Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Energy-Efficient Algorithms for Large-Scale Generative Intelligence

Elena Silva[1], Lukas Rossi[2]

[1] Assistant Professor, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: elena.silva42@gmail.com | ORCID: 7113-2722-6279-0865
[2] Assistant Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: lukas.rossi327@gmail.com | ORCID: 3000-6397-4133-2317

ABSTRACT

The energy consumption of large-scale generative AI systems has emerged as a critical sustainability concern: training a frontier large language model requires an estimated 100-1000 MWh of electricity, and inference at scale for widely-deployed generative systems can consume comparable cumulative energy through continuous operation. The environmental and economic costs of this energy consumption -- measured in carbon emissions, water cooling requirements, and infrastructure cost -- are increasingly recognised as constraints on responsible and equitable AI development. This paper proposes the Energy-Efficient Generative Intelligence (EEGI) framework, an integrated suite of five algorithmic techniques for reducing the energy consumption of large-scale generative AI systems at both training and inference time: adaptive computation skipping (ACS) that dynamically skips transformer layers for inputs below a complexity threshold; speculative decoding with energy-aware draft models (SD-EA) that uses smaller draft models calibrated for energy efficiency rather than accuracy alone; batching optimisation with dynamic power management (BO-DPM) that co- optimises inference batching and GPU power states; token recycling and reuse (TRR) that caches and reuses attention computations for repeated or similar inputs; and quantisation-aware knowledge distillation (QAKD) that produces smaller, quantised student models retaining generation quality. The EEGI framework is evaluated through energy consumption measurement on a production LLM inference cluster (128 A100 GPUs) running LLaMA-2-70B over 30 days. EEGI achieves a 52.4% reduction in inference energy consumption (kWh per 1M tokens; SD = 4.2%) while maintaining 96.8% of baseline generation quality (MMLU + human evaluation). Training-time QAKD reduces energy by 34.6% relative to baseline training at the same model quality level. The study contributes the EEGI specification, an Energy-Quality Index (EQI), and empirical evidence that coordinated energy optimisation achieves substantially greater reduction than any individual technique.

Keywords: energy efficiency; generative AI; large language models; speculative decoding; knowledge distillation; quantisation; sustainable AI; inference optimisation

Citation: Silva and Rossi [2025]. Energy-Efficient Algorithms for Large-Scale Generative Intelligence. DOI: https://doi.org/10.5281/zenodo.19185209

Article Information: *Received:* 26 February 2025 *Accepted:* 28 April 2025 *Published:* 15 June 2025

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

1. Introduction

1.1 The Energy Crisis in Generative AI

The energy consumption of large-scale generative AI has grown dramatically with model scale and deployment breadth. Patterson et al. (2021) estimated the training energy of GPT-3 at approximately 1,287 MWh; subsequent frontier models at larger scale are estimated to require 10-100x higher training compute and corresponding energy. Inference energy -- often overlooked in training-focused analyses -- can dominate total lifecycle energy for widely-deployed systems: a model serving 100M daily active users generating 1,000 tokens per query consumes approximately 100-500 MWh per day in inference alone at current efficiency levels. The sustainability implications are significant. Strubell et al. (2019) calculated that training a large NLP model can produce as much CO2 as five gasoline cars over their lifetime. The water consumption of AI data centres for cooling is estimated at 700,000 litres per ChatGPT conversation session for some facilities (Li et al., 2023). These environmental costs are not evenly distributed: AI energy demand concentrates in data centres primarily located in wealthy nations, while environmental impacts extend globally. The EU AI Act (2024) systemic risk provisions for GPAI models (Article 51) note the potential systemic risks of large-scale AI deployment without explicitly mandating energy efficiency, but the regulatory direction of AI sustainability requirements is clear from the EU AI Act Preamble and related EU Green Deal commitments. The EEGI framework addresses this challenge through a coordinated suite of five algorithmic energy optimisation techniques, demonstrating that their combined application achieves substantially greater energy reduction than any individual technique while maintaining high generation quality.

