

IBM Research

InstructLab and document AI.

A journey between IBM Research and Red Hat.

—

Dr. Michele Dolfi

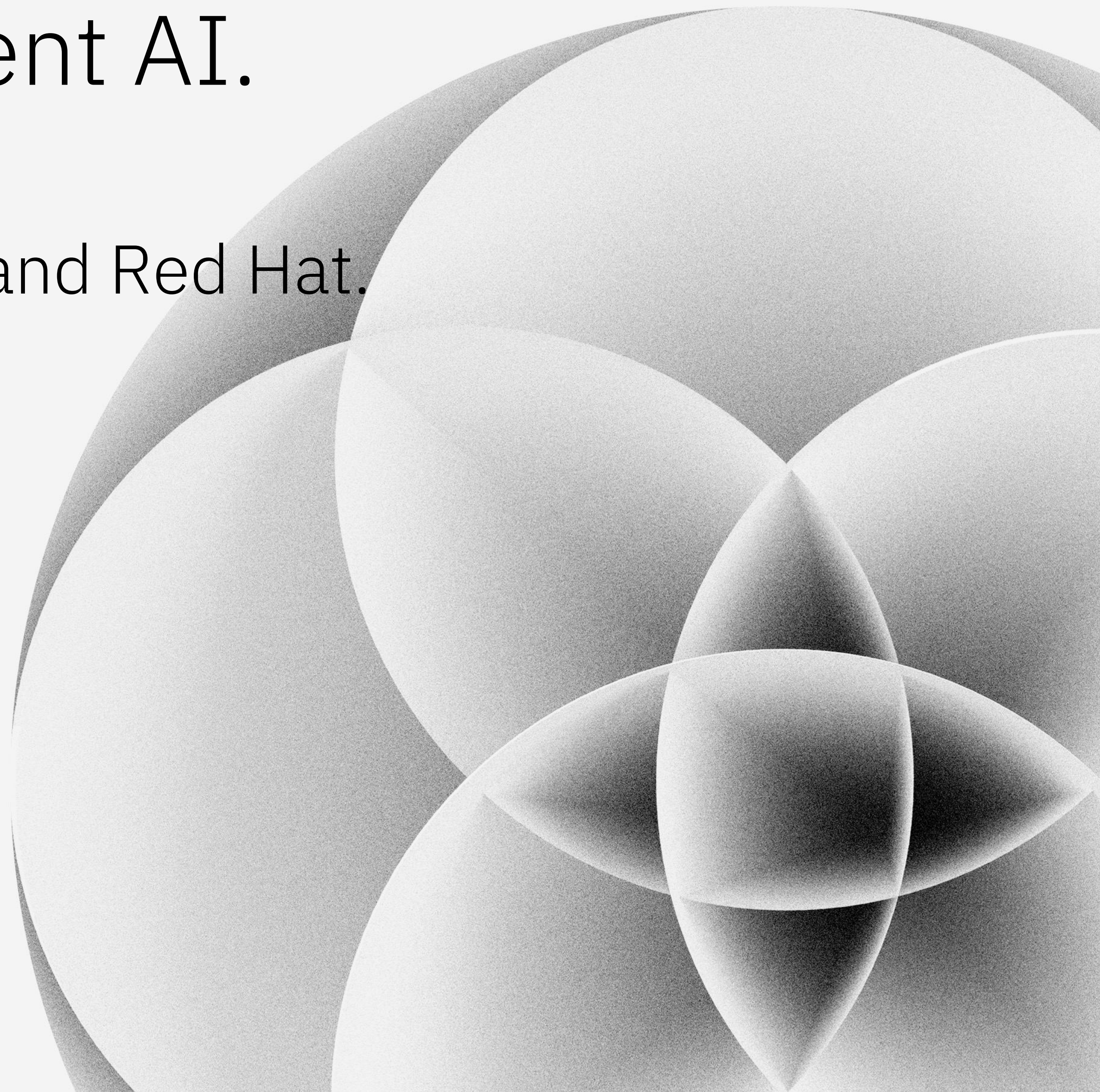
Senior Technical Staff Member | AI for Knowledge

IBM Research – Europe

Carol Chen

Principal AI Community Architect

Red Hat





InstructLab

A new community-based approach to build truly open-source LLMs

<https://github.com/instructlab/>
<https://instructlab.ai>

Open LLMs versus Open Source Software Today

Successful open-source software

- Frequent updates
- Highly structured release cycles
- Well-defined APIs
- Well-defined feature roadmaps
- Incremental contributions
- Community contributions
- Contributions can be merged and reconciled
- Almost any developer can contribute

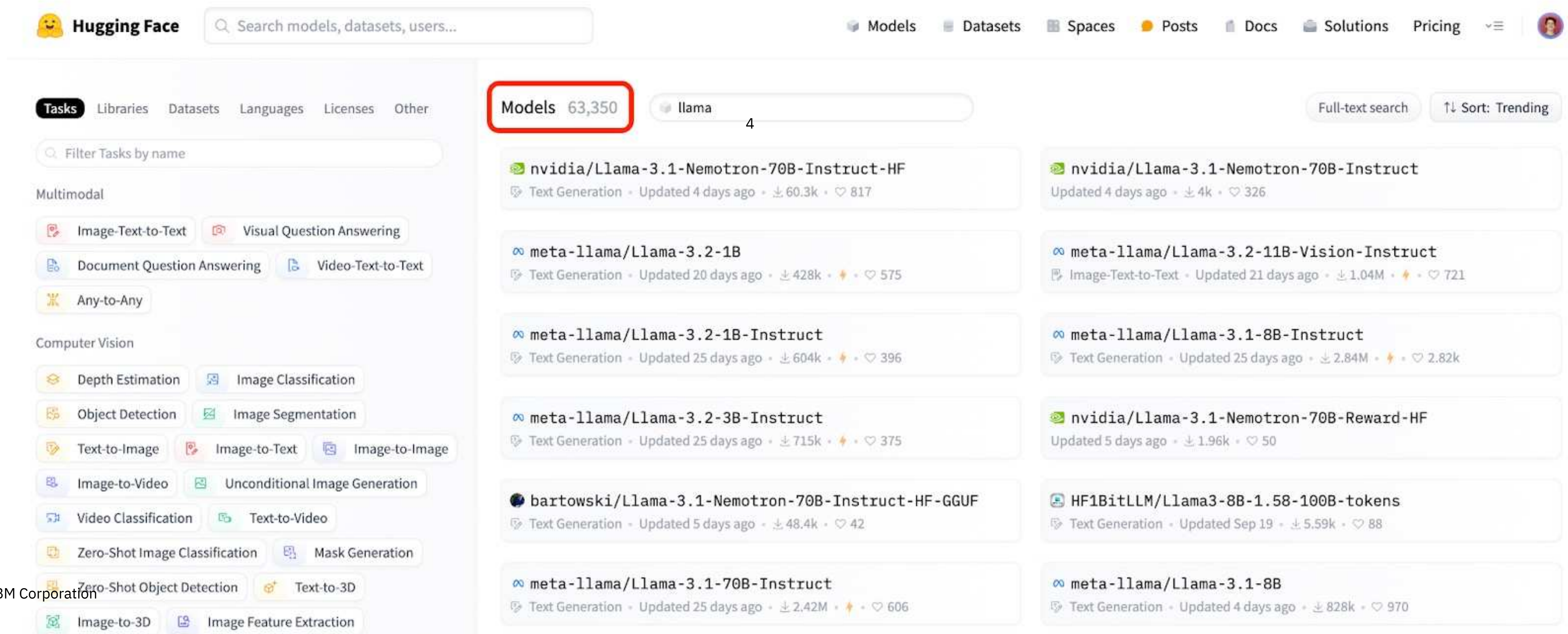
Today's open-source LLMs

- × Monolithic releases at irregular intervals
- × Releases at irregular intervals
- × Guess and try prompt engineering
- × No one knows what's coming
- × Monolithic development
- × Large single-party development
- × Mutually incompatible contributions
- × High barrier to contribution

Today's challenge:

There is no clear path to merge and combine open-source contributions

Contributions built on Llama result in... more llamas...



The screenshot shows the Hugging Face website interface. At the top, there is a search bar with the text "Search models, datasets, users...". To the right of the search bar are navigation links for "Models", "Datasets", "Spaces", "Posts", "Docs", "Solutions", and "Pricing". Below the search bar, the "Models" tab is selected, and the search results are filtered by "llama", showing 4 results. The "Models" count is highlighted with a red box. The search results are sorted by "Trending".

On the left side of the page, there are navigation tabs for "Tasks", "Libraries", "Datasets", "Languages", "Licenses", and "Other". Under "Tasks", there are sub-categories for "Multimodal" (Image-Text-to-Text, Visual Question Answering, Document Question Answering, Video-Text-to-Text) and "Computer Vision" (Depth Estimation, Image Classification, Object Detection, Image Segmentation, Text-to-Image, Image-to-Text, Image-to-Image, Image-to-Video, Unconditional Image Generation, Video Classification, Text-to-Video, Zero-Shot Image Classification, Mask Generation, Zero-Shot Object Detection, Text-to-3D, Image-to-3D, Image Feature Extraction).

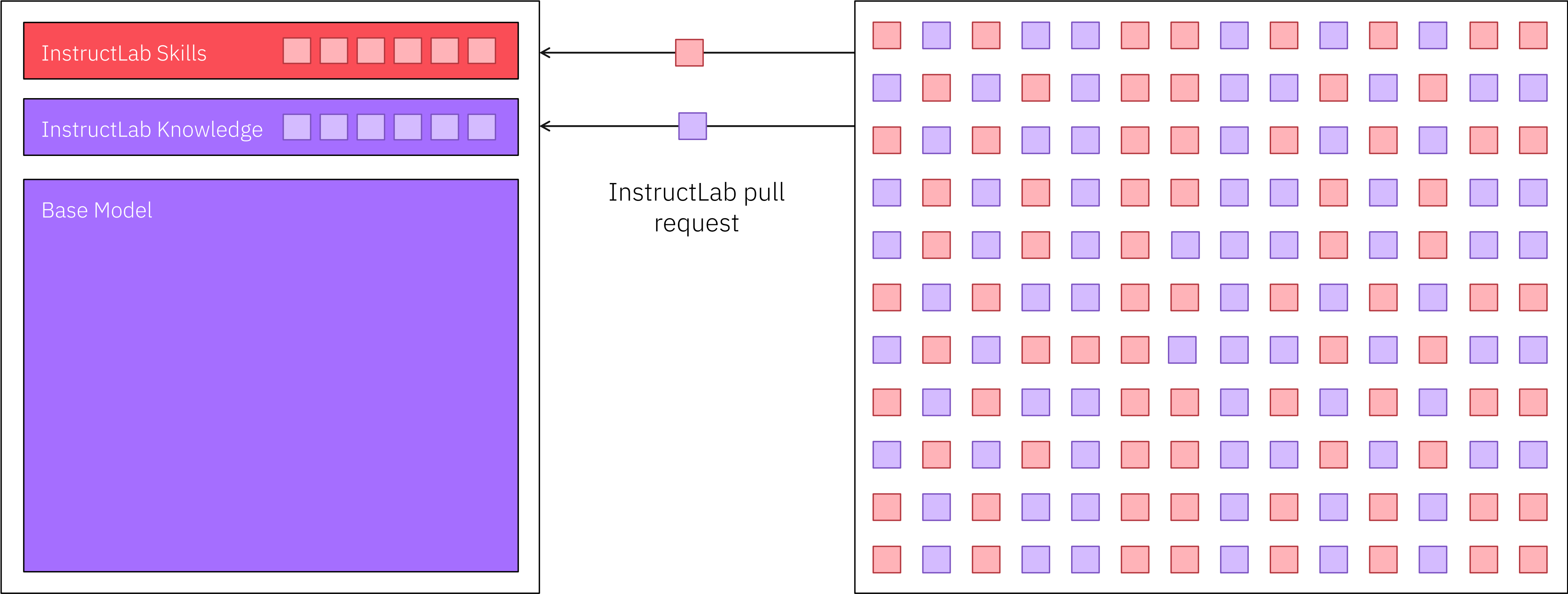
The search results for "llama" are as follows:

