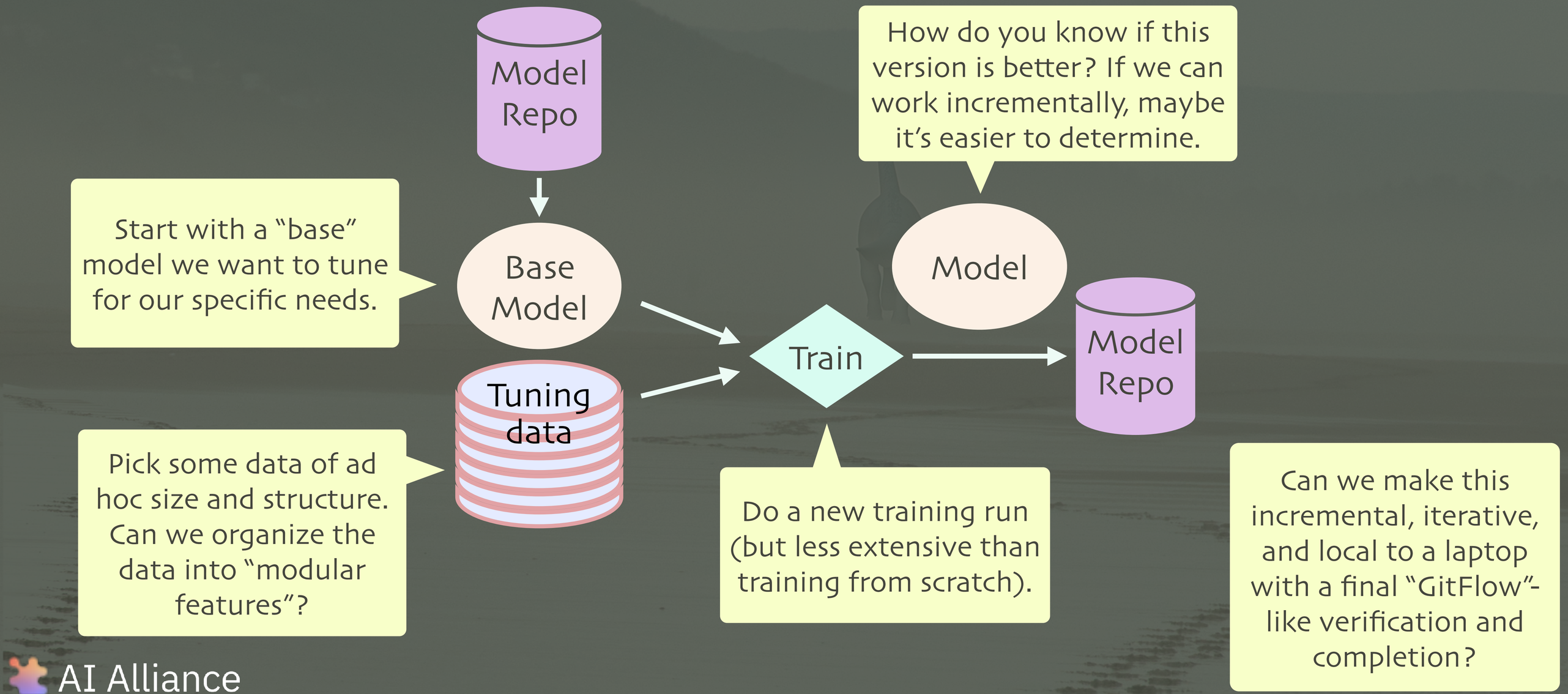
AI Alliance

What Software Developers Like

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



