Workloads and Benchmarking

Omar Alonso[1*] & Kenneth Church[2]

1: Amazon (work not related to author's position at Amazon), 2: Northeastern University

Abstract

- Workloads are a moving target
  - Chatbots/LLMs → Consequences for workloads.
  - As workloads evolve, benchmarks need to catch up.
- Ideally, performance on benchmarks should be
  - A leading indicator of customers experience
  - not a lagging indicator.
- Lessons to be learned from benchmarking exercises from other fields:
  - The SPEC benchmark was designed to help customers of CPUs decide what to buy.
  - https://www.spec.org/
- This is a challenging question because
  - Different customers have different workloads.

Examples of Benchmarks with Multiple Tasks/Tracks

- SPEC
- Many TREC Tracks over Many Years
- MSMARCO (Information Retrieval)
- GLUE (Natural Language Processing)
- CRAG KDD CUP 2024 (RAG)

- What is the purpose of multiple tasks?

An Alternative Propose: Stratified Sampling

- Like many benchmarks in our field,
  - SPEC is a collection of tasks
  - Intended to measure different needs
- Some tasks are limited by cycles
  - and some are limited by memory
- Mashey: *Summarizing performance is no mean feat.*
  - https://ieeexplore.ieee.org/document/1525995
    - geometric means >> arithmetic means
  - As a single metric
    - for different customers
    - with different workloads
- So too, our benchmarks should be designed
  - to address different workloads

The Institute for Experiential AI
Northeastern University

Additional Considerations

- Reliability and Validity
  - Reliability is about data
  - Validity presupposes a hypothesis
    - e.g., SPEC scores are leading indicators of your experience with your workloads
    - *Your mileage may vary*
- Constructive Feedback
  - An overall score is not enough
  - Benchmark should produce scores for each strata to help all parties appreciate strengths and weaknesses

Conclusions

- Goal: Reusable Benchmarks
- Suggestion: Stratified Sampling
- Desiderata
  - Reliability
  - Validity
    - Populations, Workloads, Use Cases
    - With sufficient flexibility to cover
      - Known Variability, as well as
      - Unknown Variability (future-proof)
- Benefits of studying case law (from other fields)
  - SPEC has had more impact than
    - benchmarks in our field
  - Computer Industry depends on
    - customers making informed decisions
  - Bubbles (AI winters/booms and busts):
    - consequence of less informed decisions

References

- Machey's: *Summarizing performance is no mean feat*
- Machey's presentation at ACL-2021