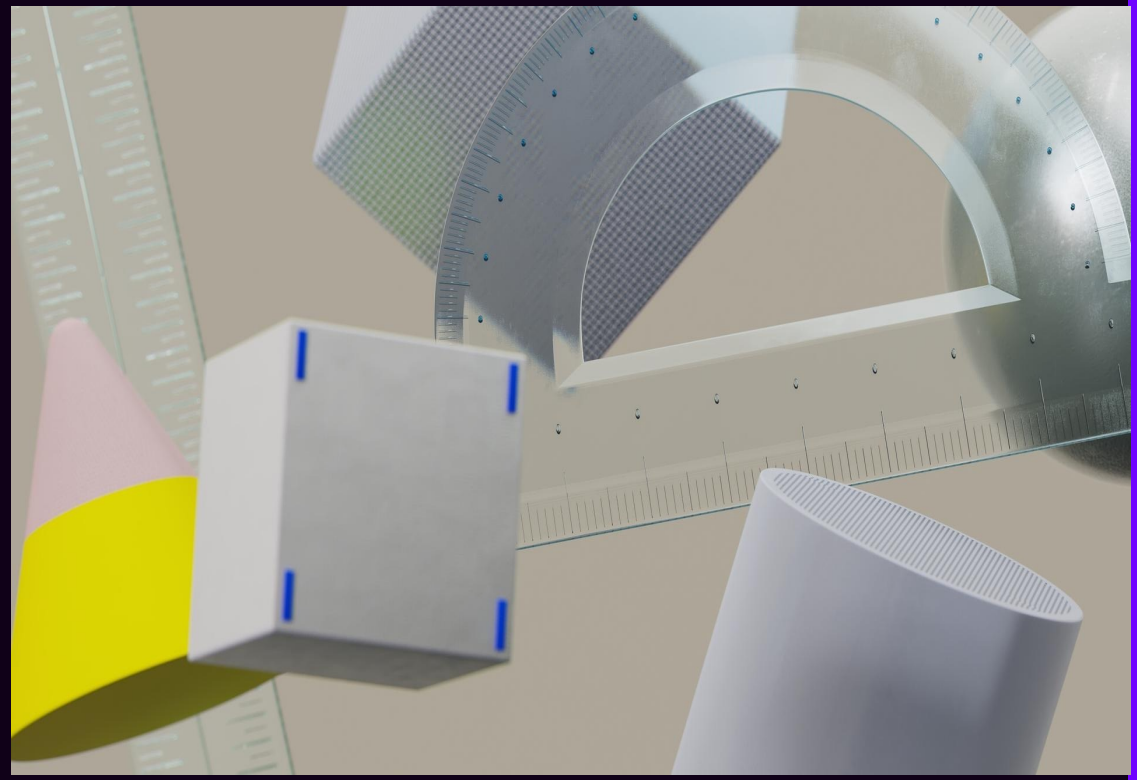
Overview of some LLM benchmarks

Lubos Steskal



En del av
EGMONT

Today’s Menu

- Introduction
- What to benchmark?
- Evaluation metrics
- Benchmarking frameworks
- ”User friendly” benchmarking
- Concluding Thoughts

Introduction and Motivation

Why benchmark models?

- Most models come with some benchmarks, but
  - These might not include scenarios relevant for our use case.
  - These might not measure the aspects we care about.
  - The underlying data structure/distribution can vary significantly from our data.
  - These are often hard to replicate.
  - Should we really trust them?

- When used in real production, it is very important to understand the reliability and limits of the models being used.

- 1 Benchmark = Task + Metric
- We want a more

Introduction and Motivation

- 1 Benchmark = Task + Metric
- The goal is to get a holistic picture of how well the model understands language
- How it can generate meaningful and coherent text, and its ability to reason, among other things
- There are several key methodological approaches to this process

How to Benchmark?

- **Task-Specific Evaluation**: In this method, the language model is evaluated based on its performance on specific tasks, such as question answering, text summarization, or machine translation. This evaluation usually involves using established datasets and metrics for each task.

- **Few-Shot Learning Evaluation**: In this approach, the language model is given a few examples of a task at inference time, then asked to complete a similar task. The performance on these tasks is then measured. This method tests the model's ability to generalize learning from a few examples to a new instance of a task.

- **Zero-Shot Learning Evaluation**: Similar to few-shot learning, but in this case, the model is not given any prior examples at the inference time. The model's performance on these tasks, not seen during training, is measured. This tests the model's ability to understand and complete tasks it was not specifically trained to perform.

