Energy-Aware Quantization for LLMs

Taarana Jammula, Krishna Karthikeya Chemudupati, Thomas Ngulube
tjammula@seas.upenn.edu, krishkc@seas.upenn.edu, tngulube@seas.upenn.edu
University of Pennsylvania, 12/8/25

Problem & Motivation

Motivation

- Transformer models (DistilBERT, GPT-2) are computationally expensive and power-hungry.
- Quantization is often used to reduce compute and model size, but its energy impact is poorly understood.
- Edge devices and datacenters care about energy per inference, not just speed.

Problem

- Existing research focuses on accuracy and latency, but rarely measures energy consumption across quantization formats. LLM inference dominates real-world compute and energy costs

Goal

- Evaluate energy, performance, and quality trade-offs across:
- FP32 (baseline)
- FP16 (native Tensor Core accelerated)
- Mixed precision and BF16 (ultimately dropped for limitations)

Model

Models

- DistilBERT — Sentiment classification (SST-2)
- GPT-2 Small — Next-token prediction (WikiText-2)

Quantization Formats Tested

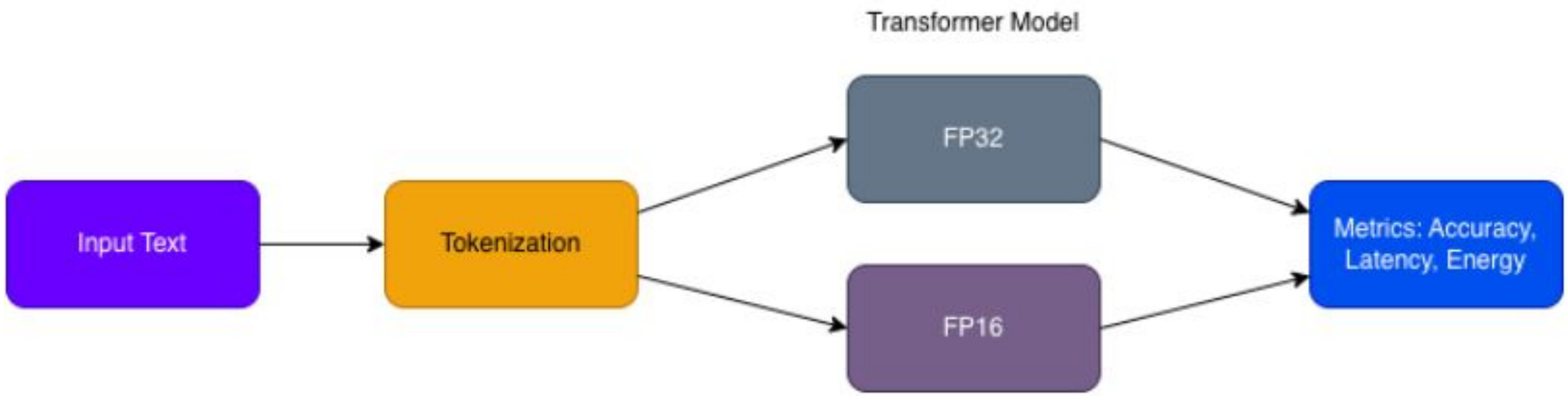
- FP32 baseline
- FP16 (full-model conversion)
- Mixed precision (initially tested; removed — training-only benefits)
- BF16 (removed — unsupported on T4 GPUs)

Model

Measurement Pipeline

- Zero-I/O tokenized dataset (preprocessed offline)
- Warmup passes to stabilize GPU clocks
- Multi-trial measurement (5 trials × 300 iterations)
- Real-time power sampling using nvidia-smi
- Accuracy, latency, throughput, energy/inference aggregated
- Energy measured using board-level nvidia-smi sampling (~20 Hz).
    - Uses mean power x runtime; captures stable load but not sub-kernel spikes."

