Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260214
Published: June 9, 2026

A Multi-stage Energy Optimization Framework for Large Language Models Under Uncertainty

Yara A. Asiri, Abdulelah A. Alathwan, Jana S. Alsaqrah, Feras A. Al Omar and Mohammed M. Alqahtani
Department of Industrial Engineering
College of Engineering, King Khalid University Abha, Saudi Arabia
444807586@kku.edu.sa, 444801249@kku.edu.sa, 445803002@kku.edu.sa,
445812558@kku.edu.sa, m.alqahtani@kku.edu.sa

Abstract

The rapid expansion of Large Language Models (LLMs) has introduced significant environmental challenges, as inference emissions now surpass training costs due to the scale of daily operational requests. Global data indicates that a single Artificial intelligence (AI) query produces ten times more carbon emissionscompared toa traditional web browsing, and every 10–50 prompts consume approximately 500ml of water for service ventilation and cooling. This study proposes a Multi-stage Optimization (MSO)framework to address those challenges under usage uncertainties. In the first stage, we employ Response Surface Methodology (RSM) to analyze the impact of multiple factors including model version, prompt length, and response quantity on total carbon output. In the second stage, we developed a decision-making policy to switch between Large Language Model (LLM) and Small Language Model(SLM) using Stochastic Programming (SP) for optimal operation of the system. Our experimental results demonstrate that the model version is the most critical factor ($P < 0.001$), with a significant interaction identified between prompt length and the number of responses. While unmanaged request distributions showed high variability in emissions, our optimization model achieved a minimized emission level of 2.0515 (e-05) by utilizing optimal settings of operational parameters. Running the system under the optimal settings can achieve up to 45% reduction in energy consumption (saving 450 kWh per cycle) and reductions up to 45% of carbon emissions. These findings offer a quantitative roadmap for green and sustainable AI deployment, that balances computational performance with ecological responsibility.

Keywords
Energy Optimization, Large Language Models, Uncertainty