What Model Tuning Is Often Like

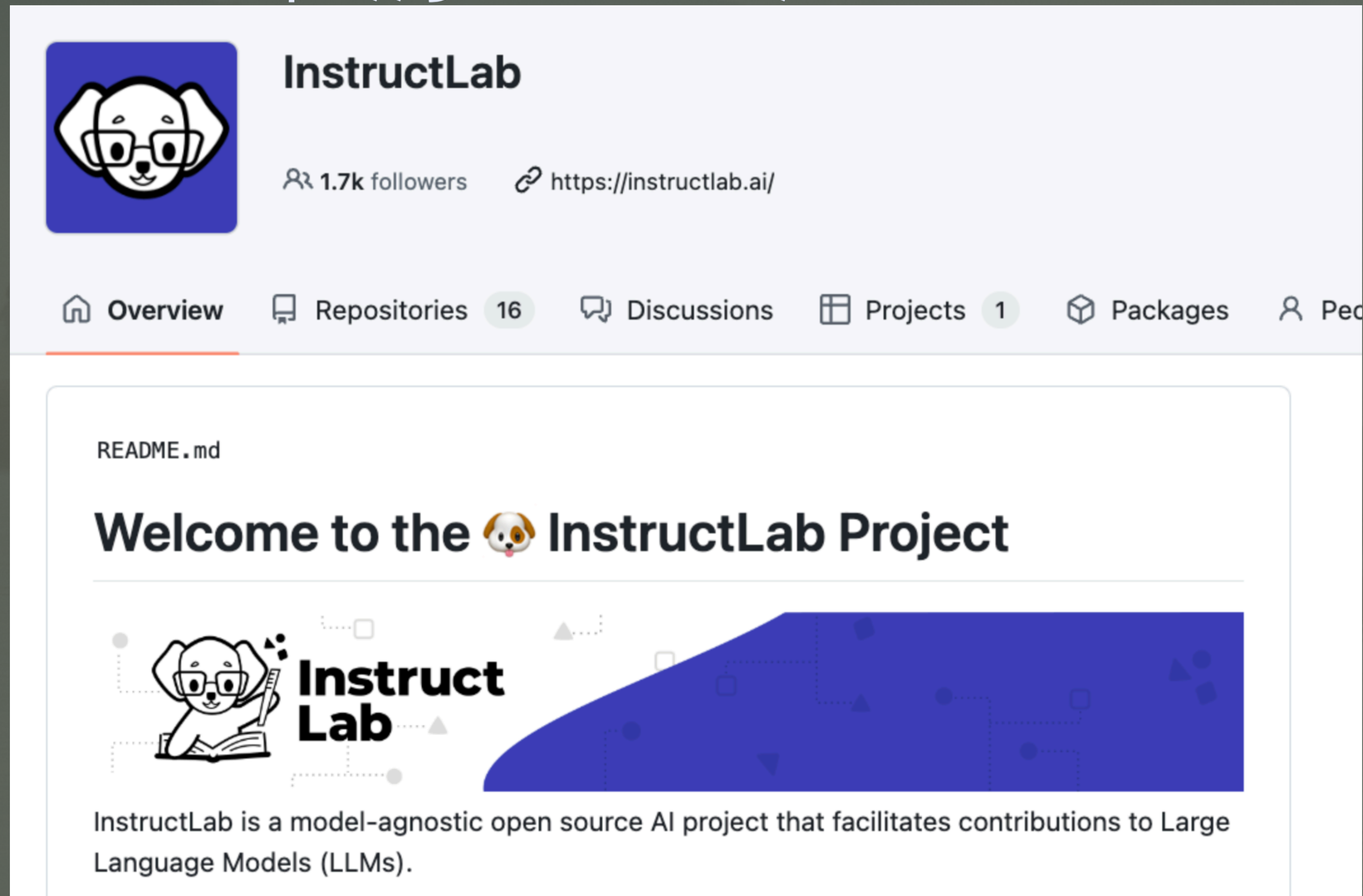


One Approach: InstructLab

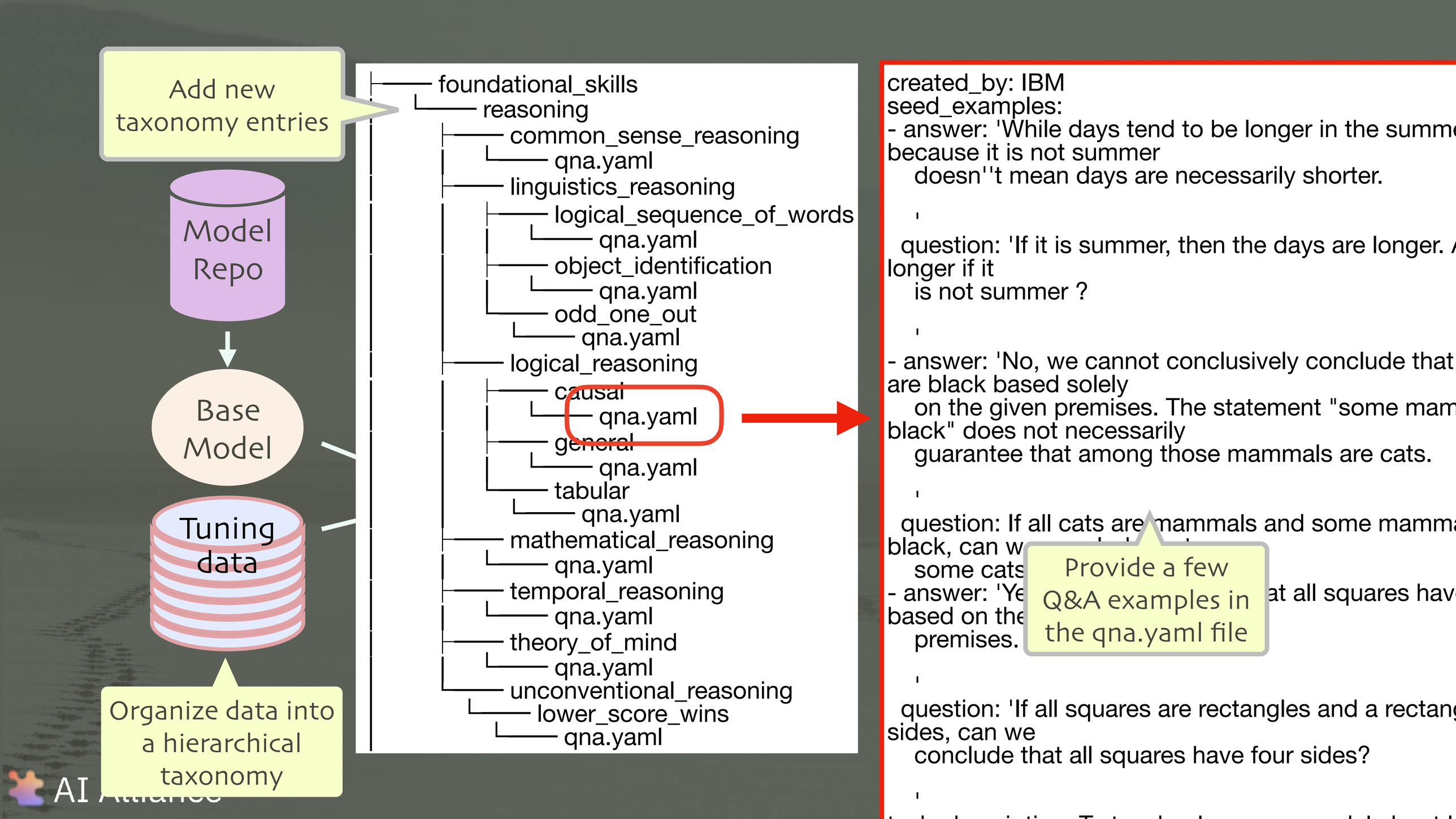
https://github.com/instructlab