- nvidia/Llama-3.1-Nemotron-70B-Instruct-HF**: Text Generation · Updated 4 days ago · ↓ 60.3k · ♥ 817
- nvidia/Llama-3.1-Nemotron-70B-Instruct**: Updated 4 days ago · ↓ 4k · ♥ 326
- meta-llama/Llama-3.2-1B**: Text Generation · Updated 20 days ago · ↓ 428k · ⚡ · ♥ 575
- meta-llama/Llama-3.2-11B-Vision-Instruct**: Image-Text-to-Text · Updated 21 days ago · ↓ 1.04M · ⚡ · ♥ 721
- meta-llama/Llama-3.2-1B-Instruct**: Text Generation · Updated 25 days ago · ↓ 604k · ⚡ · ♥ 396
- meta-llama/Llama-3.1-8B-Instruct**: Text Generation · Updated 25 days ago · ↓ 2.84M · ⚡ · ♥ 2.82k
- meta-llama/Llama-3.2-3B-Instruct**: Text Generation · Updated 25 days ago · ↓ 715k · ⚡ · ♥ 375
- nvidia/Llama-3.1-Nemotron-70B-Reward-HF**: Updated 5 days ago · ↓ 1.96k · ♥ 50
- bartowski/Llama-3.1-Nemotron-70B-Instruct-HF-GGUF**: Text Generation · Updated 5 days ago · ↓ 48.4k · ♥ 42
- HF1BitLLM/Llama3-8B-1.58-100B-tokens**: Text Generation · Updated Sep 19 · ↓ 5.59k · ♥ 88
- meta-llama/Llama-3.1-70B-Instruct**: Text Generation · Updated 25 days ago · ↓ 2.42M · ⚡ · ♥ 606
- meta-llama/Llama-3.1-8B**: Text Generation · Updated 4 days ago · ↓ 828k · ♥ 970

InstructLab offers a new way to make community contributions additive

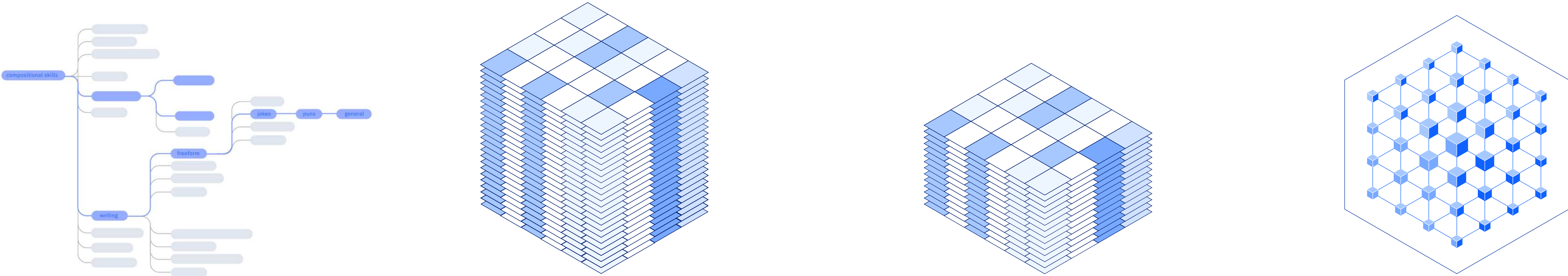
The model stack

The community can create and contribute skills recipes.



InstructLab offers a new way to make community contributions additive

Large-scale Alignment for chatBots (LAB), available on [Arxiv](#)



Taxonomy-based skill & knowledge representation

Represent any missing model knowledge or skills in a hierarchical taxonomy, providing 5+ exemplifying data points of the missing behavior per missing skill.

Synthetic data generation with teacher model

A teacher model generates a “curriculum” of millions of questions and answers across the taxonomy.

Synthetic data validation with critic model

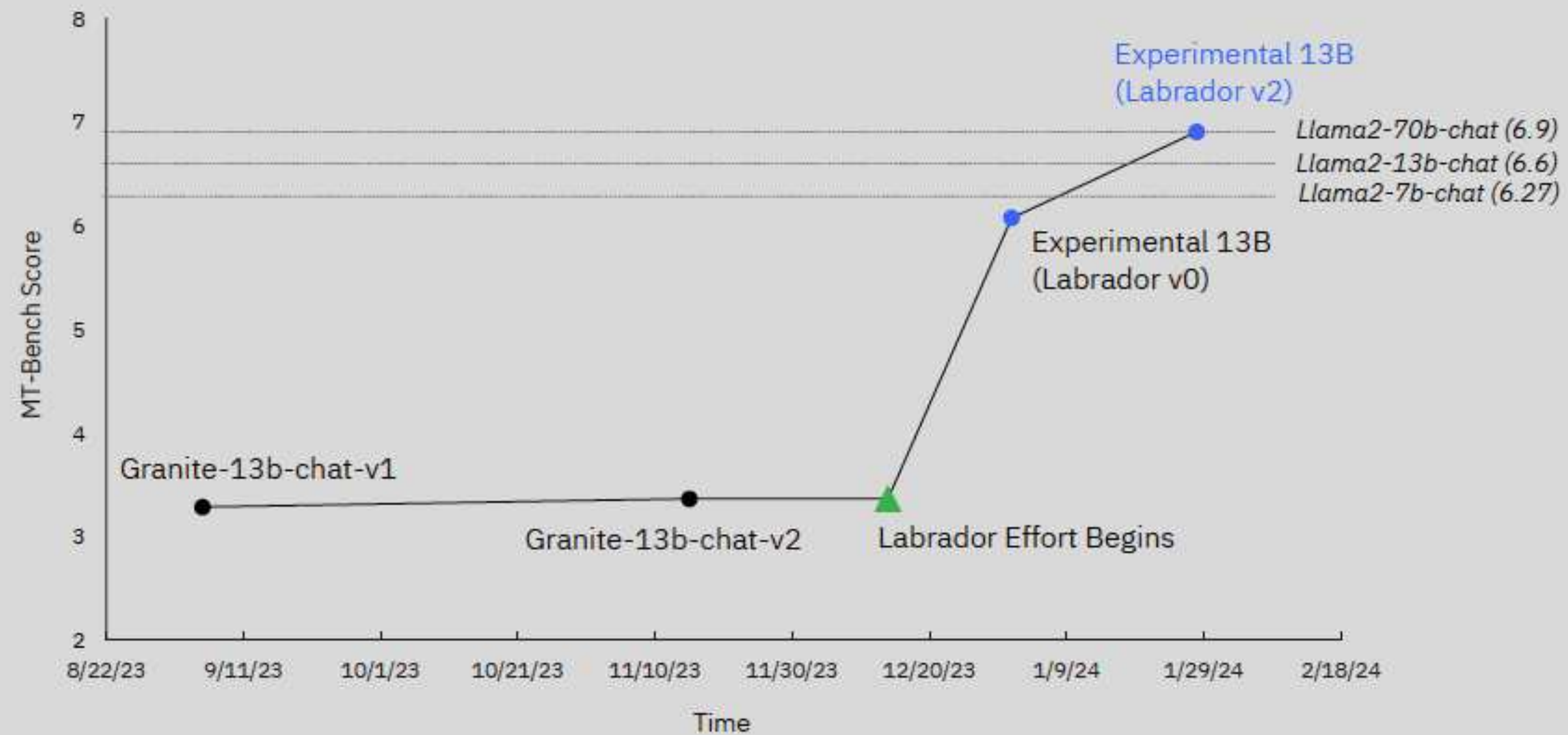
Critic models filter the questions for correctness and quality. Synthetic data is scanned for prohibited material (such as hate, abuse, profanity, and personal information).

Skill and knowledge training on top of student model(s)

The student model is trained with the curriculum using a novel training approach.

Performance improvement of Granite model using InstructLab

Granite-13b-chat-v2.1 is on par with Llama2-70b-chat on MT-Bench evaluations.





Knowledge



Foundational
Skills



Compositional
Skills

Types of Tuning Data in InstructLab

Knowledge

Adding **knowledge** using the InstructLab method means including files with facts, data, policies and more.

Examples of knowledge:

- FedRAMP policies
- The Magna Carta
- Linux `man` pages
- The 7 wonders of the world
- The teachings of Buddha

Foundational Skills

Adding **foundational skills** means adding core differences in interaction patterns.

