現在、生成 AI サービスの普及に伴い、GPU などの処理リソース不足や電力需要の増大が深刻な課題となっている。本研究では、私たちの研究グループが提案するリニアアレイ型 Coarse-Grained Reconfigurable Architecture (CGLA)の IMAX3 プロトタイプを用い、Large Language Models (LLMs)の処理性能と電力効率を評価することを目的とした。IMAX3 は Field Programmable Gate Array (FPGA)上に実装されており、CPU などの他計算基盤と比較して、処理速度と電力効率に関する有効性を検証した。本研究の主な貢献は以下の 3 点である。第一に、IMAX3 の浮動小数点演算器に 4bit 整数から単精度浮動小数点への変換テーブルを追加し、LLM ライブラリ GGML の動作を可能にした。第二に、GGML を後継ライブラリである llama.cpp に変更し、それに伴う IMAX の改良を実施した。これにより、GPU と比較した処理性能と電力効率の評価が可能となった。第三に、縮小したモデルの量子化サイズに対応し、推論速度を従来比で 10 倍以上に向上させた。これらの結果、CGLA が LLM の計算基盤として高い潜在能力を持つことを示した。将来的には、大規模化した CGLA を実現し、生成 AI を含む多様なサービスにおいて汎用かつ高効率な計算基盤を提供する可能性が期待される。