Model


Transformer Model
FP32
Input Text
Tokenization
Metrics: Accuracy, Latency, Energy
FP16

Previous approaches

What Prior Work Has Focused On

- Precision reduction for speed or throughput
- Quantization for accuracy retention
- Mobile/edge INT8 optimization
- Model compression literature
- Energy studies ≠ precision studies

Limitations of Previous Work

- Rarely include end-to-end GPU energy measurements
- Lack controlled, I/O-free evaluation
- Often exclude transformers, using CNNs instead
- Few studies examine both DistilBERT and GPT-2 across multiple precisions
- No systematic comparison on general-purpose GPUs like Tesla T4

Contributions of this work

1. First controlled, end-to-end energy evaluation of FP32 vs FP16 for transformer inference on general-purpose GPUs (T4).
2. Full evaluation across both encoder and decoder architectures
3. Energy, accuracy, latency, model size analyzed together
4. Per-layer energy profiling
5. Practical guidelines for GPU quantization

Measurement Assumptions & Limitations

- GPU power draw ≈ nvidia-smi averages at 50 ms cadence.
- Single-GPU execution (GPU 0 only).
- CPU work, data loading, and host-side overhead are excluded from energy accounting.
- CUDA synchronizations enforce accurate timing.
- Warmup removes JIT and cache-population effects, so measured times ≈ true kernel runtimes.
- Per-layer energy is first-order only.
- Energy/sample assumes stable workload.

Implications:

- Good for relative comparisons (FP16 vs FP32), not absolute hardware-level power modeling.
- Board-level telemetry reflects total GPU subsystem power (VRAM, PCIe, memory controllers), not isolated MAC unit power.

Penn Engineering

Details of the contributions

1. Zero-I/O Dataset & Evaluation Pipeline

- All datasets pre-tokenized and saved as .pt tensors
- Loaded directly to GPU, no disk access during inference
- Ensures energy reflects **compute only**, not I/O

2. GPU-Level Energy Measurement Infrastructure

- Power sampled from NVIDIA's board-level sensors
- Warmup phase stabilizes GPU clocks
- PowerLogger runs asynchronous sampling at ~10 Hz
- 5× repeated trials per precision mode (300 iterations each)
- Energy computed as: $E = P_{avg} \times t$

- Power samples represent total board power; memory controllers + background GPU subsystems are included
- Sampling at 50 ms cadence approximates stable average power; short transients are smoothed

Penn Engineering

Details of the contributions

3. Full-Model Precision Conversion

- FP32 → FP16 via .half()
- Verified stability (no overflow/NaN)
- Mixed precision removed after validation showed training-only benefits
- BF16 unsupported on T4 → removed
- INT8 attempted → fallback to FP32 kernels on T4

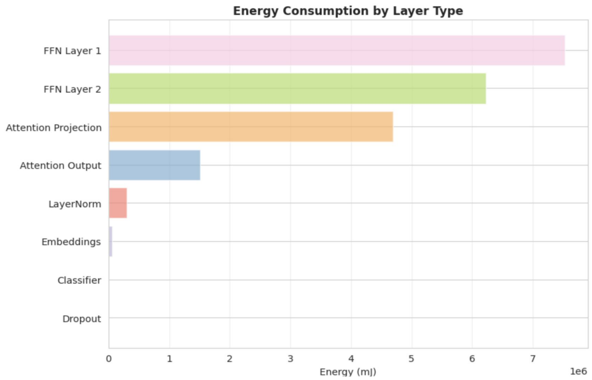
4. Per-Layer Energy Profiling

- Timed forward pass with per-module hooks
- Energy per layer computed from time share × mean GPU power
  - (first-order estimate; assumes near-constant power across layers)

Per-Layer Energy Profiling

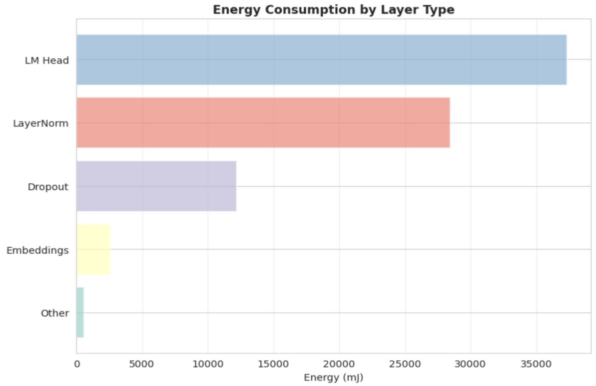
Distilbert (Encoder-Only)

- FFN layers dominate energy but have low prediction impact
- Attention layers: moderate energy, low impact
- Embeddings + LayerNorm: low energy, high importance
- Quantization opportunity: FFN + attention
- Keep high precision: embeddings, LayerNorm, classifier