How to Benchmark?

- **Fine-Tuning Evaluation**: In this approach, the model is fine-tuned on a specific task with additional task-specific training, then its performance on that task is measured. This helps to understand how well the model can adapt to specific tasks after pre-training.

- **Human Evaluation**: Finally, human evaluation plays a crucial role in benchmarking language models. This might involve humans rating the coherence, relevance, or factual correctness of the text generated by the model. Human evaluation can also involve more specific tasks, such as assessing the model's ability to generate creative stories, its

- **Bias and Fairness Evaluation**: This involves assessing the model's output for any biases or unfair portrayals based on factors like gender, race, religion, etc. This helps in understanding if the model has inadvertently learned any societal biases from its training data.

- **Safety and Robustness Evaluation**: This approach tests how well the model handles malicious input, misinformation, or adversarial attacks.

Some Common Metrics

- **Perplexity:** Perplexity is a measure of how well a language model predicts a sample. Lower perplexity indicates that the model is less ”perplexed” by the test data, meaning it’s performing better.

- **BLEU (Bilingual Evaluation Understudy) Score:** BLEU score is a metric used to evaluate the quality of machine-generated text, such as in machine translation. It measures how close the model’s output is to a human reference translation. Higher BLEU scores indicate better performance.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score:** ROUGE score is a set of metrics used to evaluate automatic summarization and machine translation. It compares the model’s output with a set of reference summaries. Higher ROUGE scores indicate better performance.

- **F1 Score:** The F1 score is the harmonic mean of precision and recall, and it’s often used in tasks such as named entity recognition and other classification tasks. A higher F1 score indicates better performance.

Some Common Metrics

- **Accuracy:** Accuracy is the proportion of correct predictions made by the model out of all predictions. It's a common metric for classification tasks.

- **Precision, Recall, and F1 Score:** These metrics are often used in information retrieval and classification tasks. Precision measures the proportion of true positive predictions among all positive predictions, recall measures the proportion of true positive predictions among all actual positives, and the F1 score is the harmonic mean of precision and recall.

- **Matthews Correlation Coefficient (MCC):** MCC is a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

- **Spearman's Rank Correlation Coefficient:** This is used to measure the strength and direction of association between two ranked variables. It's often used in tasks like semantic textual similarity.

There is more!

Traditional NLP Tasks

- Contextual question-answering
- Context-free question answering
- Reading comprehension
- Conversational question answering
- Summarization, paraphrase, text simplification
- Word sense disambiguation, coreference resolution
- Question generation, narrative understanding, dialogue system
- Memorization, morphology, translation, writing style
- Grammar, syntax, and segmentation
- ...

There is more!

Traditional NLP Tasks

- Contextual question-answering
- Context-free question answering
- Reading comprehension
- Conversational question answering
- Summarization, paraphrase, text simplification
- Word sense disambiguation, coreference resolution
- Question generation, narrative understanding, dialogue system
- Memorization, morphology, translation, writing style
- Grammar, syntax, and segmentation
- ...

Logic, Math, Code

- Algorithms, logical reasoning, implicit reasoning
- Mathematics, arithmetic, algebra, mathematical proof
- Decomposition, fallacy, negation
- Computer code, semantic parsing, probabilistic reasoning
- ...

There is more!

Traditional NLP Tasks

- Contextual question-answering
- Context-free question answering
- Reading comprehension
- Conversational question answering
- Summarization, paraphrase, text simplification
- Word sense disambiguation, coreference resolution
- Question generation, narrative understanding, dialogue system
- Memorization, morphology, translation, writing style
- Grammar, syntax, and segmentation
- ...

Logic, Math, Code

- Algorithms, logical reasoning, implicit reasoning
- Mathematics, arithmetic, algebra, mathematical proof
- Decomposition, fallacy, negation
- Computer code, semantic parsing, probabilistic reasoning
- ...

Understanding the World

- Causal reasoning, consistent identity
- Physical reasoning, common sense
- Visual reasoning
- ...

Scientific and Technical Understanding

- Biology, chemistry, physics, medicine
- Domain-specific knowledge
- ...

Pro-Social Behavior

- Alignment, social bias, racial bias, gender bias
- Religious bias, political bias, toxicity, inclusion
- Truthfulness, misconceptions, accommodation to reader
- Human-like behavior, self-awareness, emotional intelligence
- ...

