TPC Collaboration Hackathon: Data, Evaluation, and Architectures

Dates: October 9-11, 2024

Location: Argonne National Laboratory

**APS Conference Center #402, E1100, E1200, Lower Gallery**

[Full Meeting Details](https://docs.google.com/document/d/186_L7YIs-VWHelHq6b1a59DmHRHVnzUU4G8CVfI28dM/edit?usp=sharing) (group goals, logistics, registration, etc.)

**Repo to store Slides and Notes:**

**Slides and Note:** [**TPC Argonne Hackathon Oct. 9-11, 2024**](https://drive.google.com/drive/folders/1UJ60R0wY3e_us4gNfF4yQMVqHj4crqDU?usp=sharing)

Presentations: [what you have done, gap analysis, what’s next, and collaboration]

Time for sessions indicated below is in CST

# Wednesday - October 9, 2024

*Participants should have lunch on their own at the Argonne Guest House or Argonne Cafeteria*

13:00-14:00 [Lower Gallery]Plenary: Opening Goals (I. Foster, R. Underwood, F. Cappello)   
Zoom: <https://argonne.zoomgov.com/j/1606381897> passcode: 468339

* Intro about TPC, logistics, Safety (evacuation) (10 min)
* Introduction to topics
  + Data (15 min)
  + Eval (15 min)
  + Model (15 min)

14:00-15:30 Session 1

* [E1100] Eval (Skill Evaluation groups)   
  [Zoom record](https://argonne.zoomgov.com/rec/share/W7VoTCcVlEzlZ5nTHtAgWWcaJAKGpKoFwQ4L4u-3v4TbqcFYmW38FgSSSN3cnq7M.lbnpIOtLG6WiH_7N) Passcode: ST3i^pNM
  + State of the Evaluation (15 minutes each)

Notes: [TPC\_hackathon](https://docs.google.com/document/d/1-EyZs4WLuC_F5z698dLix3WMjSadj6vM8kUdT7oi42g/edit)

* + - ANL (Franck/Sandeep)
    - BSC (Javier)
    - ETH (Maciej)
  + Evaluation of open responses (30 mins.)
    - CheckEmbed - Maciej (ETH)
    - Neil Getty (ANL)
* [E1200] Data – state of the projects  
  Zoom: <https://argonne.zoomgov.com/j/1608562033>
  + State of Datasets, Text, etc… at each site
  + State of Data pipelines at each site
  + Roadmapping

15:30-16:00 Break

16:00-17:30 Session 2

* [E1100] Eval (Skill Evaluation groups)

[Zoom record](https://argonne.zoomgov.com/rec/share/W7VoTCcVlEzlZ5nTHtAgWWcaJAKGpKoFwQ4L4u-3v4TbqcFYmW38FgSSSN3cnq7M.lbnpIOtLG6WiH_7N) Passcode: ST3i^pNM

* + Present Argonne’s Multiple Choice Question (MCQ) initiative and the challenge of generating/validating scientific MCQs at scale. (30 mins.)
    - Analysis of the MCQ benchmark - Murat Keceli (ANL)
    - Coding Benchmark for Science - Eliu Huerta (ANL)
  + Present Argonne’s Multiple Choice Question (MCQ) initiative and the challenge of generating/validating scientific MCQs at scale. (30 mins.)
    - Automatic generation of benchmarks in Astrophysics - Josh Nguyen (University of Pennsylvania/ANL)
    - Automatic validation of scientific benchmarks Neil Getty (ANL)
  + Eval in the age of Inference Compute (i.e. how do we measure and compare when there are different numbers of “thinking” tokens like for o1 using the same size model) (15 mins.)
    - Sandeep Madireddy (ANL)
* [E1200] Data – Workflows for Data Hackathon   
  Zoom: https://argonne.zoomgov.com/j/1608562033
  + Scaling Data Ingest (PDFs/Text)
  + Scaling and Dedup
  + Scaling Curation
  + Scaling vectorDBs and databases for data

*Dinner in self-organized small groups*

# Thursday - October 10, 2024

08:00-08:30 [Lower Gallery] Coffee and Breakfast Snacks

08:30-09:00 Plenary

Update from day 1.

09:00-09:30 Plenary[Zoom record](https://argonne.zoomgov.com/rec/share/c0oFkE2Mc3VZlLW1C2SSfJrYz4GPpXFNUM4JP1zdt_WCVpJJqWFLyGjnJYlPtzsf.otbTEPEcsxf_2Voo) Passcode: kr66h.7+   
 **Salman Habib (15 + 15)**

09:30-10:45 [Lower Gallery] Session 3 (Eval+data) (chairs: Sandeep, Robert, Franck)

* DataSources/Eval – Narratives
  + Speaker:
    - * Clark Cucinell (ANL),
      * Azton Wells (ANL)
  + Datasets+Eval for multi-modal data (e.g. use cases and metrics)
* Model/Data adaptations for Multi-modal (tokenizer, semantic, data reduction for training, etc.)
  + Do we need and if so how to use a small set of tokenizers for training and inferences? Speaker: Neil Getty, Tanwi Mallick (ANL)
* Datasets for training this tokenizer - images, graphs, and beyond…
* Floating point tokenization in HEP: Azton Wells (ANL)