Examples of foundational skills:

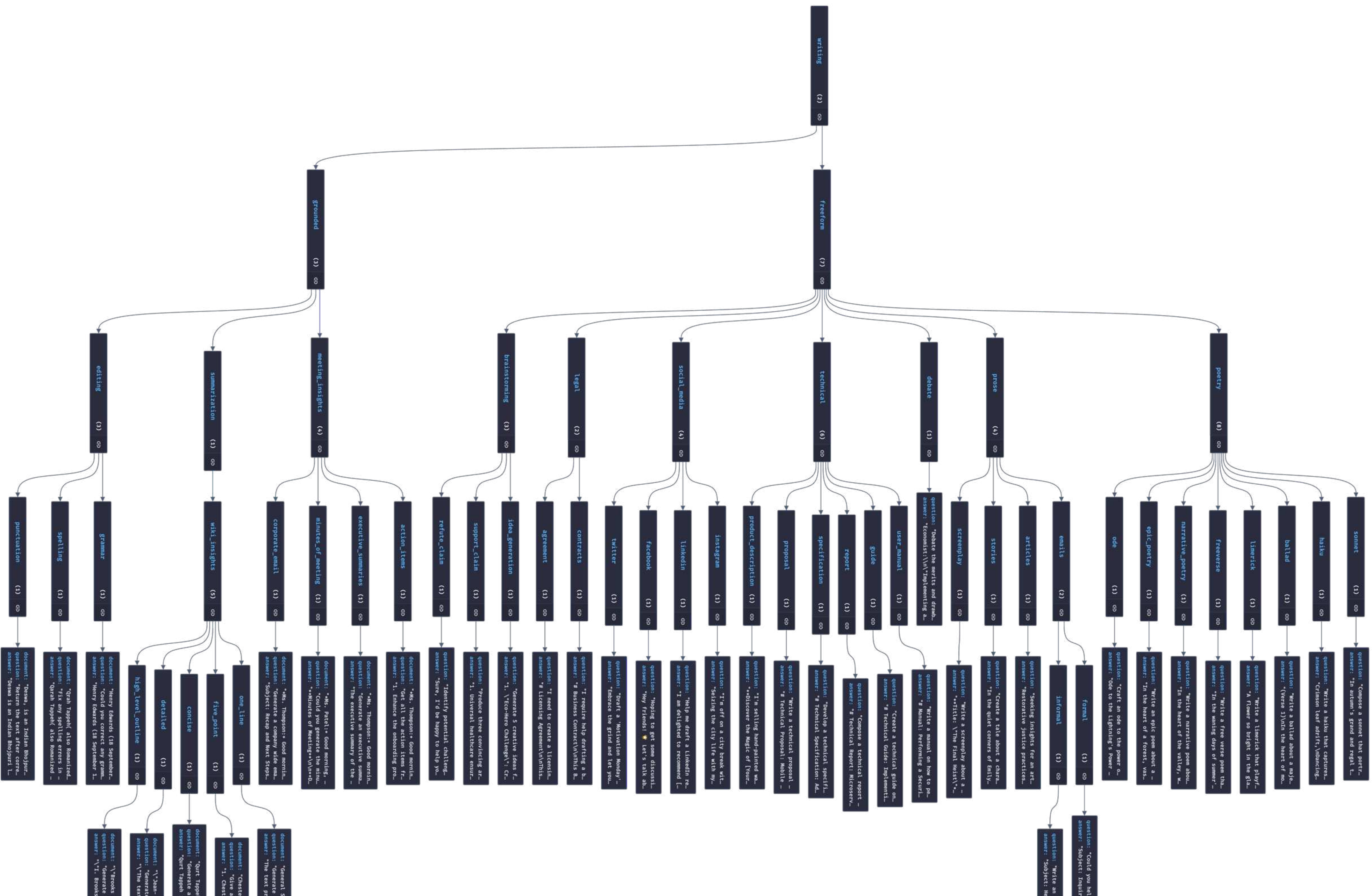
- Coding
- Math
- Writing

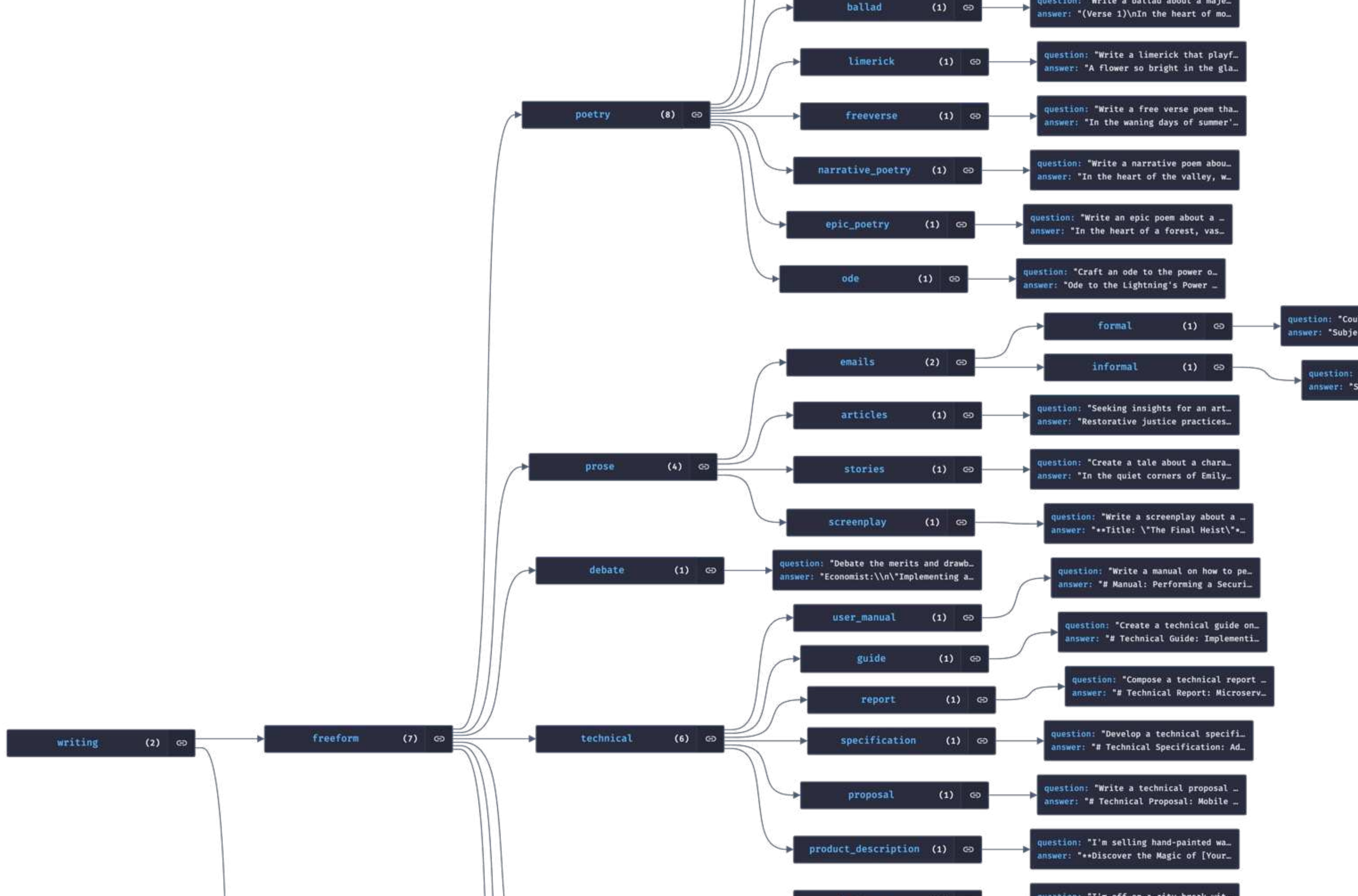
Compositional Skills

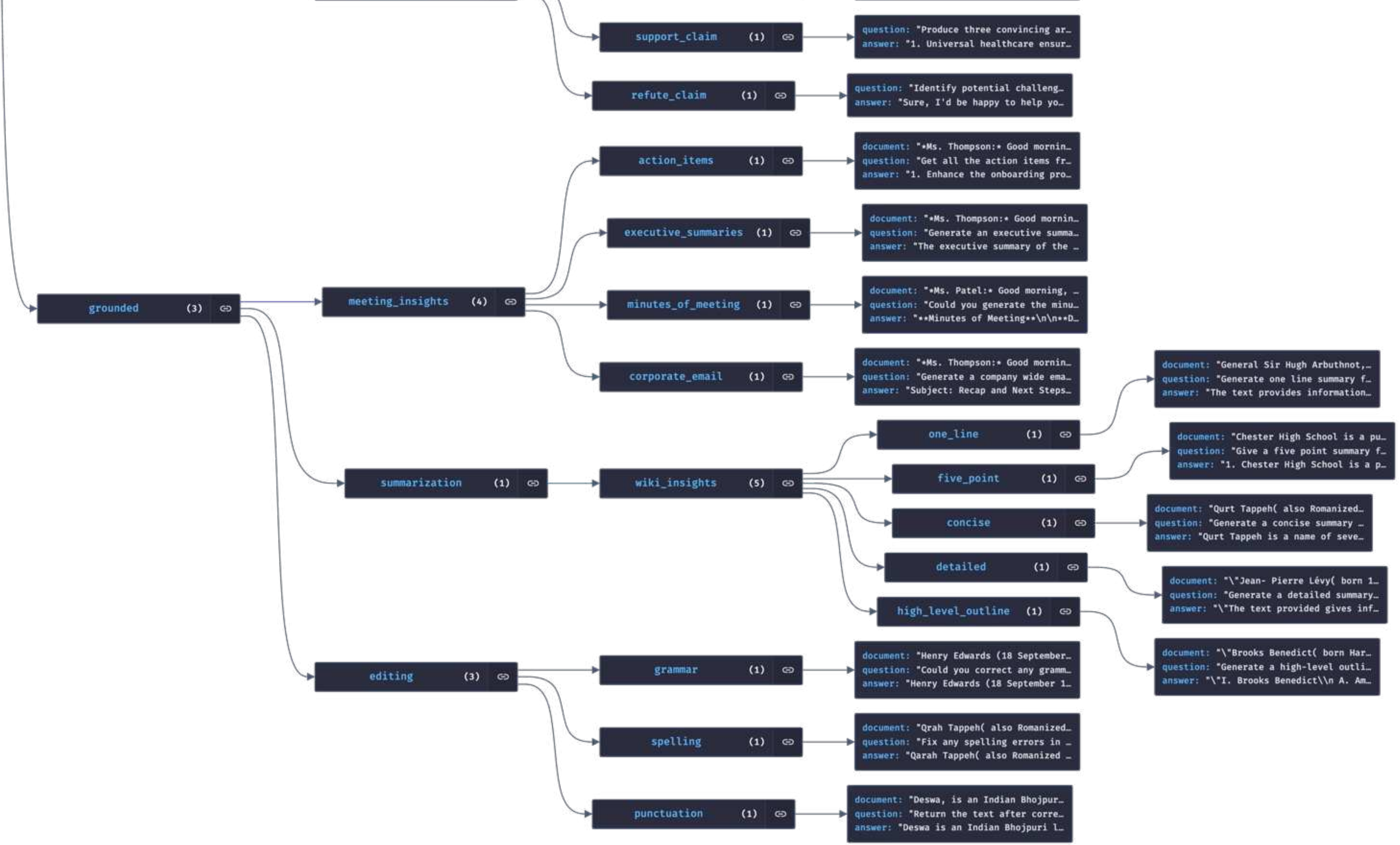
Adding **compositional skills** means adding and improving complex behaviors in the model's responses.

