🐊 Lumigator: evaluating LLMs made simple

Davide Eynard

FOSDEM, 2 Feb 2025

"I look at HuggingFace leaderboard and try a bunch of those... I feel overwhelmed and don't know which model to trust"

Engineer from a leading technology company

"I tried a couple of models... eventually I just went with GPT4 because it seemed like the safest bet"

Data scientist from customer engagement company

You...

- code
- work with AI and LLMs
- ... but are not necessarily publishing on NeurIPS!
- need to choose LLMs and spend as little time as possible

We are...

Contributors 18

+ 4 contributors

Mozilla.ai's goal:

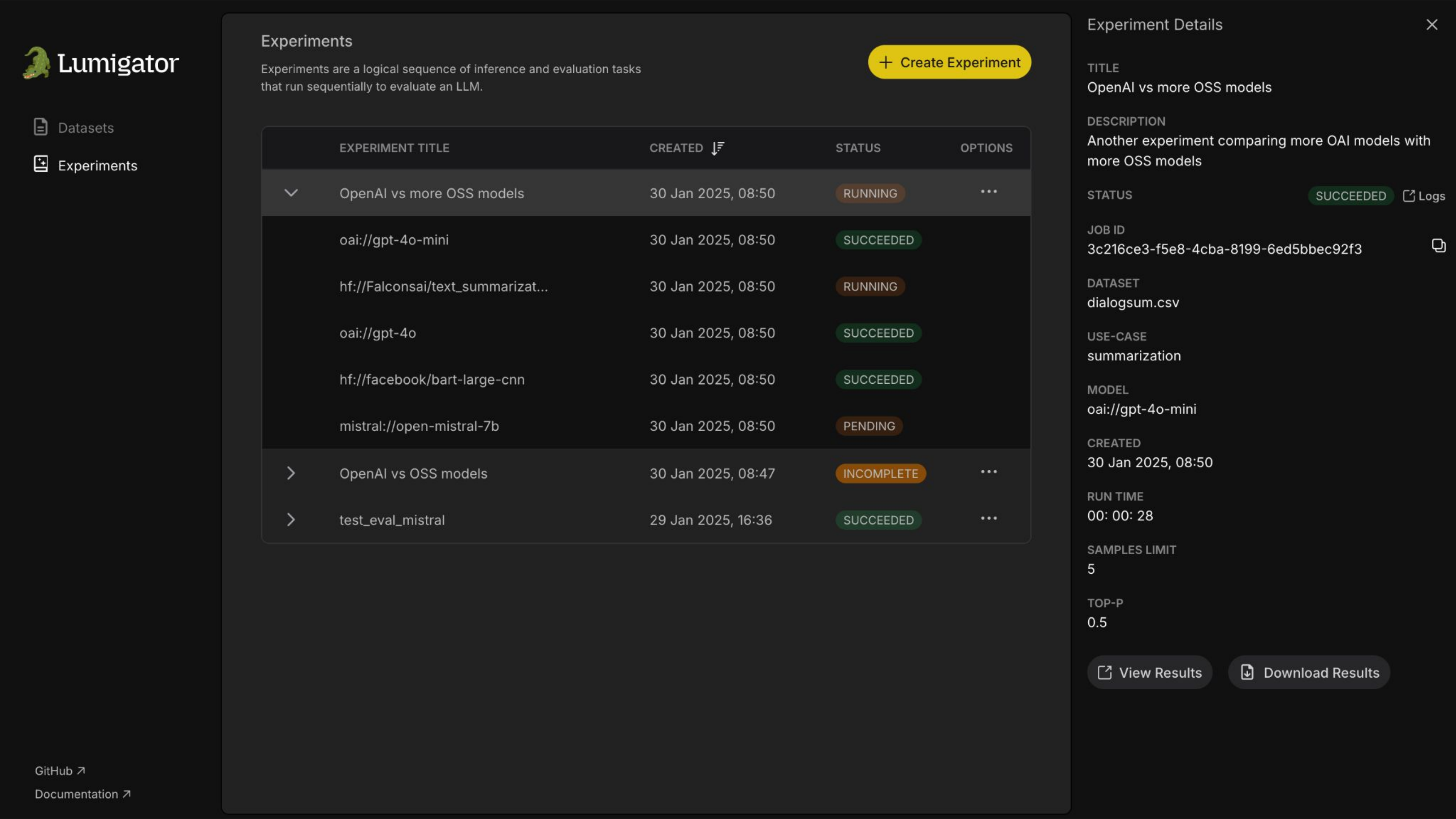
- Make AI easier and more transparent
- Understand what is broken and fix it
- Integrate existing tools
- Only build what is missing

Building Lumigator

Lumigator is a **platform** that **guides** users through the process of selecting the right language model for their **specific** needs.

Key features

- Infrastructure agnostic (local/cluster/cloud)
- Relying on existing tools / standards for interoperability
- API + SDK + UI
- Extensible: by the community for the community
- First use case: evaluation of summarization models



Experiments

Experiments are a logical sequence of inference and evaluation tasks that run sequentially to evaluate an LLM.

