

# REPRODUCIBILITY BY DESIGN

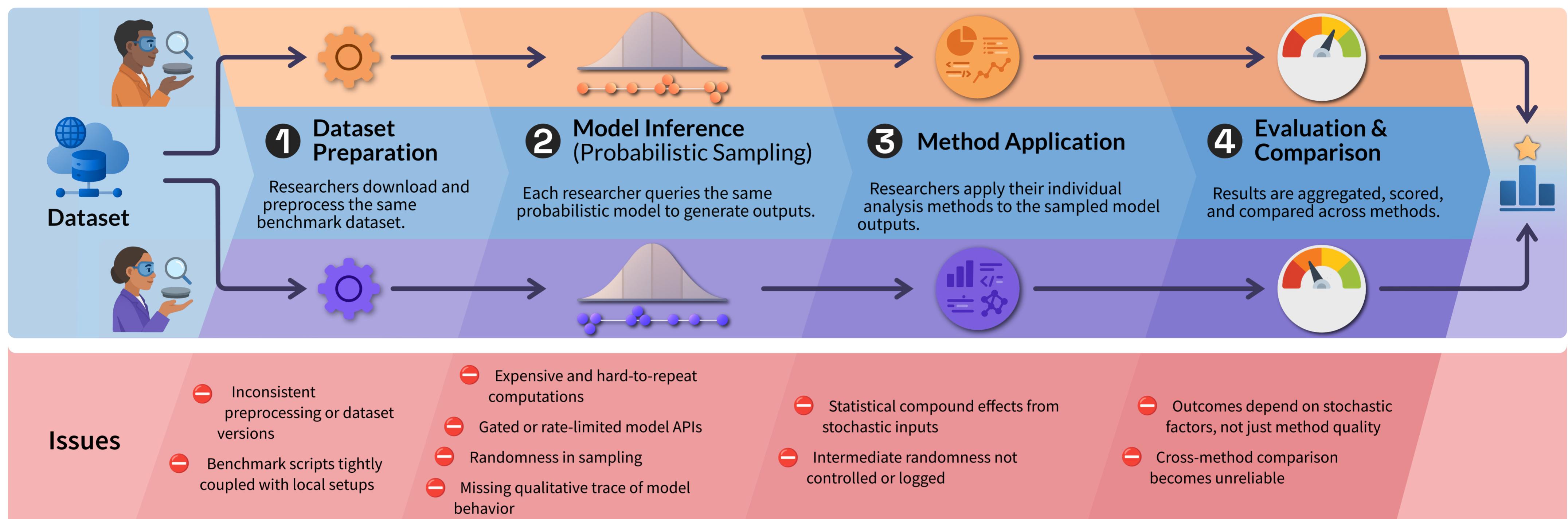
## A Modular Framework for Benchmarking Evolving Probabilistic AI Systems

Philip Müller<sup>1</sup>, Peter Steinbach<sup>1</sup>

<sup>1</sup> Helmholtz-Zentrum Dresden-Rossendorf

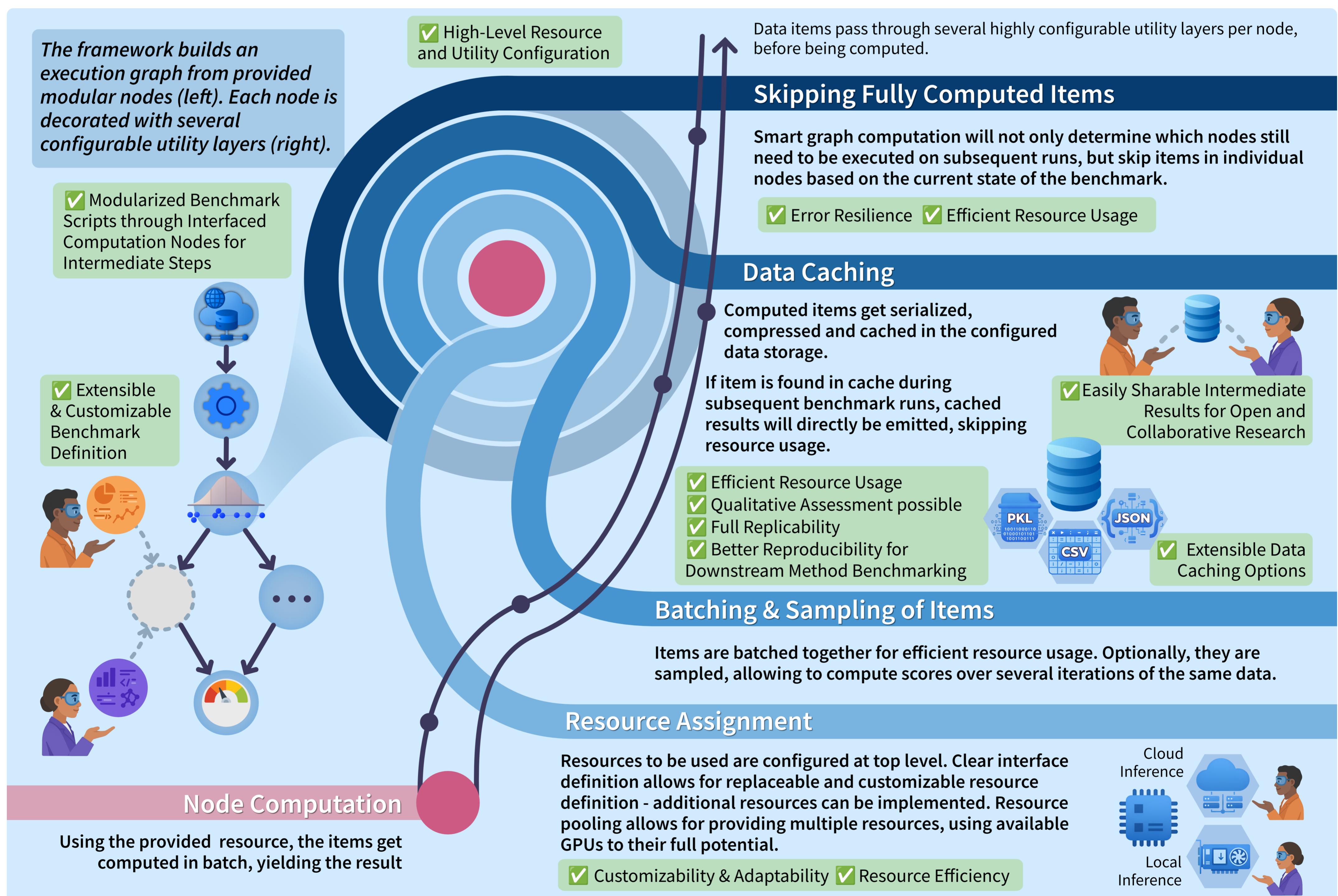
### PROBLEM STATEMENT

#### Benchmarking Pipelines Using Probabilistic Models



### SOLUTION

#### Reproducibility-Centered Framework for Probabilistic Model Benchmarking



### IMPACT STATEMENT

#### Advancing Reproducible and Transparent Benchmarking

This work strengthens trustworthy AI evaluation by enabling reproducible benchmarking for probabilistic and resource-intensive models. By decoupling benchmark logic from model execution and caching stochastic outputs, it ensures consistent comparisons across evolving systems while reducing computational cost. The framework supports longitudinal studies, shared intermediate results, and open collaboration, offering a scalable and transparent foundation for benchmarking in rapidly changing AI environments.

The Framework will be released at the end of January — star to get notified!