Examples of compositional skills:

- Act like a personality (say, Yoda)
- Convert to camel case
- Write me a limerick
- Answer riddles and trick questions







InstructLab Taxonomy on Github

The screenshot shows the GitHub repository for 'InstructLab Taxonomy'. At the top, it displays the repository name 'taxonomy' as public, with options to edit pins and watch (42). Below this, there are navigation elements for the 'main' branch, 42 branches, and 0 tags. A search bar and 'Add file' and 'Code' buttons are also visible. The main content area lists recent commits by user 'bjhargrave', including updates to 'stale_bot.yml', 'knowledge-contribution-guide.md', and 'check_yaml.py'. At the bottom, there are links for 'README', 'Code of conduct', 'Apache-2.0 license', and 'Security'. The footer features the 'InstructLab Taxonomy' logo.

This example snippet assumes the GitHub username `mairin` and shows some of the question/answer pairs present in the actual file:

```
version: 2
task_description: |
  This skill provides the ability to read a markdown-formatted table.
created_by: mairin # Use your GitHub username; only one creator supported
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.
- context: |
  | **Name** | **Date** | **Color** | **Letter** | **Number** |
  |-----|-----|-----|-----|-----|
  | George | Mar 5 | Green | A | 1 |
  | Gráinne | Dec 31 | Red | B | 2 |
  | Abigail | Jan 17 | Yellow | C | 3 |
  | Bhavna | Apr 29 | Purple | D | 4 |
  | Rémy | Sep 9 | Blue | E | 5 |
question: |
  What is Gráinne's letter and what is her color?
answer: |
  Gráinne's letter is B and her color is red.
- context: |
  | Banana | Apple | Blueberry | Strawberry |
  |-----|-----|-----|-----|
  | Yellow | Red, Green | Blue | Red |
  | Large | Medium | Small | Small |
  | Peel | Peel | No peel | No peel |
question: |
  Which fruit is blue, small, and has no peel?
answer: |
  The blueberry is blue, small, and has no peel.
```

<https://github.com/instructlab/taxonomy>

InstructLab demo

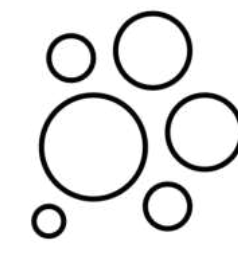
1. Find a task which was not working yet with the default model
 2. Add taxonomy (skills or taxonomy)
 3. Generate new data
 4. Train
 5. Check out the new model
- Most steps can be performed using either
 - CLI version
 - UI version

Which model?



Granite

A family of open, performant and trusted AI models to accelerate enterprise AI adoption



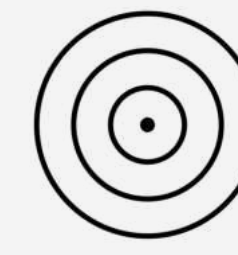
Open

- Open sourced under Apache 2.0
- Seamless deployment and control over environments, models and proprietary data
- Enhances developer efficiency



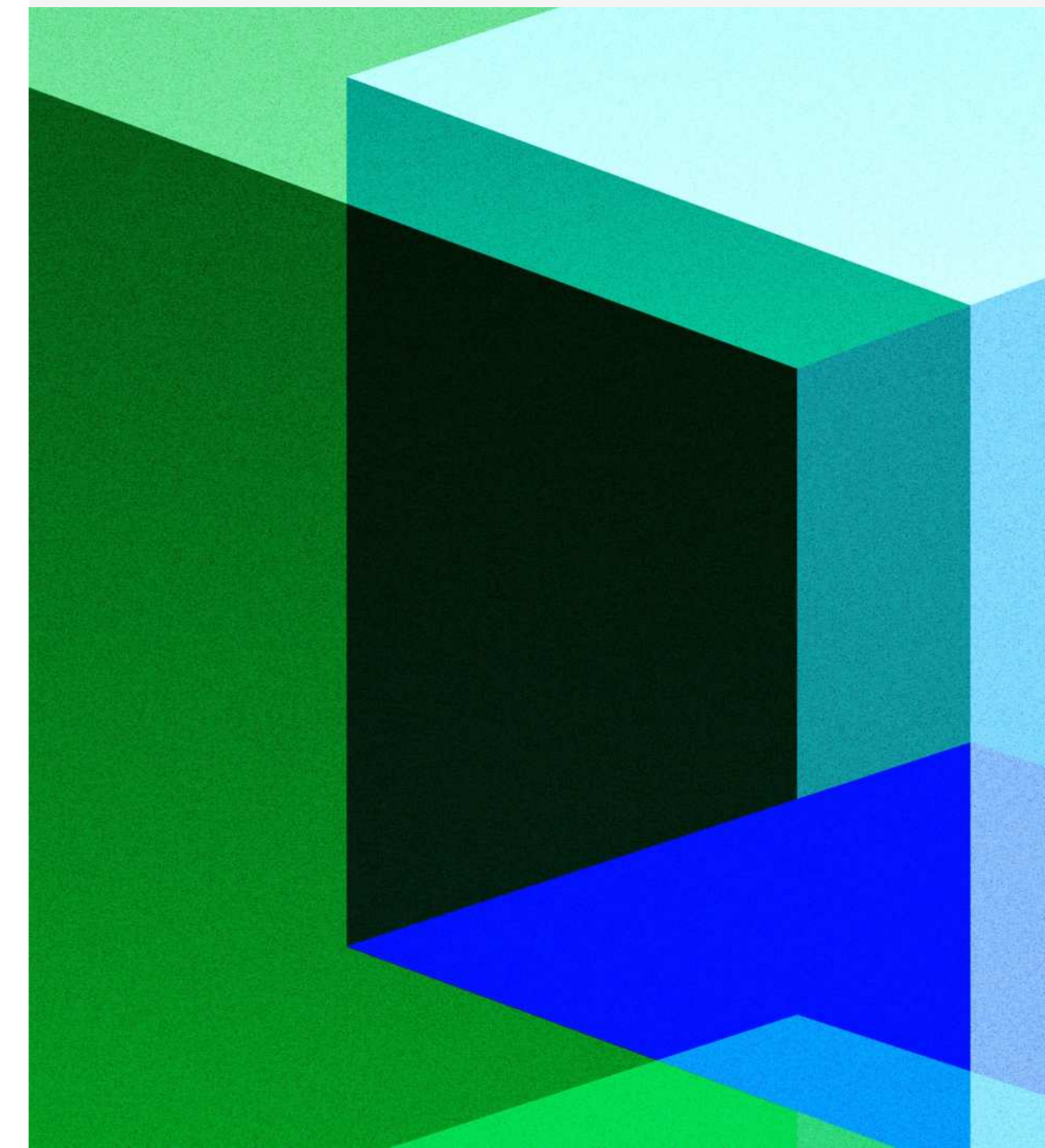
Trust

- Transparency of data, training methods
- IP indemnification
- Responsible and safe AI

