EcoDashAI: A Visual Analytics Dashboard for Multi-Criteria Evaluation of Language Models Frugality

Maxime Masson[1,*], Philippe Roose[1] and Gorka Dalmayrac-Belascain[1]

[1]LIUPPA, E2S, Université de Pau et des Pays de l'Adour, Pau, France

Abstract

The rapid proliferation of Language Models (LMs) raises growing concerns about environmental sustainability, particularly regarding energy consumption, inference cost, and hardware requirements. Selecting an appropriate model for a given task is no longer a purely technical decision: it is increasingly a sustainability trade-off. This challenge is compounded by the heterogeneity of task requirements; not all use cases demand state-of-the-art performance, and deploying high-capacity models for low-complexity tasks incurs unnecessary computational and environmental overhead. We present *EcoDashAI*, an interactive visual analytics dashboard that frames model selection as a trade-off between task-specific performance and resource efficiency, rather than optimizing for performance alone. Through a set of complementary visualization views and a unified frugality scoring framework, *EcoDashAI* enables practitioners and researchers to explore, compare, and identify the most frugal model adequate for their specific requirements.

Keywords

Large Language Models, Environmental Sustainability, Frugality, Visualization, Green AI, Information Systems

1. Introduction

The integration of Artificial Intelligence (AI) components, and in particular Large Language Models (LLMs), has accelerated considerably across all sectors of software engineering. Yet, the environmental cost of this integration remains largely invisible to system designers and practitioners. This opacity is driven by a lack of standardized energy-tracking telemetry and compounded by an industry-wide focus on optimizing for model capability rather than energy efficiency [1]. Training and serving LLMs requires substantial computational resources, which translates into significant energy consumption, carbon emissions, and economic costs [2, 1]. As information systems grow increasingly reliant on AI inference pipelines, these hidden costs compound into a systemic sustainability challenge.

Beyond the initial training phase, *inference frugality*, the capacity of a deployed model to deliver sufficient performance at minimal resource consumption, is emerging as a first-class concern in sustainable software design [1]. However, model selection remains a notoriously opaque process. Benchmark leaderboards typically surface performance metrics in isolation, fundamentally disconnected from cost, latency, energy, or hardware constraints. Consequently, practitioners are left to reconcile dozens of heterogeneous metrics across disparate sources, with no integrated framework to guide their deployment decisions. To address this critical gap, we present *EcoDashAI*, a web-based visual analytics dashboard. *EcoDashAI* operationalizes the notion of *frugality* as a multi-dimensional construct combining performance, cost, energy, latency, and memory requirements, exposing this complex design space through a rich, interactive visualization environment. The system targets two complementary user profiles: (1) software engineers seeking to select the most appropriate and sustainable model for a given deployment context, and (2) researchers studying the evolving sustainability landscape of the LM ecosystem.

This research is still in its early stages; however, to illustrate the ongoing work, a demonstration based on simulated mock data is available at https://mmasson003.perso.univ-pau.fr/frugality.

RCIS 2026: Companion Proceedings of the 20th Conference on Research Challenges in Information Science: RCIS Research Projects and Workshops, May 26-29, 2026, Toulouse, France

*Corresponding authors.

maxime.masson@univ-pau.fr (M. Masson); philippe.roose@univ-pau.fr (P. Roose); gdbelascain@iutbayonne.univ-pau.fr (G. Dalmayrac-Belascain)

0000-0001-5254-7583 (M. Masson); 0000-0002-2227-3283 (P. Roose)

2. Related Work

The environmental impact of deep learning was initially scrutinized primarily through the lens of model training, which requires massive computational bursts [2]. However, as the paradigm shifts from development to widespread deployment, inference now dominates the lifetime carbon and energy footprint of LMs [3]. Addressing this shift requires adopting an approach in which frugality is evaluated as a primary metric alongside conventional performance [1]. Consequently, selecting a model is no longer a straightforward exercise in accuracy maximization, but a complex multi-objective optimization problem. Deploying a billion-parameter model for simple text classification constitutes massive computational over-provisioning; conversely, over-optimizing for speed or memory can degrade performance below acceptable functional thresholds.