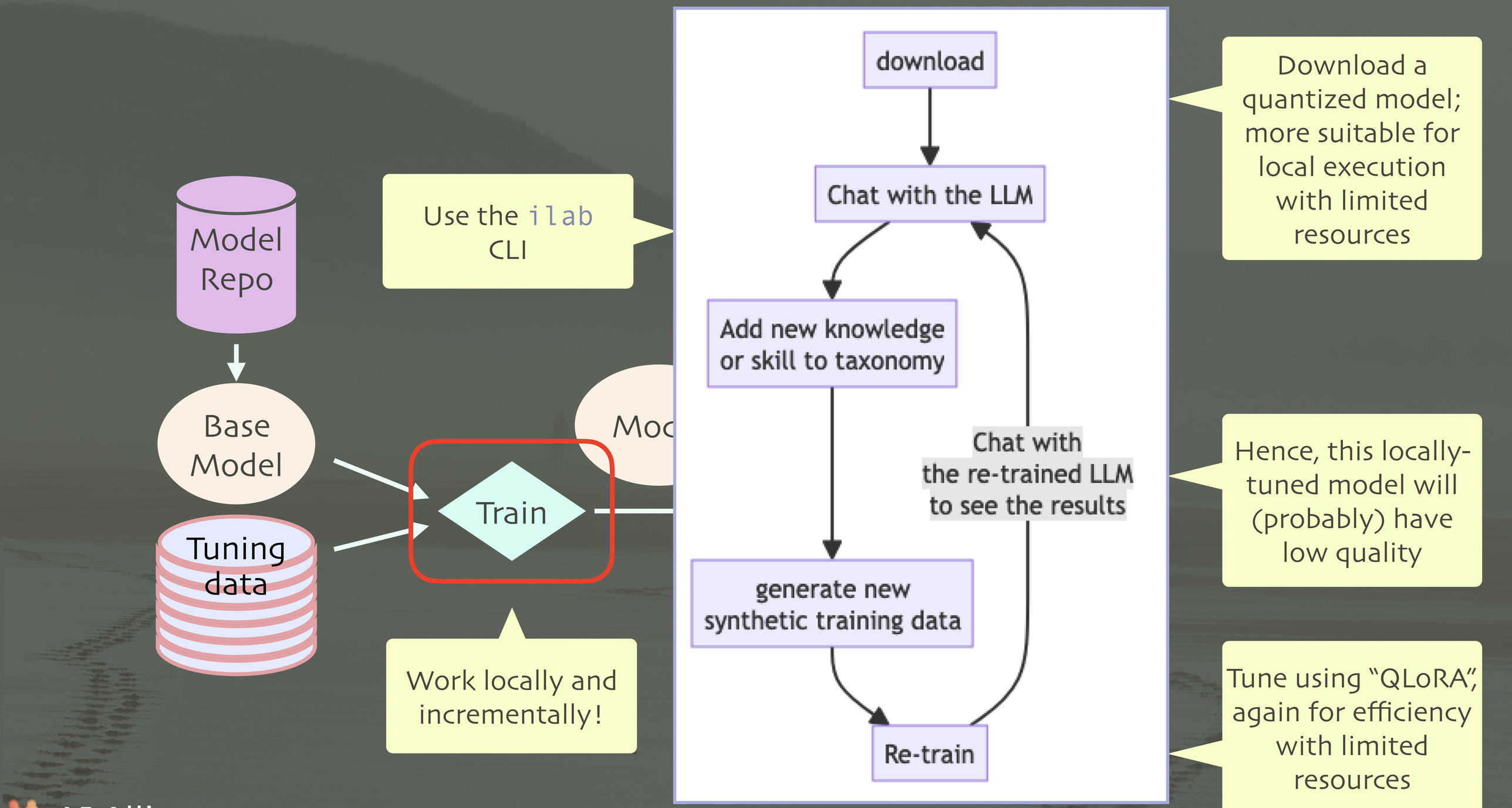
Open sourced by
IBM and Red Hat

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



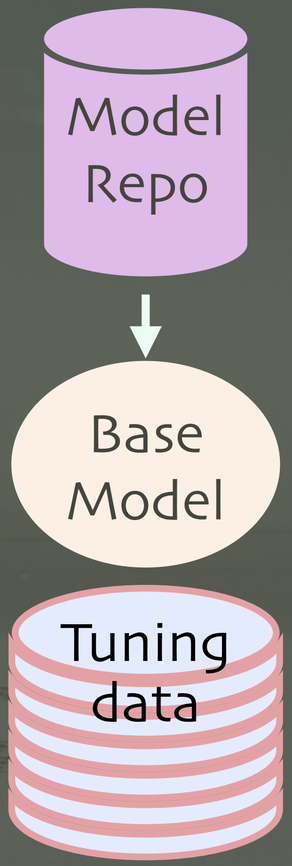
AI Alliance


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?
'
Provide a few Q&A examples in the qna.yaml file