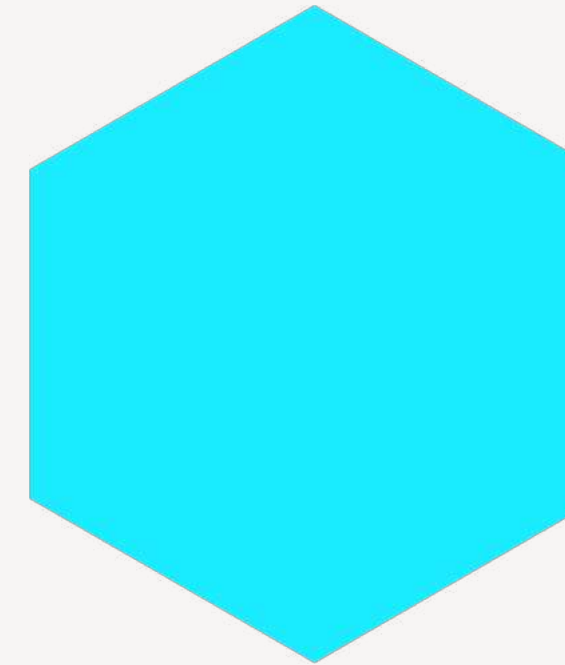
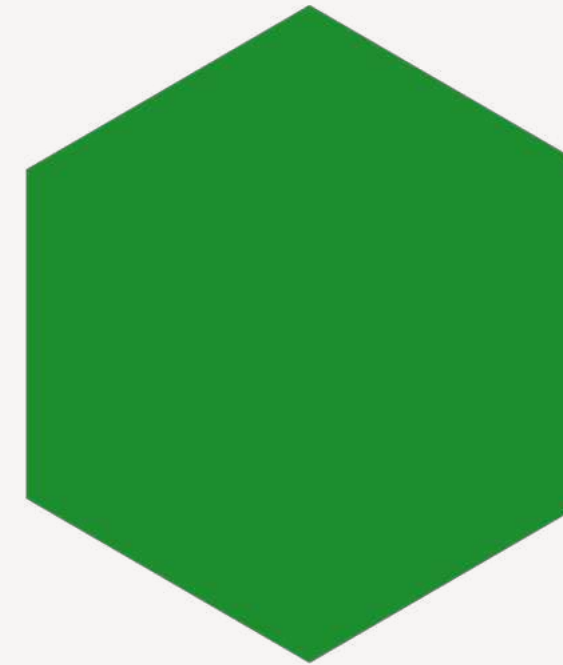
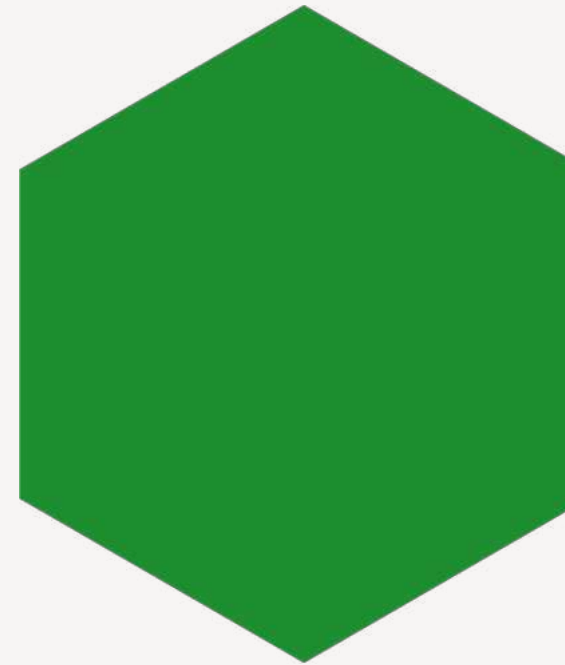
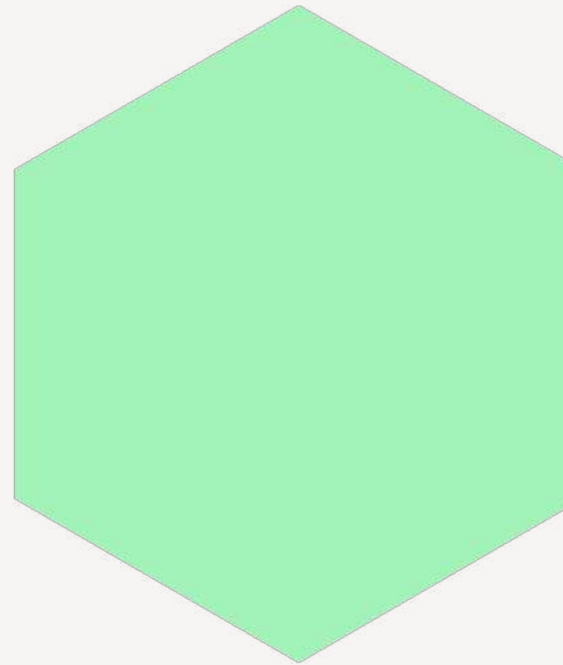
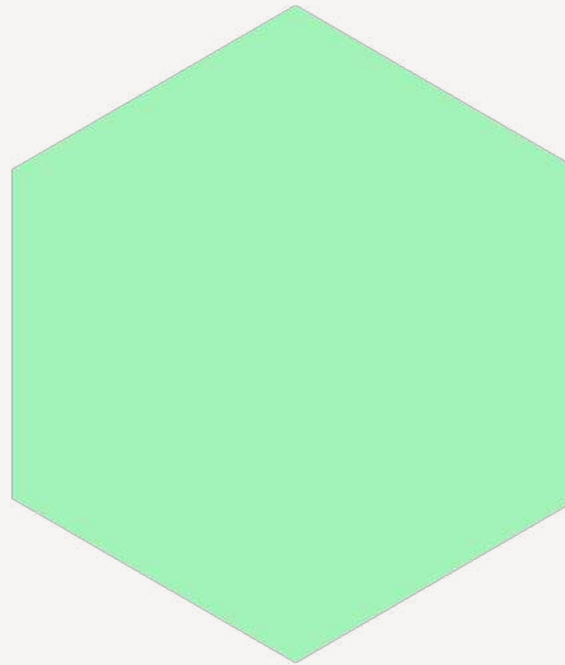
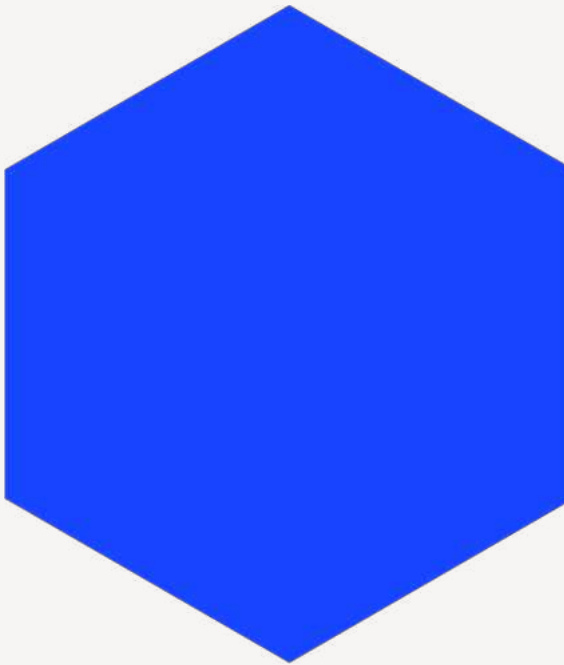
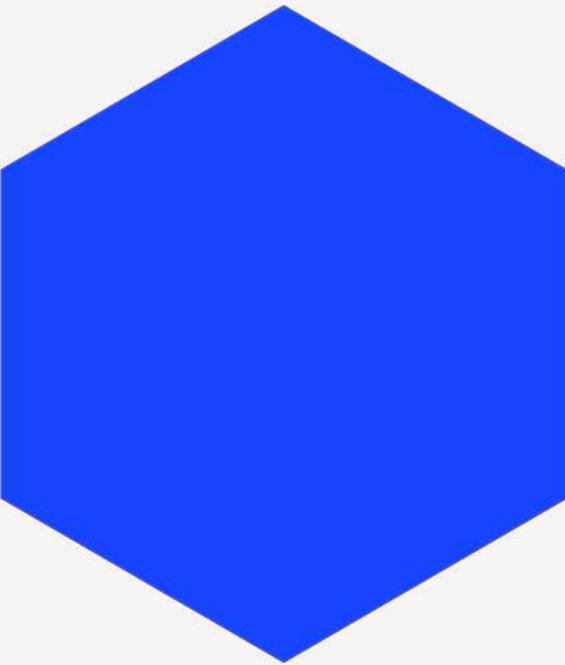


Performant

- Diverse range of fit-for-purpose models
- Designed for scalability



Granite 3.1 Models



Large Language Models
(LLMs) for enterprise

Granite-3.1-8B-Instruct
Granite-3.1-2B-Instruct

Inference-efficient
Mixture of Experts (MoE)

Granite-3.1-3B-A800M
Granite-3.1-1B-A400M

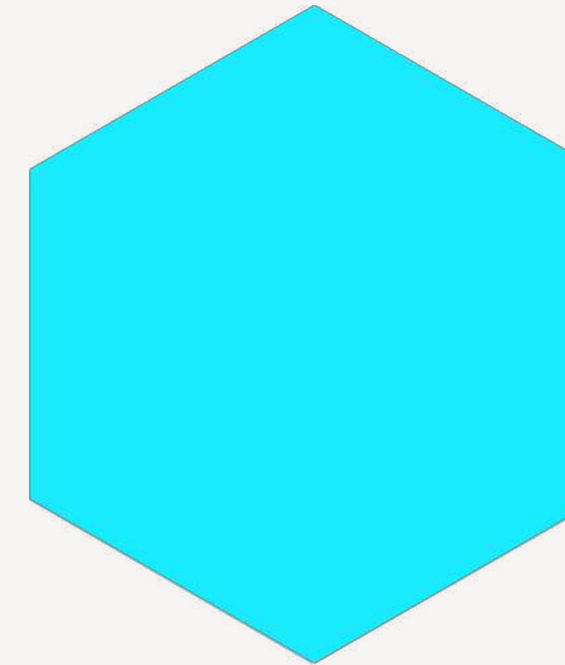
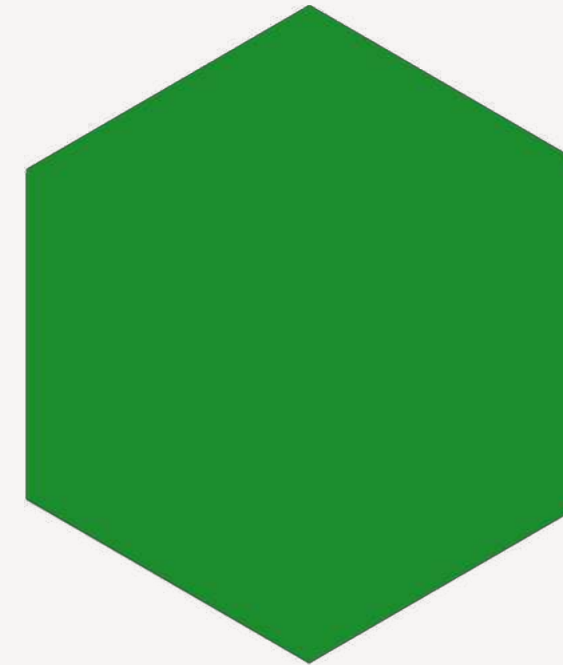
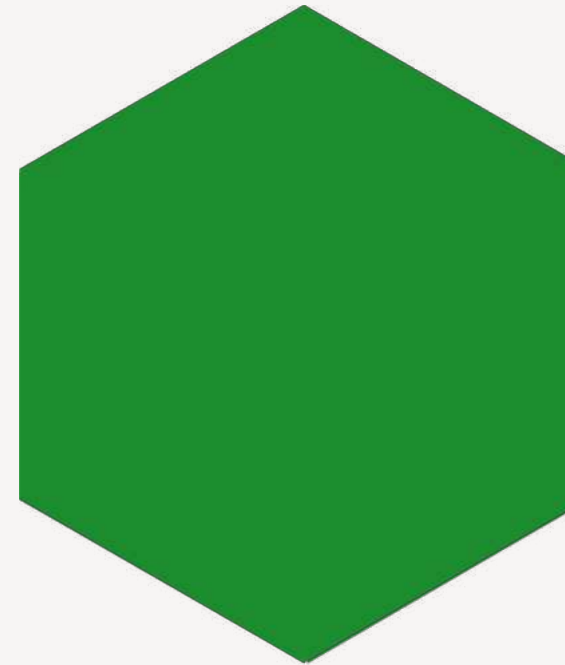
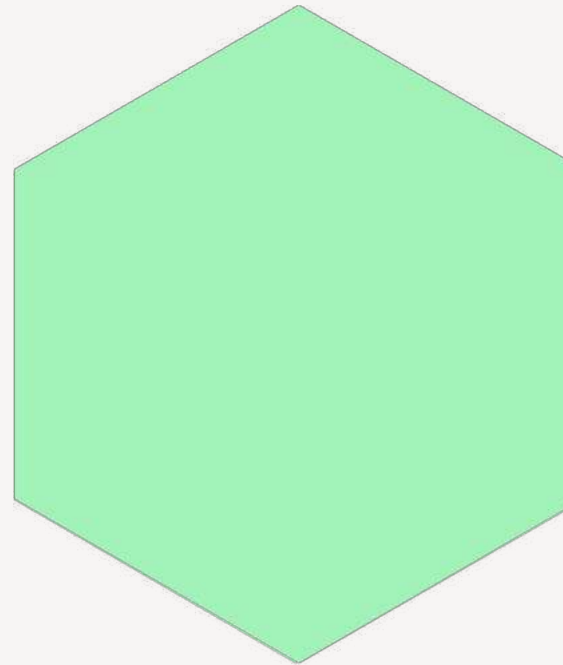
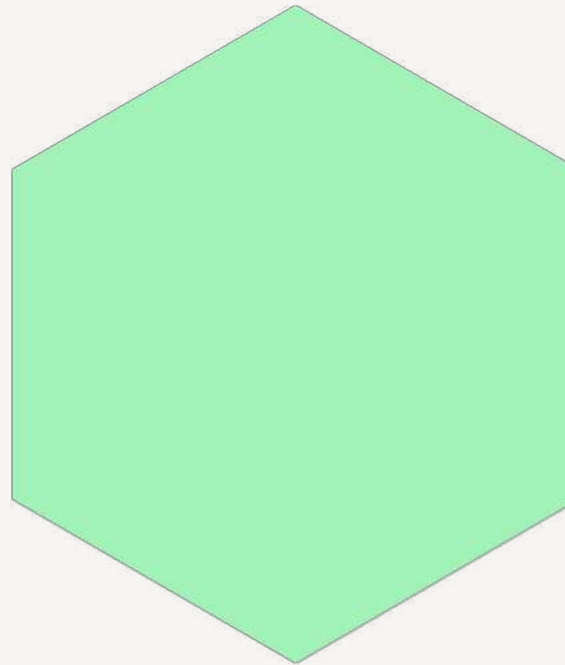
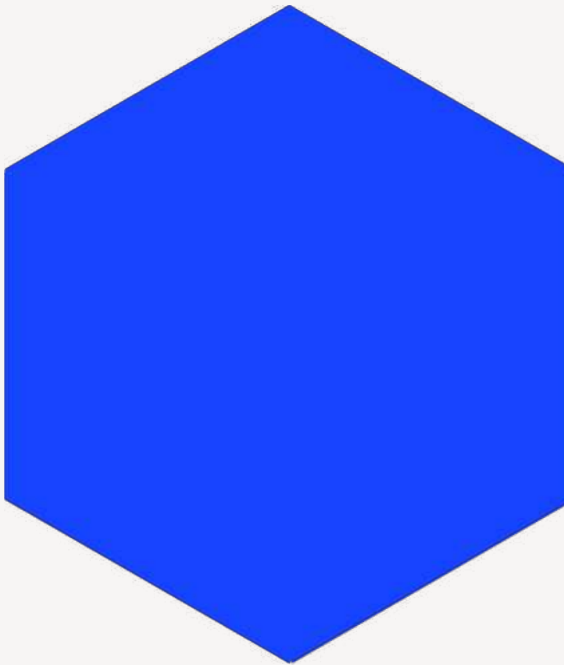
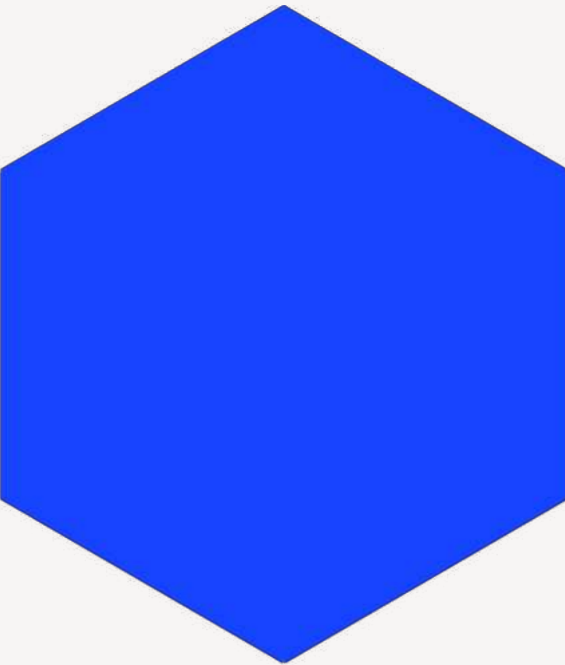
Guardrail models