10:45-11:00 Break

11:00-12:30 [Lower Gallery] Session 4 (Eval+data)

* Multi-modal RAG
  + Arvind Ramanathan (ANL)
  + Robert Underwood (ANL)
* Multi-modal Contamination/Dedup
  + Robert Underwood, Yadu, Arham

12:30-14:00 [Lower Gallery] Lunch

14:00-15:30 Session 5

* [E1100] Eval (AI Risk, Trustworthiness and Reliability):

[Zoom record](https://argonne.zoomgov.com/rec/share/ZQ91R_oA1Mf6my1HyrOJC60R4zArMchomZmvJHa7tj2fChniFs-R89YQDgh-F_6K.uwRUAnQJaHZB-PTU): Passcode: @?=i+5B%

* + Safety Evaluation and Red-teaming (perspectives, prompting techniques, and frameworks)
    - ANL (Sandeep Madireddy/Franck Cappello)
    - LLNL (Bhavya Kaikhura)
    - UChicago (Bo Li)
    - ORNL (Sudarshan Srinivasan/Maria Mahbub/ Zheming Jin)
    - LANL (Eric Michalak)
    - ML Commons (Chris Knotz)
  + Discussions (30 min)
* [E1200] Data – Workflows for Data Hackathon, continued   
  Zoom: <https://argonne.zoomgov.com/j/1608562033>
  + Scaling Data Ingest (PDFs/Text)
  + Scaling and Dedup
  + Scaling Curation
  + Scaling vectorDBs and databases for data
* [Lower Gallery] Model Architecture and Performance

Zoom: <https://zoom.us/j/95420616211> passcode: !ffZs6AzBq

* + Identify performance benchmarks to use
  + Develop steps to profile on systems such as Polaris/Aurora/Stampede
  + Select frameworks to profile

15:30-16:00 Break

16:00-17:30 Session 6 (Eval: AI Safety and Reliability)

* [E1100] Eval (AI Risk, Trustworthiness and Reliability):

[Zoom record](https://argonne.zoomgov.com/rec/share/ZQ91R_oA1Mf6my1HyrOJC60R4zArMchomZmvJHa7tj2fChniFs-R89YQDgh-F_6K.uwRUAnQJaHZB-PTU): Passcode: @?=i+5B%

Creation of proxy safety benchmarks (e.g. WMDP) for sensitive science areas [30 min]

* + - Bhavya Kaikhura (LLNL)
    - Eric Michalak (LANL)
  + Reliability Metrics, Uncertainty Quantification and Hallucination detection [30 min]
    - UQ: Kaidi Xu (Drexel)
    - Hallucination Detection: Chaowei Xiao (UW-MADISON)
  + Consideration for Risk and safety in multimodal models [30 min]
    - Multimodal safety benchmark: Chaowei Xiao (UW-MADISON)

* [E1200] Data (Joint Session with Models?) Datamix and Experimentation [https://argonne.zoomgov.com/j/1608562033]
  + Ablation study methods, continual pretraining, distilled models, etc…
* [Lower Gallery] Model Architecture and Performance

Zoom: https://zoom.us/j/95420616211 passcode: !ffZs6AzBq

* + Run scripts to collect metrics at small scale and large scale
  + Understand impact of various CCL environment variables on performance
  + Optimize performance of LLM framework on Polaris/Aurora/Stampede

*Dinner at* [*Argonne Guest House restaurant*](https://www.anlgh.org/restaurants-dining)

# Friday - October 11, 2024

08:30-09:00 [Lower Gallery] Coffee and Breakfast Snacks

09:00-10:00 Session 7

* [E1100] Eval (Gap identification, proposed direction, collaborations):

Definition of difficulty and way forward, given O1 and future (soon) OpenAI stage   
 3 (15 mins.)

* + - Franck Cappello (ANL)

Roadmapping Skills (45 minutes)

* + - Overall eval strategy: Q&As + Lab style experiments + In the wild
    - Manual MCQs
    - Automatic generation/validation
    - Evaluation of open response
  + Roadmapping Safety (30 minutes)
  + Create a charter and review potential members for an international structure similar to what has been done for large-scale science projects, with co-chairs and representation from multiple scientific domains (15 minutes)
* [E1200] Data – Workflows for Data Hackathon, continued; demo progress

Zoom: https://argonne.zoomgov.com/j/1608562033

* + Scaling Data Ingest (PDFs/Text)
  + Scaling and Dedup
  + Scaling Curation
  + Scaling vectorDBs and databases for data
* [Lower Gallery] Model Architecture and Performance

Zoom: https://zoom.us/j/95420616211 passcode: !ffZs6AzBq

* + Optimize performance of LLM framework on Polaris/Aurora/Stampede

10:00-10:30 Report/Slides preparation for the plenary session

10:30-11:00 Break

11:00-12:30 [Lower Gallery] Plenary: 30 minutes per team

* Readout - 30 min per group

12:30-14:00 [Lower Gallery] Lunch and Adjourn