AI Alliance

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




```
├── foundational_skills
│   └── reasoning
│       ├── common_sense_reasoning
│       │   └── qna.yaml
│       ├── linguistics_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.

ys are longer. A

 conclude that

nt "some mam

nals are cats.

d some mamma

all squares have

based on the given
    premises.

    '
  question: 'If all squares are rectangles and a rectang
sides, can we
    conclude that all squares 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 and unquantized base model, etc., etc.

AI Alliance

InstructLab

Cons (1/2)

- Testing!
  - Supports a combination of standard benchmarks and chat ("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…

AI Alliance

https://github.com/instructlab

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.

https://github.com/instructlab

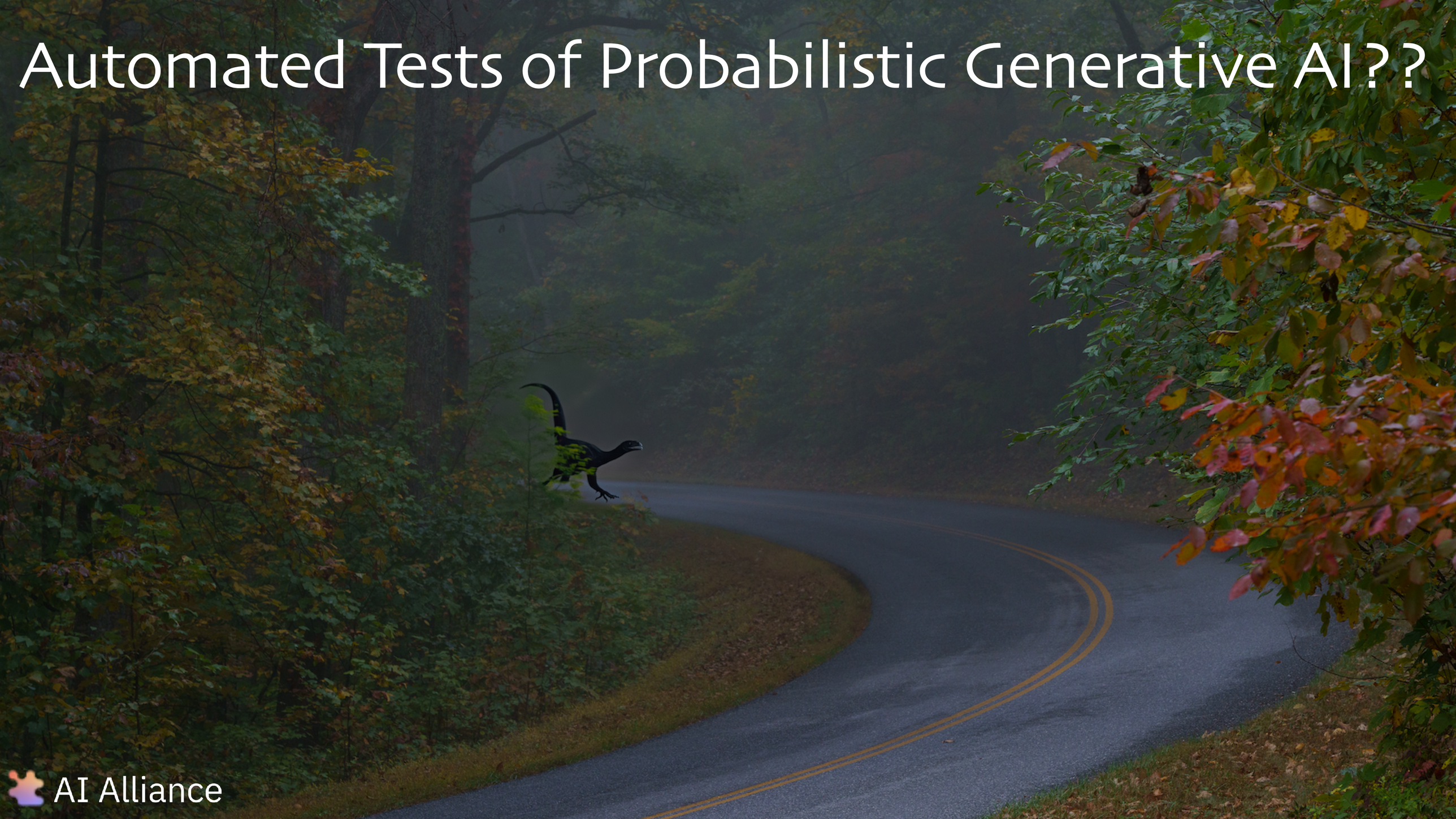
InstructLab

Pros

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

https://github.com/instructlab

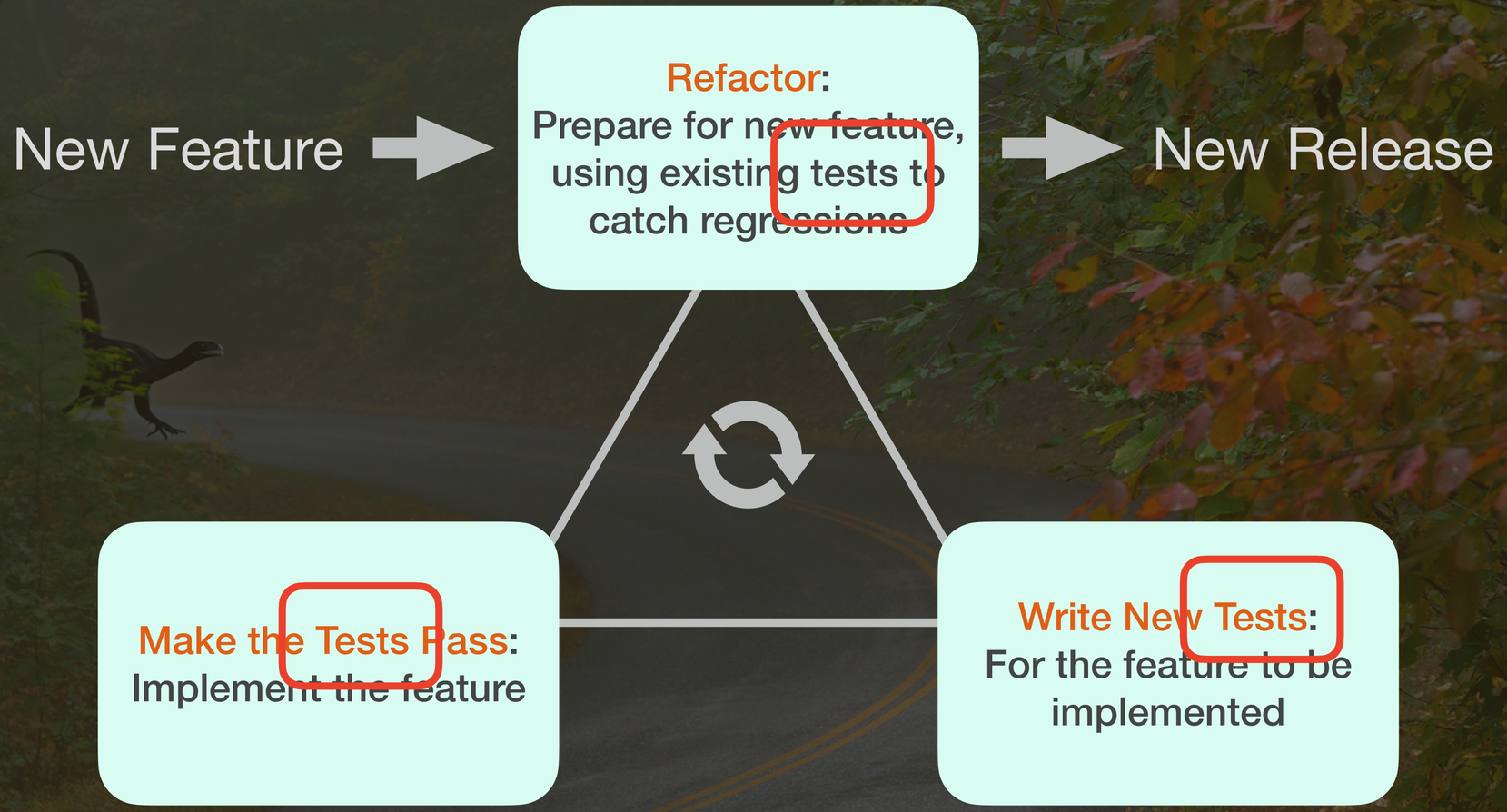
Automated Tests of Probabilistic Generative AI??



AI Alliance

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

AI Alliance

‡ Distributed systems break this clean picture.

What Do Developers Expect?

Developers expect software to be deterministic‡:

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

corre

regre

oken!

ensure

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

‡ 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?

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

‡ A tunable "temperature" controls how probabilistic.

Automated Tests of Probabilistic Generative AI??

- What about benchmarks?
  - Most existing benchmarks aren't specific enough, like good tests should be.
    - They aren't written like typical tests.
  - However, more specific, use case focused benchmarks should 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 the generated synthetic data and reject bad Q&A examples.

Automated Tests of Probabilistic Generative AI??

- Can Data Science help?
  - 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

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.