Granite-Guardian-3.1-8B
Granite-Guardian-3.1-2B

Speculative
decoding models

Granite-3.1-8B-Instruct-
Accelerator

Granite 3.1 Models



New Granite-Dense
architecture

Over 12T tokens training data
across 12 languages and 116
programming languages

Available under
Apache 2.0

Why IBM Granite?

Open

Performant

Trusted

Open

Open-sourced under the permissive Apache 2.0 license

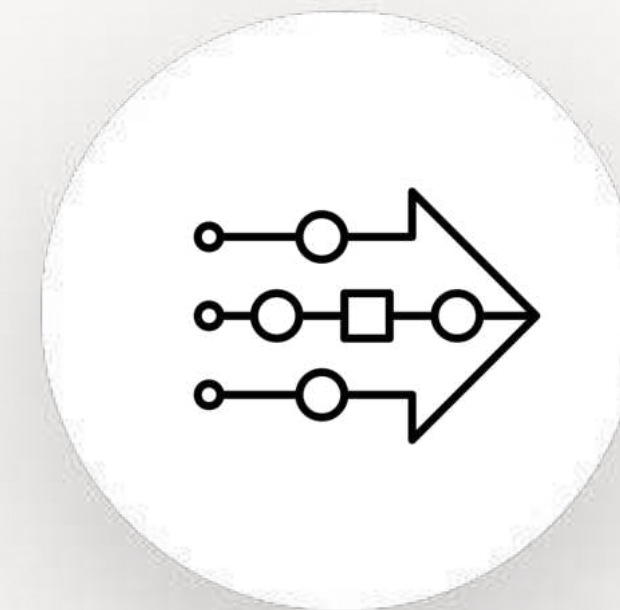
Drive developer
productivity



Build responsibly with
model transparency

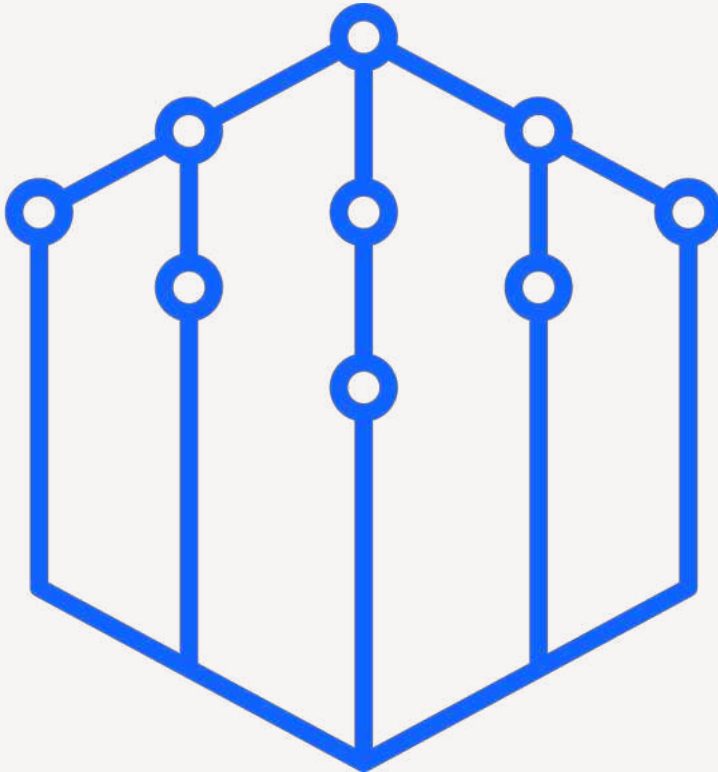


Customize models
with your data

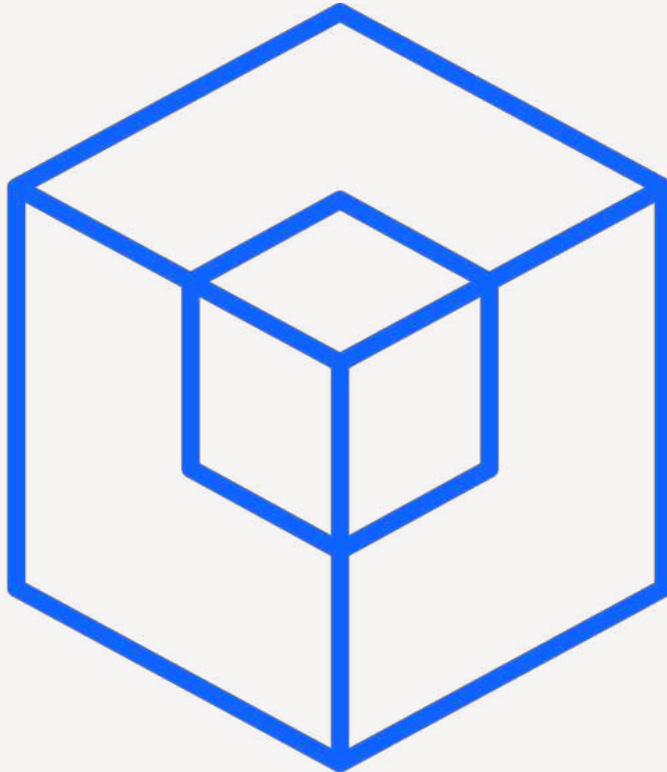


Performant

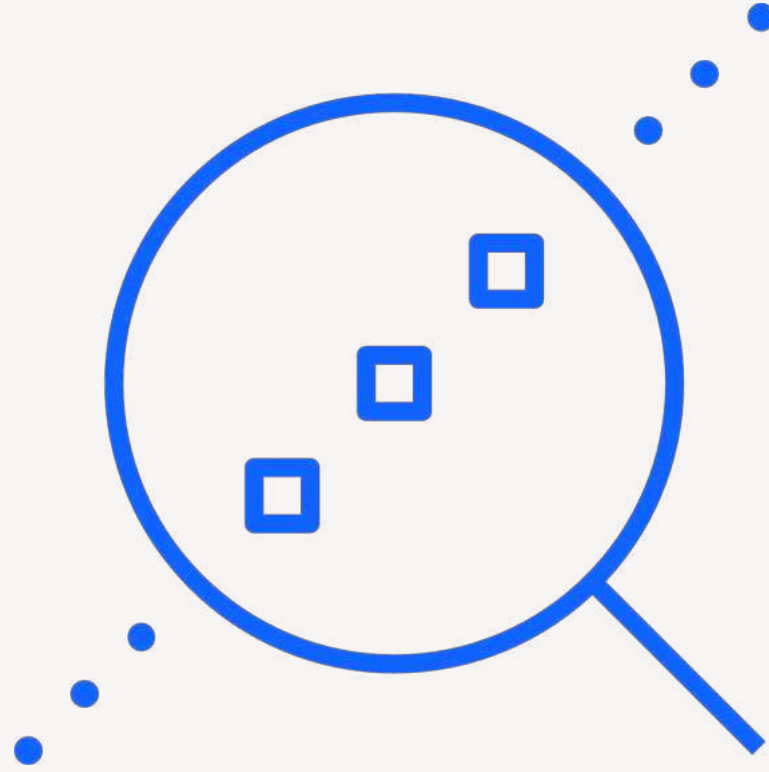
The IBM approach: fit-for-purpose models



