International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

CARBON-AWARE GREEN AI, EMERGING TECHNOLOGIES AND FUTURE DIRECTIONS

Ms. Qudsiya Basreen

Student, Department of Computer Science, Vishwa Vishwani Institute of Systems and Management, Hyderabad, Telangana, India| Email: qudsiyabasreen@gmail.com| Mobile no: 8978314943

ARTICLE INFO

KEYWORDS:

Green AI; Carbon-Aware Computing; Dynamic Model Selection; Energy-Estimation Uncertainty; Selection Overhead; Performance Constraints.

ABSTRACT

Sustainable AI deployment means making choices that balance predictive performance, energy use, and operational carbon emissions. Carbon-aware Green AI brings environmental factors into model selection. This approach supports responsible computing while maintaining the performance needed for real-world uses.

Uncertain energy estimates make model selection based only on predicted values less reliable. Evaluating candidate models and routing requests also uses energy and adds delay. These costs must be included when assessing the approach and its potential environmental benefits.

This paper examines how energy estimation uncertainty and selection overhead affect model selection under accuracy and latency constraints. It presents a conceptual framework that filters candidate models against mandatory performance requirements, compares estimated environmental costs, and accounts for uncertainty before selecting an eligible model.

The framework is illustrated through a hypothetical example of three candidate models using both nominal and uncertainty-adjusted energy estimates. Selection overhead is included in energy consumption and total latency. Operational emissions are calculated using a fixed assumed electricity carbon intensity.

The example shows that changing the assumed level of energy uncertainty can change the preferred eligible model. Both selection energy and latency are included, without implying that a latency limit is exceeded. These hypothetical

observations do not establish actual energy savings, validate deployment performance, or prove the framework's novelty.

INTRODUCTION

Artificial intelligence systems increasingly support prediction, classification and automated decision-making across many applications. Their value must be considered alongside the resources required to develop and operate them. Green AI emphasises computational efficiency and environmental sustainability rather than evaluating progress only through predictive performance. A systematic review identifies energy monitoring, benchmarking, deployment and accuracy–energy trade-offs as important research directions (Verdecchia et al., 2023). Energy consumption and operational carbon emissions are related but distinct: emissions depend on electricity consumption and the carbon intensity of the supply. Sustainable AI deployment therefore requires attention to predictive performance, resource requirements and operating conditions.

Model selection offers an approach to managing these considerations during inference. Instead of using the same model for every request, a system can select among candidates with different performance and resource requirements. Research on production context, rather than any model being universally preferable.

A practical challenge is determining how much confidence to place in estimated resource costs. Energy estimation involves assumptions and measurement choices that require careful interpretation (Verdecchia et al., 2023).Uncertain estimates may affect comparisons between candidate models, particularly when their expected resource costs are close. Model selection itself also consumes resources: evaluating candidates, routing requests and loading models can introduce additional energy consumption and delay. These overheads must be considered when assessing the potential benefits of dynamic selection (Cruciani & Verdecchia, 2025). Together, these concerns motivate examining how energy-estimation uncertainty and selection overhead influence decisions, instead of treating nominal estimates as exact or assuming that adaptive selection necessarily reduces environmental impact.

This paper presents a conceptual framework for carbon-aware model selection under accuracy and

information-extraction ensembles examines static and dynamic selection while incorporating resource consumption into selection decisions (Nijkamp et al., 2024). Energy-aware model cascading and routing also consider task requirements, predictive performance and energy consumption when determining which model should execute a task (Cruciani & Verdecchia, 2025). These approaches highlight potential trade-offs between accuracy, execution time and energy use. However, their suitability depends on the workload and deployment rankings mirror energy rankings. The framework provides an explicit decision structure while acknowledging its illustrative scope. Empirical evaluation remains necessary to assess real-world performance, calibrate uncertainty assumptions and compare the approach with existing methods. The example does not establish deployment benefits or demonstrate novelty.

latency constraints. It first excludes candidates that fail minimum workload accuracy or maximum end-to-end latency requirements, including selection overhead. Eligible candidates are compared using adjusted energy estimates and associated operational emissions. A hypothetical example involving three models illustrates nominal estimates, upward energy adjustments, overhead accounting and sensitivity analysis. The purpose is to examine how different assumptions can change the preferred eligible model, not to demonstrate actual savings. Because electricity carbon intensity is fixed across candidates, emissions

