Red Hat Summit

Connect

InstructLab

Introducing a new community based approach to truly open source LLMs

Henrik Løvborg
Tech Sales Leader Denmark
Red Hat

Red Hat

*I am not one trying to predict the future of technology, but I think **this** is a safe prediction. AI won't be built by a single vendor. It isn't going to revolve around a single monolithic model. Your choice of where to run AI will be everywhere, and it's going to be based on open source*

Matt Hicks
CEO, Red Hat

Source:
Red Hat Summit '24 Keynote

Red Hat

Henrik Løvborg

Tech Sales Leader Denmark
Red Hat

Why do we care about LLMs?

The potential of AI/ML

AI/ML tools such as chatGPT are causing seismic change in the enterprise.

Adoption rates of **100 million users in less than 2 months** demonstrate the rapid acceptance and adoption of AI/ML.[1]

The potential of AI/ML

The investment in AI/ML will grow exponentially and those who bring intelligent applications to market faster will win.

"The reality is, **AI offers solutions to everything we are facing at the moment.** AI can be a source for fast-tracking digital transformation journeys, enable cost savings in times of staggering inflation rates, and support automation efforts in times of labor shortages."

Rasmus Andsbjerg
Associate Vice President, Data & Analytics at IDC

Challenges with LLMs?

Red Hat

Challenges with LLMs

Model size

Often pretty big, making them demanding and costly to host

Efficiency

Even the big ones will have its limitations leading to hallucinations and probable loss of faith and goodwill