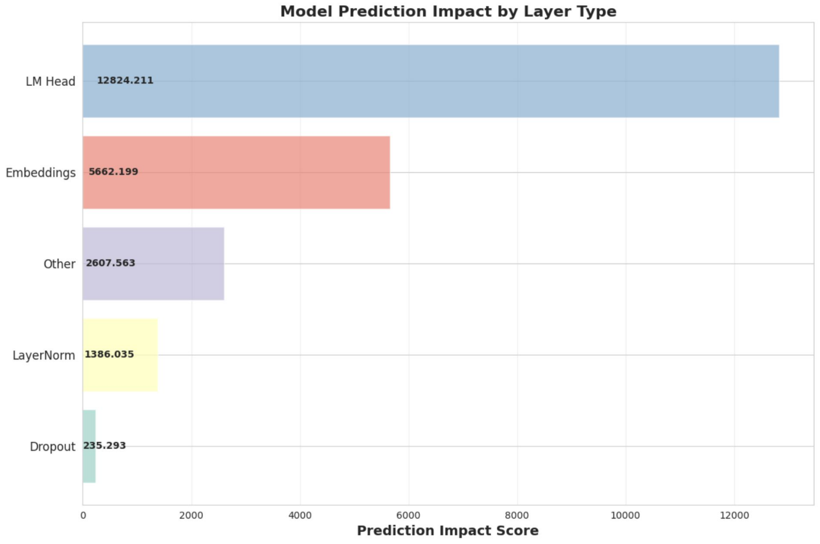
GPT-2 (Decoder-Only)

- LM Head = highest energy and highest impact
- Embeddings + LayerNorm: also high impact
- Dropout/residual: low energy, low impact
- Quantization opportunity: only internal attention/MLP layers
- Avoid quantizing: LM Head + embeddings

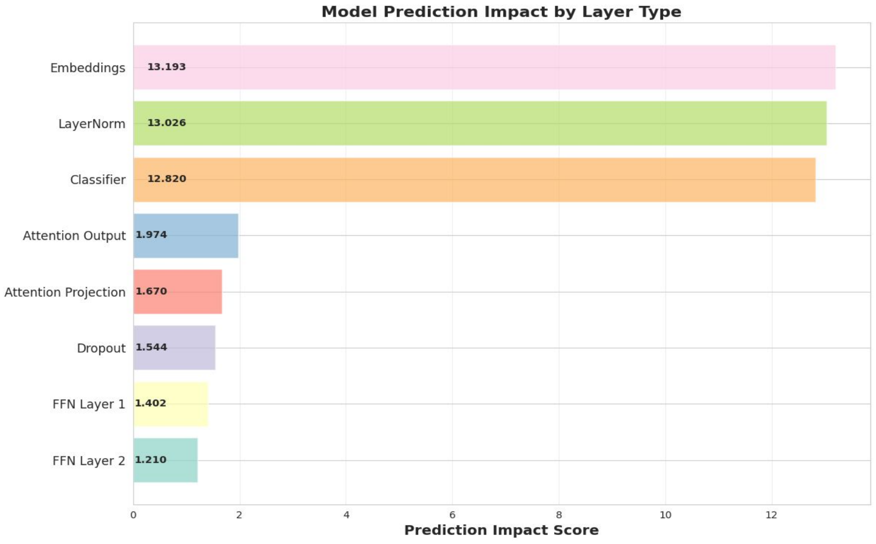


Penn Engineering

Per-Layer Prediction Impact Profiling



GPT-2



Distilbert

Penn Engineering

Results and Analysis **DistilBERT**

Task

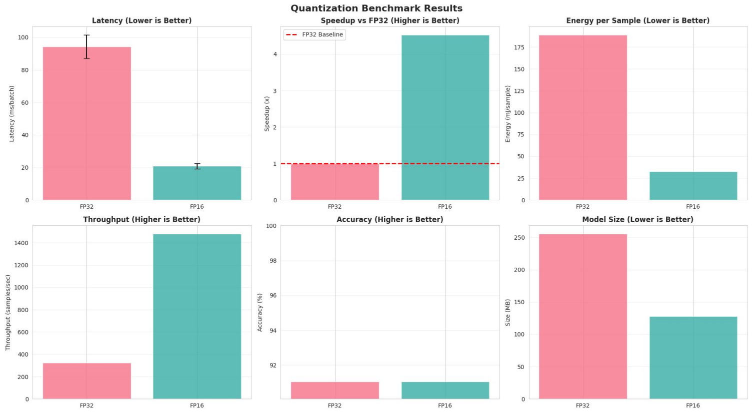
- SST-2 sentiment classification (50-sample evaluation set)