REVIEW OF LITERATURE

Green AI and Environmental Assessment

Verdecchia et al. (2023) conducted a systematic review of Green AI research to examine how the field defines, measures and evaluates the environmental impact of artificial intelligence. The review identifies monitoring, model benchmarking, deployment, energy estimation and accuracy–energy trade-offs as relevant research areas. It also distinguishes energy efficiency from carbon footprint and broader environmental impacts. This distinction is important because electricity consumption alone does not fully electricity use and the carbon intensity of the electricity supply. The review therefore supports considering environmental costs alongside predictive performance, while making clear which parts of an AI system are measured, estimated or excluded.

Energy-Aware Model Selection

Nijkamp et al. (2024) examined model selection in DocQMiner, a production information-extraction

represent the environmental consequences of an AI system; operational emissions depend on both

selection, which identifies a subset for a specific queried property.The study also considered energy-aware selection strategies that incorporate resource consumption alongside predictive performance. Their work shows that resource considerations can be included in model-selection decisions within a production application. However, the authors used CPU time as a proxy for energy consumption because direct power measurements were unavailable in their production monitoring environment. Therefore, their reported resource reductions should not be interpreted as directly measured electricity savings or carbon reductions. Their evaluation also used two legal-domain datasets, so broader generalisation to other workloads and deployment conditions remains uncertain.

system. They compared static selection, which identifies a model subset for a domain, with dynamic

Dynamic Selection, Uncertainty and Overhead

Cruciani and Verdecchia (2025) investigated energy-aware model cascading and routing. Cascading executes models sequentially until a confidence requirement is met or a fallback model is used, whereas routing selects one model before inference. Their proof-of-concept analysis considered predictive accuracy, execution time and estimated energy consumption, highlighting the trade-offs involved in adaptive model selection. The study also recognises the costs of the selection process, including selection time, the energy requirements of the routing component and model-loading considerations. Selection-energy overhead was not separately reported because it was considered small and difficult to measure accurately. The authors also distinguish energy consumption from carbon footprint, especially when candidate models use electricity supplies with different carbon intensities.

Table 1. Summary of Reviewed Literature

Research paper	Research focus	Method or approach	Main findings	Limitations relevant to your topic
Verdecchia et al. (2023)	Green AI and the environmental impact of AI	Systematic literature review of 98 primary studies	Identifies monitoring, benchmarking, deployment, energy estimation and accuracy–energy trade-offs as key Green AI research areas. Distinguishes energy efficiency from carbon footprint and broader environmental impacts.	Does not provide a specific model-selection framework. Highlights challenges in consistently measuring energy consumption, emissions and environmental impact.
Nijkamp et al. (2024)	Static and dynamic model selection for production information extraction	Analysis of model-selection strategies in DocQMiner, including energy-aware selection. Uses CPU	Shows that resource consumption can be incorporated into model-selection decisions and reports resource reductions for selected model subsets.	Uses CPU time rather than direct power measurements, so the results do not directly measure electricity consumption

		time as a proxy for energy consumption.		or carbon emissions. The evaluation uses two legal-domain datasets.
Cruciani and Verdecchia (2025)	Energy-aware dynamic model selection, cascading and routing	Proof-of-concept analysis comparing cascading and routing using predictive accuracy, execution time and estimated energy consumption.	Highlights trade-offs between accuracy, execution time and estimated energy use. Recognises that selection and routing introduce additional time, energy and model-loading costs.	Selection-energy overhead is not separately reported because it is considered small and difficult to measure. Energy estimates do not fully capture differences in carbon footprints across electricity supplies.

RESEARCH GAP

The reviewed studies show that energy-aware model selection, accuracy–energy trade-offs and selection overhead are established concerns in Green AI research. They also demonstrate that measurement assumptions and deployment conditions affect how energy and environmental comparisons should be interpreted. However, the reviewed studies do not provide a combined conceptual treatment of energy-estimation uncertainty,

selection overhead and mandatory accuracy and latency constraints.