	EXPERIMENT TITLE	CREATED	STATUS	OPTIONS
˅	OpenAI vs more OSS models	30 Jan 2025, 08:50	RUNNING	···
	oai://gpt-4o-mini	30 Jan 2025, 08:50	SUCCEEDED	
	hf://Falconsai/text_summarizat...	30 Jan 2025, 08:50	RUNNING	
	oai://gpt-4o	30 Jan 2025, 08:50	SUCCEEDED	
	hf://facebook/bart-large-cnn	30 Jan 2025, 08:50	SUCCEEDED	
	mistral://open-mistral-7b	30 Jan 2025, 08:50	PENDING	
›	OpenAI vs OSS models	30 Jan 2025, 08:47	INCOMPLETE	···
›	test_eval_mistral	29 Jan 2025, 16:36	SUCCEEDED	···

Experiment Details

TITLE
OpenAI vs more OSS models

DESCRIPTION
Another experiment comparing more OAI models with more OSS models

STATUS
SUCCEEDED

JOB ID
3c216ce3-f5e8-4cba-8199-6ed5bbec92f3

DATASET
dialogsum.csv

USE-CASE
summarization

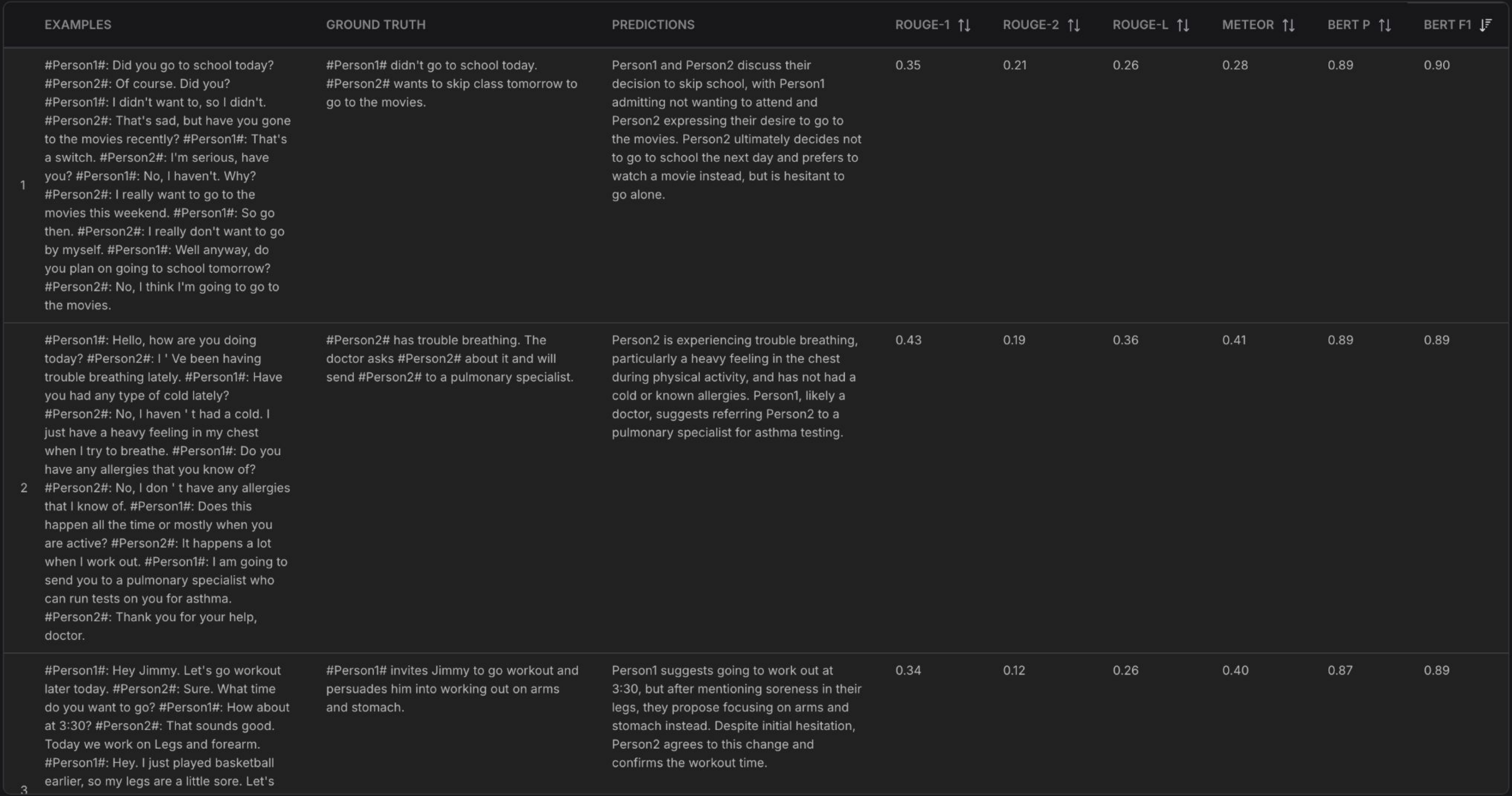
MODEL
oai://gpt-4o-mini

CREATED
30 Jan 2025, 08:50

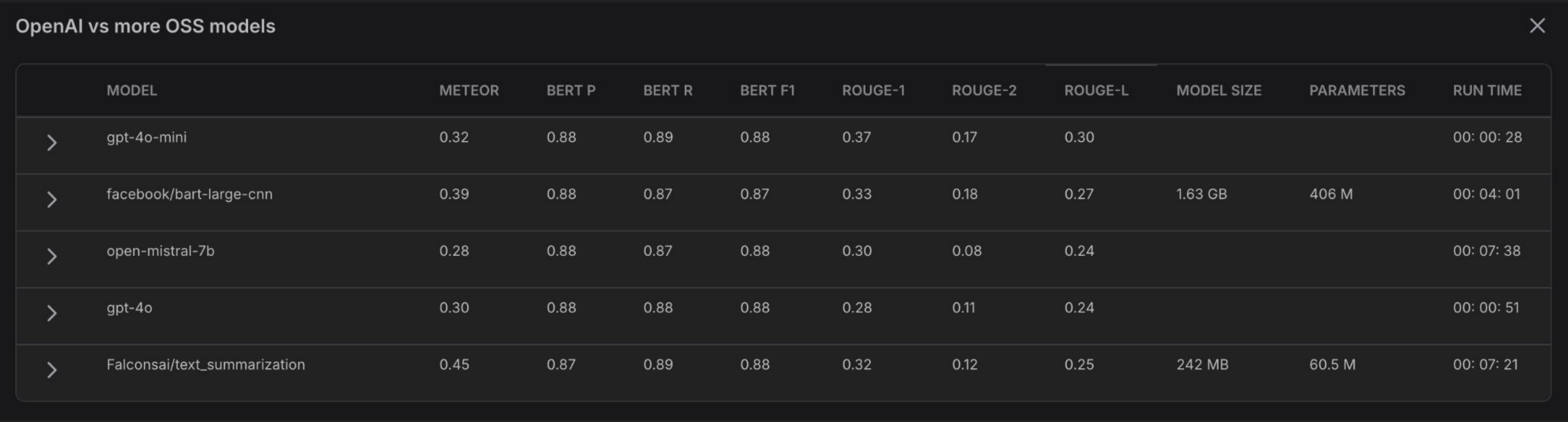
RUN TIME
00: 00: 28

SAMPLES LIMIT
5

TOP-P
0.5



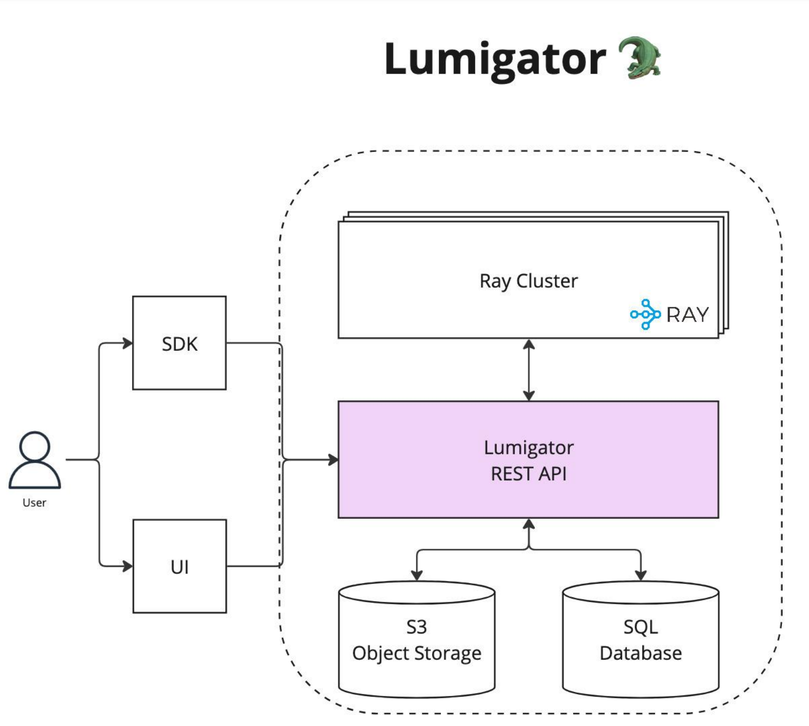
	EXAMPLES	GROUND TRUTH	PREDICTIONS	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT P	BERT F1
1	#Person1#: Did you go to school today? #Person2#: Of course. Did you? #Person1#: I didn't want to, so I didn't. #Person2#: That's sad, but have you gone to the movies recently? #Person1#: That's a switch. #Person2#: I'm serious, have you? #Person1#: No, I haven't. Why? #Person2#: I really want to go to the movies this weekend. #Person1#: So go then. #Person2#: I really don't want to go by myself. #Person1#: Well anyway, do you plan on going to school tomorrow? #Person2#: No, I think I'm going to go to the movies.	#Person1# didn't go to school today. #Person2# wants to skip class tomorrow to go to the movies.	Person1 and Person2 discuss their decision to skip school, with Person1 admitting not wanting to attend and Person2 expressing their desire to go to the movies. Person2 ultimately decides not to go to school the next day and prefers to watch a movie instead, but is hesitant to go alone.	0.35	0.21	0.26	0.28	0.89	0.90