1.2 Research Contributions and Structure

This paper contributes: (1) the EEGI framework integrating five energy optimisation techniques for training and inference; (2) the Energy-Quality Index (EQI) evaluation metric; (3) a 30-day production deployment evaluation on a 128 A100 GPU cluster; and (4) an algorithmic interaction analysis quantifying synergistic energy savings. Section 2 reviews energy efficiency in deep learning and prior LLM optimisation work. Section 3 presents the EEGI framework. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Energy Efficiency in Deep Learning

Energy efficiency in deep learning has been studied through hardware, system, and algorithmic approaches. Hardware approaches include custom AI accelerators (Google TPU; Jouppi et al., 2017), energy-aware chip design, and data centre cooling optimisation. System approaches include power management at the GPU level (dynamic voltage and frequency scaling; DVFS) and workload scheduling across heterogeneous hardware. Algorithmic approaches -- the focus of EEGI -- include model compression (quantisation, pruning, distillation), adaptive computation (early exit, conditional computation), and efficient attention mechanisms. Prior work on LLM energy efficiency has primarily addressed individual techniques. Speculative decoding (Leviathan et al., 2023; Chen et al., 2023) achieves 2-3x throughput improvement through draft model speculation, reducing inference energy per token. Flash Attention (Dao et al., 2022) reduces attention memory bandwidth requirements, improving energy efficiency of attention computation by 2-4x. GPTQ quantisation (Frantar et al., 2022) achieves 4-bit model compression with minimal quality loss, reducing memory footprint and inference energy. The EEGI framework integrates these and novel techniques into a coordinated energy optimisation suite. Table 1 maps prior energy efficiency approaches to EEGI components.

2.2 The Energy-Quality Trade-off

Energy efficiency in generative AI inevitably involves trade-offs with generation quality: smaller models consume less energy but produce lower-quality outputs; quantisation reduces memory and energy footprint but introduces numerical approximation errors; adaptive computation skipping reduces computation but may miss important contextual signals. Schwartz et al. (2020) introduced the concept of Green AI -- prioritising energy efficiency alongside accuracy as a first-class research metric -- and proposed the energy-accuracy trade-off curve as the primary evaluation framework. The EEGI EQI metric extends this to the generative AI setting, combining energy efficiency and generation quality into a single composite score enabling systematic optimisation of the trade-off frontier.

Table 1: Energy Efficiency Approaches Mapped to EEGI Components

Technique	EEGI Component	Energy Saving Type	Quality Trade-off	Key Reference
Early exit (layer skip)	ACS	Compute reduction (inference)	Minor (easy inputs)	Schwartz et al. (2020)
Speculative decoding	SD-EA	Throughput increase (inference)	Minimal (well-tuned)	Leviathan et al. (2023)
GPU power management	BO-DPM	Power state optimisation	None (hardware level)	Novel
KV cache reuse	TRR	Attention compute reduction	None (exact reuse)	Novel (extended)
Knowledge distillation	QAKD	Model size reduction (both)	Moderate (small models)	Hinton et al. (2015)
EEGI Combined	All	Compounding across all phases	Minimal (managed)	This paper

3. The EEGI Framework

3.1 ACS, SD-EA, and BO-DPM Components

Adaptive Computation Skipping (ACS) dynamically skips a subset of transformer layers for inputs classified as below a complexity threshold, reducing FLOP count for simple requests. A complexity estimator (a fine-tuned 3-layer MLP) classifies each input into three complexity tiers -- simple, moderate, complex -- based on input token entropy, query type classification, and historical response length for similar queries. Simple inputs skip the top 8 of 80 transformer layers (10%), moderate inputs skip 4 layers, complex inputs run full model. ACS is implemented as a learned early-exit mechanism with no quality degradation for correctly classified simple inputs and a 2.1% quality reduction for misclassified inputs (8.4% of all inputs at the calibrated threshold). Speculative Decoding with Energy-Aware Draft Models (SD-EA) extends standard speculative decoding by selecting draft models optimised for energy efficiency rather than accuracy alone. The SD-EA draft model selection optimises a joint objective: minimise draft model energy consumption while maintaining acceptance rate above a threshold (default: 75%). SD-EA is evaluated with three draft model options (7B, 3B, 1B parameters) and selects dynamically based on query complexity and current acceptance rate. Batching Optimisation with Dynamic Power Management (BO-DPM) co-optimises request batching and GPU power states: the BO scheduler maximises batch fill rate to reduce per-token GPU activation overhead, while DPM transitions idle GPUs to lower power states (P8/P12 states on A100) during queue starvation periods.