Your data



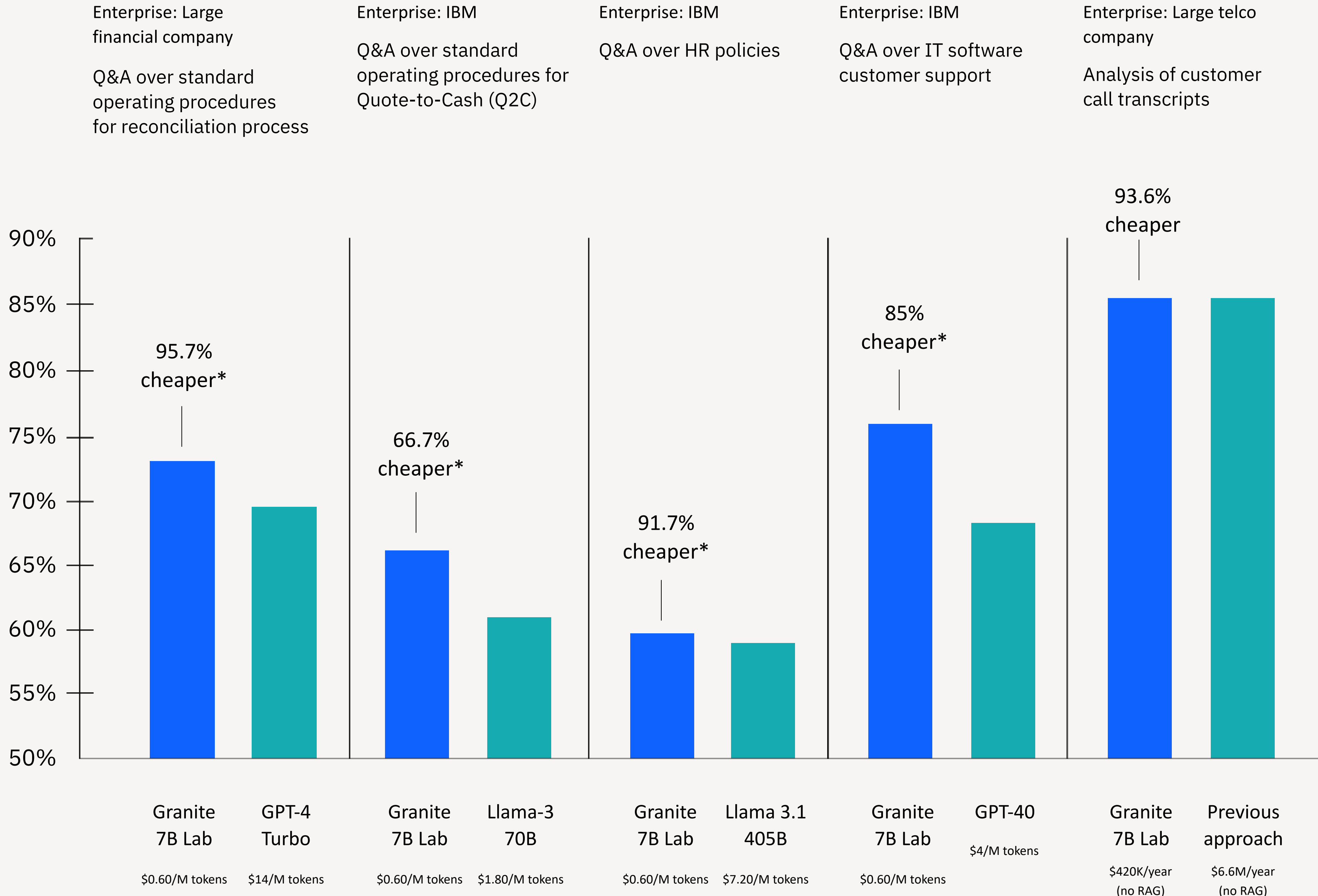
The right model



Targeted use case
fine tuning

Customizing an LLM with InstructLab

Enterprise data enables smaller, targeted, optimized models to deliver state-of-the-art performance at lower costs.



*SaaS cost per million tokens (assuming blend of 80% inout, 20% output), <https://www.ibm.com/products/watsonx-ai/foundation-models>, <https://openai.com/api/pricing/>

Trusted

Ensure safe and responsible AI



Safety training of models

Innate safety training and simulations to protect our models



Intellectual Property (IP) Indemnification

Client protection, accuracy, and trust while using your enterprise data



Guardrails

Secure data and mitigate risks across a variety of prompts and responses

InstructLab demo

1. Find a task which was not working yet with the default model
 2. Add taxonomy (skills or taxonomy)
 3. Generate new data
 4. Train
 5. Check out the new model
- Most steps can be performed using either
 - CLI version
 - UI version



InstructLab demo

Initial test of the student model

Initially the student model doesn't know about the IBM Granite models.

```
$ ilab model chat
system
Welcome to InstructLab Chat w/ MERLINITE-7B-LAB-Q4_K_M.GGUF
>>> What is IBM Granite? [S][default]
merlinite-7b-lab-Q4_K_M.gguf
IBM Granite is a powerful and versatile tool that combines a programming environment with an application server, making it suitable for both web development and application hosting. It offers a user-friendly interface for creating, testing, and deploying applications, as well as features such as:
1. Ease of use: The IBM Granite IDE provides a visual editor that simplifies the creation of web pages, forms, and other elements.
2. Integration with Java: It offers seamless integration with Java technology, allowing developers to leverage the vast ecosystem of Java libraries and frameworks.
3. Support for open standards: IBM Granite supports various open standards, including HTML5, CSS3, and JavaScript, ensuring compatibility and interoperability with other web technologies.
4. Scalability: IBM Granite can handle high-traffic websites and applications due to its modular architecture and efficient resource management.
5. Security: It includes built-in security features, such as authentication and authorization mechanisms, to protect sensitive data and ensure secure communication between components.
6. Performance: IBM Granite provides advanced caching and load balancing techniques to optimize the performance of web applications and improve user experience.
7. Community support: IBM Granite benefits from an active developer community that shares resources, best practices, and tutorials, fostering a collaborative learning environment.
To get started with IBM Granite, you can visit the official website (<https://www.ibm.com/software/products/en/granite>) to download the software, explore the documentation, and access the community forums for assistance. Additionally, various online resources, such as tutorials and video courses, are available to help you master IBM Granite and develop impressive web applications.
If you have any specific questions about IBM Granite or need guidance on a particular feature, feel free to ask!
elapsed 7.229 seconds
```


InstructLab demo

Initial test of the student model

Initially the student model doesn't know about the IBM Granite models.

Add new knowledge

The taxonomy is enhanced with knowledge from the relative Wikipedia page.

The screenshot shows the InstructLab web application interface. At the top, there is a navigation bar with the InstructLab logo and the name 'InstructLab' on the left, and a user profile for 'Michele Dolfi' on the right. Below the navigation bar is a sidebar menu with the following items: 'Dashboard', 'Contribute' (with a dropdown arrow), 'Skill', 'Knowledge', 'Playground' (with a dropdown arrow), 'Chat', and 'Custom Model Endpoints'. The main content area is titled 'Dashboard' and features a section for 'My Submissions' with the subtitle 'View and manage your taxonomy contributions.?' Below this is a large light gray area containing the InstructLab logo (a cartoon dog wearing glasses and holding a pencil) and the text 'InstructLab'. Underneath the logo is a 'Welcome to InstructLab' message, followed by a paragraph explaining that InstructLab is a tool for advancing generative AI through community collaboration and open-source principles. Below the paragraph is a line of text: 'To get started, contribute a skill or contribute knowledge.' At the bottom of this section are three blue buttons: 'Contribute Skill', 'Contribute Knowledge', and 'Chat with the Models'. Below the buttons is a link: 'View the Project on Github' with a GitHub icon.

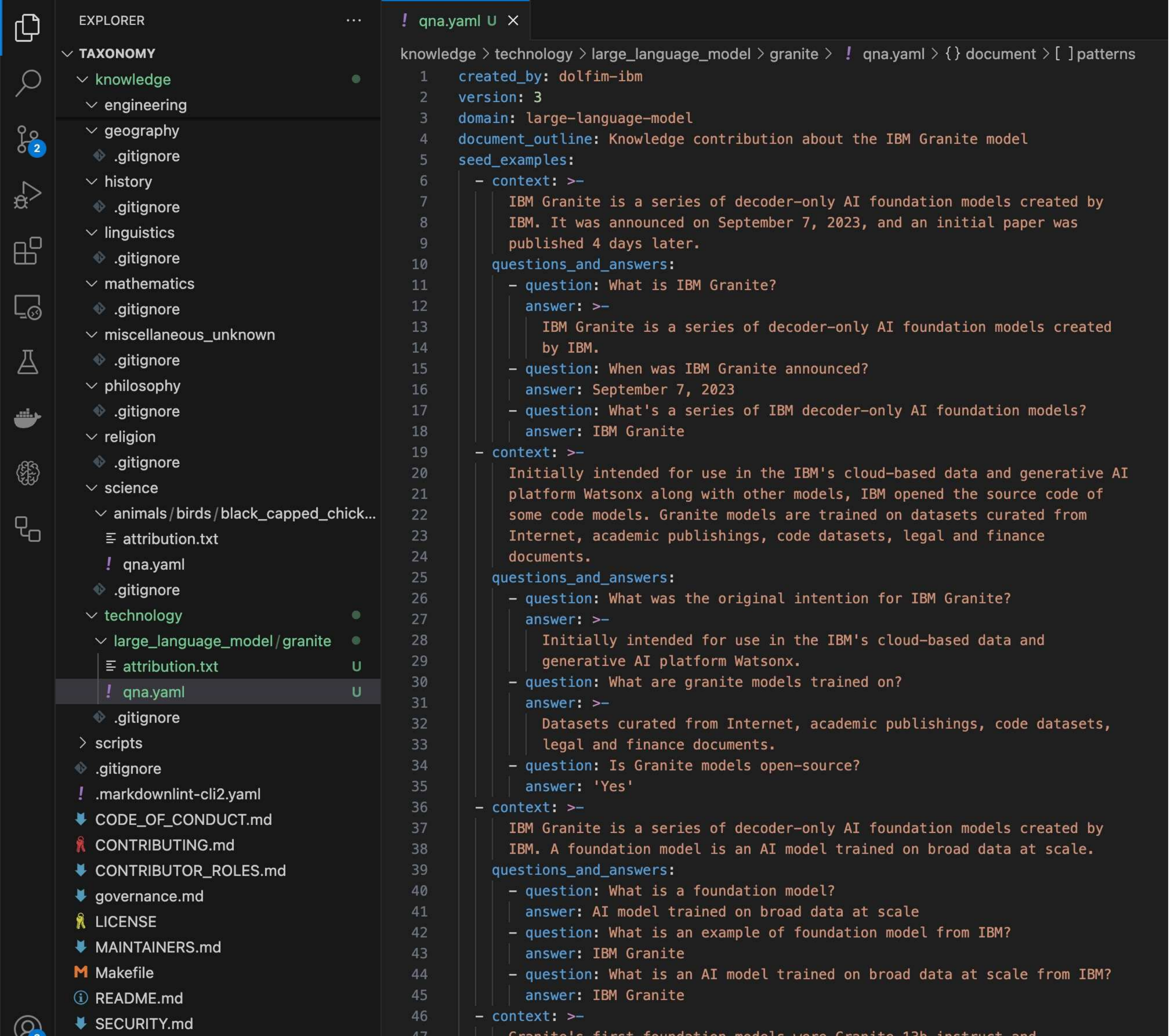
InstructLab demo

Initial test of the student model

Initially the student model doesn't know about the IBM Granite models.