OBJECTIVES OF THE STUDY

The objectives of this study are:

- To examine how predictive accuracy, estimated energy consumption and end-to-end latency can influence model-selection decisions in a carbon-aware Green AI framework.
- To examine how different assumptions about energy-estimation uncertainty may affect the selection of an eligible model.
- To account for the additional energy and latency associated with selecting

The present study addresses this focused question by proposing a decision sequence that filters models using minimum-accuracy and maximum end-to-end latency requirements, applies stated energy-adjustment assumptions, includes selection energy and latency, and compares estimated operational emissions among eligible models. A sensitivity analysis examines whether changing an assumed energy adjustment changes the selected model. The contribution is a conceptual framework and worked illustration, not experimentally demonstrated energy or emissions savings.

or routing a request to a selected model.

- To compare estimated operational emissions among models that satisfy the required minimum-accuracy and maximum end-to-end latency constraints.
- To illustrate decision sensitivity through a hypothetical three-model example by examining whether changes in assumed energy-estimation adjustments alter the preferred model.

These objectives support the development of a conceptual decision sequence that first filters candidate models using mandatory accuracy and latency requirements, then includes selection overhead and compares estimated operational emissions.

aware model selection under accuracy, energy and latency constraints. The framework is illustrated using hypothetical model characteristics and assumed operating conditions. This approach allows the study to examine how energy-estimation uncertainty, selection overhead and mandatory performance requirements may affect model-selection decisions in a transparent and reproducible manner.

Model-Selection Procedure

The proposed framework begins by screening candidate models against mandatory performance requirements. Each candidate model must satisfy a minimum predictive-

The study focuses on a hypothetical worked example and does not claim experimentally demonstrated energy savings, emissions reductions or validated deployment performance.

RESEARCH METHODOLOGY

Research Design

This study adopts a conceptual research approach supported by a hypothetical worked example. It does not involve experiments on real AI models, direct energy measurements or deployment testing. Instead, the study develops a decision framework for carbon-

accuracy requirement and a maximum end-to-end latency requirement. Models that fail either constraint are excluded before environmental costs are considered. This sequencing ensures that a model is not selected solely because it has a favourable

estimated energy profile while failing required accuracy or latency standards. Only eligible models proceed to the energy, latency and emissions comparison stage.

Energy-Estimation Approach

Because actual energy consumption may differ from initial estimates, the framework applies assumed upward adjustments to nominal energy estimates. These adjustments represent possible uncertainty arising from factors such as hardware variation, workload differences and measurement limitations. The adjusted energy estimate is calculated by multiplying the nominal energy estimate by an uncertainty factor. The adjustments used in this study are illustrative sensitivity parameters rather than statistically calibrated confidence intervals. They are used to examine whether different assumptions about energy uncertainty change the relative ranking of eligible models.

Selection Overhead

The framework includes the additional energy and latency associated with selecting or routing a request to a selected model. Selection overhead may arise from evaluating routing conditions, executing a selection mechanism and loading or directing a request to the chosen model. In the hypothetical example, selection is assumed to add a fixed amount of energy and latency to each request. Total energy is calculated as the sum of adjusted model energy and selection energy. Total latency is calculated as the sum of inference latency and selection latency.

CONCEPTUAL BACKGROUND

Green AI and Sustainable Computing

Green AI refers to the development and use of artificial intelligence systems that consider environmental costs alongside predictive performance. Sustainable computing seeks to reduce the resource consumption and environmental impact of computing systems throughout their operation. In the context of

Including this overhead provides a more complete representation of the resource requirements of a dynamic selection process.

Carbon-Emission Calculation

Estimated operational carbon emissions are calculated using total energy consumption and an assumed electricity carbon intensity. Total energy consumption in watt-hours is converted into kilowatt-hours and multiplied by the assumed carbon intensity, expressed in

grams of CO_2 equivalent per kilowatt-hour. The example assumes a common carbon intensity for all candidate models. Therefore, differences in estimated emissions follow differences in total energy consumption. The calculation illustrates operational emissions only and does not include training emissions, hardware manufacturing or other life-cycle impacts.

