Introducing a Green Leaderboard for Sustainable Risk Prediction in Streaming NLP Shared Tasks.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Arturo Montejo-Ráez
Computer Science Department, SINAI, CEATIC, University of Jaén
Campus Las Lagunillas, 23071, Jaén, Spain
{amarmol, ammunoz, amontejo}@ujaen.es

Abstract

Current NLP shared-task evaluations predominantly rank systems by predictive performance, overlooking computational efficiency and environmental impact. This limitation is particularly critical in streaming and early risk detection scenarios, where models operate continuously, and resource consumption accumulates over time. We propose a sustainability-aware evaluation framework for streaming NLP tasks by introducing the Green Early Detection Score (GED), which integrates classification performance, detection timeliness, and carbon emissions. We also present an energy-based variant tailored to on-device early risk detection settings where energy consumption per inference is a key constraint. Applying these metrics to three editions (2023-2025) of the MentalRiskES shared task, we construct the first Green Leaderboard for early risk detection. Our results show that sustainability-aware ranking substantially reshapes system positions, highlighting efficient models that remain undervalued under performance-only evaluation.

Keywords: Natural Language Processing, Code Carbon, Energy Consumption, Environmental Impact, Large Language Models, Machine Learning

1. Introduction

The rapid advancement of Natural Language Processing (NLP) has led to increasingly powerful models, but this progress has come at a growing computational and environmental cost. Recent studies have shown that training large language models (LLMs) can emit amounts of CO_2 comparable to multiple transatlantic flights, raising serious concerns about the sustainability of current AI development practices (Strubell et al., 2019; Bender et al., 2021). As a result, the NLP community has begun to reflect on the environmental footprint of modern models. However, evaluation practices in shared tasks and competitions still overwhelmingly prioritize predictive performance, often relegating efficiency and sustainability indicators to secondary metadata.

Traditional leaderboards in NLP shared tasks rank systems exclusively according to a single predictive metric. In classification tasks, systems are ordered by Macro-F1, while regression tasks typically rely on RMSE. This single-metric paradigm creates implicit incentives to maximize performance regardless of computational efficiency, resource consumption, or deployment feasibility. These limitations are particularly pronounced in streaming scenarios such as early risk detection, where systems continuously process incoming data and must produce timely predictions. In such settings, computational and energy costs accumulate over time, yet current evaluation frameworks fail to account for these critical dimensions.

The MentalRiskES shared task (Mármol-Romero et al., 2023) constitutes a notable step forward in this regard. Focused on the early detection of mental health risks from Spanish social media, MentalRiskES has required participating teams to report detailed environmental indicators, including energy consumption and carbon emissions, alongside predictive metrics since 2023 (Mármol-Romero et al., 2024). This initiative has resulted in a unique multi-year dataset in which systems are evaluated not only in terms of accuracy and timeliness, but also with respect to their computational footprint. Nevertheless, despite the availability of this information, official rankings continue to be based solely on predictive performance, potentially favoring systems that achieve marginal gains at disproportionate computational or environmental cost.

Importantly, many early risk detection systems are explicitly designed for deployment on resource-constrained devices, such as smartphones or wearable platforms, where continuous monitoring, low latency, and limited battery capacity are key constraints. In these on-device scenarios, energy consumption per inference is often a more relevant and actionable metric than aggregate carbon emissions, as it directly impacts battery life, usability, and real-world feasibility. This observation motivates the need for evaluation metrics that explicitly account for energy efficiency and timeliness alongside predictive performance.

To address these limitations, we propose a sustainability-aware evaluation framework for streaming NLP tasks. We introduce the **Green**

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 144–153
May 12, 2026

Early Detection Score (GED), a composite metric that integrates predictive effectiveness, detection timeliness, and carbon emissions, and its energy-oriented variant, specifically tailored to on-device deployment scenarios. By applying these metrics to re-rank systems submitted to three editions of the MentalRiskES shared task, we construct the first *Green Leaderboard for early risk detection*. This leaderboard demonstrates how incorporating efficiency-aware criteria can significantly alter system rankings and encourage the development of models that are not only accurate and timely, but also practical and sustainable.

2. Related Work

The NLP and AI community has only recently begun systematically studying the energy and carbon costs of modern models. Early foundational work has shown that training large NLP models incurs substantial energy usage and financial costs, and recommends that researchers report resource metrics (e.g., training time) alongside accuracy (Strubell et al., 2019; Hershcovich et al., 2022). This set the stage for "Green AI" efforts: Schwartz et al. (2020) surveyed multiple efficiency measures (CO2 emissions, electricity, FLOPs, etc.) and argued for explicitly treating performance-compute trade-offs as a metric, for example by plotting accuracy against training size or energy use. In this spirit, Bender et al. (2021) famously warned of the dangers of "ever-larger" language models. They urge weighing the environmental and financial costs first in model design, rather than blindly scaling up, recommending more careful data curation and exploring research beyond ever-larger models. In short, these works emphasize that, in NLP and AI, carbon costs should be reported and minimized as part of standard practice.