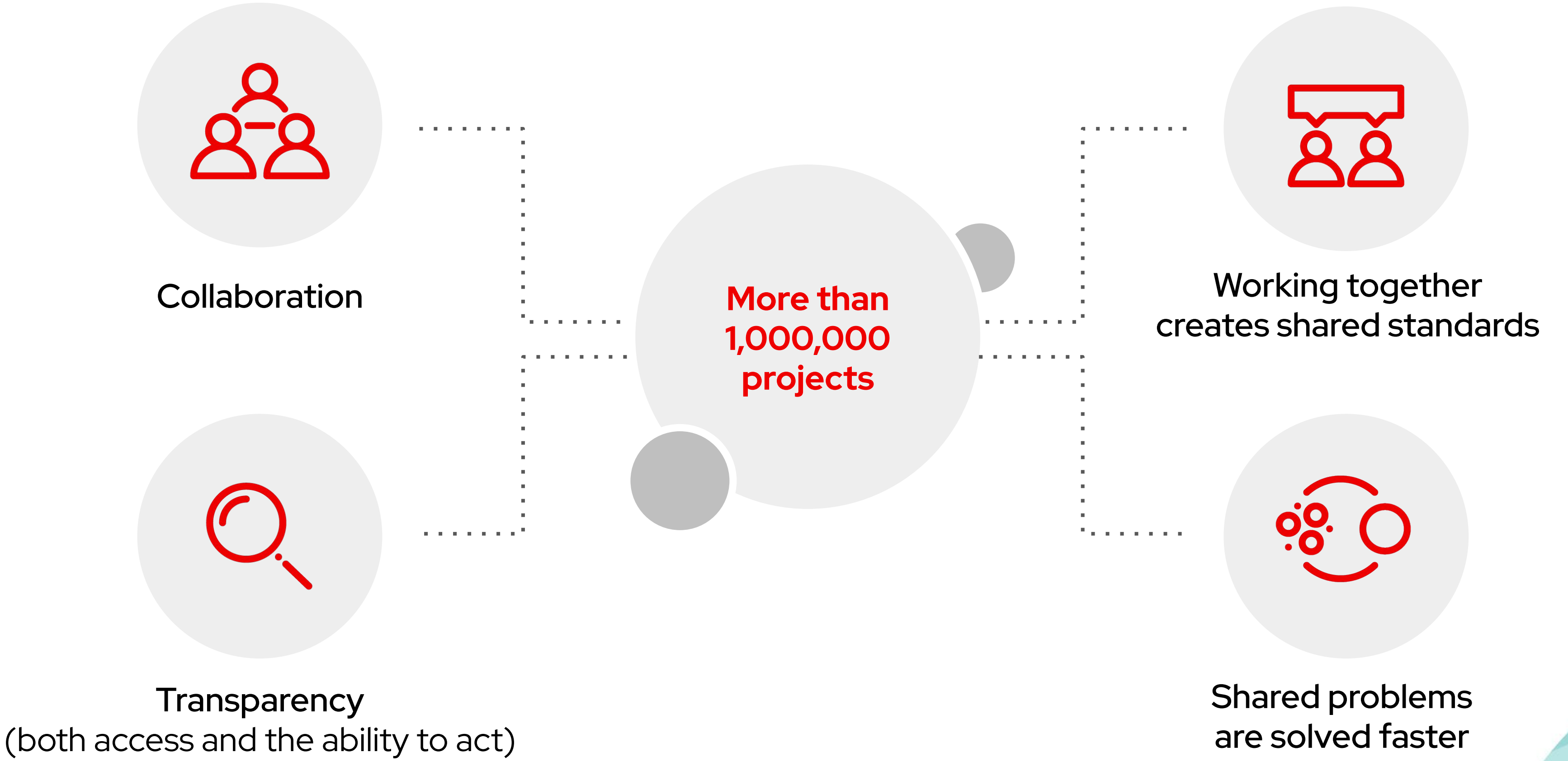
Training

Training them is often pretty complicated and has requires quite large setups

Why do we care about Open Source?

Red Hat

Benefits of Open Source

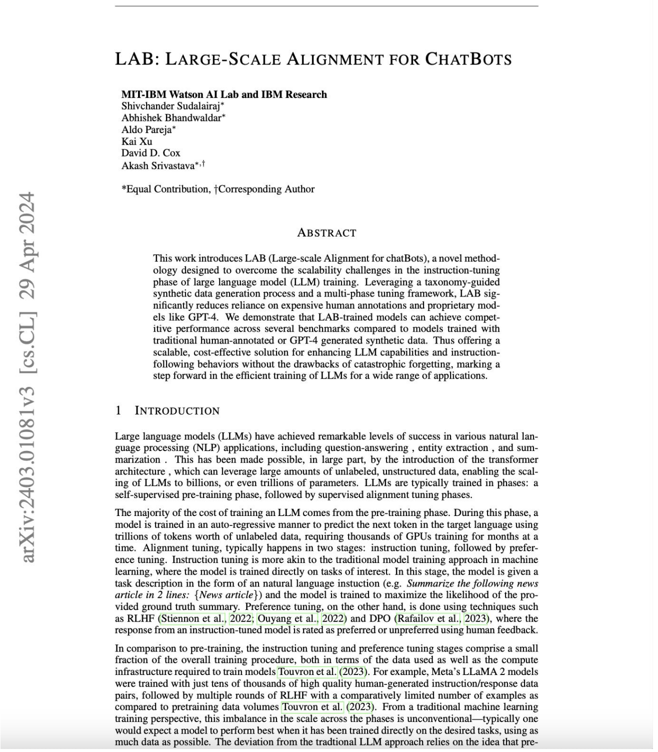


Red Hat

What is InstructLab?

Red Hat

Large-scale **A**lignment of chat **B**ots



arXiv:2403.01081v3 [cs.CL] 29 Apr 2024

LAB: LARGE-SCALE ALIGNMENT FOR CHATBOTS

MIT-IBM Watson AI Lab and IBM Research
Shivchander Sudalairaj*
Abhishek Bhandwaldar*
Aldo Pareja*
Kai Xu
David D. Cox
Akash Srivastava*,†

*Equal Contribution, †Corresponding Author

ABSTRACT

This work introduces LAB (Large-scale Alignment for chatBots), a novel methodology designed to overcome the scalability challenges in the instruction-tuning phase of large language model (LLM) training. Leveraging a taxonomy-guided synthetic data generation process and a multi-phase tuning framework, LAB significantly reduces reliance on expensive human annotations and proprietary models like GPT-4. We demonstrate that LAB-trained models can achieve competitive performance across several benchmarks compared to models trained with traditional human-annotated or GPT-4 generated synthetic data. Thus offering a scalable, cost-effective solution for enhancing LLM capabilities and instruction-following behaviors without the drawbacks of catastrophic forgetting, marking a step forward in the efficient training of LLMs for a wide range of applications.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable levels of success in various natural language processing (NLP) applications, including question-answering , entity extraction , and summarization . This has been made possible, in large part, by the introduction of the transformer architecture , which can leverage large amounts of unlabeled, unstructured data, enabling the scaling of LLMs to billions, or even trillions of parameters. LLMs are typically trained in phases: a self-supervised pre-training phase, followed by supervised alignment tuning phases.