Mechanics of Interaction with Model

- Self-play, self-evaluation, multiple choice, free response
- Game play, repeated interaction, non-language
- Numerical response, show work, zero-shot, one-shot, many-shot
- ...

Targeting Common Language Model Technical Limitations

- Context length, multi-step, out of distribution
- Instructions, tokenization, paragraph
- ...

Other Tasks

- Analogical reasoning, creativity, linguistics, sufficient information
- Riddle, low-resource language, non-English, non-language
- Multilingual, cheating, example task, json, programmatic
- ...

Understanding Humans

- Theory of mind, emotional understanding
- Social reasoning, gender prediction
- Intent recognition, humor, figurative language
- ...

There is more!

Traditional NLP Tasks

- Contextual question-answering
- Context-free question answering
- Reading comprehension
- Conversational question answerin
- Summarization, paraphrase, text simplification
- Word sense disambiguation, coreference resolution
- Question generation, narrative understanding, dialogue system
- Memorization, morphology, transl writing style
- Grammar, syntax, and segmentati
- ...

Logic, Math, Code

- Algorithms, logical reasoning, implicit

Scientific and Technical Understanding

- Biology, chemistry, physics, medicine
- Domain-specific knowledge
- ...

Pro-Social Behavior

- Alignment, social bias, racial bias, gender

s,

el

n-

o-

shot, one-shot, many-shot

- ...

Targeting Common Language Model Technical Limitations

- Context length, multi-step, out of distribution
- Instructions, tokenization, paragraph
- ...

Other Tasks

- Analogical reasoning, creativity, linguistics, sufficient information
- Riddle, low-resource language, non-English, non-language
- Multilingual, cheating, example task, json, programmatic
- ...

Understanding Humans

- Theory of mind, emotional understanding
- Social reasoning, gender prediction
- Intent recognition, humor, figurative language
- ...

Benchmarking frameworks to the rescue

- Standardize benchmarks
- Run many benchmarks on many models
- Central ”source of truth”
- Reproducibility
- Transparency
- Modularity

Some Benchmarking Frameworks

BIG-Bench

- Beyond the Imitation Game Benchmark.
- Collaborative benchmark intended to probe large language models and extrapolate their future capabilities.
- more than 200 tasks.
- By Google
- https://github.com/google/BIG-bench/blob/main/docs/doc.md
- https://github.com/google/BIG-bench

Language Model Evaluation Harness

- Project provides a unified framework to test generative language models on a large number of different evaluation tasks.
- Small subset of tasks used in HuggingFace Open LLM Leaderboard
- Actively used andextended by the community
- https://github.com/EleutherAI/lm-evaluation-harness

HELM

- Holistic Evaluation of Language Models
- https://crfm.stanford.edu/helm/latest/
- Holistic structure beyond just Task+Metric
- Contains online leaderboard

Some More Benchmarking Frameworks

GLUE/superGLUE

- General Language Understanding Evaluation benchmark
- A nine-task benchmark using diverse, pre-existing datasets for sentence understanding.
- A public leaderboard and dashboard for tracking and visualizing model performance.
- SuperGLUE, a new benchmark styled after GLUE with more difficultimproved resources
- https://gluebenchmark.com/