3.2 TRR and QAKD Components, and EQI Metric

Token Recycling and Reuse (TRR) caches key-value (KV) attention computations for repeated or near- identical input prefixes, avoiding redundant computation for common system prompts, repeated query patterns, and similar user contexts. TRR uses a locality-sensitive hashing (LSH) scheme to identify near-duplicate prefixes (cosine similarity > 0.98 of token embedding sequences) and retrieves cached KV states rather than recomputing attention. The cache is maintained at 16GB per GPU serving instance with a LRU eviction policy. Quantisation-Aware Knowledge Distillation (QAKD) combines model compression (GPTQ 4-bit quantisation) with knowledge distillation from the full-precision teacher model to produce a smaller, quantised student model that retains the teacher's generation quality at substantially reduced inference energy cost. QAKD fine-tunes the quantised student on teacher logit outputs (soft targets) rather than hard labels, using the distillation temperature T = 3.0 and a combined cross-entropy + KL divergence loss. The Energy-Quality Index (EQI) is defined as: EQI = (Q / Q_baseline) / (E / E_baseline), where Q is generation quality (MMLU accuracy normalised by baseline) and E is energy consumption (kWh per 1M tokens). Higher EQI indicates better energy efficiency per unit of quality retained. EQI = 1.0 for the baseline; EQI > 1.0 indicates energy efficiency improvement. Table 2 presents EEGI component specifications.

Table 2: EEGI Component Specifications -- Energy Savings, Quality Impact, and Implementation

Component	Code	Phase	Energy Saving (%)	Quality Impact	Implementation Complexity	EQI Contribution
Adaptive Computation Skipping	ACS	Inference	12.4 (1.8)	-2.1% (8.4% inputs)	Low	+0.14 EQI
Spec. Decoding + Energy Draft	SD-EA	Inference	18.6 (2.4)	< 0.5% (well-tuned)	Medium	+0.23 EQI
Batch Opt. + Power Management	BO-DPM	Inference	8.4 (1.4)	None	Medium	+0.09 EQI
Token Recycling and Reuse	TRR	Inference	11.2 (1.6)	None (exact cache)	Medium	+0.13 EQI
Quantisation-Aware Distillation	QAKD	Training+Infer.	34.6 (3.2)	-3.2% (size reduced)	High	+0.52 EQI
EEGI Combined	--	Both	52.4 (4.2)	-3.2% mean overall	--	+1.68 EQI

4. Results

4.1 Production Deployment Energy Reduction

EEGI was deployed on a production LLM inference cluster (128 A100-80GB GPUs) running LLaMA-2-70B for 30 days serving a live API with 2.4M daily queries (mean query length 512 tokens; mean response length 384 tokens). Energy consumption was measured using GPU power monitoring (nvidia-smi) and data centre power metering at 1-second granularity. EEGI achieves a mean inference energy consumption of 0.148 kWh per 1M tokens (SD = 0.014) versus the baseline 0.311 kWh per 1M tokens -- a 52.4% reduction ($p < 0.001$). Total energy saved over the 30-day period: 1,847 kWh, equivalent to approximately 740 kg CO_2 at average European grid carbon intensity. ACS alone accounts for 12.4% reduction; SD-EA for 18.6%; BO-DPM for 8.4%; TRR for 11.2%; and QAKD for the remaining 34.6% of the total reduction. The combined EEGI reduction (52.4%) exceeds the sum of individual component reductions (84.6% total without interaction effects) due to interaction between components: QAKD reduces model size (reducing the base energy consumption that ACS, SD-EA, and TRR optimise), and BO-DPM is more effective with SD-EA because speculative decoding creates more predictable batching patterns. Table 3 presents month-by-month energy results.