2	#Person1#: Hello, how are you doing today? #Person2#: I ' Ve been having trouble breathing lately. #Person1#: Have you had any type of cold lately? #Person2#: No, I haven ' t had a cold. I just have a heavy feeling in my chest when I try to breathe. #Person1#: Do you have any allergies that you know of? #Person2#: No, I don ' t have any allergies that I know of. #Person1#: Does this happen all the time or mostly when you are active? #Person2#: It happens a lot when I work out. #Person1#: I am going to send you to a pulmonary specialist who can run tests on you for asthma. #Person2#: Thank you for your help, doctor.	#Person2# has trouble breathing. The doctor asks #Person2# about it and will send #Person2# to a pulmonary specialist.	Person2 is experiencing trouble breathing, particularly a heavy feeling in the chest during physical activity, and has not had a cold or known allergies. Person1, likely a doctor, suggests referring Person2 to a pulmonary specialist for asthma testing.	0.43	0.19	0.36	0.41	0.89	0.89
3	#Person1#: Hey Jimmy. Let's go workout later today. #Person2#: Sure. What time do you want to go? #Person1#: How about at 3:30? #Person2#: That sounds good. Today we work on Legs and forearm. #Person1#: Hey. I just played basketball earlier, so my legs are a little sore. Let's	#Person1# invites Jimmy to go workout and persuades him into working out on arms and stomach.	Person1 suggests going to work out at 3:30, but after mentioning soreness in their legs, they propose focusing on arms and stomach instead. Despite initial hesitation, Person2 agrees to this change and confirms the workout time.	0.34	0.12	0.26	0.40	0.87	0.89



OpenAI vs more OSS models

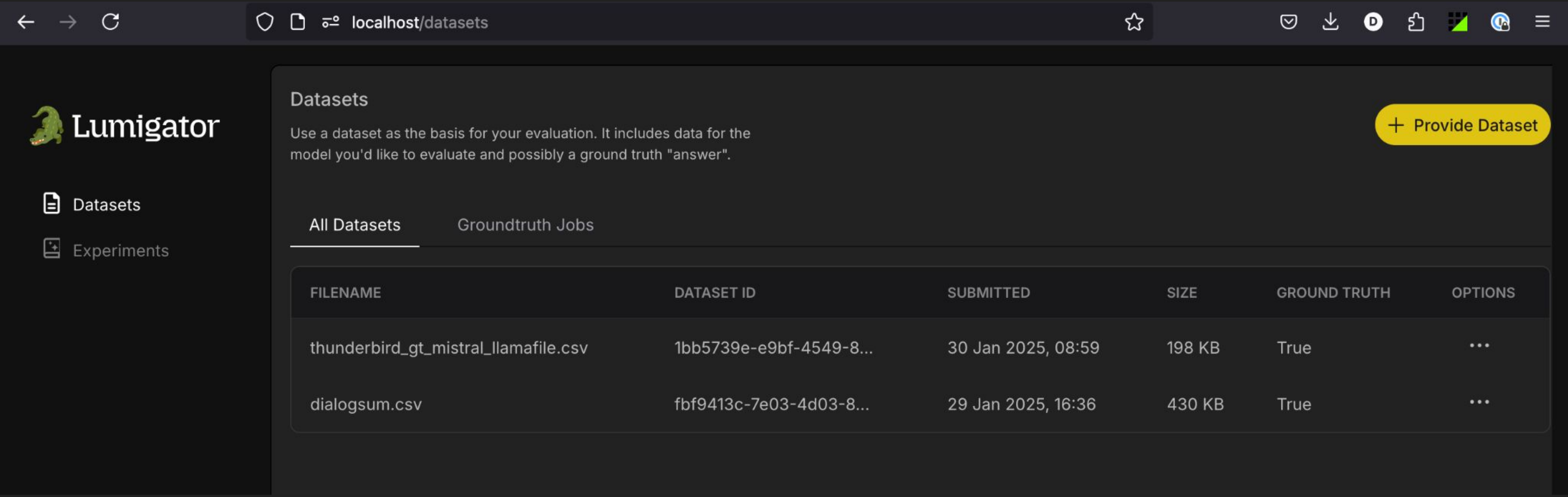
	MODEL	METEOR	BERT P	BERT R	BERT F1	ROUGE-1	ROUGE-2	ROUGE-L	MODEL SIZE	PARAMETERS	RUN TIME
>	gpt-4o-mini	0.32	0.88	0.89	0.88	0.37	0.17	0.30			00: 00: 28
>	facebook/bart-large-cnn	0.39	0.88	0.87	0.87	0.33	0.18	0.27	1.63 GB	406 M	00: 04: 01
>	open-mistral-7b	0.28	0.88	0.87	0.88	0.30	0.08	0.24			00: 07: 38
>	gpt-4o	0.30	0.88	0.88	0.88	0.28	0.11	0.24			00: 00: 51
>	Falconsai/text_summarization	0.45	0.87	0.89	0.88	0.32	0.12	0.25	242 MB	60.5 M	00: 07: 21

Architecture



Mozilla.ai

Web UI

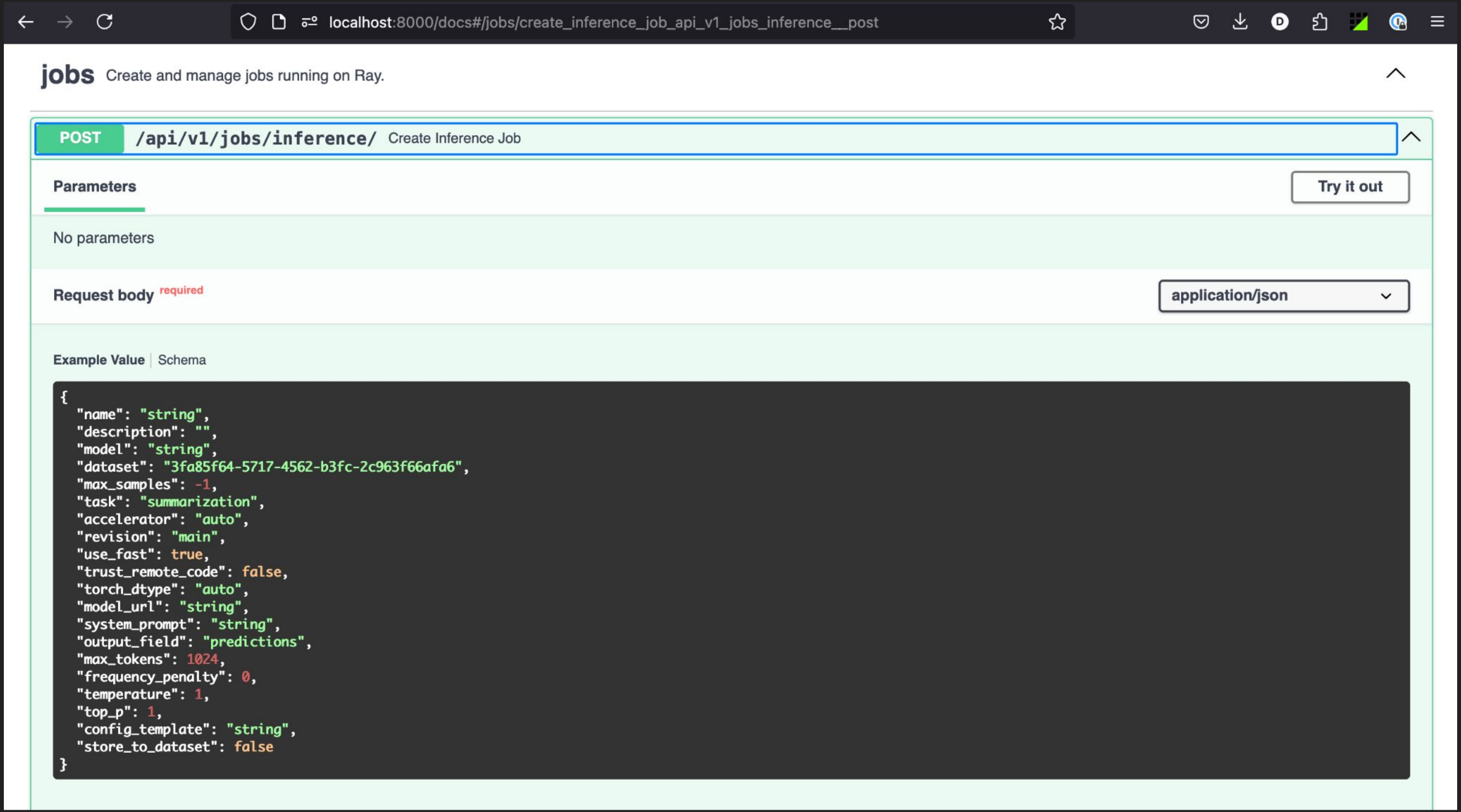


Datasets

Use a dataset as the basis for your evaluation. It includes data for the model you'd like to evaluate and possibly a ground truth "answer".

FILENAME	DATASET ID	SUBMITTED	SIZE	GROUND TRUTH	OPTIONS
thunderbird_gt_mistral_llamafile.csv	1bb5739e-e9bf-4549-8...	30 Jan 2025, 08:59	198 KB	True	···
dialogsum.csv	fbf9413c-7e03-4d03-8...	29 Jan 2025, 16:36	430 KB	True	···

REST API



localhost:8000/docs#/jobs/create_inference_job_api_v1_jobs_inference__post

jobs Create and manage jobs running on Ray.

`/api/v1/jobs/inference/` Create Inference Job

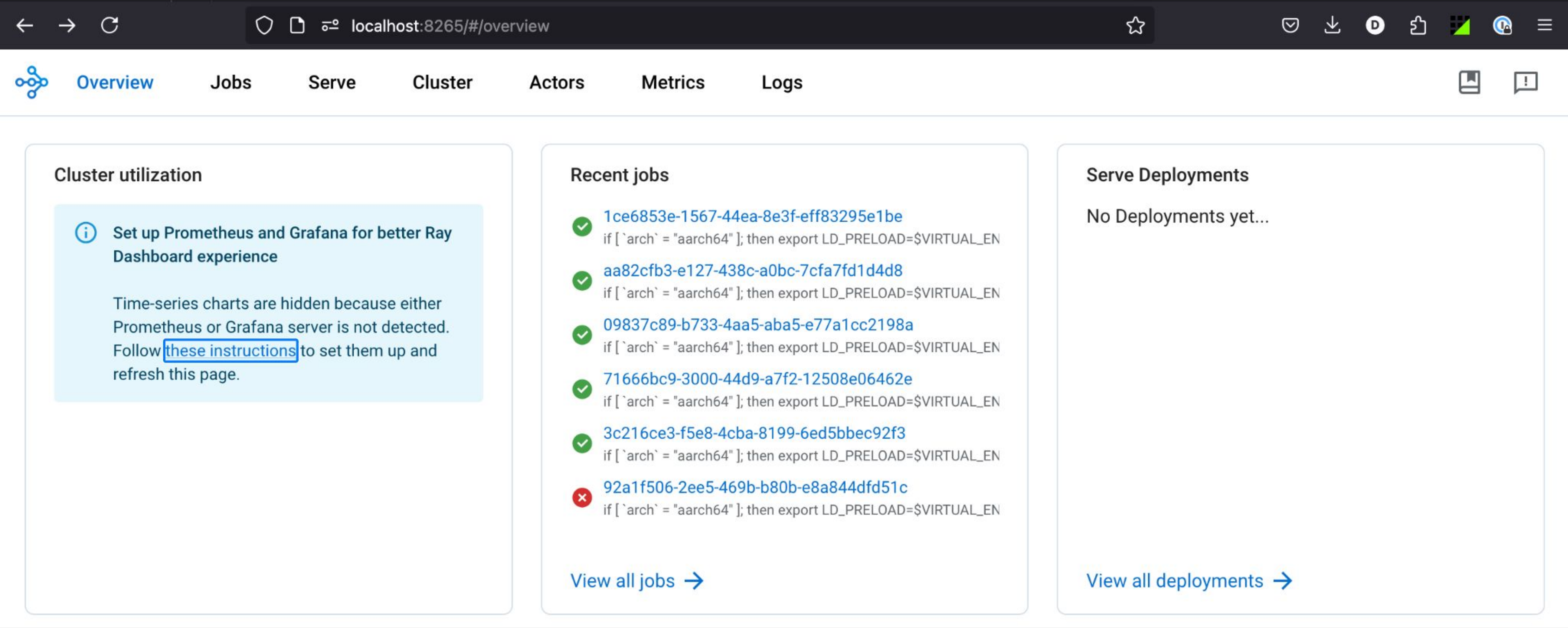
No parameters

Request body required

application/json
Example Value | Schema

```json
{
  "name": "string",
  "description": "",
  "model": "string",
  "dataset": "3fa85f64-5717-4562-b3fc-2c963f66afa6",
  "max_samples": -1,
  "task": "summarization",
  "accelerator": "auto",
  "revision": "main",
  "use_fast": true,
  "trust_remote_code": false,
  "torch_dtype": "auto",
  "model_url": "string",
  "system_prompt": "string",
  "output_field": "predictions",
  "max_tokens": 1024,
  "frequency_penalty": 0,
  "temperature": 1,
  "top_p": 1,
  "config_template": "string",
  "store_to_dataset": false
}
```

Ray backend



Cluster utilization

Set up Prometheus and Grafana for better Ray Dashboard experience

Time-series charts are hidden because either Prometheus or Grafana server is not detected. Follow these instructions to set them up and refresh this page.

Recent jobs

- ✓ 1ce6853e-1567-44ea-8e3f-eff83295e1be
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN
- ✓ aa82cfb3-e127-438c-a0bc-7cfa7fd1d4d8
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN
- ✓ 09837c89-b733-4aa5-aba5-e77a1cc2198a
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN
- ✓ 71666bc9-3000-44d9-a7f2-12508e06462e
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN
- ✓ 3c216ce3-f5e8-4cba-8199-6ed5bbec92f3
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN
- ✗ 92a1f506-2ee5-469b-b80b-e8a844dfd51c
  if [`arch` = "aarch64"]; then export LD_PRELOAD=$VIRTUAL_EN

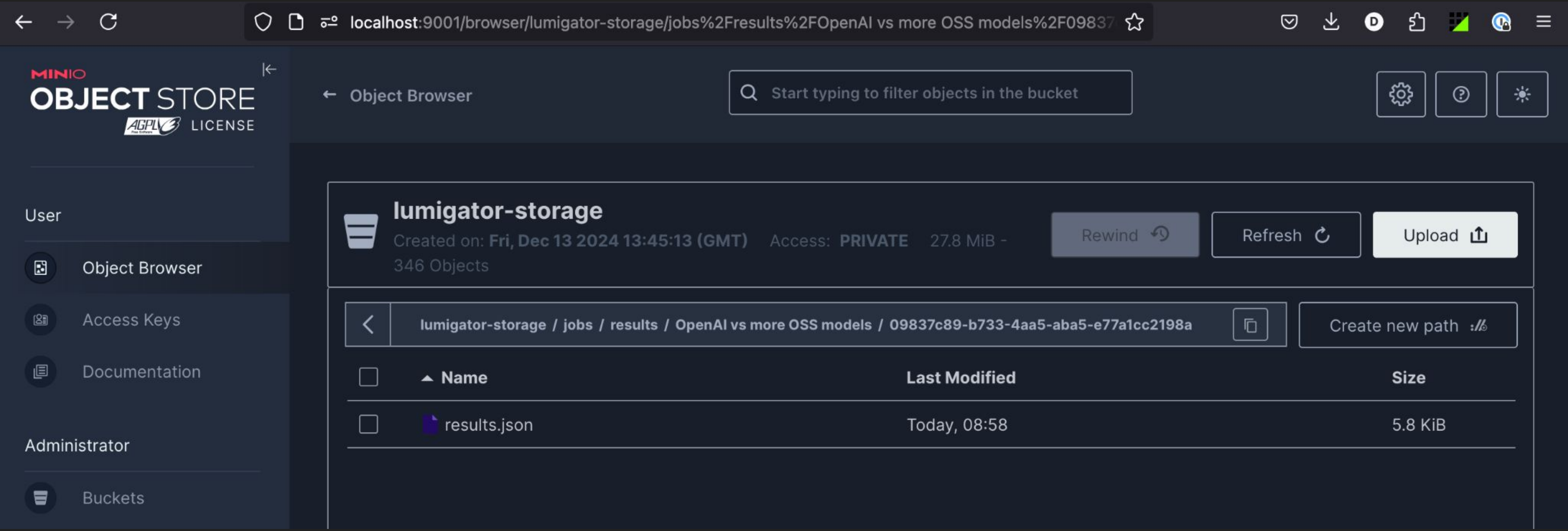
View all jobs →

Serve Deployments

No Deployments yet...

View all deployments →

Object Store



lumigator-storage

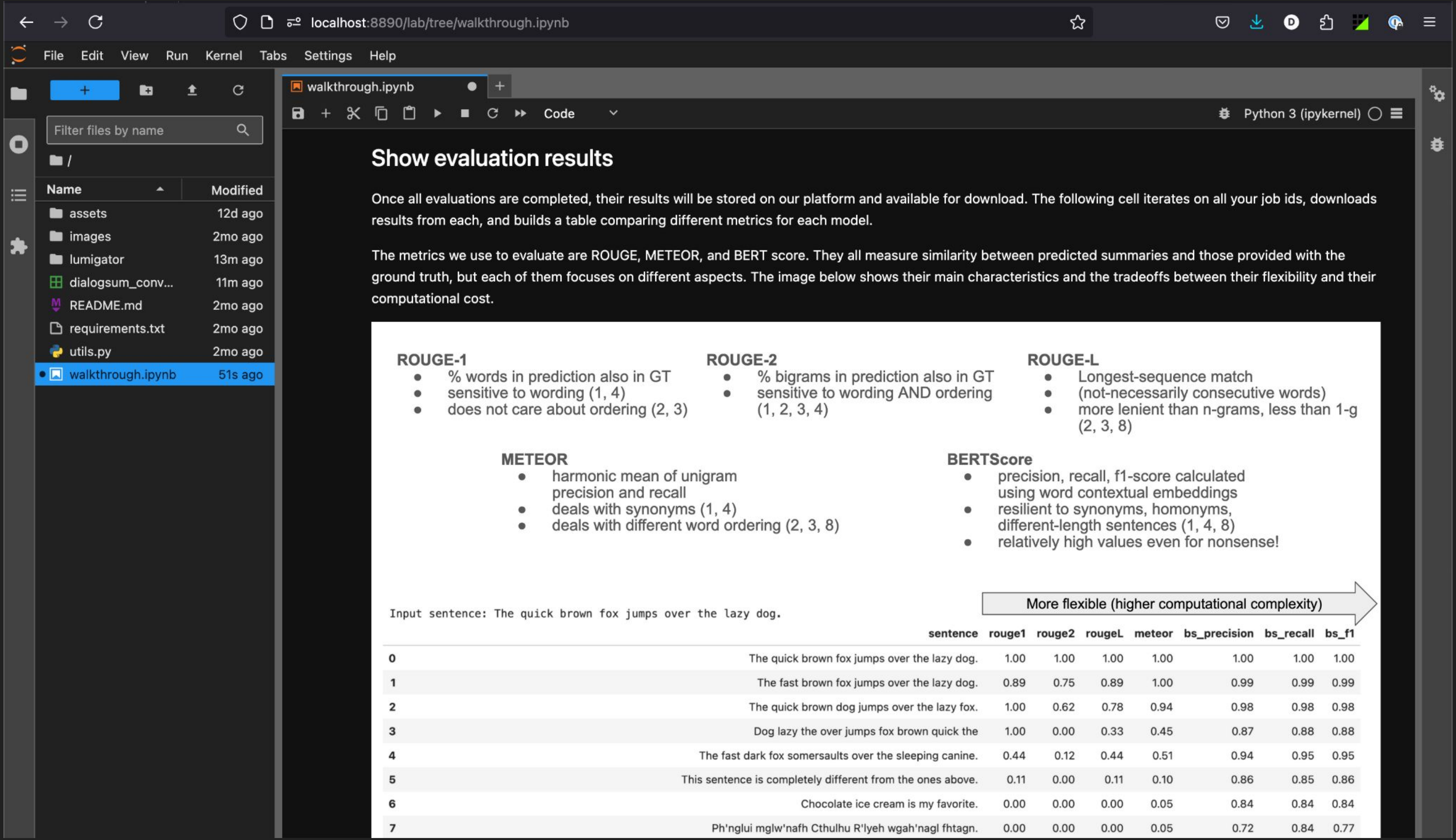
Created on: **Fri, Dec 13 2024 13:45:13 (GMT)** Access: **PRIVATE** 27.8 MiB - 346 Objects

lumigator-storage / jobs / results / OpenAI vs more OSS models / 09837c89-b733-4aa5-aba5-e77a1cc2198a

☐	Name	Last Modified	Size
☐	results.json	Today, 08:58	5.8 KiB

Mozilla.ai

SDK



Show evaluation results

Once all evaluations are completed, their results will be stored on our platform and available for download. The following cell iterates on all your job ids, downloads results from each, and builds a table comparing different metrics for each model.

The metrics we use to evaluate are ROUGE, METEOR, and BERT score. They all measure similarity between predicted summaries and those provided with the ground truth, but each of them focuses on different aspects. The image below shows their main characteristics and the tradeoffs between their flexibility and their computational cost.

ROUGE-1
- % words in prediction also in GT
- sensitive to wording (1, 4)
- does not care about ordering (2, 3)

ROUGE-2
- % bigrams in prediction also in GT
- sensitive to wording AND ordering (1, 2, 3, 4)

ROUGE-L
- Longest-sequence match
- (not-necessarily consecutive words)
- more lenient than n-grams, less than 1-g (2, 3, 8)

METEOR
- harmonic mean of unigram precision and recall
- deals with synonyms (1, 4)
- deals with different word ordering (2, 3, 8)

BERTScore
- precision, recall, f1-score calculated using word contextual embeddings
- resilient to synonyms, homonyms, different-length sentences (1, 4, 8)
- relatively high values even for nonsense!

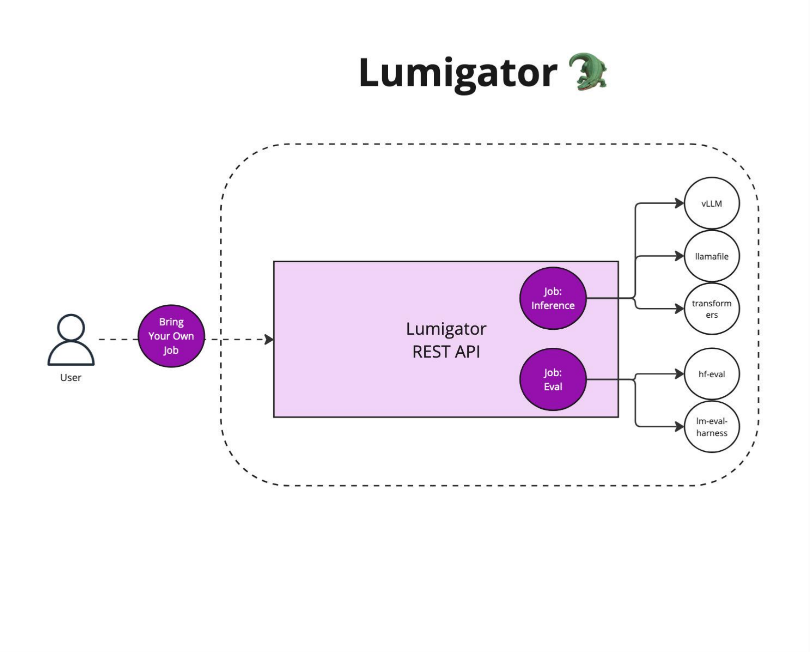
More flexible (higher computational complexity)

```
Input sentence: The quick brown fox jumps over the lazy dog.
```

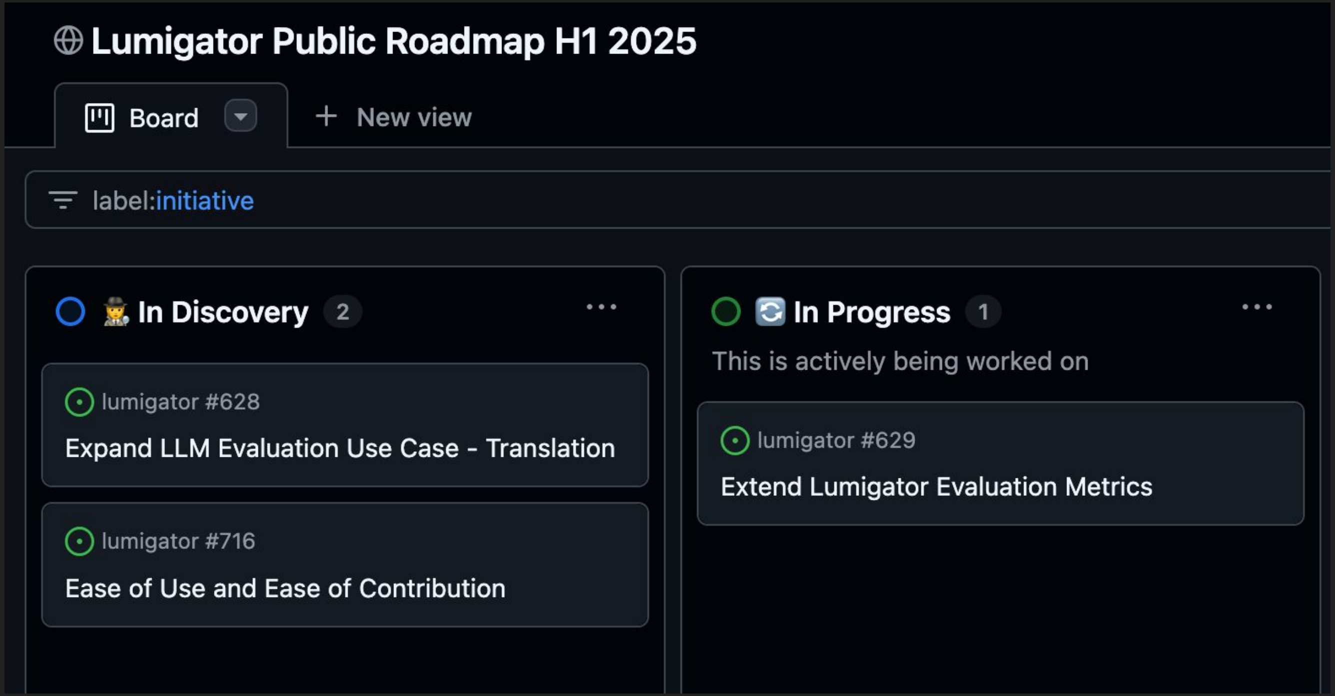
	sentence	rouge1	rouge2	rougeL	meteor	bs_precision	bs_recall	bs_f1
0	The quick brown fox jumps over the lazy dog.	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	The fast brown fox jumps over the lazy dog.	0.89	0.75	0.89	1.00	0.99	0.99	0.99
2	The quick brown dog jumps over the lazy fox.	1.00	0.62	0.78	0.94	0.98	0.98	0.98
3	Dog lazy the over jumps fox brown quick the	1.00	0.00	0.33	0.45	0.87	0.88	0.88
4	The fast dark fox somersaults over the sleeping canine.	0.44	0.12	0.44	0.51	0.94	0.95	0.95
5	This sentence is completely different from the ones above.	0.11	0.00	0.11	0.10	0.86	0.85	0.86
6	Chocolate ice cream is my favorite.	0.00	0.00	0.00	0.05	0.84	0.84	0.84
7	Ph'nglui mglw'nafh Cthulhu R'lyeh wgah'nagl fhtagn.	0.00	0.00	0.00	0.05	0.72	0.84	0.77

Choose your own component

- Different classes of supported models
  - HuggingFace Transformers
  - APIs
  - Local models (e.g. llamafile, vLLM) through OpenAI-compatible APIs
- Different type of jobs
  - Simple (e.g. annotation) as well as composite (e.g. inference+evaluation)
- Different access
  - UI, Python SDK, CURL calls
- Different deployments
  - Local, cluster, cloud



What next?

- 

In Discovery 2

lumigator #628
Expand LLM Evaluation Use Case - Translation

lumigator #716
Ease of Use and Ease of Contribution

In Progress 1

This is actively being worked on

lumigator #629
Extend Lumigator Evaluation Metrics

Wanna know more?

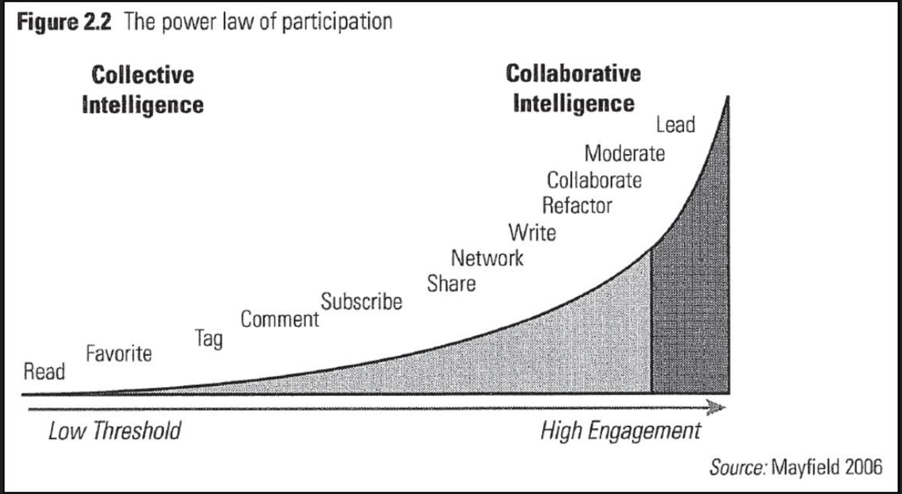
- This talk
- Code
- Documentation
- Roadmap

Wanna contribute?

- Try it out
- Submit a bug/feature request
- Check our documentation
- Pick up a good first issue
- Add a new feature :-)

See our CONTRIBUTING.md

Figure 2.2 The power law of participation



Source: Mayfield 2006

Thank you!

@mala@fosstodon.org

https://github.com/mozilla-ai/lumigator