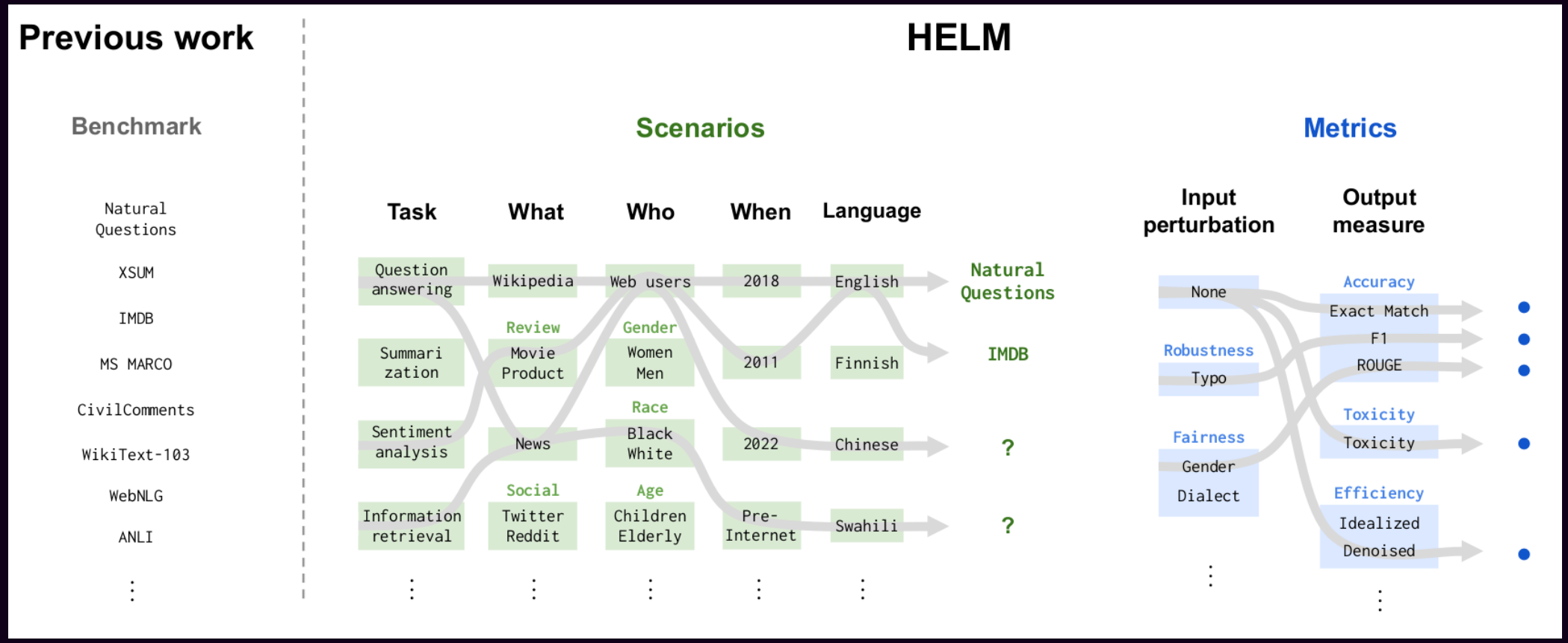
OpenAI Evals

- framework for evaluating LLMs (large language models) or systems built using LLMs as components. It also includes an open-source registry of challenging evals.
- https://github.com/openai/evals

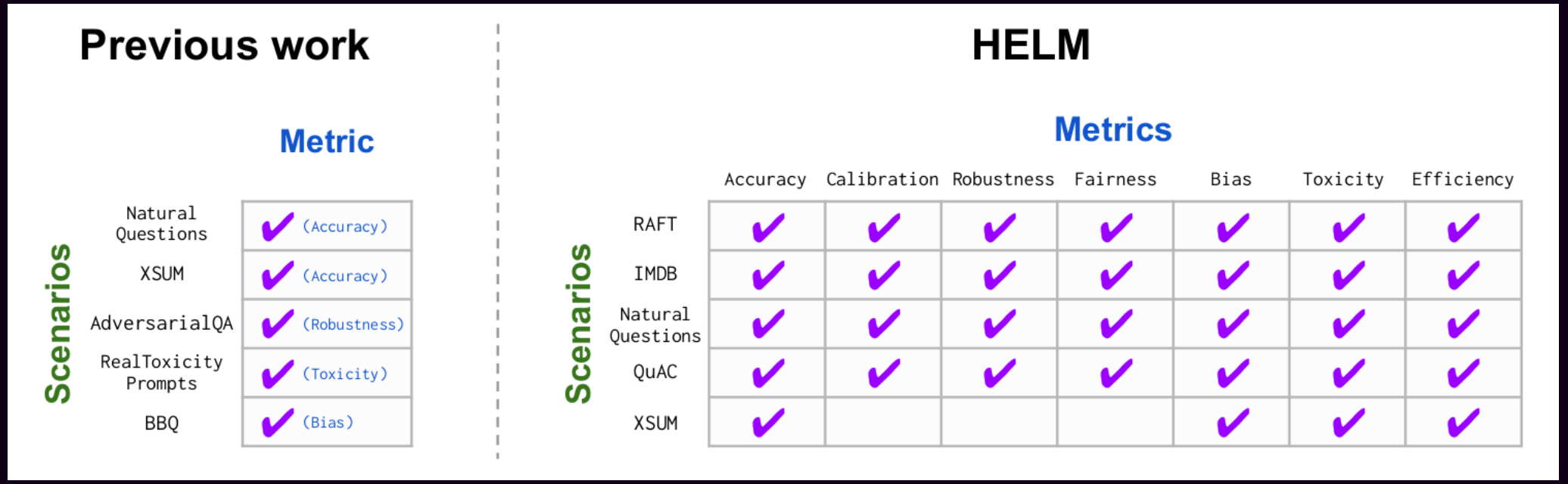
Plan Benchmark

- Large Language Models Still Can't Plan（A Benchmark for LLMs on Planning and Reasoning about Change）
- Identified weaknesses in BIG Bench
- https://ar5iv.labs.arxiv.org/html/2206.10498
- https://github.com/karthikv792/gpt-plan-benchmark

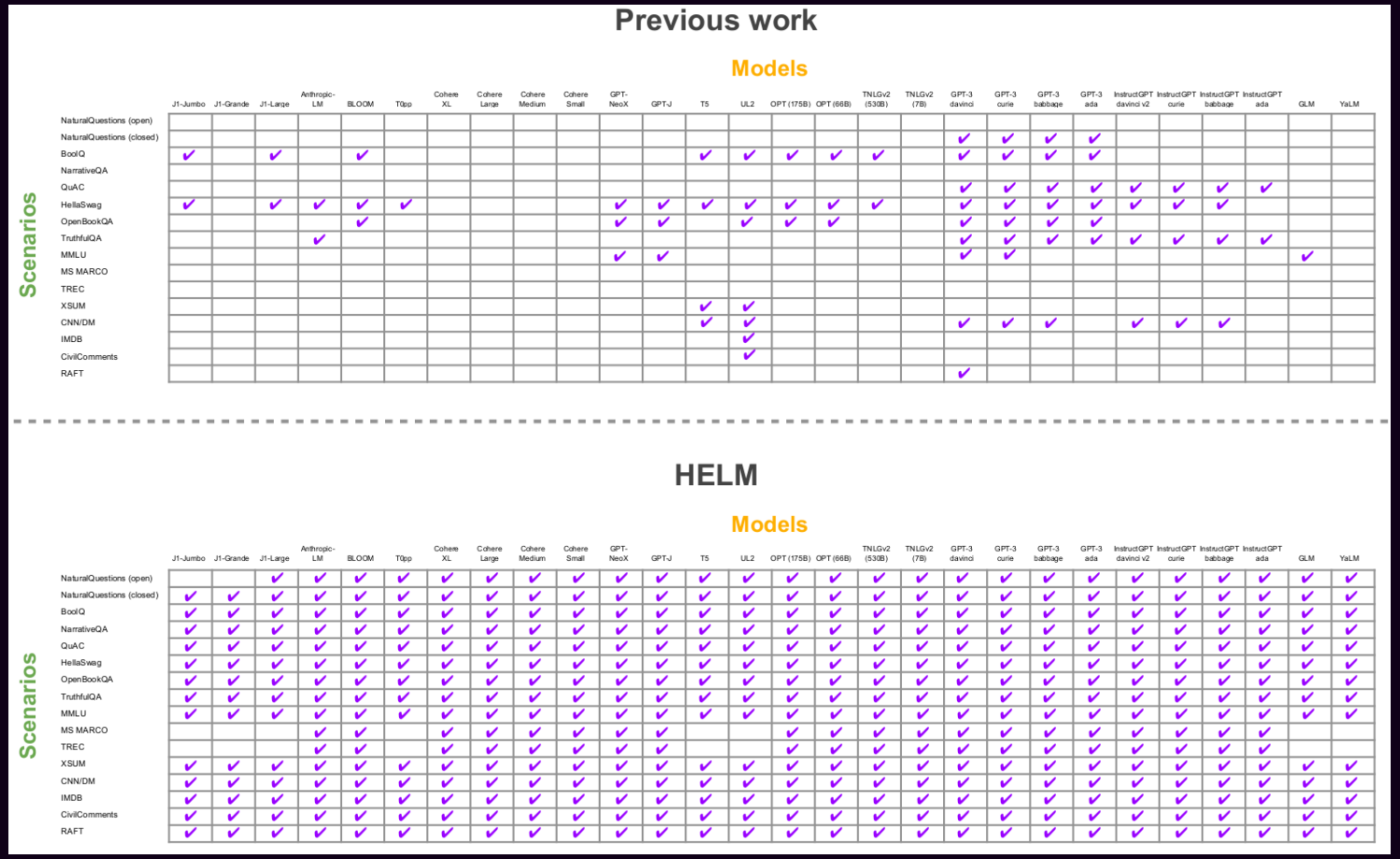
HELM (Holistic Evaluation of Language Models)



