SeedAI: Sustainable Data and Energy Efficient AI Model Training Framework

*Srihith Chennareddy[1], Vamshi Krishna Appala[2]

1. BHS USA, 2. Microsoft USA

Training large-scale AI models, including Large Language Models (LLMs) and Generative AI (GenAI) systems, requires substantial computational resources and extensive datasets, leading to high energy consumption and huge costs. To address this, we propose SeedAI, a sustainable data and energy efficient AI model training framework that uses stratified adaptive sampling to identify an optimal yet highly impactful subset of training data. By dividing datasets into meaningful strata and dynamically prioritizing the most important and unique samples, SeedAI significantly reduces the size of the training dataset required for AI models. We also integrate a self-adaptive test suite to ensure comprehensive evaluation of model behaviour throughout training. Our experimental results show that SeedAI can reduce the size of model training datasets by up to 62% without impacting model accuracy or performance. By enabling more efficient use of energy, computational, and storage resources, this framework offers a way toward sustainable, scalable, and cost-effective AI model training. SeedAI is adaptable across a wide range of AI model training use-cases, supporting environmentally conscious AI model development.

Keywords: sustainable AI, data and energy efficient AI, stratified adaptive sampling, scalable AI model training