Add new knowledge

The taxonomy is enhanced with knowledge from the relative Wikipedia page.



InstructLab demo

Initial test of the student model

Initially the student model doesn't know about the IBM Granite models.

Add new knowledge

The taxonomy is enhanced with knowledge from the relative Wikipedia page.

✔ Let the teacher improve the model

The teacher model is generating synthetic data based on the examples, which is used to train the student model.

```
$ ilab model chat --model instructlab-granite-7b-lab-trained-llm-granite
system
Welcome to InstructLab Chat w/ INSTRUCTLAB-GRANITE-7B-LAB-Q4_K_M.GGUF

>>> What is IBM Granite? [S][default]
instructlab-granite-7b-lab-Q4_K_M.gguf
IBM Granite is a series of highly scalable, secure, and cost-effective AI models developed by Red Hat and IBM Research, based on the Granite-7b-base language model. These models are designed to be adaptable and extensible, enabling developers to create intelligent applications with a wide range of use cases. The models can be fine-tuned for specific tasks and industries, providing clients with customized solutions tailored to their needs.










Some key features of IBM Granite include:

1. Adaptable and extensible: The open-source nature of the Granite-7b-base models allows developers to modify, extend, and integrate them seamlessly into their projects.
2. Scalability: IBM Granite models can handle large amounts of data and perform complex computations, making them suitable for applications requiring high scalability.
3. Security: The models are designed with security in mind, ensuring that sensitive data remains protected during processing and transmission.
4. Cost-effective: By leveraging open-source models, organizations can save on licensing fees and reduce the time and resources required for model development.
5. Versatility: IBM Granite models can be applied to a wide variety of industries and domains, such as healthcare, finance, education, and customer service, among others.

In summary, IBM Granite is a powerful and flexible AI platform that enables developers to create intelligent applications with ease, while also providing robust security features and cost savings.

elapsed 5.792 seconds
```

Introducing Docling

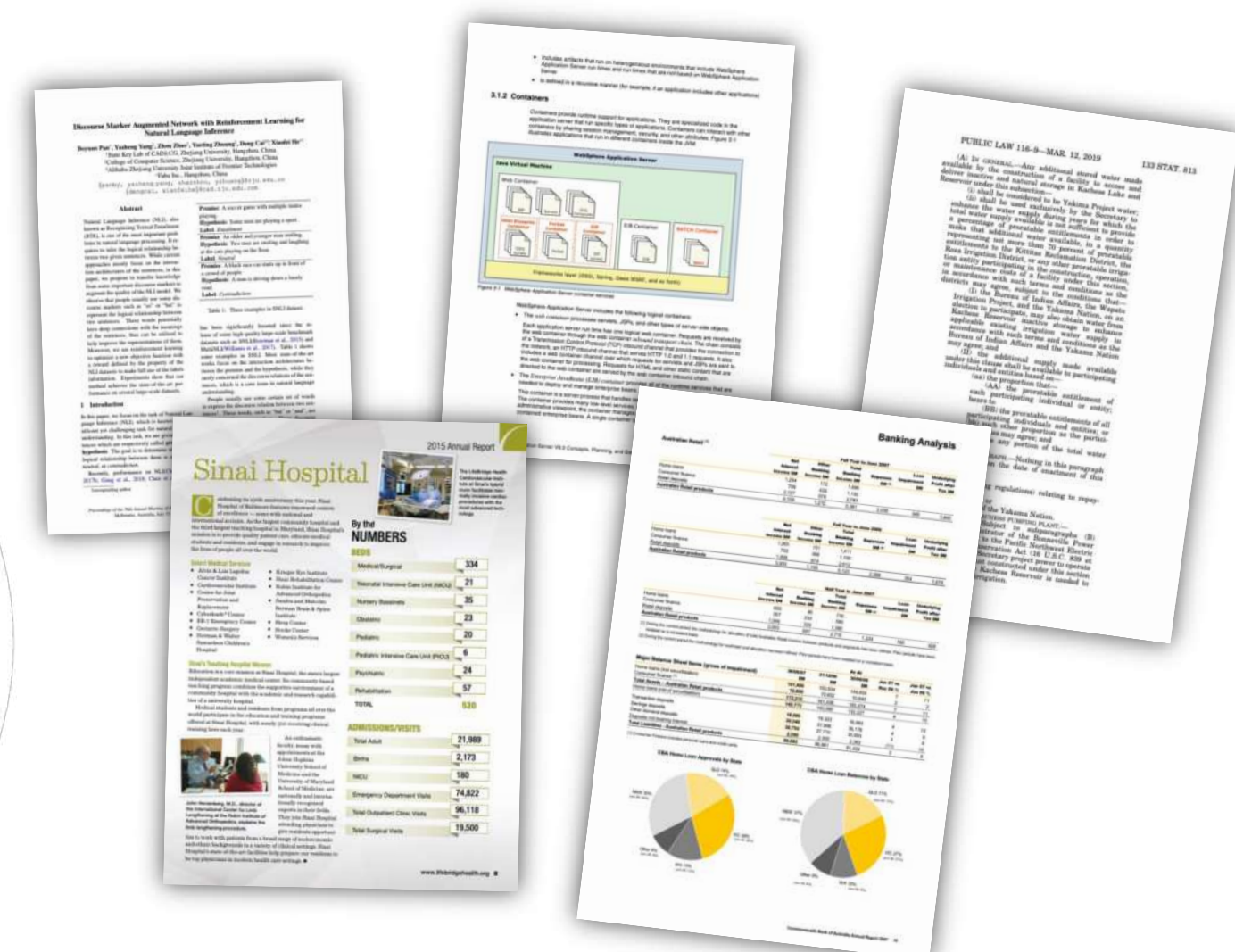
-  Reads popular document formats (PDF, DOCX, PPTX, XLSX, Images, HTML, AsciiDoc & Markdown) and exports to HTML, Markdown and JSON
-  Advanced PDF document understanding including page layout, reading order & table structures
-  Unified, expressive DoclingDocument representation format
-  OCR support for scanned PDFs
-  Easy integration with LLM app / RAG frameworks like  LlamaIndex,  LangChain,  Bee agent framework
-  Simple and convenient CLI



Raised +10k Github stars in a week, adoption of Docling is going viral!



Allows for PDF in knowledge since ilab 0.21, included in RHEL AI 1.3



```
pip install docling

# a single document to markdown
docling https://arxiv.org/pdf/2408.09869.pdf

# a folder of documents to markdown and json
docling --to json --to md ./inputs/
```

<https://github.com/DS4SD/docling>

Docling is for developers!

```
files = ["test/data/2206.01062.pdf", ...]

doc_converter = DocumentConverter()
results = doc_converter.convert_all(files)

for conv_result in results:
    if conv_result.status == ConversionStatus.SUCCESS:
        doc = conv_result.document

        # save results
        doc.save_as_json("...")

        # inspect tables to pandas dataframe...
        for table in doc.tables:
            df_table = table.export_to_dataframe()
            ...

        # inspect the images
        for pic in doc.pictures:
            pil_img = pic.get_image()
            pil_img.show()
            ...
```

The screenshot shows the API reference page for the Docling Document class. The page title is "Docling Document". Below the title, it states: "This is an automatic generated API reference of the DoclingDocument type." The page is organized into sections: "mod doc" (Package for models defined by the Document type), "Classes:" (a list of classes including DoclingDocument, DocumentOrigin, DocItem, DocItemLabel, ProvenanceItem, GroupItem, GroupLabel, NodeItem, PageItem, and FloatingItem), and "Table of contents" (a list of attributes and methods for the DoclingDocument class, such as body, furniture, groups, key_value_items, name, origin, pages, pictures, schema_name, tables, texts, version, add_group, add_heading, add_list_item, add_page, add_picture, add_table, add_text, and add_title).



..as simple as it gets

```
! qna.yaml U x  attribution.txt U
knowledge > technology > large_language_model > granite > ! qna.yaml > document_outline
5  seed_examples:
46  - context: >-
51  questions_and_answers:
56  - question: What are the latest models variations from Granite?
58  - context: >-
59  On May 6, 2024, IBM released the source code of four variations of Granite
60  Code Models under Apache 2, an open source permissive license.
61  questions_and_answers:
62  - question: When has IBM released Granite Models as open source?
63  answer: May 6, 2024
64  - question: What are the open source license for IBM Granite models?
65  answer: Apache 2
66  - question: >-
67  How many variations has IBM released as open source for Granite on
68  6-May?
69  answer: Four
70  document:
71  repo: https://github.com/dolfim-ibm/taxonomy-knowledge-docs
72  commit: f82016ee5187852adac9e917f83c24861801db64
73  patterns:
74  - *.pdf
75
76
```

Simply point the qna.yaml to the PDF knowledge file

Coming soon: support for docx, pptx, html, etc

InstructLab Experiences



InstructLab

Laptop / desktop



InstructLab

BYO Server/VM (upstream)



Red Hat Enterprise Linux AI

Appliance / BYO
Server/VM

Community Project

- **SDG** using a simplified implementation of the **Self-Instruct** methodology
- Fine-Tuning using the **LoRA** technique
- Uses quantized models as base resulting in fine-tuning using the **QLoRA** technique
- BYO models, configurations, drivers for each hardware accelerator

Community Project

- **SDG 1.0** using the original **simple flow** from the **LAB** methodology
- Fine-Tuning of LLMs using full resolution or quantized models
- Use **multi-stage fine-tuning** as per **LAB** methodology
- BYO configurations and drivers for each hardware accelerator

Enterprise Product

- **SDG 1.5** using **agentic flows** on top of the original **LAB** methodology
- Production-grade fine-tuning using full resolution models
- Use **multi-stage fine-tuning** as per **LAB** methodology
- OOB models & optimizations per hardware accelerators



InstructLab

SDG refers to Synthetic Data Generation

Self-instruct refers to the technique described in the paper [Self-Instruct: Aligning LM with Self Generated Instructions](#)

LAB refers to the technique described in the paper [LAB: Large-Scale Alignment for ChatBots](#)

LoRA (Low-Rank Adaptation) is a technique for efficient fine-tuning of LLMs using full-resolution base models

QLoRA (Quantized Low-Rank Adaptation) is a LoRA technique for efficient fine-tuning of quantized base models

Red Hat AI platforms

Generative AI and MLOps capabilities for building flexible, trusted AI solutions at scale