The majority of the cost of training an LLM comes from the pre-training phase. During this phase, a model is trained in an auto-regressive manner to predict the next token in the target language using trillions of tokens worth of unlabeled data, requiring thousands of GPUs training for months at a time. Alignment tuning, typically happens in two stages: instruction tuning, followed by preference tuning. Instruction tuning is more akin to the traditional model training approach in machine learning, where the model is trained directly on tasks of interest. In this stage, the model is given a task description in the form of an natural language instuction (e.g. *Summarize the following news article in 2 lines:* {*News article*}) and the model is trained to maximize the likelihood of the provided ground truth summary. Preference tuning, on the other hand, is done using techniques such as RLHF (Stiennon et al., 2022; Ouyang et al., 2022) and DPO (Rafailov et al., 2023), where the response from an instruction-tuned model is rated as preferred or unpreferred using human feedback.

In comparison to pre-training, the instruction tuning and preference tuning stages comprise a small fraction of the overall training procedure, both in terms of the data used as well as the compute infrastructure required to train models Touvron et al. (2023). For example, Meta's LLaMA 2 models were trained with just tens of thousands of high quality human-generated instruction/response data pairs, followed by multiple rounds of RLHF with a comparatively limited number of examples as compared to pretraining data volumes Touvron et al. (2023). From a traditional machine learning training perspective, this imbalance in the scale across the phases is unconventional—typically one would expect a model to perform best when it has been trained directly on the desired tasks, using as much data as possible. The deviation from the tradtional LLM approach relies on the idea that pre-

Red Hat

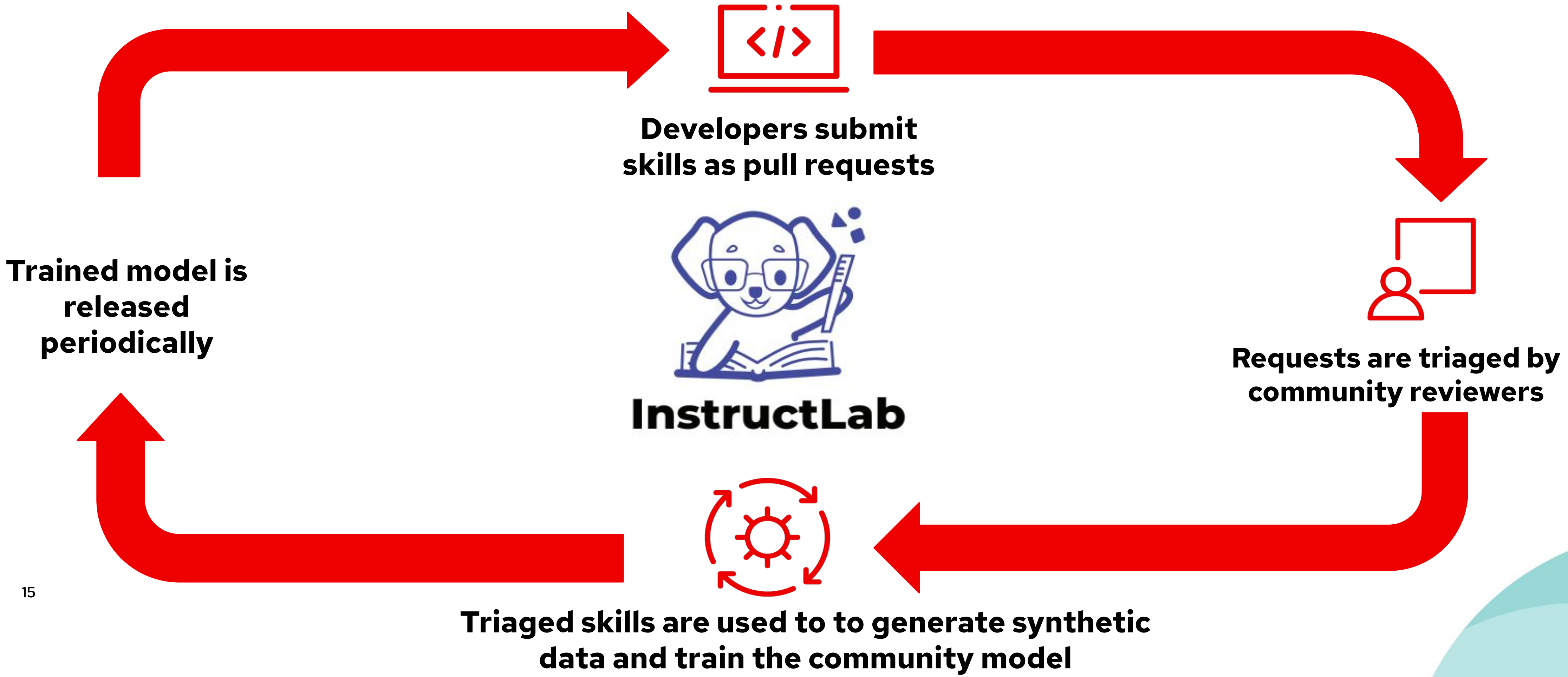
Taxonomy

```
- knowledge
  - science
    - astronomy
- foundational_skills
  - spelling
- composite
  - writing
```

```yaml
seed_examples:
  - context: |
      | **Breed**       | **Size**       | **Barking** | **Energy** |
      |----------------|---------------|-------------|------------|
      | Afghan Hound   | 25-27 in      | 3/5         | 4/5        |
      | Labrador       | 22.5-24.5 in | 3/5         | 5/5        |
      | Cocker Spaniel | 14.5-15.5 in | 3/5         | 4/5        |
      | Poodle (Toy)   | <= 10 in      | 4/5         | 4/5        |
    question: |
      Which breed has the most energy?
    answer: |
      The breed with the most energy is the Labrador.
```