Because frugality is inherently a multidimensional construct [4], we present a taxonomy of common inference frugality metrics in Figure 1. They were selected according to three primary criteria: function, observability, and stakeholder utility. First, each metric corresponds to a distinct real-world deployment constraint (e.g., economic budget, power consumption limits, or user experience thresholds). Second, these metrics are architecturally agnostic, meaning they can be measured via standard model calls without requiring access to the proprietary training process. Finally, every metric provides actionable data for specific organizational stakeholders, such as budget owners or infrastructure leads.

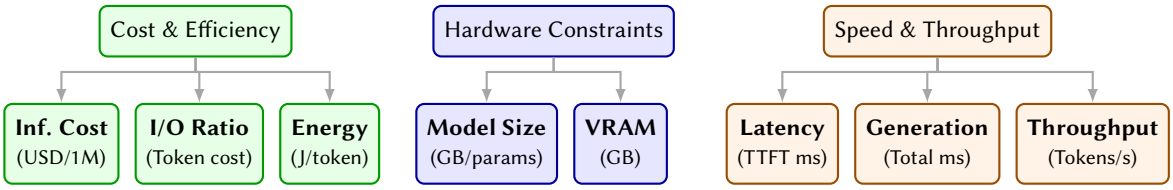

Figure 1: Taxonomy of LM inference frugality dimensions frequently identified in the literature.

Despite the critical nature of this trade-off, current evaluation ecosystems remain highly fragmented. Traditional resources such as the LMSYS Chatbot Arena [5] and the HuggingFace Open LLM Leaderboard [6] are predominantly performance-centric, treating model size as a mere categorical filter rather than a continuous cost variable. Efficiency-focused tools such as LLM-Perf [7] benchmark latency and throughput, but do so in isolation from task-specific capabilities or environmental costs. To our knowledge, no existing platform jointly visualizes this trade-off space across a broad model landscape in a task-aware manner.

3. Platform Requirements

Based on the systemic gaps identified in the literature, specifically the fragmentation of performance and efficiency metrics [2], and the lack of task-aware evaluation frameworks, we define a set of requirements for an ideal multi-criteria selection system. Model selection is treated as a dynamic, context-dependent, and multi-objective decision-making process. Requirements fall into 3 categories:

Analytical Requirements: The literature establishes that frugality is not reducible to a single variable [1], such as model size. Therefore, an effective system must be capable of evaluating cost, energy, latency, memory, throughput [2], and potentially other frugality metrics simultaneously (R1). To make these heterogeneous metrics mathematically comparable, the system necessitates a robust normalization layer (R4). Furthermore, since prior works emphasize the balance between capability and resource consumption, users must be able to explicitly visualize the Pareto frontier (R2) to identify instances of unnecessary computational over-provisioning. Finally, the literature highlights a need for explicit substitutability analysis [8] (R6) to help users discover lighter, equivalent models.

Contextual Requirements: Model deployment does not occur in a vacuum. A prevalent gap in current benchmarking practices is the implicit assumption of uniform task complexity [9]. An ideal framework must be inherently task-aware, recognizing that not all use cases demand state-of-the-art performance

(R3). Beyond individual deployment tasks, researchers require the ability to explore the macro-level landscape of the LLM ecosystem (R5), including provider footprints and parameter-count distributions, to accurately trace broader sustainability trends across the industry.

Usability Requirements: To transition Green AI from a theoretical academic concern to a practical software engineering standard [1], the synthesis of these multi-dimensional metrics must be accessible to system architects and practitioners, not just machine learning specialists. This mandates presenting complex frugality reasoning in an intuitive, visual format rather than raw, static data tables (R8). Because different organizational stakeholders (e.g., budget owners vs. infrastructure leads) prioritize criteria differently, the system must support interactive, dynamic reconfiguration of metric weights and thresholds (R7). These conceptual requirements, directly mapping to the shortcomings of the current evaluation ecosystem, are summarized in Table 1.

Literature Gap	Derived Requirement
Isolated metrics	R1: Multi-dimensional frugality tracking; R4: Unified normalization layer.
Overemphasis on peak performance	R2: Pareto frontier visualization; R6: Substitutability analysis.
Uniform task assumptions	R3: Task-aware capability evaluation.
Fragmented model landscape	R5: Macro-level ecosystem exploration.
Opaque selection processes	R7: Dynamic weight/threshold configuration; R8: Intuitive visual analytics.

Table 1: Mapping of literature gaps to the requirements for a frugality-aware evaluation system.

4. System Architecture and Data Pipeline

To satisfy the requirements outlined above, we propose a pipeline structured into four sequential stages: data collection, normalization, metric categorization, and aggregation.

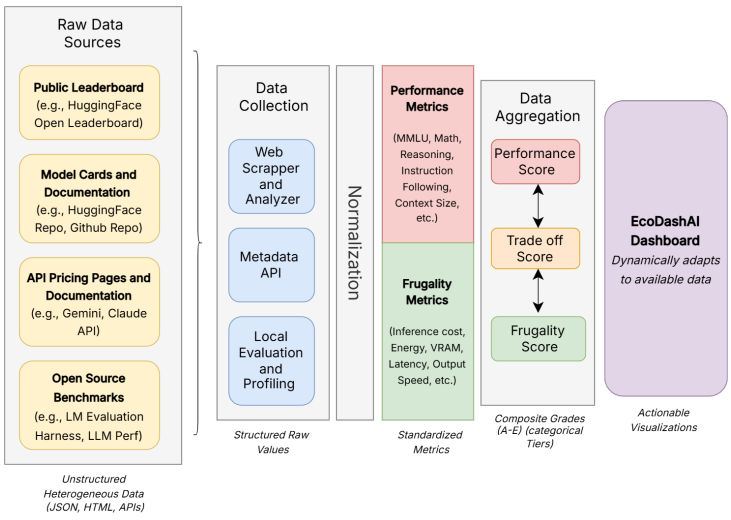

Figure 2: Overview of the *EcoDashAI* system architecture, illustrating the data flow from raw data sources to the final dashboard scores.

This pipeline, illustrated in Figure 2, culminates in the *EcoDashAI* dashboard. The system first aggregates model metadata from a diverse set of *Raw Data Sources*. These include public leaderboards (e.g., HuggingFace Open Leaderboard [6]), model cards, API pricing pages (e.g., Gemini, Claude), and open-source benchmarks (e.g., LM Evaluation Harness [10], LLM Perf [7]). This data is ingested during the *Data Collection* phase via web scrapers, metadata APIs, and local evaluation profiling.

Once collected, the disparate raw values pass through a *Normalization* layer, mapping them to a standard unified scale (e.g., 0 to 1) to enable cross-dimensional comparison. These normalized data points are then classified into two core metric groups. *Performance Metrics* capture capabilities such as MMLU [11], mathematics, reasoning, instruction following, and context window size. *Frugality Metrics*

cover operational efficiencies such as inference cost, estimated energy consumption, VRAM usage, latency, and throughput. Finally, in the *Data Aggregation* stage, the categorized metrics are synthesized into two high-level representations: a *Performance Score* (P) and a *Frugality Score* (F). The Frugality Score is computed as a weighted aggregate of normalized metrics, $F = \sum_{i=1}^{n} w_i \cdot \hat{m}_i$, where n denotes the total number of evaluated metrics, $\hat{m}_i$ is the i-th normalized metric, and w_i is its corresponding user-configurable weight. F is then mapped to a letter grade (A–E) for interpretability. The Performance Score P is derived analogously using the corresponding performance metric set. Finally, P and F are jointly evaluated to produce a *Trade-off Score*.

5. The EcoDashAI Dashboard

Translating the aggregated scores into actionable insights, *EcoDashAI*[1] provides six complementary visualization views. These views are organized around two distinct analytical goals: *landscape exploration* (understanding the overall model ecosystem) and *task-focused selection* (identifying the best model for a specific use case). Global filters are available and consistently applied across all views.

The *Efficiency Landscape* is a configurable scatter plot positioning models by frugality (X) and performance (Y), with bubble size encoding parameter count and color encoding a categorical attribute; a dashed Pareto frontier highlights non-dominated models. *Constellations* [12] attracts models toward peers sharing a grouping property, with clusters coalescing around labeled gravity centers to expose macro-level ecosystem trends. The *Similarity Graph* [12] renders the pairwise similarity matrix as a weighted graph where edge thickness encodes similarity score S, with per-metric sliders controlling dimensional contributions and edge hovering revealing a direct comparison panel. The *Market Ecosystem* treemap [13] groups models by family with area encoding parameter count, prompting reflection on whether a family's performance justifies its resource footprint. *Gravitational Efficiency* plots models by distance D from the *Singularity*, the theoretical ideal of maximum performance at zero frugality cost, with dotted arcs connecting models of equivalent D to reveal performance-vs-frugality trade-offs. Finally, *Prompt Analysis* presents a parallel coordinates chart styled as a transit map, with axes split into green/red zones via threshold sliders; AUTO-ANALYZE infers thresholds from a natural-language description while the right panel ranks models by composite *curve score*.

6. Demonstration Scenario

The demonstration will be conducted live at the workshop using the publicly accessible prototype instance of *EcoDashAI* (https://mmasson003.perso.univ-pau.fr/frugality/). We propose a scenario structured around the practical workflow of a software engineer evaluating models, detailed in Table 2.

Phase	User Action	Key Observation
Step 1	*Efficiency Landscape*: swap X/Y axes, modify color/size encoding.	Reveals open-source clustering vs. proprietary long tail and the shifting Pareto frontier.
Step 2	*Constellations*: set color to license type, select models, compare via heatmap and side panel.	Visualizes open/closed divide (green/red) within clusters and provider footprints.
Step 3	*Similarity Graph*: maximize MMLU weight, zero cost/latency.	Identifies performance-equivalent pairs; observe graph restructuring as cost weights are reintroduced.
Step 4	*Market Ecosystem* treemap: group by model family.	Highlights dominant families by parameter count; prompts sustainability discussion.
Step 5	*Gravitational Efficiency*: analyze models relative to the *Singularity*.	Contrasts models on equal arcs with opposite frugality–performance trade-offs.
Step 6	*Prompt Analysis*: enter prompt, click AUTO-ANALYZE.	Column gradients adapt to requirements; identify models in green zones and best performance–sustainability trade-off.

Table 2: Step-by-step demonstration walkthrough for the workshop.

[1] **Live demonstration available at:** https://mmasson003.perso.univ-pau.fr/frugality/ (simulated mock data are used)

7. Conclusion

We have presented *EcoDashAI*, an interactive visual analytics dashboard designed for the multi-criteria frugality evaluation of Large Language Models. By offering six complementary visualization views and a unified scoring framework, the system empowers software engineers and researchers to rigorously reason about the performance–sustainability trade-off during model selection. The planned demonstration will invite participants to explore the dashboard hands-on and engage in discussions about sustainability metrics, visualization design choices, and the broader challenge of embedding environmental considerations into AI-driven information systems. Future work aims to finalize the automated data collection and normalization pipeline, transitioning the system from a simulated prototype to a live platform continuously updated with the latest model benchmarks. Another key direction is the development of a prompt analysis system capable of automatically identifying the model that offers the best performance–sustainability trade-off for a given prompt.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green ai, Comm. of the ACM 63 (2020) 54–63.

[2] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, in: Proceedings of the 57th annual meeting of the ACL, 2019, pp. 3645–3650.

[3] K. Lottick, S. Susai, S. A. Friedler, J. P. Wilson, Energy usage reports: Environmental awareness as part of algorithmic accountability, in: NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2019.

[4] J. Violos, K.-C. Diamanti, I. Kompatsiaris, S. Papadopoulos, Frugal machine learning for energy-efficient and resource-aware artificial intelligence, in: Artificial Intelligence, Data and Robotics: Foundations, Transformations and Future Directions, Springer, 2026, pp. 175–198.

[5] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al., Chatbot arena: An open platform for evaluating llms by human preference, in: Forty-first International Conference on Machine Learning, 2024.

[6] O. L. Leaderboard, Open llm leaderboard - a hugging face space by open-llm-leaderboard, 2026. URL: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/.

[7] R. P. Ilyas Moutawwakil, Llm-perf leaderboard, https://huggingface.co/spaces/optimum/llm-perf-leaderboard, 2023.

[8] G. Menghani, Efficient deep learning: A survey on making deep learning models smaller, faster, and better, ACM Computing Surveys 55 (2023) 1–37.

[9] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2022).

[10] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, The language model evaluation harness, 2024. URL: https://zenodo.org/records/12608602. doi:10.5281/zenodo.12608602.

[11] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, arXiv preprint arXiv:2009.03300 (2020).

[12] T. M. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement, Software: Practice and experience 21 (1991) 1129–1164.

[13] B. Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach, ACM Transactions on graphics (TOG) 11 (1992) 92–99.