To quantify these effects, several tools and frameworks have been developed. Bannour et al. (2021) surveyed six popular carbon-tracking tools (CarbonTracker, Experiment-Impact-Tracker, Green Algorithms, ML-CO2 Impact, etc.) and applied them to NLP experiments. They demonstrate that tools differ in scope and methodology but all aim to measure kWh and CO2 of training runs. Lannelongue et al. (2021) went further by providing a simple online calculator that estimates carbon footprint from compute hardware, runtime, and location; they applied it to NLP tasks. Their calculator[1] and open methodology have been adopted widely, illustrating how to generalize carbon accounting across domain. In practice, researchers now often use tools like CodeCarbon (Schmidt et al., 2021) to record emissions during experiments. These measurement techniques are now being applied within shared evaluation campaigns. Mármol-Romero et al. (2024) analyze the MentalRiskES shared task on Spanish mental health risk detection, where organizers required participants to submit carbon-efficiency metrics alongside predictions. This study empirically correlates submitted CO2 emissions with model architecture and performance, demonstrating how a shared-task framework can reveal the ecological footprint of diverse NLP systems. In short, a growing number of NLP competitions now include energy reporting, enabling analyses of model emissions at scale (Vallecillo-Rodríguez et al., 2024).

Beyond tools, researchers have also articulated conceptual frameworks for "sustainable AI". Van Wynsberghe (2021) defines Sustainable AI as AI that is compatible with maintaining environmental resources for current and future generations. More concretely, Bolón-Canedo et al. (2024) review the notion of "Green AI", distinguishing green-by-AI (using AI to aid environmental applications) from green-in-AI (making AI itself more efficient). They highlight strategies like designing energy-efficient models, using renewable-energy data centers, and adding regulatory incentives. Similarly, Kaack et al. (2022) argues for aligning AI with climate change mitigation, calling for high-level commitments and policies to ensure AI development reduces rather than worsens emissions. In practice, many researchers promote concrete best practices (e.g. using efficient hardware, pruning, distillation) and transparency: reporting climate-related performance is urged as a way to drive improvements (Hershcovich et al., 2022; Henderson et al., 2020).

3. Methodology

3.1. Task Setting

MentalRiskES is formulated as an online early risk detection task. For each user u, systems receive a temporally ordered sequence of messages:

$$\{m_{u,1}, m_{u,2}, \ldots, m_{u,T}\}$$

At each round t, the system must emit a prediction regarding whether the user is at risk of a given mental disorder. Predictions are produced incrementally, simulating a real-time monitoring scenario where decisions must be made under partial information.

Let t_u^* denote the first round at which a system correctly detects the disorder for user u. If the system never detects the disorder, t_u^* is undefined and penalized according to the official evaluation protocol. MentalRiskES evaluates systems along two main dimensions:

[1] https://www.green-algorithms.org/

- **Predictive performance**, measured using Macro-F1.
- **Detection timeliness**, measured using Early Risk Detection Error (ERDE) (Losada and Crestani, 2016). (ERDE). ERDE combines correctness and delay: false negatives receive the maximum penalty, while correct positive detections become increasingly costly as the system consumes more messages before emitting the alert. In MentalRiskES, ERDE30 uses a delay parameter of 30 messages, so late but correct detections are penalized more strongly than early correct alerts.

3.2. Environmental Impact Measurement

In addition to predictive metrics, MentalRiskES requires participants to report environmental metadata for each prediction round, measured using CodeCarbon (Courty et al., 2024). For each system and prediction round, they collect:

- CO_2 emissions in kilograms of CO_2 equivalent.
- Energy consumption in kilowatt-hours.
- Hardware configuration, including CPU, GPU, and RAM usage.

3.3. Data Collection

We analyze all systems submitted to the MentalRiskES shared tasks during the 2023, 2024, and 2025 editions. For each system, we extract:

- Predictive metrics: Macro-F1 and ERDE30 (metrics used for the official ranking).
- Environmental metrics: average and cumulative CO_2 emissions and total energy consumption.
- System characteristics: model family (machine learning, deep learning, or LLMs) and reported hardware configuration.

This results in a unified dataset covering more than 170 systems evaluated under identical streaming conditions. Data from the 2023 and 2024 editions are based on the MentalRiskES dataset (Mármol Romero et al., 2024), while data from the 2025 edition correspond to the PRECOM dataset (Álvarez-Ojeda et al., 2025).

4. Green Early Detection Evaluation Framework

To integrate predictive effectiveness, detection timeliness, and computational efficiency into a single evaluation framework, we adopt a normalized multi-criteria approach rather than a single ad-hoc scalar metric. This design choice follows established practices in Green AI and multi-objective evaluation, where heterogeneous dimensions are first normalized and then combined transparently.

Predictive Effectiveness Score We define a unified predictive effectiveness score that jointly captures classification performance and detection timeliness:

$$P = \text{Macro-F1} \cdot (1 - \text{ERDE}_{30}) \tag{1}$$

This formulation ensures that systems are rewarded only when they achieve both high predictive accuracy and early detection. Systems that detect risks late or inconsistently are penalized, even if their final classification performance is strong. This reflects the clinical motivation of early intervention scenarios, where timeliness is a first-class requirement.

The resulting score is bounded in the interval $[0, 1]$, facilitating comparison and aggregation with other normalized indicators.

Efficiency Score To account for computational efficiency, we normalize environmental indicators across all evaluated systems. Let X denote an efficiency-related metric, such as mean CO_2 emissions per prediction or mean energy consumption per prediction. We compute a normalized efficiency score as:

$$E_X = 1 - \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where $X_{\min}$ and $X_{\max}$ correspond to the minimum and maximum observed values across all submissions. This transformation yields an efficiency score in $[0, 1]$, where higher values indicate more efficient systems.

Importantly, efficiency scores are *not comparable across different years*, as each edition of the shared task relies on distinct datasets, experimental conditions, and submission pools. Consequently, normalization is performed independently for each year to ensure fair and meaningful comparisons only among systems evaluated under the same conditions.

This normalization strategy offers three advantages. First, it avoids mixing heterogeneous physical units directly with predictive metrics. Second, it constrains all scores to a common bounded range [0, 1], making them directly composable with the predictive effectiveness score P without requiring further scaling; scores derived from raw physical units (e.g., kWh or kg CO2eq) would otherwise dominate the composite metric due to differences

in magnitude. Third, it allows efficiency to be evaluated relative to the actual design space explored by participants in a given year, which is particularly appropriate in shared-task settings. We acknowledge that this relative normalization means scores depend on the composition of a given year's submission pool; a system's efficiency score may shift if the pool changes. This is a deliberate design choice: the goal is to rank systems fairly within a shared experimental context, not to assign absolute efficiency values independent of competition conditions.

Green Early Detection Score We combine predictive effectiveness and efficiency into a GED score using a weighted linear aggregation:

$$\text{GED} = \alpha P + \beta E_X, \quad \text{with } \alpha + \beta = 1 \tag{3}$$

The weights α and β explicitly control the trade-off between predictive effectiveness and efficiency. In our analysis, we consider two configurations:

- Performance-oriented Green Score: $\alpha = 0.7$, $\beta = 0.3$
- Balanced Green Score: $\alpha = 0.5$, $\beta = 0.5$

By making the weighting scheme explicit, this framework avoids hidden assumptions and allows organizers and practitioners to adapt the evaluation to different deployment priorities. Importantly, the resulting score remains bounded, interpretable, and suitable for leaderboard-based ranking. We emphasize that the GED is not intended to replace traditional performance metrics, but to complement them by making efficiency an explicit and actionable evaluation dimension.

4.1. Energy and Carbon-Aware Instantiations

Sustainability constraints vary substantially depending on the intended deployment context. In centralized evaluation settings, aggregate carbon emissions are often the primary concern, while in continuous and on-device monitoring scenarios, energy consumption per inference directly impacts battery life, latency, and usability.

To reflect these differences, we instantiate the proposed framework using two complementary efficiency indicators.

Carbon-Aware Green Score For evaluating global environmental impact, we define a carbon-aware Green Early Detection Score by setting $X = C$, where C denotes the mean CO_2-equivalent emissions per prediction:

$$E_{\text{CO}_2} = 1 - \frac{C - C_{\min}}{C_{\max} - C_{\min}} \tag{4}$$

$$\text{GED}_{\text{CO}_2} = \alpha P + \beta E_{\text{CO}_2} \tag{5}$$

This score favors systems that achieve strong early detection performance while minimizing carbon emissions, encouraging environmentally responsible model design at the system level.

Energy-Aware Green Score For on-device and real-time deployment scenarios, we instead consider mean energy consumption per prediction E as the relevant efficiency indicator:

$$E_{\text{energy}} = 1 - \frac{E - E_{\min}}{E_{\max} - E_{\min}} \tag{6}$$

$$\text{GED}_{\text{energy}} = \alpha P + \beta E_{\text{energy}} \tag{7}$$

This formulation directly rewards models that provide high-quality early detection under strict energy constraints. Notably, systems optimized for low carbon emissions are not necessarily optimal in terms of per-inference energy usage, making this distinction critical for realistic deployment assessment.

By maintaining a unified evaluation framework and varying only the efficiency indicator, we ensure methodological consistency while enabling context-aware sustainability evaluation. Rather than proposing a single universal metric, our approach supports multiple, transparent instantiations aligned with different operational priorities.

5. Green Leaderboard

The Green Leaderboard re-ranks all participating systems from the 2023, 2024, and 2025 editions of the MentalRiskES shared task using the proposed GED metric, incorporating both predictive performance and environmental efficiency. The following observations can be drawn from Tables 1, 2, and 3.

Comparing the official rankings, which are based solely on Macro-F1, with the GED-based rankings reveals notable differences. Several systems that did not occupy top positions in the traditional leaderboard achieve leading positions under GED, highlighting the importance of considering sustainability and efficiency alongside predictive effectiveness. This trend is particularly evident for lightweight machine learning (ML) models and compact neural architectures, which maintain competitive early detection performance while consuming significantly less energy and producing lower CO_2 emissions.

For most teams, GED_CO2 and GED_energy values are closely aligned, indicating that systems optimized for carbon efficiency are generally also energy-efficient per prediction. However, in more

complex tasks (e.g., Task 2c in 2023), discrepancies between GED_CO2 and GED_energy are observed. Some systems achieve high carbon efficiency but relatively lower energy efficiency, demonstrating that these two indicators capture complementary aspects of environmental performance.

Yearly Trends

- **2023:** CO_2 and energy rankings are largely aligned in Task 1a, whereas differences become more pronounced in Tasks 2c and 3a, highlighting the amplification of efficiency effects in complex tasks.
- **2024:** Teams such as ELiRF-UPV and UnibucAI consistently appear at the top of the Green Leaderboard, demonstrating that medium-sized models can compete with larger architectures when sustainability is considered.
- **2025:** New teams such as MCDI and PUXai appear among the top-ranked systems. Variability between GED_CO2 and GED_energy increases, indicating that optimizing solely for one environmental indicator does not necessarily maximize the overall GED score.

6. Sustainability Analysis

This section analyzes the relationship between predictive effectiveness, model complexity, and resource consumption across all systems submitted to the MentalRiskES shared tasks from 2023 to 2025. Rather than focusing on individual rankings, we examine global trends that emerge when sustainability indicators are considered jointly with performance metrics.

Performance vs. Energy Consumption Figure 1 illustrates the relationship between Macro-F1 and mean energy consumption per prediction across all submitted systems. A clear trade-off emerges: while some high-performing systems achieve strong Macro-F1 scores, they do so at substantially higher energy costs, often spanning several orders of magnitude.

Notably, the Pareto frontier reveals that competitive performance can be obtained with relatively low energy consumption. Several lightweight ML models and hybrid approaches (e.g., Transformer + ML) lie on or near the frontier, achieving Macro-F1 values comparable to larger transformer-based systems while consuming significantly less energy per inference. This suggests diminishing returns in performance when increasing model complexity beyond a certain point, particularly in continuous streaming settings.

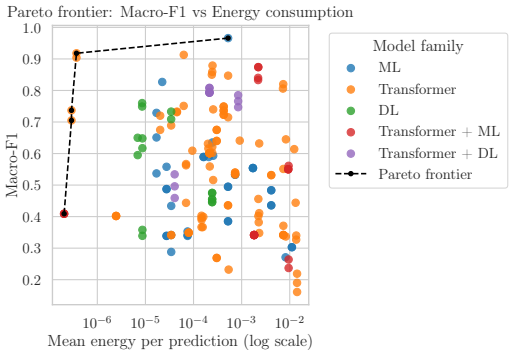

Figure 1: Pareto frontier between predictive performance (Macro-F1) and mean energy consumption per prediction (log scale) across all MentalRiskES submissions from 2023 to 2025. Points are grouped by model family. The frontier highlights systems that achieve competitive performance under strict energy constraints.

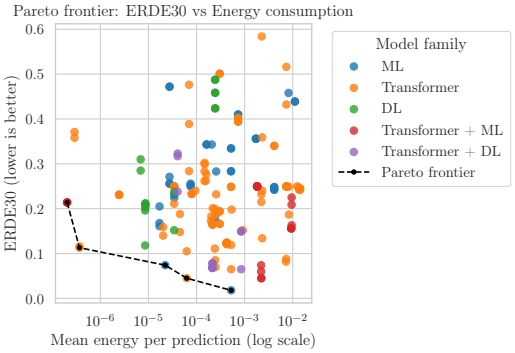

Figure 2: Pareto frontier between detection timeliness (ERDE30, lower is better) and mean energy consumption per prediction (log scale). The figure shows that early risk detection can be achieved without resorting to energy-intensive models.

Figure 2 further examines energy consumption in relation to ERDE30. Systems with lower energy usage tend to exhibit a wider range of timeliness behavior, but the Pareto frontier again highlights that early detection performance does not require energy-intensive architectures. Importantly, several energy-efficient systems achieve low ERDE30 values, indicating that prompt detection is compatible with strict energy constraints.

Performance vs. Emissions Figures 3 and 4 present analogous trends when mean CO_2 emissions per prediction are considered instead of energy consumption. As with energy, emissions span multiple orders of magnitude, reflecting substantial variability in hardware choices and computational strategies across teams.

Table 1: Top 10 Green Leaderboard for all tasks (2023) showing both GED_CO2 and GED_energy. The "Official Rank" column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while "CO2 Rank" and "Energy Rank" refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	P ↑	GED_CO2 ↑	GED_energy ↑
Task 1a									
1	1	1	CIMAT-NLP-GTO	0	**0.966**	**0.018**	**0.949**	**0.952**	**0.953**
2	2	3	UNSL	1	0.913	0.045	0.872	0.909	0.909
4	3	2	UMUTeam	0	0.918	0.113	0.814	0.870	0.870
3	4	5	VICOM-nlp	2	0.879	0.070	0.817	0.870	0.867
5	5	4	UMUTeam	1	0.904	0.116	0.799	0.859	0.859
6	6	6	VICOM-nlp	1	0.859	0.085	0.786	0.848	0.845
7	7	8	CIMAT-NLP-GTO	1	0.847	0.065	0.792	0.843	0.843
8	8	9	plncmm	0	0.827	0.074	0.766	0.836	0.836
9	9	7	VICOM-nlp	0	0.850	0.111	0.756	0.827	0.824
11	11	12	NLP-UNED	0	0.760	0.118	0.670	0.769	0.769
Task 2a									
1	1	2	UNSL	1	**0.733**	0.148	**0.625**	**0.736**	**0.736**
2	2	5	SINAI-SELA	0	0.720	**0.140**	0.619	0.733	0.733
3	3	4	TextualTherapists	1	0.729	0.161	0.612	0.728	0.728
4	4	3	UNSL	0	0.731	0.188	0.594	0.715	0.715
5	5	7	SINAI-SELA	1	0.675	0.159	0.568	0.697	0.697
6	6	8	TextualTherapists	0	0.651	0.168	0.542	0.679	0.679
7	8	12	VICOM-nlp	2	0.631	0.173	0.522	0.664	0.661
8	7	11	CIMAT-NLP-GTO	0	0.635	0.175	0.524	0.661	0.662
9	9	9	NLP-UNED	1	0.648	0.207	0.514	0.660	0.660
10	11	15	VICOM-nlp	1	0.616	0.183	0.503	0.651	0.648
Task 2c									
1	1	1	NLP-UNED	1	**0.358**	0.203	**0.285**	**0.500**	**0.500**
2	2	2	NLP-UNED	0	0.339	0.211	0.267	0.487	0.487
3	3	3	plncmm	0	0.288	0.232	0.221	0.454	0.454
4	4	4	I2C-UHU	0	0.232	**0.198**	0.186	0.422	0.419
5	5	5	SPIN	1	0.219	0.242	0.166	0.291	0.117
6	6	6	SPIN	0	0.190	0.245	0.143	0.276	0.101
7	7	7	SPIN	2	0.161	0.245	0.122	0.260	0.085
Task 3a									
1	1	1	CIMAT-NLP-GTO	2	**0.740**	**0.188**	**0.601**	**0.715**	**0.715**
2	2	2	NLP-UNED	1	0.650	0.285	0.465	0.625	0.625
3	3	5	CIMAT-NLP-GTO	0	0.593	0.283	0.425	0.592	0.592
4	4	4	NLP-UNED	0	0.595	0.310	0.411	0.587	0.587
5	5	6	CIMAT-NLP-GTO	1	0.516	0.232	0.396	0.572	0.572
6	6	8	UPM	0	0.402	0.231	0.309	0.516	0.516
6	6	8	UPM	2	0.402	0.231	0.309	0.516	0.516
6	6	8	UPM	1	0.402	0.231	0.309	0.516	0.516
9	9	3	CIMAT-NLP	0	0.614	0.250	0.460	0.342	0.361
10	10	7	CIMAT-NLP	1	0.444	0.247	0.334	0.234	0.254

The Macro-F1 vs. emissions plot (Figure 3) shows that high predictive performance is not exclusive to high-emission systems. Several low-emission models achieve performance levels close to the best-performing systems, again forming a well-defined Pareto frontier. This indicates that environmentally efficient solutions remain competitive in terms of accuracy.

Similarly, Figure 4 demonstrates that early detection timeliness does not systematically improve with higher emissions. In fact, many of the lowest-ERDE systems operate at relatively low emission levels, reinforcing the conclusion that carbon efficiency and early detection objectives are not inherently conflicting.

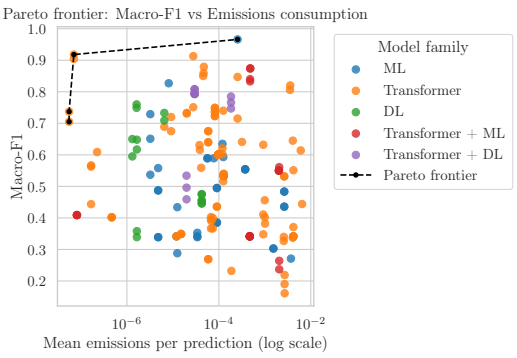

Figure 3: Pareto frontier between predictive performance (Macro-F1) and mean CO_2 emissions per prediction (log scale). Several low-emission systems achieve performance comparable to high-emission models, illustrating the trade-off between accuracy and environmental impact.

Temporal Evolution of Model Families Figure 5 shows the evolution of model family adoption across the 2023–2025 editions of the shared task. While transformer-based architectures remain prominent throughout all years, their dominance decreases slightly over time, coinciding with an increased presence of hybrid and lightweight machine learning approaches.

This trend suggests a gradual shift toward more efficiency-aware system design. In particular, the growing proportion of Transformer + ML systems indicates that teams increasingly combine representational power with computational efficiency, likely motivated by both sustainability concerns and deployment constraints.

Table 2: Top 10 Green Leaderboard for all tasks (2024) showing both GED_CO2 and GED_energy. The "Official Rank" column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while "CO2 Rank" and "Energy Rank" refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	P ↑	GED_CO2 ↑	GED_energy ↑
Task 1									
1	1	1	ELiRF-UPV	2	**0.874**	**0.045**	**0.835**	**0.845**	**0.824**
2	2	4	UnibucAI	0	0.808	0.078	0.745	0.819	0.816
3	3	5	UnibucAI	1	0.795	0.069	0.740	0.816	0.812
4	4	6	UnibucAI	2	0.793	0.068	0.739	0.815	0.812
5	6	3	ELiRF-UPV	0	0.833	0.060	0.783	0.809	0.789
6	7	2	ELiRF-UPV	1	0.840	0.074	0.778	0.806	0.785
7	5	7	UNED-GELP	0	0.785	0.065	0.734	0.799	0.791
8	8	9	Ixa-Med	1	0.749	0.124	0.656	0.753	0.748
9	9	11	Ixa-Med	2	0.736	0.121	0.647	0.746	0.741
10	10	8	UNED-GELP	2	0.766	0.149	0.652	0.741	0.733
Task 2									
1	1	1	ELiRF-UPV	0	**0.874**	**0.045**	**0.835**	**0.845**	**0.824**
2	2	2	UnibucAI	2	0.808	0.078	0.745	0.819	0.816
4	4	3	UnibucAI	1	0.793	0.068	0.739	0.815	0.812
3	3	3	UnibucAI	0	0.793	0.068	0.739	0.815	0.812
5	5	5	Ixa-Med	1	0.749	0.124	0.656	0.753	0.748
6	6	6	Ixa-Med	2	0.736	0.121	0.647	0.746	0.741
7	7	7	Ixa-Med	0	0.723	0.124	0.633	0.737	0.732
8	8	8	UMUTeam	2	0.675	0.166	0.563	0.689	0.686
9	9	9	UMUTeam	0	0.640	0.194	0.516	0.656	0.653
10	10	10	UC3M-DAD	0	0.601	0.165	0.502	0.643	0.645
Task 3									
1	1	1	UnibucAI	0	**0.534**	0.238	**0.407**	**0.583**	**0.584**
2	2	2	UnibucAI	1	0.496	0.317	0.339	0.536	0.536
3	3	5	V team	0	0.409	**0.214**	0.321	0.525	0.525
3	3	5	V team	2	0.409	**0.214**	0.321	0.525	0.525
3	3	5	V team	1	0.409	**0.214**	0.321	0.525	0.525
6	6	3	UnibucAI	2	0.459	0.323	0.311	0.516	0.516
7	7	4	UNED-GELP	0	0.456	0.215	0.358	0.470	0.490
8	8	8	UNED-GELP	1	0.402	0.232	0.309	0.435	0.456
9	9	9	UNED-GELP	2	0.382	0.584	0.159	0.329	0.350

Table 3: Top 10 Green Leaderboard for all tasks (2025) showing both GED_CO2 and GED_energy. The "Official Rank" column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while "CO2 Rank" and "Energy Rank" refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	P ↑	GED_CO2 ↑	GED_energy ↑
Task 1									
1	1	2	UNSL	0	0.563	0.284	**0.403**	**0.582**	**0.580**
2	2	1	UNSL	2	**0.567**	0.389	0.346	0.542	0.541
3	3	8	UC3Mental	1	0.495	0.334	0.329	0.524	0.510
4	4	15	UC3Mental	2	0.436	**0.249**	0.327	0.523	0.509
5	5	5	ELiRF-UPV	2	0.534	0.394	0.324	0.517	0.498
6	6	4	ELiRF-UPV	1	0.540	0.402	0.323	0.517	0.497
7	9	6	ELiRF-UPV	0	0.533	0.410	0.314	0.511	0.491
8	7	18	PUXai	2	0.396	0.283	0.284	0.494	0.494
9	8	17	PUXai	0	0.403	0.302	0.281	0.492	0.492
10	19	19	UC3Mental	0	0.385	0.283	0.276	0.487	0.473
Task 2									
1	1	4	HULAT_UC3M	1	0.558	**0.271**	**0.407**	**0.584**	**0.585**
3	3	1	MCDI	0	**0.589**	0.343	0.387	0.567	0.565
2	2	1	MCDI	1	**0.589**	0.343	0.387	0.567	0.565
4	4	1	MCDI	2	**0.589**	0.343	0.387	0.565	0.564
5	5	9	UC3Mental	1	0.495	0.334	0.329	0.524	0.510
6	6	15	UC3Mental	2	0.436	0.249	0.327	0.523	0.509
8	8	5	ELiRF-UPV	1	0.540	0.402	0.323	0.517	0.497
7	7	6	ELiRF-UPV	2	0.534	0.394	0.324	0.517	0.498
9	10	7	ELiRF-UPV	0	0.533	0.410	0.314	0.511	0.491
10	9	17	PUXai	2	0.399	0.277	0.289	0.497	0.497

Biases Induced by Ranking Figure 6 illustrates the model family composition of the Top-10 systems ranked exclusively by Macro-F1. Across all years, this ranking strategy disproportionately favors transformer-based and deep learning architectures, despite their often higher energy and emission costs.

In contrast, Figures 7 and 8 show the Top-10 composition under the carbon-based and energy-based GED rankings, respectively. When sustainability criteria are incorporated, the leaderboard becomes markedly more diverse, with a substantial increase in ML and hybrid models.

The effect is particularly pronounced under the energy-based GED, where lightweight models consistently replace more computationally intensive systems in the Top-10. This highlights how performance-only evaluation implicitly biases shared-task outcomes toward resource-heavy solutions, whereas sustainability-aware metrics expose a broader and more realistic design space.

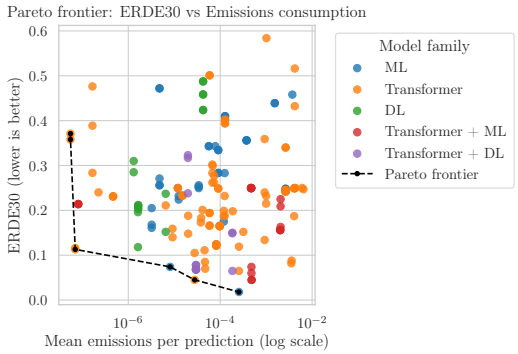

Figure 4: Pareto frontier between detection timeliness (ERDE30, lower is better) and mean CO_2 emissions per prediction (log scale). Results indicate that improved timeliness does not necessarily require higher carbon emissions.

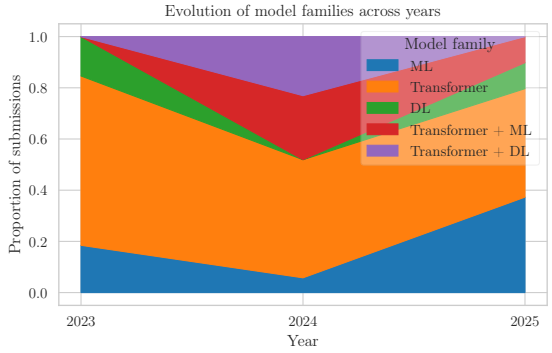

Figure 5: Temporal evolution of model family distribution across the 2023, 2024, and 2025 editions of the MentalRiskES shared task. While transformer-based models remain prevalent, hybrid and lightweight approaches gain prominence over time.

6.1. Implications for Streaming NLP Evaluation

Overall, the sustainability analysis demonstrates that evaluation metrics directly shape the types of models that are incentivized and rewarded. Performance-only leaderboards systematically favor computationally intensive systems, even when their advantages over more efficient alternatives are marginal. By contrast, sustainability-aware metrics reveal a broader and more realistic design space, in which multiple architectures achieve competitive early detection performance under strict resource constraints.

These findings support the inclusion of explicit efficiency-aware criteria in streaming NLP evaluations. Making sustainability visible at the ranking level encourages responsible model development and helps align shared-task outcomes with real-

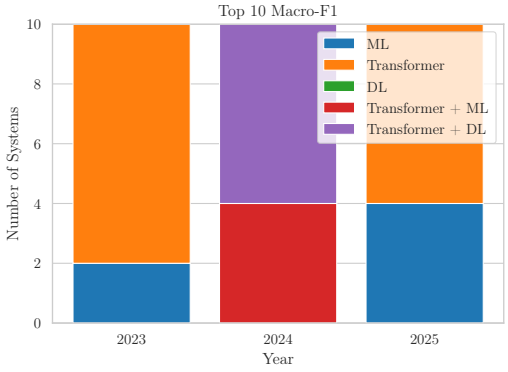

Figure 6: Model family composition of the Top-10 systems ranked exclusively by Macro-F1 for each year. Performance-only ranking disproportionately favors computationally intensive architectures.