InstructLab cycle

Open source community project for GenAI model development



How do you use InstructLab?

What to do next?

Ideas for InstructLab

What do I do with it?

Business processes

Tell it about your organization, customer service guidelines, business processes etc and let chat help you.

Code reviews

Feed it code and your comments to it, to let it know how you review code (or anything else for that matter) and let it help you review!

You know best!

Since it's so easy to write these augmentations, what better way to tap into the creative minds of your own company!

Take it to production

InstructLab

STEP 1

Learn and experiment via limited desktop-scale training method (qlora) on small datasets.

Laptop / desktop

Red Hat Enterprise Linux AI

STEP 2

Production-grade model training using full synthetic data generation, teacher and critic models. Tooling focused on scriptable primitives.

Server / VM

Red Hat OpenShift AI

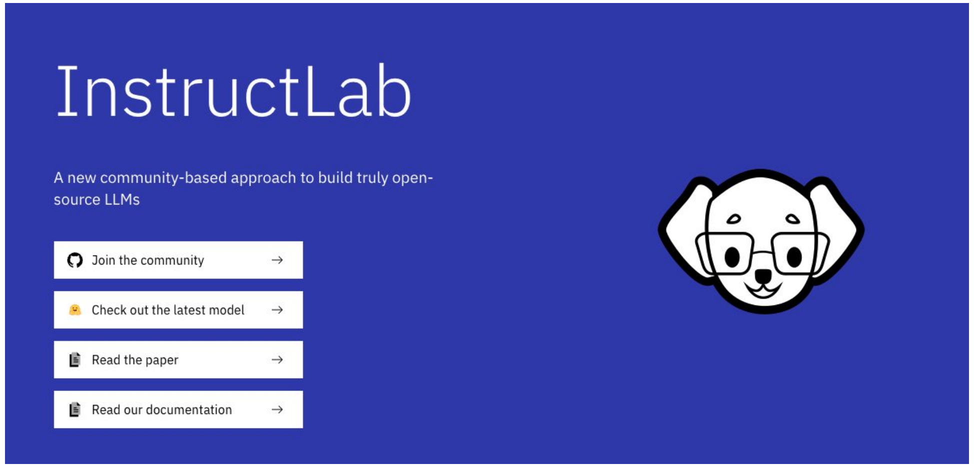
STEP 3

Production-grade model training as in RHEL AI, using full power of Kubernetes scaling, automation, and MLOps services.

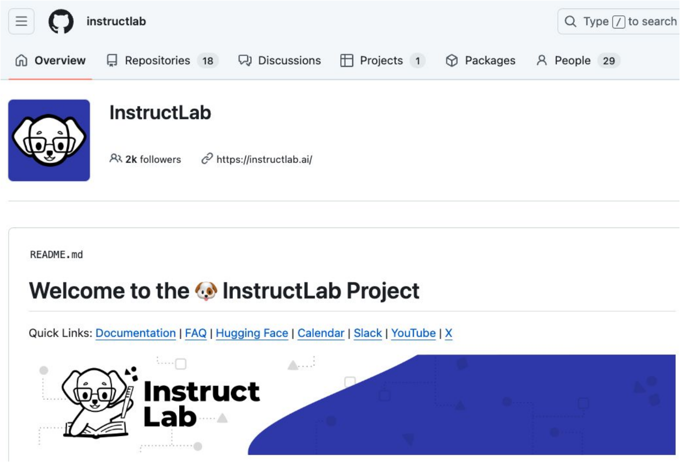
Cluster

contribute!

instructlab.ai



github.com/instructlab



Red Hat Summit

Connect

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat