ELSEVIER

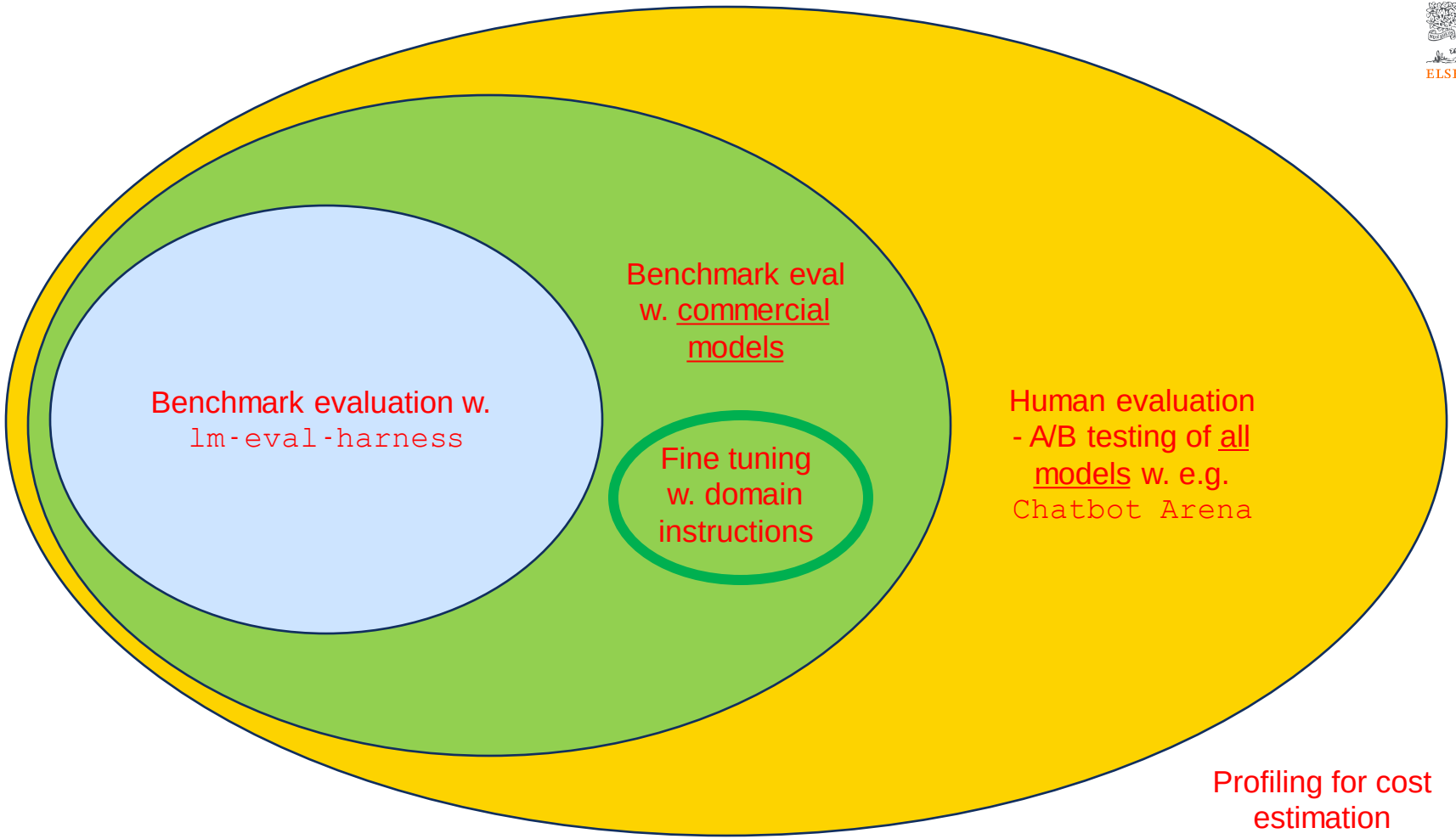
CC BY SA

NLP4Chemistry

Exploring the Fringe

Camilo Thorne, Saber Akhondi

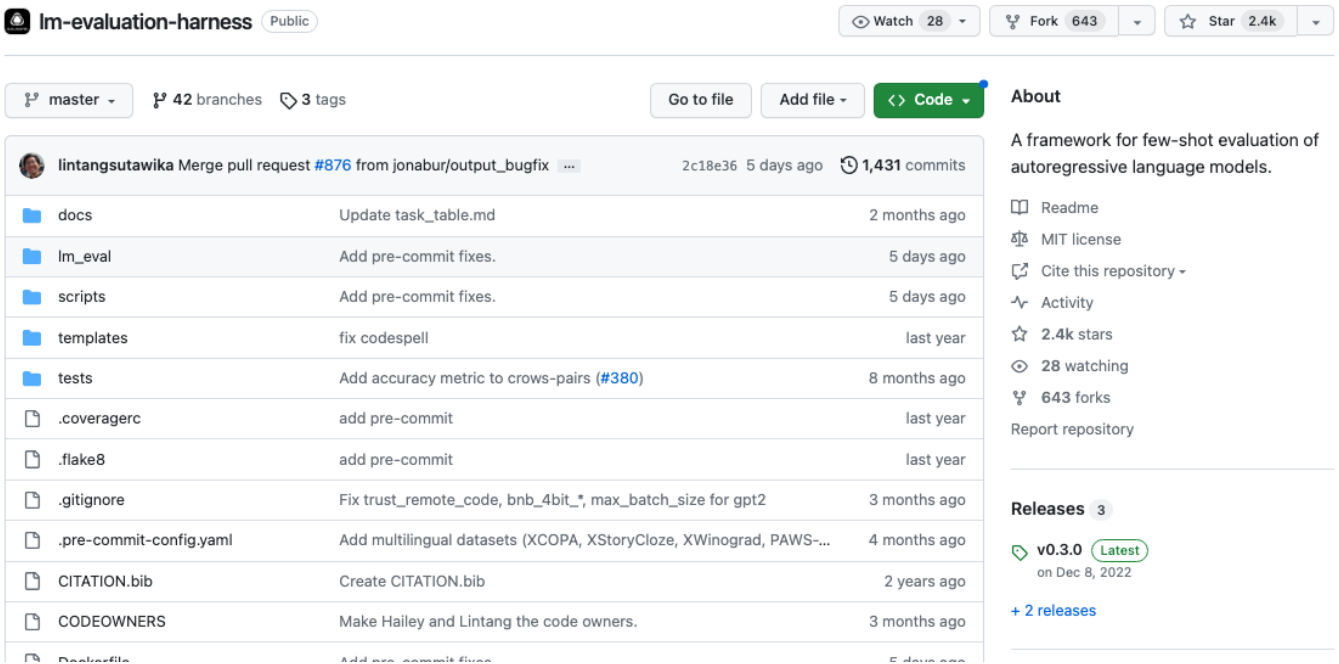
May 2024 - COLING/LREC '24

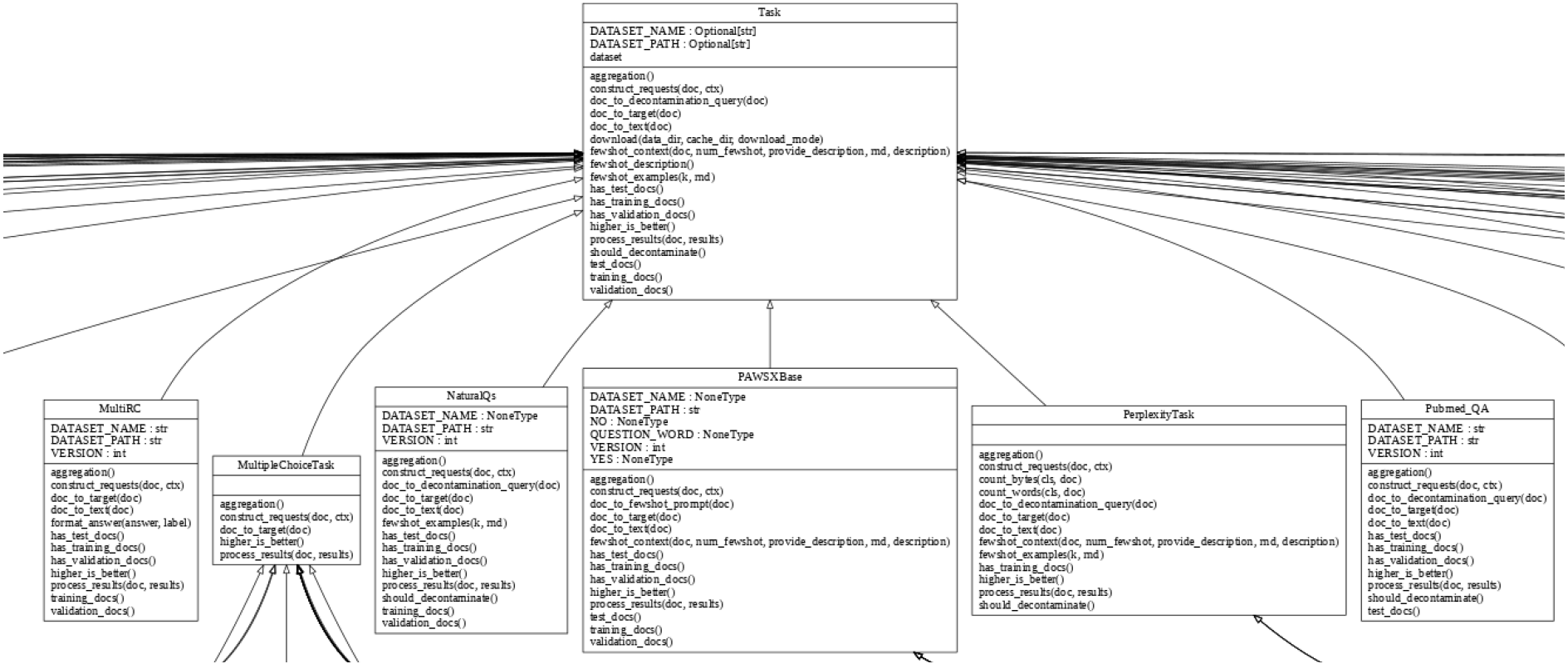

Benchmark evaluation w.
lm-eval-harness
Benchmark eval
w. commercial
models
Fine tuning
w. domain
instructions
Human evaluation
- A/B testing of all
models w. e.g.
Chatbot Arena
Profiling for cost
estimation

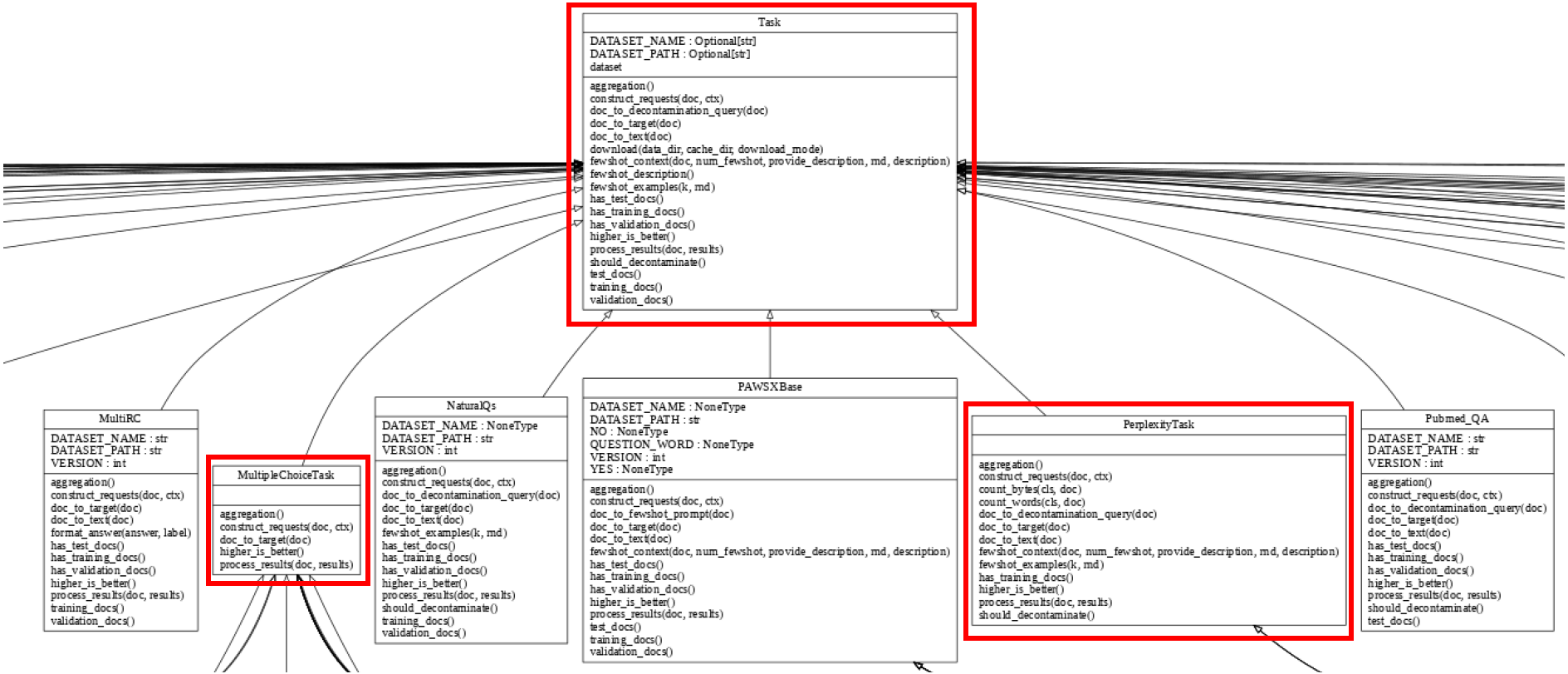
Benchmark Evaluation

Benchmark evaluation w. `lm-eval-harness` [2022]

https://github.com/EleutherAI/lm-evaluation-harness




Task
DATASET_NAME : Optional[str]
DATASET_PATH : Optional[str]
dataset
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
download(data_dir, cache_dir, download_mode)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
fewshot_description()
fewshot_examples(k, rnd)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
test_docs()
training_docs()
validation_docs()
MultiRC
DATASET_NAME : str
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_target(doc)
doc_to_text(doc)
format_answer(answer, label)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
training_docs()
validation_docs()
MultipleChoiceTask
aggregation()
construct_requests(doc, ctx)
doc_to_target(doc)
higher_is_better()
process_results(doc, results)
NaturalQs
DATASET_NAME : NoneType
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_examples(k, rnd)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
training_docs()
validation_docs()
PAWSXBase
DATASET_NAME : NoneType
DATASET_PATH : str
NO : NoneType
QUESTION_WORD : NoneType
VERSION : int
YES : NoneType
aggregation()
construct_requests(doc, ctx)
doc_to_fewshot_prompt(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
test_docs()
training_docs()
validation_docs()
PerplexityTask
aggregation()
construct_requests(doc, ctx)
count_bytes(cls, doc)
count_words(cls, doc)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
fewshot_examples(k, rnd)
has_training_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
Pubmed_QA
DATASET_NAME : str
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
test_docs()


Task
DATASET_NAME : Optional[str]
DATASET_PATH : Optional[str]
dataset
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
download(data_dir, cache_dir, download_mode)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
fewshot_description()
fewshot_examples(k, rnd)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
test_docs()
training_docs()
validation_docs()
MultiRC
DATASET_NAME : str
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_target(doc)
doc_to_text(doc)
format_answer(answer, label)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
training_docs()
validation_docs()
MultipleChoiceTask
aggregation()
construct_requests(doc, ctx)
doc_to_target(doc)
higher_is_better()
process_results(doc, results)
NaturalQs
DATASET_NAME : NoneType
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_examples(k, rnd)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
training_docs()
validation_docs()
PAWSXBase
DATASET_NAME : NoneType
DATASET_PATH : str
NO : NoneType
QUESTION_WORD : NoneType
VERSION : int
YES : NoneType
aggregation()
construct_requests(doc, ctx)
doc_to_fewshot_prompt(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
test_docs()
training_docs()
validation_docs()
PerplexityTask
aggregation()
construct_requests(doc, ctx)
count_bytes(cls, doc)
count_words(cls, doc)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
fewshot_context(doc, num_fewshot, provide_description, rnd, description)
fewshot_examples(k, rnd)
has_training_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
Pubmed_QA
DATASET_NAME : str
DATASET_PATH : str
VERSION : int
aggregation()
construct_requests(doc, ctx)
doc_to_decontamination_query(doc)
doc_to_target(doc)
doc_to_text(doc)
has_test_docs()
has_training_docs()
has_validation_docs()
higher_is_better()
process_results(doc, results)
should_decontaminate()
test_docs()

```python
import datasets

from lm_eval.base import PerplexityTask
from lm_eval.utils import escaped_split

class ICSRPerplexity(PerplexityTask):
    '''
    We re-use here the ICSR data, but we throw away the labels.
    We ask the model to generate the abstracts, and measure LM perplexity.
    '''
    VERSION = "0.1.0"
    DATASET_PATH = "icsr"
    DATASET_NAME = "icsr-perplexity"

    def __init__(self, data_dir=None, cache_dir=None, download_mode=None):
        self.data_dir = data_dir
        self.dataset = datasets.load_dataset(
            path=self.DATASET_PATH,
            name=self.DATASET_NAME,
            data_dir=data_dir,
            cache_dir=cache_dir,
            download_mode=download_mode,
            field="data",
        )
        self._training_docs = None
        self._fewshot_docs = None

    def download(self, data_dir=None, cache_dir=None, download_mode=None):
        raise TypeError("cannot download an arbitrary JSON dataset")

    def has_validation_docs(self):
        return False

    def has_training_docs(self):
        return False

    def has_test_docs(self):
        return True

    def test_docs(self):
        return map(self._process_doc, self.dataset["test"])

    def _process_doc(self, doc):
        return doc['Abstract']
```

- Extend task (sub) class
- Implement custom constructor
- Implement or override super-class methods

```
MODEL_REGISTRY = {
    "hf":                       gpt2.HFLM,
    "hf-causal":                gpt2.HFLM,
    "hf-causal-experimental":   huggingface.AutoCausalLM,
    "hf-seq2seq":               huggingface.AutoSeq2SeqLM,
    "gpt2":                     gpt2.GPT2LM,
    "gpt3":                     gpt3.GPT3LM,
    "anthropic":                anthropic_llms.AnthropicLM,
    "textsynth":                textsynth.TextSynthLM,
    "dummy":                    dummy.DummyLM,
}
```

Datasets need to be formatted into HF `Dataset` JSON format

Models and datasets by default fetched from HF remote servers

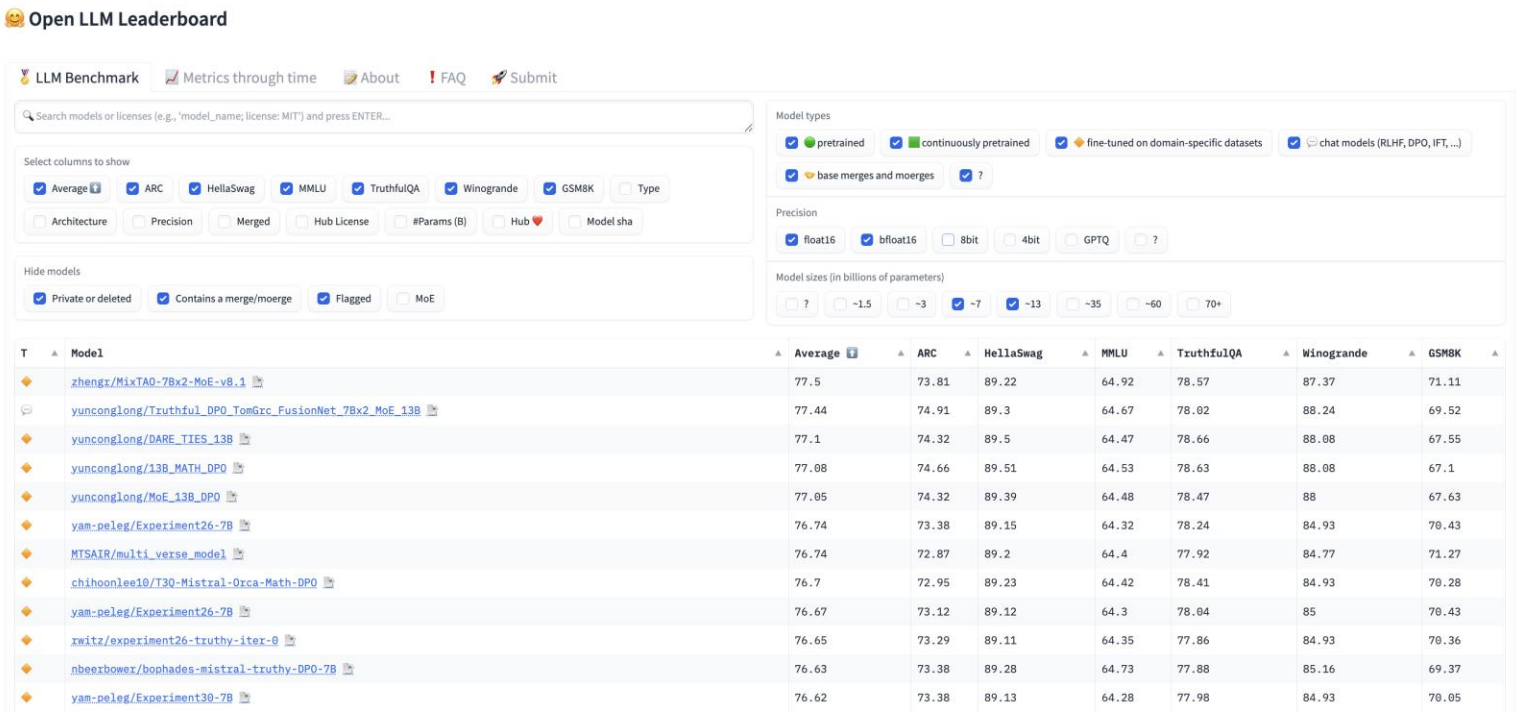
But they can also be deployed locally, provided HF formats are satisfied

Library supports:

- HF-supported **decoder-only** models + architectures
- HF-supported **encoder-decoder** models + architectures
- OpenAI legacy models (GPT3, GPT4)
- Anthropic Claude v1

lm-eval-harness leaderboard [2022-]

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard



T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
🔶	zhengr/MixTAO-7Bx2-MoE-v8.1	77.5	73.81	89.22	64.92	78.57	87.37	71.11
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3	64.67	78.02	88.24	69.52
🔶	yunconglong/DARE_TIES_13B	77.1	74.32	89.5	64.47	78.66	88.08	67.55
🔶	yunconglong/13B_MATH_DPO	77.08	74.66	89.51	64.53	78.63	88.08	67.1
🔶	yunconglong/MoE_13B_DPO	77.05	74.32	89.39	64.48	78.47	88	67.63
🔶	yam-peleg/Experiment26-7B	76.74	73.38	89.15	64.32	78.24	84.93	70.43
🔶	MTSAIR/multi_verse_model	76.74	72.87	89.2	64.4	77.92	84.77	71.27
🔶	chihoonlee10/T3Q-Mistral-Orca-Math-DPO	76.7	72.95	89.23	64.42	78.41	84.93	70.28
🔶	yam-peleg/Experiment26-7B	76.67	73.12	89.12	64.3	78.04	85	70.43
🔶	rwitz/experiment26-truthy-iter-0	76.65	73.29	89.11	64.35	77.86	84.93	70.36
🔶	nbeerbower/bophades-mistral-truthy-DPO-7B	76.63	73.38	89.28	64.73	77.88	85.16	69.37
🔶	yam-peleg/Experiment30-7B	76.62	73.38	89.13	64.28	77.98	84.93	70.05

ChemLLM [2023]

- Open release of chemistry LLM evaluation benchmarks
- "Promptified" version of known benchmarks
  - USPTO-50k
  - PubChem
  - ...

Ability	Task	Task Type	Dataset	#ICL candidates	#test	Evaluation Metrics
Understanding	Name Prediction	Generation	PubChem	500	100	Accuracy
	Property Prediction	Classification	BBBP, HIV, BACE, Tox21, ClinTox	2053, 41127, 1514, 8014, 1484	100	Accuracy, F1 score
Reasoning	Yield Prediction	Classification	Butchward-Hartwig, Suzuki-Miyaura	3957, 5650	100	Accuracy
	Reaction Prediction	Generation	USPTO-Mixed	409035	100	Accuracy, Validity
	Reagents Selection	Ranking	Suzuki-Miyaura	5760	100	Accuracy
	Retrosynthesis	Generation	USPTO-50k	40029	100	Accuracy, Validity
	Text-Based Molecule Design	Generation	ChEBI-20	26407	100	BLEU, Exact Match, etc
Explaining	Molecule Captioning	Generation	ChEBI-20	26407	100	BLEU, Chemists, etc

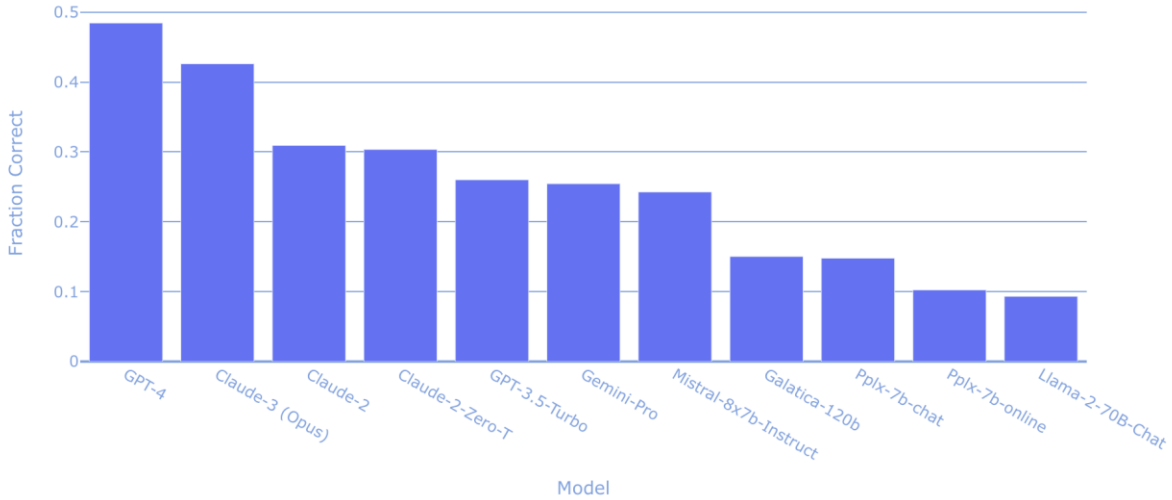
https://github.com/ChemFoundationModels/ChemLLMBench

ChemBench [2024]

- Similar to ChemLLM, but focused on pharmaceutical domain
- Evaluation of open source models + leaderboard

https://lamalab-org.github.io/chem-bench/leaderboard/

Relies on completion "accuracy" for benchmarking!



Challenges of metrics

Benchmark metrics focus too much on lexical match, at the expense of topical relevancy!

```
doc_id: 3
prompt_0: "Question: Pyridine synthesis via hetero-Diels-Alder reaction of 1,2,4-triazines and dienophiles (e.g. enamine) followed by extrusion of N2 What is the name of the reaction? Answer: Boger Thermal Cycloaddition Question: How to synthesis L-valeric acid? Answer: The Marckwald Asymmetric Synthesis is a well-known reaction to synthesise L-valeric acid.It's a Chiral synthesis of L-valeric acid by pyrolysis of brucine salt of racemic α-methyl-α-ethylmalonic acid Question: What is a Clauson-Kaas Synthesis? Answer: Synthesis of pyrrole derivative by the condensation between a primary aliphatic or aromatic amine and 2,5-dialkoxy-tetrahydrofuran in the presence of an acid catalyst Question: Which reaction type is able to convert trialkyl phosphite and alkyl halide to phosphonate? Answer: until"
logit_0: "The reaction type that is able to convert trialkyl phosphite and alkyl halide to phosphonate is a Wacker reaction"
truth: " The Arbuzov Reaction is a well-known reaction to synthesise phosphonate.It's a Formation of a phosphonate from a trialkyl phosphite and an alkyl halide"
```

Completion1: "The reaction type that is able to convert trialkyl phosphite and alkyl halide to phosphonate is a Wacker reaction"
Completion2: "The Arbuzov Reaction" [closer as it captures the subject!]

Truth: "The Arbuzov Reaction is a well-known reaction to synthesise phosphonate. It's a Formation of a phosphonate from a trialkyl phosphite and an alkyl halide"

Completion	BLEURT	BLEU	ROUGE-1	ROUGE-2
Completion 1	0.166	0.137	0.533	0.186
Completion 2	-0.644	0.001	0.207	0.148

Human Evaluation

Human evaluation [2023]

- Benchmark evaluation cannot cover all the quality dimensions of LLMs
- Human evaluations rely on A/B testing:
  - Pits two generations of two (blinded) models for same input
  - Humans express a preference (A is better, B is better, tie)

Two broad methods:

1. ELO scores
2. Votes w. IRR

The ELO score is a ranking score derived from chess. It is typically measured and updated after every single match. In the context of LLM "gamified" A/B testing it quantifies and aggregates user preferences towards a given model.

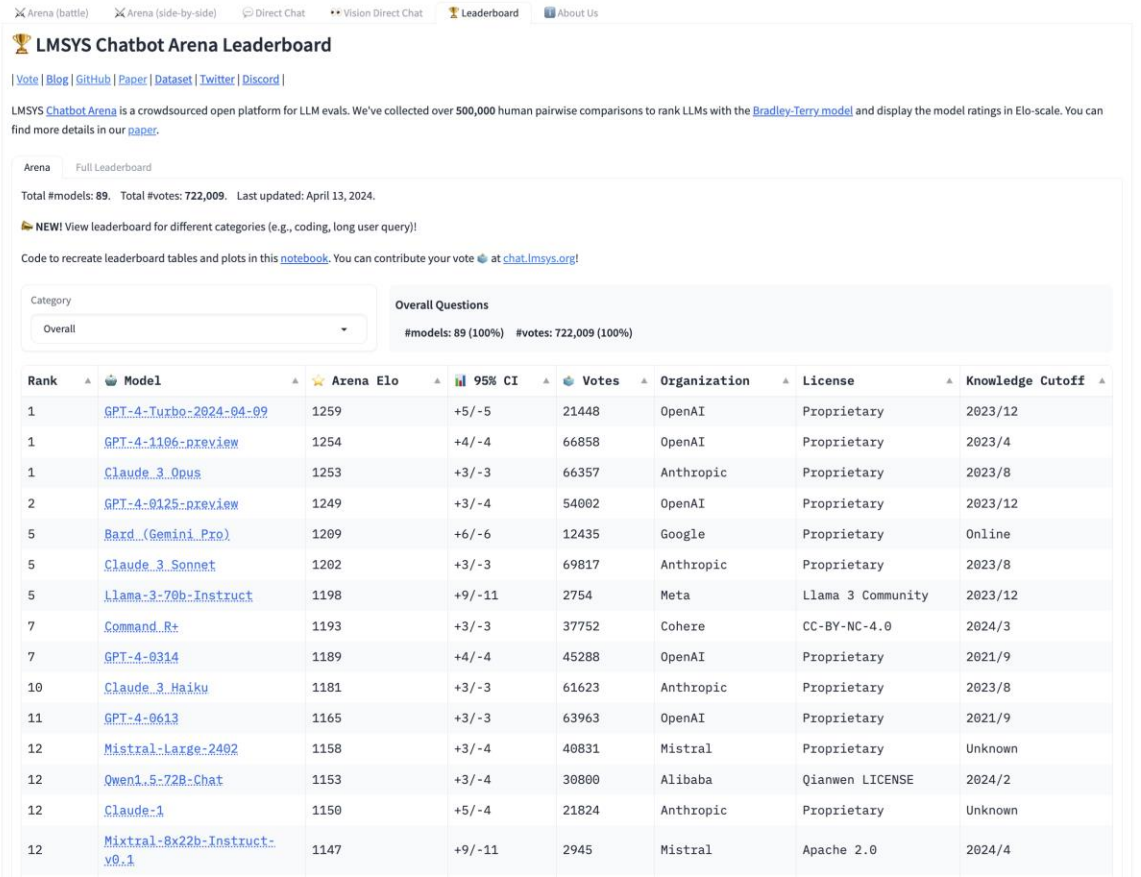
In what follows, $E_A \in [0,1]$ is the expected likelihood of A winning, whereas S_A is the "observed" likelihood with $S_A = 1$ if A wins, $S_A = 0$, if A loses and $S_A = 0.5$. Let ELO_A be the initial ELO score of A. After playing a game the new ELO score of A is

$$\text{ELO}'_A = \text{ELO}_A + k \cdot (S_A - E_A)$$

where k denotes a scaling factor that determines how much influence each particular match can have on the overall ELO rating of the player.

Chatbot arena leaderboard [2023-]

https://chat.lmsys.org/?leaderboard



Arena (battle) | Arena (side-by-side) | Direct Chat | Vision Direct Chat | Leaderboard | About Us

LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Arena | Full Leaderboard

Total #models: **89**. Total #votes: **722,009**. Last updated: April 13, 2024.

NEW! View leaderboard for different categories (e.g., coding, long user query)!

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

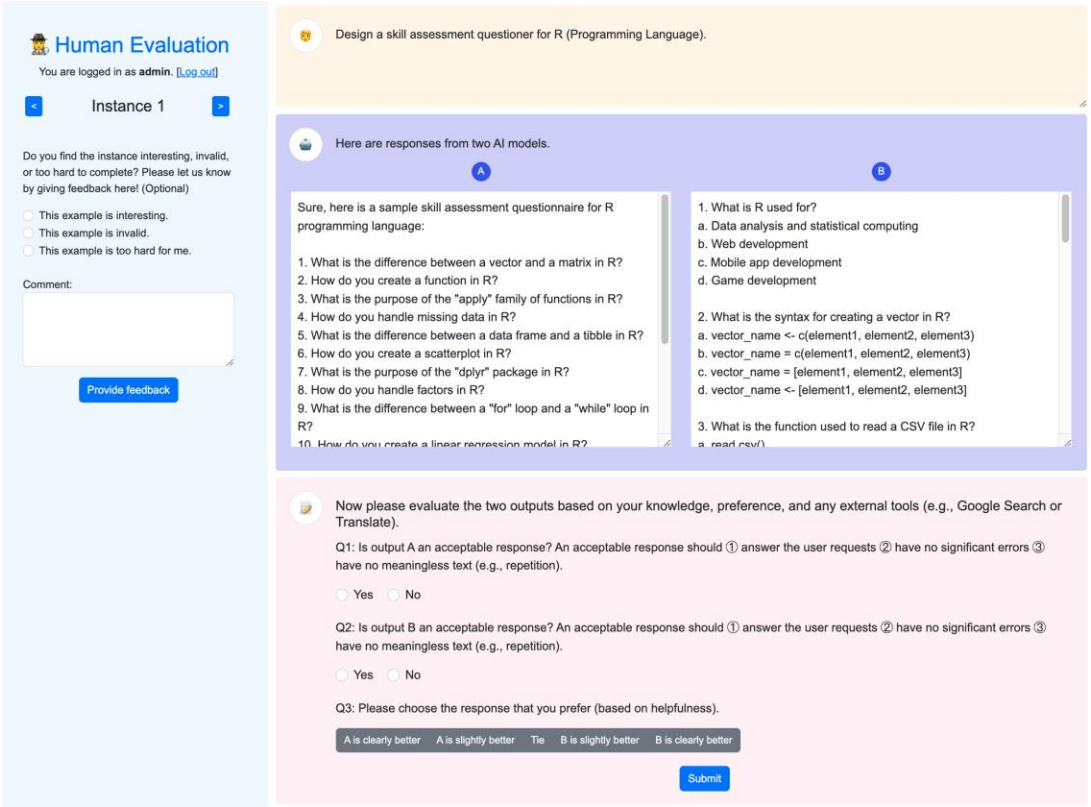
Overall Questions

#models: 89 (100%) #votes: 722,009 (100%)

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+5/-5	21448	OpenAI	Proprietary	2023/12
1	GPT-4-1106-preview	1254	+4/-4	66858	OpenAI	Proprietary	2023/4
1	Claude 3 Opus	1253	+3/-3	66357	Anthropic	Proprietary	2023/8
2	GPT-4-0125-preview	1249	+3/-4	54002	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+6/-6	12435	Google	Proprietary	Online
5	Claude 3 Sonnet	1202	+3/-3	69817	Anthropic	Proprietary	2023/8
5	Llama-3-70b-Instruct	1198	+9/-11	2754	Meta	Llama 3 Community	2023/12
7	Command R+	1193	+3/-3	37752	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	45288	OpenAI	Proprietary	2021/9
10	Claude 3 Haiku	1181	+3/-3	61623	Anthropic	Proprietary	2023/8
11	GPT-4-0613	1165	+3/-3	63963	OpenAI	Proprietary	2021/9
12	Mistral-Large-2402	1158	+3/-4	40831	Mistral	Proprietary	Unknown
12	Qwen1.5-72B-Chat	1153	+3/-4	30800	Alibaba	Qianwen LICENSE	2024/2
12	Claude-1	1150	+5/-4	21824	Anthropic	Proprietary	Unknown
12	Mixtral-8x22b-Instruct-v0.1	1147	+9/-11	2945	Mistral	Apache 2.0	2024/4

Human evaluation – A/B testing

https://github.com/allenai/open-instruct/tree/main/human_eval

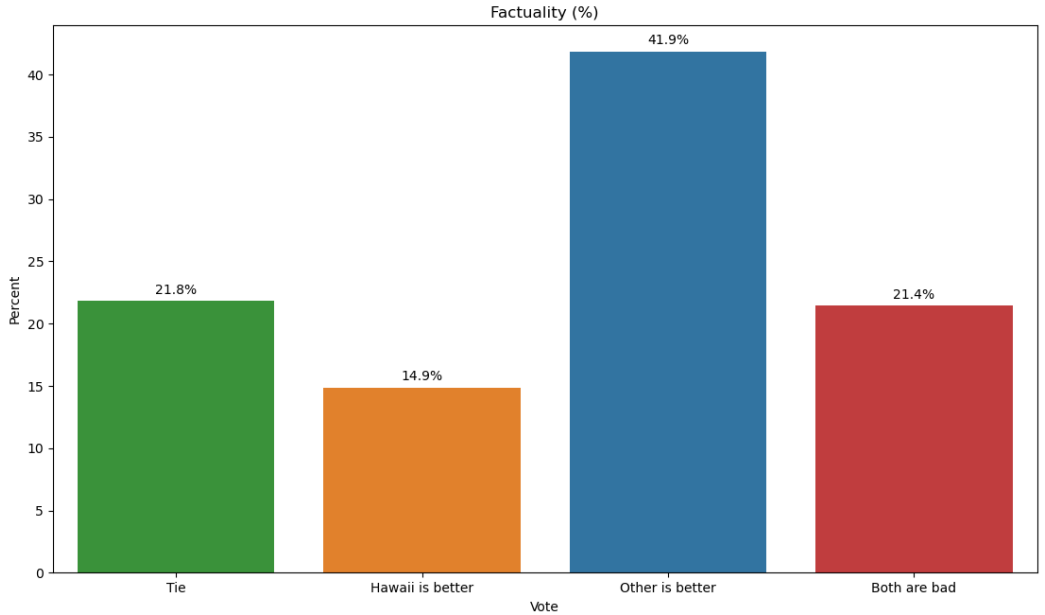


Human evaluation – factuality [2024]

- ✓ Elsevier analysed the factuality of LLMs on custom LLS input prompts (NER, RE, open and closed book QA)
- ✓ >500 prompts judged by >40 content experts
- ✓ Alpha score / IRR > 0.6

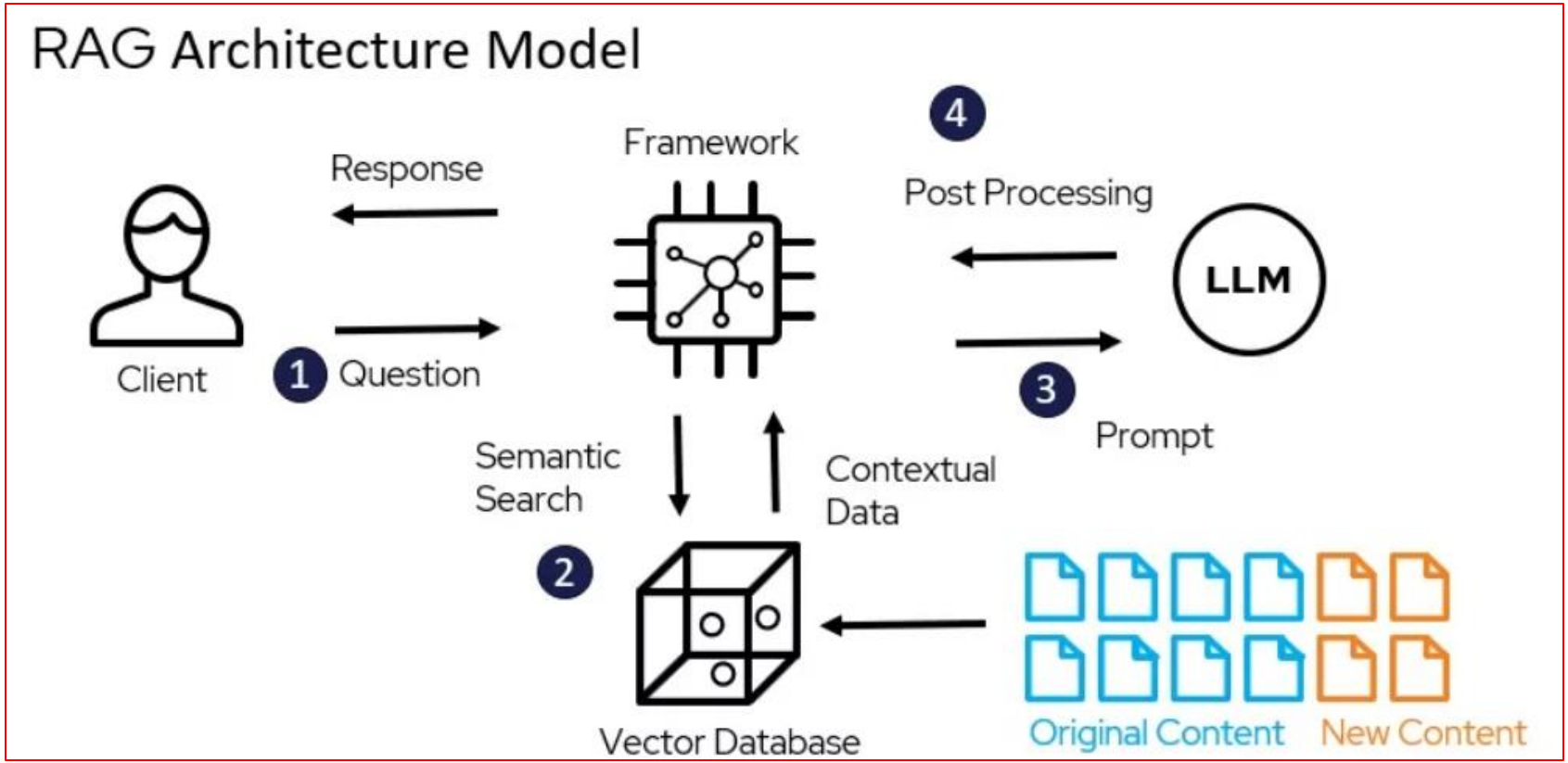
Models

A) Hawaii: LSS-specific model (8b)
B) Other: LSS-fine tuned llama2 (7b), llama-2-chat (7b), GPT-3.5-turbo (175b)



RAG Systems

Retrieval augmented generation [2022-]



References

- 2022 - EleutherAI: Evaluating LLMs - https://www.eleuther.ai/projects/large-language-model-evaluation
- 2022 - Holistic Evaluation of Language Models (HELM) - https://arxiv.org/abs/2211.09110
- 2023 - How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources - https://arxiv.org/abs/2306.04751
- 2023 - ChemNLP: A Natural Language-Processing-Based Library for Materials Chemistry Text Data - https://pubs.acs.org/doi/10.1021/acs.jpcc.3c03106
- 2024 - Are large language models superhuman chemists? - https://arxiv.org/abs/2404.01475
- 2024 - Retrieval-Augmented Generation for Large Language Models: A Survey - https://arxiv.org/abs/2312.10997

Thank you!