Formats Tested

- FP32 → FP16

Key Findings

- Energy reduced by ~5.4%
- Latency improved by ~5%
- Model size reduced by 2×
- No accuracy drop → DistilBERT is numerically stable in FP16
- FP16 benefits from **T4 Tensor Cores**



Format	Latency(ms/batch)	Throughput	Energy/sample	Accuracy	Model Size
FP32	~94 ms	~324.29/s	190 mJ	91.06%	255.41 MB
FP16	~20 ms	~1480.05/s	32.66 mJ	91.06%	127.71 MB

*Energy/sample computed as mean GPU power × runtime ÷ batch_size; assumes stable average power across iterations

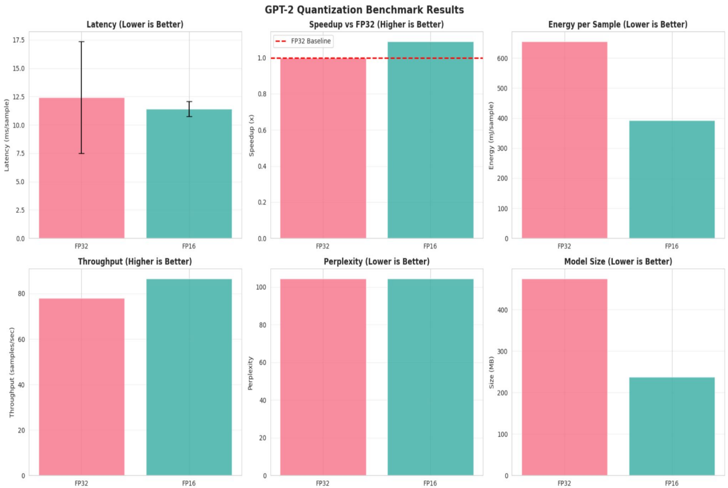
Results and Analysis **GPT2**

Task

- SST-2 sentiment classification (50-sample evaluation set)