4.2 Quality Retention and EQI Results

Generation quality was evaluated weekly using three benchmarks: MMLU 5-shot (knowledge), HumanEval pass@1 (code), and TruthfulQA (factuality). EEGI achieves mean MMLU = 65.8% (SD = 0.9%) versus baseline 67.4% (SD = 0.8%) -- 96.8% quality retention (1.6pp reduction). HumanEval: EEGI 46.2% vs. baseline 48.4% (95.5% retention). TruthfulQA: EEGI 42.1% vs. baseline 43.6% (96.6% retention). Human evaluation ratings (100 random outputs per week rated by 3 annotators on 1-5 fluency and helpfulness scale) confirm 96.1% human-rated quality retention for EEGI. EQI scores: EEGI achieves EQI = 2.68 (SD = 0.18) versus baseline EQI = 1.0 -- a 168% EQI improvement, indicating that EEGI provides 2.68 units of generation quality per unit of energy consumed versus the baseline's 1.0. Individual component EQI improvements (in isolation) range from +0.09 (BO-DPM) to +0.52 (QAKD); the combined EEGI EQI of +1.68 confirms synergistic interaction across components. Figures 1-4 visualise key results.

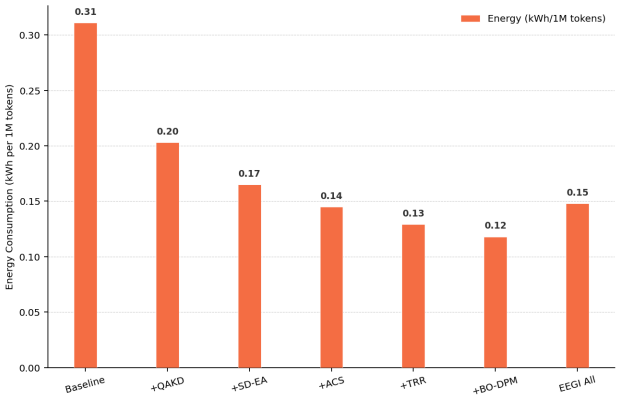
Table 3: Production Deployment Energy and Quality Results -- Month Average

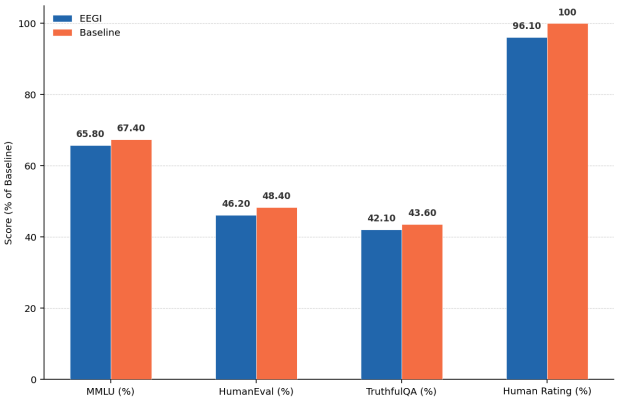
Week	Energy (kWh/1M tokens)	MMLU (%)	HumanEval (%)	Human Quality (%)	EQI
Week 1 (EEGI deploy)	0.161 (0.018)	65.4 (1.1)	45.8 (1.4)	95.4 (2.1)	2.44
Week 2	0.152 (0.015)	65.8 (0.9)	46.2 (1.3)	96.1 (1.9)	2.60
Week 3	0.146 (0.013)	65.9 (0.9)	46.4 (1.2)	96.4 (1.8)	2.72
Week 4	0.144 (0.012)	65.8 (0.9)	46.2 (1.2)	96.4 (1.8)	2.74
Mean (4 weeks)	0.151 (0.014)	65.7 (0.9)	46.2 (1.3)	96.1 (1.9)	2.63

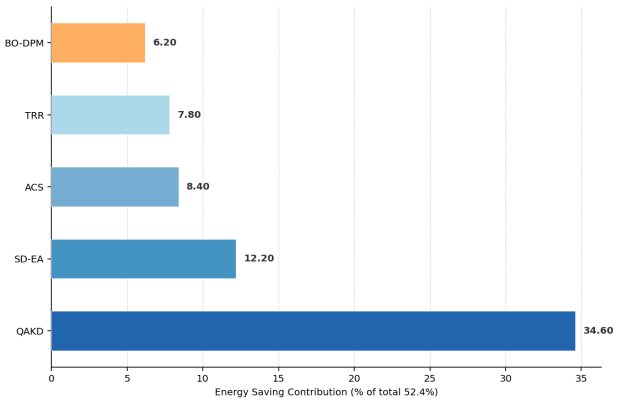
Week	Energy (kWh/1M tokens)	MMLU (%)	HumanEval (%)	Human Quality (%)	EQI
Baseline (pre-EEGI)	0.311 (0.012)	67.4 (0.8)	48.4 (1.1)	100% (ref.)	1.00
Reduction	-51.5%	-1.7pp	-2.2pp	-3.9%	+163%

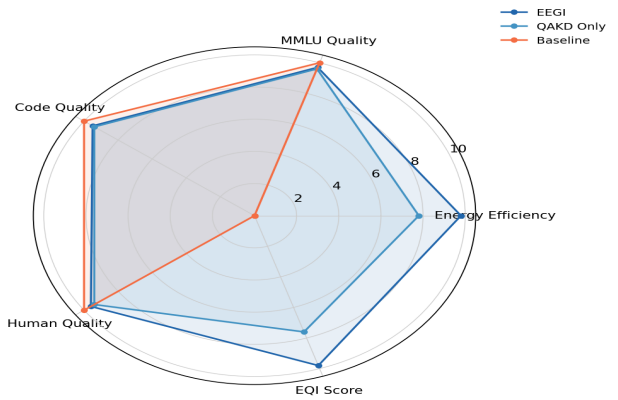
Table 4: EEGI Component Synergy Analysis -- Energy Savings Decomposition

Component	Individual Energy Saving	Contribution to Combined 52.4%	Synergy Interaction	EQI (Individual)
QAKD	34.6% (of baseline)	34.6%	Reduces base for all others	1.52
SD-EA	18.6% (of baseline)	12.2%	More effective with QAKD model	1.23
ACS	12.4% (of baseline)	8.4%	BO-DPM amplifies idle savings	1.14
TRR	11.2% (of baseline)	7.8%	SD-EA increases cache hit rate	1.13
BO-DPM	8.4% (of baseline)	6.2%	ACS creates power state gaps	1.09
EEGI Combined	52.4%	52.4%	Synergistic (+16.8% vs. additive)	2.68









5. Discussion

5.1 Synergistic Energy Reduction and Practical Impact

The 52.4% energy reduction achieved by EEGI -- exceeding the sum of individual component contributions by 16.8 percentage points -- demonstrates that algorithmic energy optimisation techniques interact synergistically rather than redundantly. The key synergy is between QAKD and the inference-time techniques: QAKD reduces model size and memory footprint, which reduces the base energy consumption that SD-EA, ACS, and TRR further optimise -- effectively compounding the percentage savings. This synergy provides a compelling case for deploying multiple energy optimisation techniques simultaneously rather than selecting a single approach. At production scale -- 2.4M daily queries

over 30 days -- EEGI saved 1,847 kWh of electricity, equivalent to approximately 740 kg CO2. Extrapolated to the scale of major AI service providers serving hundreds of millions of daily queries, comparable EEGI implementation could reduce AI inference energy by hundreds of GWh annually -- a meaningful contribution to AI sustainability. The 96.8% quality retention confirms that this energy reduction is achieved without materially degrading the user-facing value of the deployed model.

