# **TPC hackathon session 1 notes**

# **Day 1 session 1: State of the projects (Data)**

1. Who is the team
   1. AGPT team intro
      1. Ian - co-lead
      2. Robert - co-lead
      3. Arham Khan - deduplication
      4. Kyle - PDF parsing
      5. Alexander Brace - PDF parsing
      6. Nick,—continuous pretraining for the SC24 runs, bio naratives
      7. Brad – legal team
      8. Yadu – dedup
      9. Ozan - PDF parsing team
      10. Murat -> in eval today
      11. Tanwi -> in eval today
      12. Clark, Mallick – narratives for BIO
   2. Presentation: Brad Ullrick—Legal department
      1. MOU (memorandum of understanding)
      2. barriers and workarounds with dealing with publishers
         1. Data sharing
         2. time-consuming negotiation
   3. Introduction: Junchao Zhang—Software engineer at MCS AT ANL, PETSc team (PDE solver)
      1. preparing PETSc data (arxiv)—papers, attached files, figures
      2. application: PDE applications
   4. Presentation: Carlo—PDF parsing for 10 months
      1. presentation about accuracy across different PDF parsers: Nougat, Marker, PyMuPDF, Tesseract, GROBID, pypdf  
           
         discussion 1: old PDFs may or may not have embedded metadata text which if it exists may or may not be valuable
      2. Discussion 2: figures in PDFs and ViTs
   5. Introduction: Tom Perterka—Computer Scientist at ANL
      1. Large-scale models on HPCs with a focus on Particles (LAMMPS) and Unstructured Grid ()
   6. Introduction: Jorge Palomar—Barcelona Supercomputing Center, Spain
      1. Language models in Spain
      2. multilinguistic LLMs
   7. Introduction: Mathew Dearing—Software engineer at ANL and PhD in UIUC
      1. generative AI on HPCs
      2. use LLMs to generate parallel code
   8. Rob Latham—ANL
      1. I/O for simulation and modeling codes
      2. Helping with RAG as used in Narratives for generation from scientific simulation to data
         1. template-level
         2. connect raw data and text
         3. multi-modal retrieval (e.g. find relevant papers to simulation) (RAG could be a bottleneck due to speed)
   9. Hayden Hoschouer
2. What progress have you made this year?
   1. parsed 14 datasets containing primarily scientific text
   2. executed agreements with ACM and ASM
   3. in-progress agreements with AAAS, JAMA, LoC
   4. developed a best-in-class adaptive parsing approach
   5. developed a scalable deduplication framework
   6. Improved parsing for narratives

Discussion around data format and naratives

* 1. Discussion 1: format of parsed data—.jsonl w/ id, metadata, etc
  2. Discussion 2: clarification of narratives
  3. Discussion 3: clarification of features and relation extraction for filling narrative templates
  4. Discussion 4: OpenAI’s RLHF, textbook (high-quality) training data for 6B model

Discussion around building narratives for unstructured grid and particle applications:

* Preprocessing:
  + Thresholding
  + Gradients
  + FFT
* Level 1 features:
  + Minima, maxima, extreema
  + Streamlines
  + Particles -> Density
* Level 2 features:
  + Tracking positions level 1 features of over time
  + Tracking merging/biforcation of level 1 features over time
  + Tracking/identifying regime changes
* How do we link these to high level scientific questions. We do not think that extensive numbers in the prompt is likey to work. We gave the example of Clarks