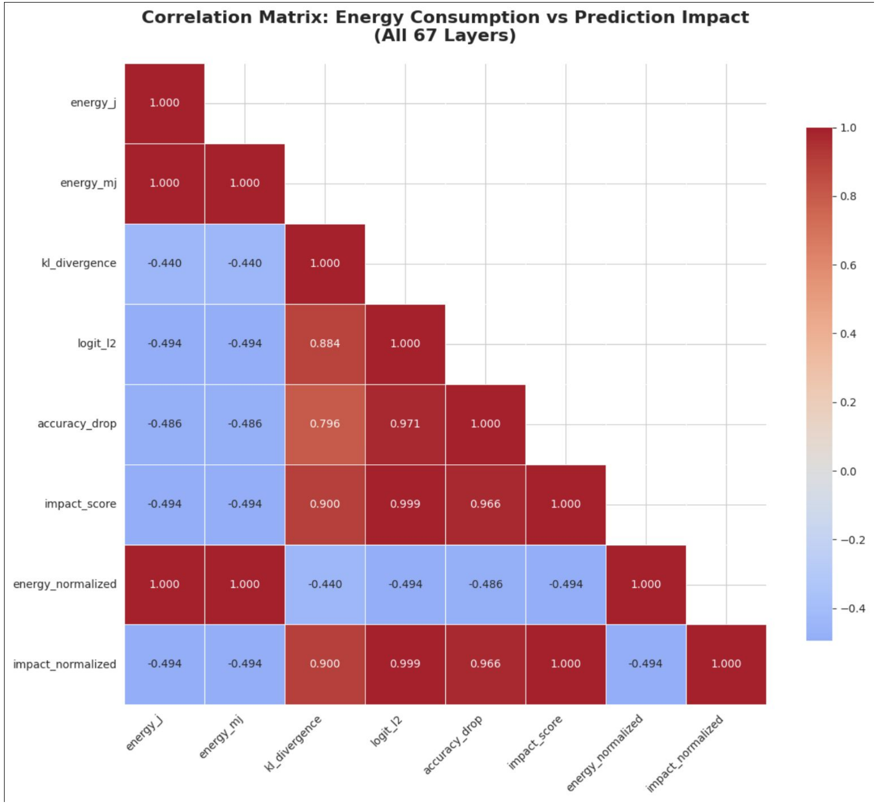
Key Findings

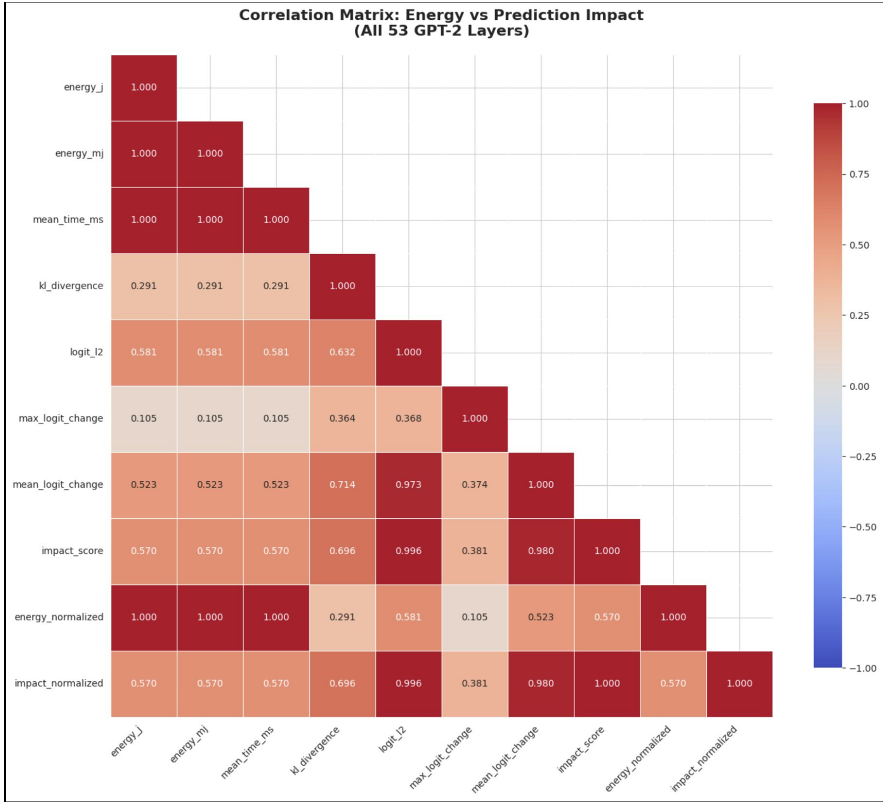
- **36% energy reduction** going to FP16
- **2× model size reduction**
- **No quality loss** (perplexity stable at 217)
- Mixed precision worse due to added overhead
- Strong FP16 acceleration due to **Tensor Cores**



Format	Latency	Speedup	Energy/sample	Perplexity	Model Size
FP32	12.4 ms	1.0x	650 mJ	105	480 MB
FP16	11.4 ms	1.09x	390 mJ	105	240 MB

Correlation Matrices





Penn Engineering

Conclusions

1. FP16 Produces Real Energy Savings

- DistilBERT: FP16 gives 5.9× energy reduction, 4.7× latency speedup
- GPT-2: FP16 gives 1.67× energy reduction, 1.09× latency speedup
- Zero accuracy degradation
- 2× smaller model size → deployability benefits
- Energy results are robust because average GPU power remains stable under steady load even though nvidia-smi cannot resolve short spikes.

2. Hardware Matters

- T4 GPU has Tensor Cores → FP16 acceleration
- BF16 unsupported → removed
- INT8 requires TensorRT + engine conversion → too complex for course scope
- Board-level power methods capture holistic GPU load but cannot attribute power to specific kernels

3. Mixed Precision is Not Useful for Inference

- Introduces overhead from casting + autocast
- Faster for training but slower for inference
- Recommendation: **Use full FP16 for inference workloads**

Shortcomings & Future Work

Limitations

- Small Evaluation Sets
- Restricted Quantization Formats
- Single GPU Platform
- Inference-Only Study
- Board-level telemetry has low temporal resolution; cannot isolate per-kernel or per-layer power directly.
- Per-layer energy is time-weighted approximation, not hardware-measured power.

Future Work

- INT8 Deployment using TensorRT
- Test on Multiple GPUs
- Selective / Per-Layer Quantization
- 4-bit Quantization (QLoRA-style)
- Larger Evaluation Sets
- Real-World Deployment Benchmarks