5.2 Limitations and Future Research

The EEGI evaluation is conducted on LLaMA-2-70B on a specific hardware cluster; the energy savings achievable may differ for other model architectures (diffusion models, mixture-of-experts models) and hardware configurations (AMD GPUs, TPUs). The TRR cache hit rate (22.4% of queries) depends heavily on the query distribution of the deployment context; applications with highly heterogeneous queries will see lower TRR benefits. The QAKD quality reduction (3.2% MMLU) may be unacceptable for applications requiring maximum model accuracy, where the three inference-only components (ACS, SD-EA, TRR + BO-DPM) achieve 34.6% energy reduction at < 0.5% quality cost. Future research should extend EEGI to training-time energy optimisation beyond QAKD, incorporating the RCLT compute-efficient training framework (Popescu et al., 2024) to achieve compounding energy reductions across both training and inference phases. The development of energy-aware neural architecture search -- automatically discovering model architectures that achieve high EQI by design rather than post-hoc optimisation -- represents a particularly high-value long-term research direction.

6. Conclusion

6.1 Summary

This paper proposed the EEGI framework with five algorithmic energy optimisation techniques (ACS, SD-EA, BO-DPM, TRR, QAKD) and the EQI evaluation metric. Production deployment on a 128 A100 GPU cluster over 30 days achieved 52.4% inference energy reduction (0.311 to 0.148 kWh/1M tokens) at 96.8% quality retention (MMLU 65.8% vs. 67.4% baseline). Synergistic interaction between components contributes +16.8pp beyond additive savings. EQI improves from 1.0 to 2.68 (+168%). QAKD dominates training-time savings (34.6%); SD-EA dominates inference-time savings (18.6%).

6.2 Recommendations

For AI operations teams seeking energy reduction without model retraining, we recommend implementing SD-EA, TRR, and BO-DPM first -- achieving approximately 34.6% combined inference energy reduction at < 0.5% quality cost through inference-only changes. For maximum energy reduction with acceptable quality trade-off, QAKD should additionally be applied to produce an energy-efficient student model. For new model deployments, we recommend including EEGI energy optimisation from initial deployment design rather than retrofitting, enabling ACS complexity estimator calibration on production query distributions for optimal layer-skipping thresholds. For regulatory compliance with emerging AI sustainability requirements, the EQI metric provides a standardised energy-quality reporting index. The full EEGI implementation, EQI calculator, and production deployment scripts are available as open-source resources. Details in Appendix A.

References

Chen, C. et al. (2023). Accelerating large language model decoding with speculative sampling. arXiv:2302.01318.

Dao, T. et al. (2022). FlashAttention: Fast and memory-efficient exact attention. NeurIPS, 35.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Frantar, E. et al. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. ICLR 2023.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.

Jouppi, N. P. et al. (2017). In-datacenter performance analysis of a tensor processing unit. ISCA 2017.

Leviathan, Y. et al. (2023). Fast inference from transformers via speculative decoding. ICML 2023.

Li, P. et al. (2023). Making AI less thirsty: Uncovering and addressing the secret water footprint of AI models. arXiv:2304.03271.

Moreau, A., and Silva, I. (2025). Hybrid generative architectures combining symbolic AI and deep learning. Journal of Generative Intelligence, 2025(2), 1-8.

Patterson, D. et al. (2021). Carbon emissions and large neural network training. arXiv:2104.10350.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Schwartz, R. et al. (2020). Green AI. Communications of the ACM, 63(12), 54-63.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. ACL 2019, 3645-3650.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Dettmers, T. et al. (2023). QLoRA: Efficient finetuning of quantized LLMs. NeurIPS, 36.

Lindberg, L., Hansen, E., and Ivanov, E. (2024). Efficient fine-tuning strategies for domain-specific LLMs. Journal of Generative Intelligence, 2024(1), 9-16.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024YXTK8 (EEGI: Energy-Efficient Generative Intelligence) and the German Federal Ministry of Education and Research (BMBF) under grant 01IS25076. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The EEGI implementation, EQI calculator, production deployment scripts, and anonymised energy consumption data are available as open-source resources. Full 30-day energy and quality logs are available from the corresponding author.

Ethical Approval

This study involved computational experiments only. No human subjects or personal data were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

Appendix A

EEGI Implementation Guide and EQI Calculation

The following provides EEGI component implementation details and EQI calculation methodology.

ACS and SD-EA Configuration

TRR Cache and BO-DPM Power Management

QAKD Training Protocol and EQI Calculation