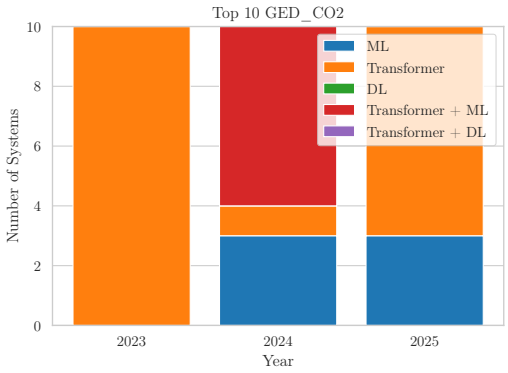

Figure 7: Model family composition of the Top-10 systems ranked using the carbon-based Green Early Detection Score (GED_{CO2}). Incorporating emissions into the ranking increases the representation of lightweight and hybrid models.

world deployment requirements, particularly in sensitive domains such as mental health monitoring.

7. Conclusion

In this paper, we introduced a sustainability-aware evaluation framework for streaming NLP tasks, addressing the limitations of performance-only leaderboards in early risk detection scenarios. We proposed the GED Score, a flexible and transparent metric that jointly accounts for predictive performance, detection timeliness, and environmental efficiency, with both carbon- and energy-aware instantiations.

By applying this framework to three consecutive editions (2023–2025) of the MentalRiskES shared task, we constructed the first Green Leaderboard for early mental health risk detection. Our analysis shows that incorporating sustainability criteria substantially reshapes system rankings, frequently

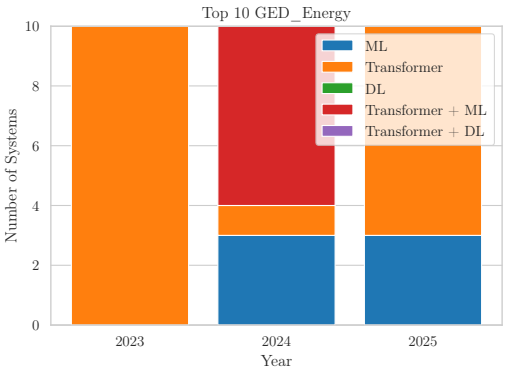

Figure 8: Model family composition of the Top-10 systems ranked using the energy-based Green Early Detection Score (GED_{Energy}). Energy-aware evaluation further amplifies the presence of efficient model families.

elevating lightweight and hybrid approaches that remain competitive in predictive effectiveness while dramatically reducing energy consumption and CO_2 emissions.

Importantly, we show that evaluation metrics directly influence the types of models incentivized in shared-task settings. Performance-only evaluation implicitly favors resource-intensive architectures, whereas sustainability-aware metrics expose a broader and more realistic design space aligned with real-world deployment constraints. This is particularly critical in sensitive domains such as mental health, where systems are expected to operate continuously, often on resource-constrained devices.

We argue that sustainability-aware evaluation is both feasible within existing shared-task infrastructures and necessary to encourage responsible model development. As future work, we plan to extend the proposed framework by incorporating additional efficiency indicators, such as inference latency and memory footprint, and by exploring normalization strategies that account for heterogeneous hardware configurations. We believe that making efficiency an explicit and first-class evaluation dimension is essential for advancing environmentally responsible and practically deployable NLP systems.

8. Acknowledgements

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by *Junta de Andalucía*.

9. Bibliography

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools. In *Proceedings of the second workshop on simple and efficient natural language processing*, pages 11–21.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos. 2024. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599:128096.

Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. 2024. mlco2/codecarbon: v2.4.1.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in nlp research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494.

Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-del Arco, María Dolores Molina-González, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña-López, and Arturo Montejo-Raéz. 2023. Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish. *Procesamiento del Lenguaje Natural*, 71:329–350.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, and Arturo Montejo-Ráez. 2024. Environmental impact measurement in the mentalriskes evaluation campaign. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability@ LREC-COLING 2024*, pages 61–72.

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. 2021. Codecarbon: estimate and track carbon emissions from machine learning computing. *Cited on*, 20.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.

María Estrella Vallecillo-Rodríguez, María Victoria Cantero-Romero, Isabel Cabrera-de Castro, Luis Alfonso Ureña-López, Arturo Montejo-Ráez, and María Teresa Martín-Valdivia. 2024. Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish. *Procesamiento del Lenguaje Natural*, 73.

Aimee Van Wynsberghe. 2021. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218.

10. Language Resource References

Álvarez-Ojeda, Pablo and Cantero-Romero, María Victoria and Semikozova, Anastasia and Montejo-Ráez, Arturo. 2025. *The precom-sm corpus: Gambling in spanish social media*.

Mármol Romero, Alba María and Moreno-Muñoz, Adrián and Plaza-Del-Arco, Flor Miriam and Molina-González, M. Dolores and Montejo-Ráez, Arturo. 2024. *MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish*. ELRA and ICCL.