A Comparative Analysis of the Energy Demands and Runtime of Large Language Models when Deployed on User-Side Devices

H.U. Ahamed, A.L.A.D.T.T. Aturusinghe, D.M.K.N. Dewmini, M.G.S.K. Jayarathne*

Sri Lanka Institute of Information Technology (SLIIT) Kandy Uni, Kandy, 20000, Sri Lanka
**sonalijayarathne@outlook.com*

Large Language Models (LLMs) are statistical algorithms trained on massive amounts of data to predict the most appropriate response to a user prompt. LLMs requires a substantial supply of resources to function and despite the increase in the number of LLMs locally deployed, studies about LLM energy consumption are limited to those run at data centers, with research on locally deployed LLM energy consumption lacking. This study aims to bridge that gap by providing important information on the energy consumption and runtime of LLMs deployed on user-side devices via a comparison of the energy consumption and runtime efficiency of selected locally deployed LLMs across different prompt categories on consumer-grade hardware. For this task, a personal laptop having 16.0GB RAM and an Intel(R) UHD Graphics GPU was configured with Ollama for local LLM deployment and HWiNFO64 for hardware monitoring. To ensure the relevancy of the research, the study uses the 5 prompt categories most used in locally deployed LLMs (Factual, Reasoning, Long Generation, Code Generation and summarization), however future studies will use more novel prompts with increasing complexity. For proper evaluation we chose models developed by different companies (DeepSeek-R1(7B), Gemma 3(4B), and Phi-4 Mini (3.8B)). Using Ollama, each model was run 45 times, with 3 tests per prompt. During each test, HWiNFO64 recorded average power use and runtime, and these average values were compared across models and categories. The findings indicate that throughout the test cases, an overall average of 23.34 Watts (W) was consumed by all three LLMs at a mean runtime of 37.81 seconds. Furthermore, results reveal that DeepSeek-R1 (7B) had the greatest power demand with an average of 29.89 W, while Gemma 3(4B) recording the lowest demand for power with an average of 18.72 W needed. The results denote that there exists a relationship between model architecture, prompt category and PC system specifications that affects energy usage. By elaborating on this relationship and offering key insights, this study provides opportunities for future LLM development in consideration of user-device capabilities and to propose a reproducible LLM benchmark for consumer grade hardware to support future research.

Keywords: large language models, energy consumption, client-side LLM deployment, runtime efficiency, energy efficiency, model runtime