HELM (Holistic Evaluation of Language Models)

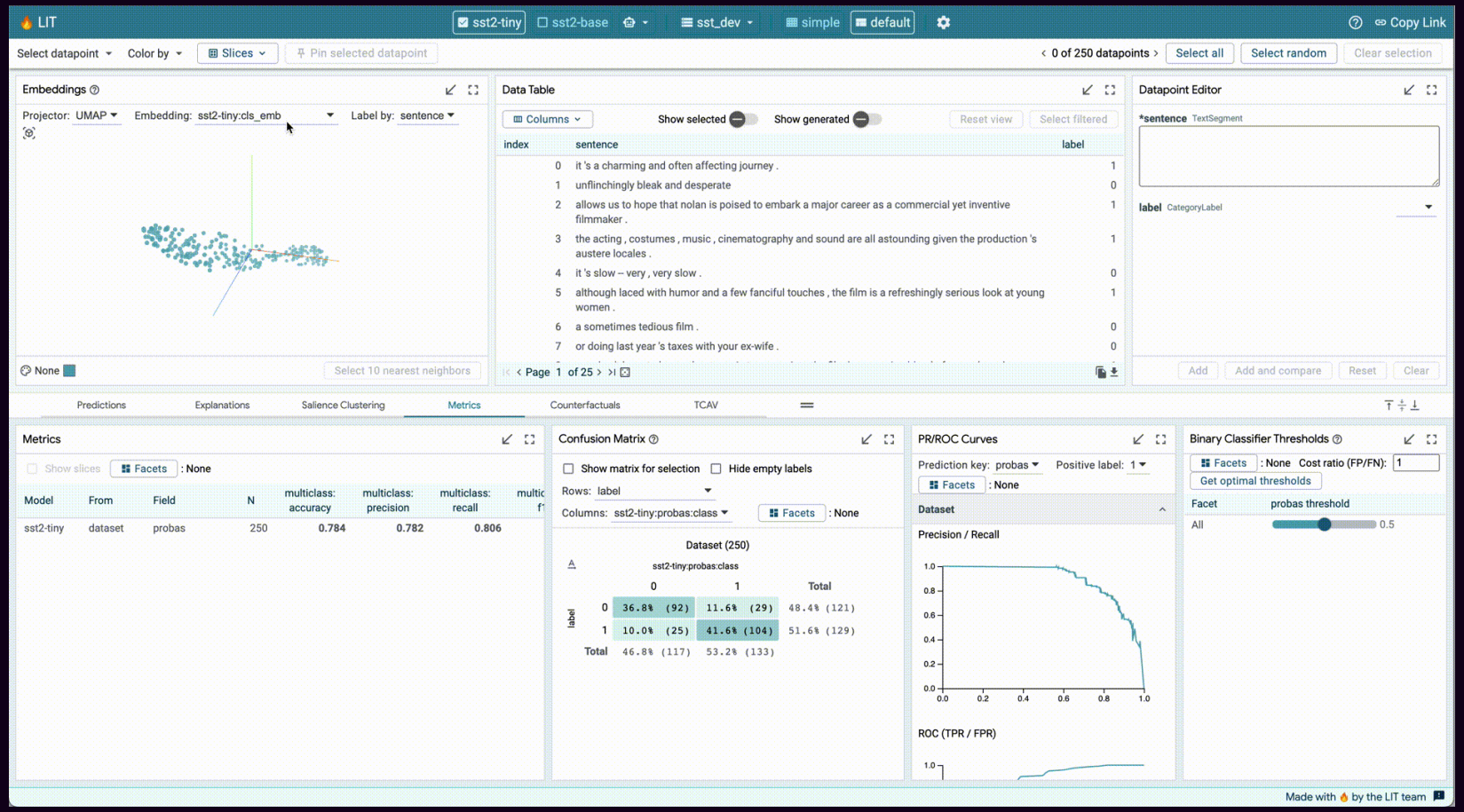


HELM (Holistic Evaluation of Language Models)



Human in the Loop Model Evaluation - The Learning Interpretability Tool

- https://ai.googleblog.com/2020/11/the-language-interpretability-tool-lit.html
- https://pair-code.github.io/lit/



Concluding Thoughts

- It is important for us to define understand which elements of LLM are most important to us as an industry sector
- Explore if there exist good enough benchmarks
- If not to provide them into the open benchmarking environment.