Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

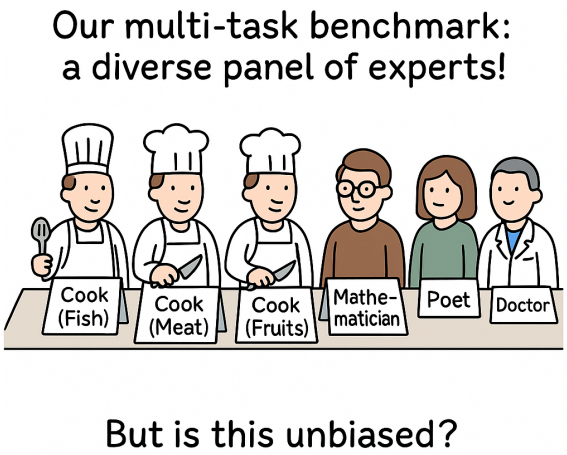
Distributed Computing

Prof. R. Wattenhofer

Measuring Task Similarity for Unbiased Benchmark Aggregation

You're hosting a "Best Chef" contest across all world cuisines. But it turns out that 60% of the judging criteria are just different takes on "making pasta". Guess who's walking away with the trophy? The Italian specialist, of course!

That's exactly the kind of hidden bias shaping some of our most popular AI benchmarks. Benchmarks like GLUE and SuperGLUE assume each task contributes equally to the final score. Yet, it is well recognized that some tasks, e.g., CoLA (linguistic acceptability) and SST-2 (sentiment analysis), are measuring nearly identical capabilities. In such a case, adding "similar" tasks doesn't improve evaluation – it amplifies some skills while drowning out others. This issue becomes of utmost importance as benchmarks carry higher stakes, shaping research priorities, funding allocations, and decisions about model deployment.



In this project, you will explore how similar benchmark tasks really are, and what happens when we account for that similarity. The goal is to develop ways to measure task distances and apply recently introduced clone-proof weighting functions that ensure similar tasks share their influence rather than amplify each other. You will then analyze how these adjustments reshape benchmark scores and model rankings. The outcome is both theoretical – understanding task similarity – and empirical – revealing its impact on popular benchmarks.

Requirements: Strong motivation, ability to work independently, and interest in theoretical and empirical aspects of evaluation in ML. Solid background in machine learning and statistics. Good programming skills in Python. Coming up with your own ideas is highly appreciated.

Interested? Please reach out for more details!
Describe your motivation for the project in a few sentences, along with any relevant courses or prior projects (personal or academic).

- Damien Berriaud: dberriaud@ethz.ch, ETZ G61.3