AI, this includes electricity consumption, execution time and the carbon emissions associated with running models. Green AI does not require sacrificing accuracy entirely; rather, it encourages decision-makers to understand the trade-offs between performance, resource use and environmental impact. This study applies these ideas to model selection, where different candidate models may offer different levels of accuracy, energy consumption and latency.

Carbon-Aware Model Selection

Carbon-aware AI model selection considers not only which model performs well but also which eligible model is likely to produce lower operational emissions. A carbon-aware approach compares models after they satisfy required performance constraints. It uses estimated energy consumption and an assumed electricity carbon intensity to compare operational emissions. This study focuses on selecting among candidate models rather than designing new models. The selection process is conceptual and uses hypothetical values to illustrate how accuracy, latency, energy uncertainty and selection overhead may affect the final decision.

Accuracy and Latency Constraints

Accuracy and latency are treated as mandatory requirements in the proposed framework. Minimum accuracy ensures that a selected model meets the required level of predictive performance. Maximum end-to-end latency ensures that the model can respond within an acceptable time limit. End-to-end latency includes both inference latency and the latency added by the selection or routing process. A model with low estimated energy consumption should not be selected if it fails either requirement. Therefore, environmental efficiency is considered only after candidate models have passed the mandatory accuracy and latency filters.

Energy-Estimation Uncertainty

Energy-estimation uncertainty recognises that estimated energy consumption may differ from the energy actually consumed during deployment. Differences may arise from hardware characteristics, workload variation, model implementation and measurement limitations. In this study, uncertainty is represented through assumed upward adjustments to nominal energy estimates. These adjustments are illustrative sensitivity parameters rather than statistically calibrated confidence intervals. They allow the study to examine whether changing energy assumptions changes the preferred model among otherwise eligible candidates.

Operational Carbon Emissions

Operational carbon emissions refer to the emissions associated with running an AI model, rather than emissions from training, hardware manufacturing or disposal. They are estimated by converting total energy consumption into kilowatt-hours and multiplying it by the carbon intensity of the electricity supply. Carbon intensity represents the amount of CO_2 equivalent emitted per kilowatt-hour of electricity consumed. In the hypothetical example, a fixed carbon intensity is assumed for all candidate models. Therefore, differences in estimated emissions follow differences in total energy consumption.

ANALYSIS AND DISCUSSION

Purpose and Assumptions

This section demonstrates how the proposed carbon-aware model-selection framework can be applied using a hypothetical example involving three candidate AI models. The example examines how predictive accuracy, energy-estimation uncertainty, selection overhead and end-to-end latency influence model selection. It also examines whether changing an energy-estimation assumption can change the preferred model.

The illustrative assumptions are a minimum predictive accuracy of 90%, a maximum end-to-end latency of 150 milliseconds, an electricity carbon intensity of 400 gCO_2e/kWh, selection energy overhead of 0.0005 Wh per request and selection latency overhead of 5 milliseconds per request.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Figure 1. Proposed Architecture for Carbon-Aware Green AI



All numerical values are hypothetical and are not measurements collected from real AI systems. The purpose is to illustrate the framework's decision sequence, not to establish real-world energy savings or emissions reductions.

Candidate Models and Performance Constraints

Table 2 presents the hypothetical characteristics of three candidate models. The models differ in predictive accuracy, nominal energy consumption per request and inference latency.

Table 2. Hypothetical characteristics of candidate models

Model	Accuracy (%)	Nominal energy (Wh/request)	Inference latency (ms)
A	96%	0.0080	180
B	94%	0.0050	120
C	91%	0.0030	90

Model A has the highest predictive accuracy but an inference latency of 180 milliseconds, which exceeds the maximum permitted end-to-end latency of 150 milliseconds. Including selection overhead raises its total latency to 185 milliseconds. Therefore, Model A is excluded before environmental costs are compared. Models B and C satisfy the minimum-accuracy requirement and remain within the latency limit after selection overhead is included. This demonstrates why the framework filters models using mandatory performance requirements before comparing environmental costs.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Figure 2. Illustrative Model-Selection and Sensitivity Analysis



Energy-Estimation Uncertainty

Nominal energy estimates may differ from the energy actually consumed during deployment because of hardware characteristics, workload variation and measurement limitations. To represent this uncertainty, the framework applies assumed upward adjustments to the nominal energy estimates of the eligible models.

$$E_{adjusted} = E_{nominal} \times (1 + u)$$

Here, $E_{adjusted}$ is the adjusted energy estimate, $E_{nominal}$ is the nominal energy estimate, and u is the assumed upward-adjustment factor expressed as a decimal.

For Model B, a 10% adjustment is applied:

$$E_{adjusted,B} = 0.0050 \times (1 + 0.10) = 0.0055 \text{ Wh/request}$$

For Model C, a 30% adjustment is applied:

$$E_{adjusted,C} = 0.0030 \times (1 + 0.30) = 0.0039 \text{ Wh/request}$$

Table 3. Adjusted energy estimates

Model	Nominal energy (Wh/request)	Upward adjustment (%)	Adjusted energy (Wh/request)
B	0.0050	10%	0.0055
C	0.0030	30%	0.0039

These adjustments are illustrative sensitivity parameters rather than statistically calibrated uncertainty intervals.

Selection Overhead and Total Energy

Dynamic model selection or request routing can introduce additional computational costs, including evaluating routing conditions, selecting a model and directing a request to the selected model. To avoid underestimating total resource requirements, the framework includes both selection energy and selection latency.

$$E_{total} = E_{adjusted} + E_{overhead}$$

$$L_{total} = L_{inference} + L_{overhead}$$

For Model B:

$$E_{total,B} = 0.0055 + 0.0005 = 0.0060 \text{ Wh/request}$$

$$L_{total,B} = 120 + 5 = 125 \text{ ms}$$

For Model C:

$$E_{total,C} = 0.0039 + 0.0005 = 0.0044 \text{ Wh/request}$$

$$L_{total,C} = 90 + 5 = 95 \text{ ms}$$

Table 4. Total energy and latency after selection overhead

Model	Adjusted energy (Wh/request)	Selection energy overhead (Wh/request)	Total energy (Wh/request)	Total latency (ms)
B	0.0055	0.0005	0.0060	125
C	0.0039	0.0005	0.0044	95

Both models remain eligible because their total latencies are below 150 milliseconds. Under the initial assumptions, Model C has lower estimated total energy consumption than Model B.

Estimated Operational Carbon Emissions

Estimated operational carbon emissions are calculated using total energy consumption and the assumed electricity carbon intensity.

$$C_{operational} = (E_{total} \times I) / 1{,}000$$

Here, $C_{operational}$ is estimated emissions in grams of CO_2 equivalent per request, E_{total} is total energy in watt-hours per request, and I is electricity carbon intensity in grams of CO_2 equivalent per kilowatt-hour. Dividing by 1,000 converts watt-hours to kilowatt-hours.

For Model B:

$$C_B = (0.0060 \times 400) / 1{,}000 = 0.00240 \text{ gCO}_2\text{e/request}$$

For Model C:

$$C_C = (0.0044 \times 400) / 1{,}000 = 0.00176 \text{ gCO}_2\text{e/request}$$

The percentage difference is calculated as:

$$((0.00240 - 0.00176) / 0.00240) \times 100 \approx 26.7\%$$

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

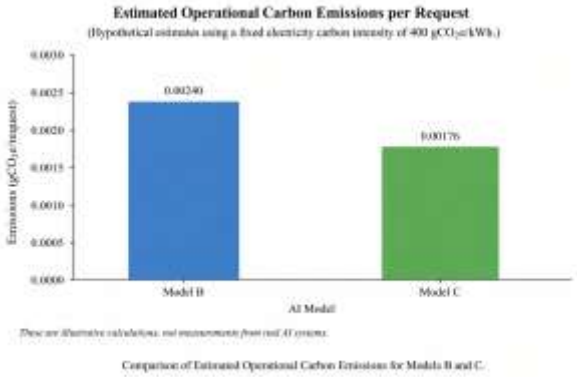


Table 5. Estimated operational emissions

Model	Total energy (Wh/request)	Carbon intensity (gCO_2e/kWh)	Estimated emissions (gCO_2e/request)
B	0.0060	400	0.00240
C	0.0044	400	0.00176

Under the initial assumptions, Model C has approximately 26.7% lower estimated operational emissions than Model B. This is a hypothetical comparison based on assumed energy values and a fixed carbon intensity; it is not a measured environmental benefit.

Sensitivity Analysis

The preferred model may change when energy-estimation assumptions change. To examine this possibility, Model C's assumed upward adjustment is increased from 30% to 100%, while Model B retains its 10% adjustment.

For Model C, the revised adjusted energy is:

$$\mathbf{E_{adjusted},C = 0.0030 \times (1 + 1.00) = 0.0060\ Wh/request}$$

After adding selection energy overhead:

$$\mathbf{E_{total},C = 0.0060 + 0.0005 = 0.0065\ Wh/request}$$

The revised operational emissions are:

$$\mathbf{C_C = (0.0065 \times 400) / 1{,}000 = 0.00260\ gCO_2e/request}$$

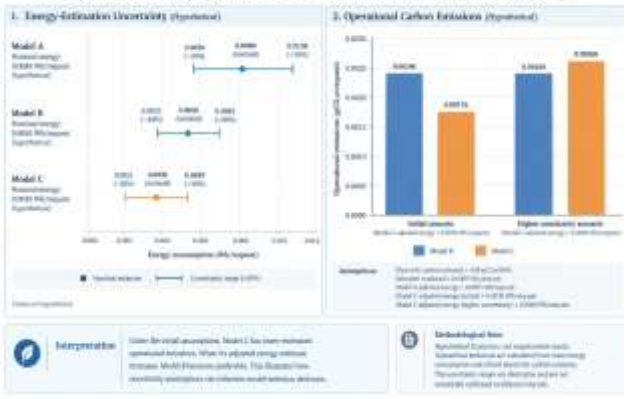
Table 6. Sensitivity to energy-adjustment assumptions

Model	Assumed adjustment (%)	Total energy (Wh/request)	Estimated emissions (gCO_2e/request)
B	10	0.0060	0.00240
C	100	0.0065	0.00260

Both models continue to satisfy the accuracy and latency constraints. However, Model B now has lower estimated energy consumption and operational emissions than Model C. Therefore, the preferred model changes from

Model C under the initial assumptions to Model B under the revised assumptions.

Figure 3. Sensitivity of Operational Carbon Emissions to Energy-Estimation Uncertainty



Interpretation and Limitations

The example demonstrates three main points. First, mandatory accuracy and latency requirements determine which models are eligible for selection. Model A is excluded because it fails the latency constraint, even though it has the highest accuracy. Second, selection overhead should be included in resource calculations because it adds both energy consumption and latency to the system. Third, energy-estimation assumptions can change the preferred model: Model C is preferred initially, whereas Model B becomes preferable when Model C's assumed energy adjustment increases.

However, the results are illustrative. The example uses hypothetical values, fixed carbon intensity and simplified energy adjustments. It does not use direct energy measurements, calibrated uncertainty estimates or real deployment workloads. Empirical validation using measured energy consumption, representative workloads and observed latency is required before drawing conclusions about practical environmental benefits. The framework illustrates a potential approach to carbon-conscious AI deployment but does not establish actual energy savings, emissions reductions or validated deployment performance.

MAJOR FINDINGS

The hypothetical example demonstrates that mandatory performance requirements should be applied before environmental costs are compared. Model A is excluded because its inference latency exceeds the maximum end-to-end latency constraint, even though it has the highest predictive accuracy. Models B and C satisfy both the minimum-accuracy and maximum-latency requirements and therefore remain eligible for further evaluation.

The analysis shows that energy-estimation uncertainty can influence model selection. Under the initial assumptions, Model C has lower adjusted total energy consumption and

lower estimated operational emissions than Model B. However, when Model C's assumed energy adjustment increases from 30% to 100%, its total energy and estimated emissions exceed those of Model B. The preferred model therefore changes from Model C to Model B, indicating that model-selection decisions may be sensitive to energy-estimation assumptions.

The example also shows the importance of including selection overhead. Selection adds assumed energy consumption and latency to each request. Although both eligible models remain within the latency constraint, excluding this overhead would provide an incomplete estimate of the total resource requirements of a dynamic selection process.

Finally, the findings show that estimated operational emissions depend on total energy consumption and the assumed electricity carbon intensity. Because the example assumes a fixed carbon intensity for all models, the emissions ranking follows the total-energy ranking. These findings are illustrative rather than experimentally validated. Further empirical research using measured energy consumption, realistic workloads and observed latency is required to assess the framework's practical effectiveness.

SUGGESTIONS OR RECOMMENDATIONS

Based on the proposed framework and the hypothetical analysis, several recommendations are made for researchers and developers seeking to improve the environmental efficiency of artificial intelligence systems.

First, AI model selection should apply mandatory predictive-accuracy and end-to-end latency constraints before comparing energy consumption and estimated operational carbon emissions. This approach

ensures that environmental considerations are incorporated without compromising essential performance requirements.

Second, developers should measure energy consumption under representative workloads where feasible rather than relying exclusively on nominal estimates. Evaluations should account for relevant hardware characteristics, workload variation and system utilisation. When direct measurements are unavailable, estimation methods, underlying assumptions and associated limitations should be documented transparently.

Third, energy-estimation uncertainty should be examined through sensitivity analysis. Comparing alternative estimation assumptions can reveal whether a model-selection decision remains stable or changes under different conditions. Such analysis can help researchers identify decisions that require more reliable energy measurements before deployment.

Fourth, evaluations of dynamic model-selection systems should include the additional energy consumption and latency introduced by model selection and request routing. Developers should assess whether these overheads affect overall system efficiency and whether the selected model continues to satisfy the required performance constraints.

Fifth, future research should evaluate the proposed framework using real AI models, measured energy consumption, representative workloads and observed end-to-end latency. Where appropriate, researchers should investigate variations in electricity carbon intensity across locations and times to develop more context-sensitive estimates of operational emissions.

Finally, future evaluations should compare the proposed framework with a clearly defined baseline, such as selecting the lowest-energy model among candidates that satisfy the same accuracy and latency requirements. Such comparisons could help establish whether explicitly accounting for energy-estimation uncertainty and selection overhead improves decision quality under realistic deployment conditions.

CONCLUSION

This study examined a conceptual framework for carbon-aware AI model selection that considers predictive accuracy, energy-estimation uncertainty, selection overhead, end-to-end latency and estimated operational

carbon emissions. The proposed decision sequence first filters candidate models according to mandatory accuracy and latency requirements and then compares the environmental costs of the eligible models.

The hypothetical three-model example illustrates how energy-estimation assumptions affect model-selection decisions. Under the initial assumptions, Model C produces lower estimated operational emissions than Model B. However, when the assumed energy adjustment for Model C increases, Model B becomes the preferred option. The example also demonstrates why the energy consumption and latency associated with model selection and request routing should

These recommendations provide a clear direction for further research into AI model-selection strategies that balance predictive performance, computational efficiency and environmental considerations.

be included in the overall evaluation.The framework provides a structured way to reason about the trade-offs between performance requirements and environmental considerations. However, the numerical results are illustrative and do not establish actual energy savings, emissions reductions or deployment effectiveness. The analysis also uses a fixed electricity carbon intensity and simplified energy-adjustment assumptions.

Future research should evaluate the framework using real AI models, measured energy consumption, representative workloads and observed latency. Such validation would help determine whether accounting for uncertainty and selection overhead improves model-selection decisions in practical deployments. Overall, this study presents a conceptual approach to supporting more transparent and environmentally informed AI model selection.

REFERENCES

1. Cruciani, E., & Verdecchia, R. (2025). Choosing to be green: Advancing Green AI via dynamic model selection. In Proceedings of the 2nd Workshop on Green-Aware Artificial Intelligence (Green-AI 2025). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-4165/short2.pdf

2. Nijkamp, N., Sallou, J., van der Heijden, N., & Cruz, L. (2024). Green AI in action: Strategic model selection for ensembles in production. In Proceedings of the 1st ACM International Conference on AI-Powered Software (pp. 50–58). Association for Computing Machinery. https://doi.org/10.1145/3664646.3664763
3. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831
4. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355
5. Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. WIREs Data Mining and Knowledge Discovery, 13(4), Article e1507. https://doi.org/10.1002/widm.1507