Talk Is Cheap, Energy Is Not: Towards a Green, Context-Aware Metrics Framework for Automatic Speech Recognition

Maria Ulan[1] (✉), Erik Johannes Husom[2], and Jeriek Van den Abeele[3]

[1] RISE Research Institutes of Sweden, Gothenburg, Sweden maria.ulan@ri.se
[2] SINTEF Digital, Oslo, Norway erik.johannes.husom@sintef.no
[3] Telenor Research & Innovation, Fornebu, Norway
jeriek-van-den.abeele@telenor.com

Abstract. Automatic Speech Recognition (ASR) systems are increasingly deployed across diverse computing environments, from cloud servers to edge devices. While accuracy has traditionally been the primary evaluation metric, the inference efficiency of these systems, including energy consumption, memory usage, and hardware utilisation, significantly impacts their practical usability. This paper introduces a novel benchmarking framework that assesses ASR models during inference from both performance and sustainability perspectives. We introduce a multi-metric evaluation approach quantifying Word Error Rate (WER), Real-Time Factor (RTF), Energy Per Audio Second (EPAS), inference latency, GPU Memory Efficiency (GME), and Hardware Utilisation Rate (HUR). Our framework includes configurable weighting schemes tailored for various deployment scenarios: balanced general-purpose evaluation, resource-constrained environments, high-throughput batch inference, and real-time processing. To demonstrate the utility of the framework, we benchmark several state-of-the-art ASR architectures (Whisper, Wav2Vec2, HuBERT, WavLM, UniSpeech, and SpeechT5) in both FP16 and FP32 precision on NVIDIA Jetson AGX Orin hardware. The proposed methodology supports researchers and practitioners in making informed model selection decisions based on context-specific inference requirements. By illuminating performance–consumption trade-offs, the metrics framework can help to reduce computational costs and the carbon footprint of ASR systems, while maintaining acceptable accuracy.

Keywords: Green Machine Learning · Sustainability · Automatic Speech Recognition · Benchmarks

1 Introduction

Automatic speech recognition (ASR), the conversion of acoustic speech signals to text, is key for enhancing human–machine interaction. Advances in hardware, algorithms, and data have given rise to ASR models with impressive accuracy, as typically measured by the Word Error Rate (WER). Consequently, a variety of speech-based applications is nowadays deployed across the edge–cloud continuum, from cloud-hosted virtual customer service agents and large-scale transcription services, to on-device applications like offline voice translation on smartphones and voice command recognition in smart speakers, and increasingly, in wearables and ultralow-power IoT devices. The deployment of AI models closer to the edge is driven by a growing demand for enhanced privacy, reduced latency, and increased energy efficiency.

The surging popularity of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has increased scrutiny to the energy consumption of AI systems [9, 40]. While most Green AI research efforts focus on the training phase [39], Google data from 2019 to 2021 indicated that about 60% of the energy consumption for their machine learning systems came from the inference phase [30], and this was before the boost in public LLM adoption triggered by OpenAI's ChatGPT release. Although a typical ASR inference task may not require as much energy as, for instance, image generation or captioning [22], when performed at scale, the environmental impact can become substantial. For example, transcribing a whole workday's worth of calls for every agent in a customer service centre accumulates a considerable energy cost. After all, ASR pipelines are computationally demanding: they typically consist of an acoustic model for inferring phoneme sequences from audio signals, a lexical model describing word pronunciations, and a language model estimating the probability of word sequences for enhanced transcription accuracy.

While various ASR model architectures reach high WER scores on benchmark datasets [36], in practice model selection requires a holistic view, going beyond accuracy. Real-world ASR deployment requires considering often underemphasised factors like inference latency, computational efficiency, memory usage, and energy consumption. As ASR systems become gradually more embedded in resource-constrained and environment-sensitive contexts, these dimensions of inference efficiency directly impact the practical usability, sustainability, and operational costs of ASR solutions. Therefore, understanding and balancing trade-offs between performance and resource use is essential to select the appropriate ASR models for specific deployment scenarios.

This paper presents a new multi-metric framework for evaluating the ASR systems in the inference phase, aiming to provide practitioners with a straightforward methodology for assessing which ASR model is most suitable for deployment in specific usage contexts. Our approach considers WER accuracy alongside sustainability- and efficiency-oriented metrics: Real-Time Factor, Inference Latency, Energy Per Audio Second, GPU Memory Efficiency and Hardware Utilisation Rate. We introduce metric weighting schemes to accommodate diverse deployment scenarios, including resource-constrained environments, and real-time or batch processing. Finally, we benchmark several state-of-the-art ASR architectures (Whisper, Wav2Vec2, HuBERT, WavLM, UniSpeech, and SpeechT5) on NVIDIA Jetson AGX Orin hardware, illustrating how the framework supports informed, context-aware model selection that balances accuracy, performance, and sustainability.

2 Related Work

The environmental impact of artificial intelligence has gained significant attention following seminal research by Strubell et al. [37], which documented the substantial carbon emissions associated with training large neural networks. Schwartz et al. [34] introduced the concept of "Green AI" to contrast "Red AI," which prioritises accuracy and capability over efficiency. They argue for incorporating energy usage and computational cost as primary evaluation metrics alongside traditional performance measures, introducing a more holistic evaluation paradigm. These studies were some of the first of a growing body of research on sustainable computing practices within AI development.

While many studies focused primarily on the training stage [25, 29, 30, 37], large-scale deployment and use of AI models may lead to substantially larger inference costs due to the cumulative impact of numerous inference requests over the lifetime of a deployed model. Wu et al. [44] demonstrated that both training and inference phases contribute significantly to the overall carbon footprint of machine learning applications, with relative proportions varying across different use cases and implementation scenarios. This variability highlights the importance of context-specific energy analyses rather than generalised approaches to environmental impact assessment. Luccioni et al. [20] called for expanding the analysis of environmental impacts across the entire ML lifecycle to include the costs during deployment and inference. Investigations into the inference phase have identified that task-specific models typically demonstrate better energy efficiency compared to multi-purpose alternatives used for the same tasks, encouraging the use of specialised models over general-purpose large models when task requirements permit [21].

The quantification of AI systems' energy usage requires robust monitoring frameworks, which have evolved considerably in recent years. Henderson et al. [12] made important contributions by establishing a framework for environmental accountability through systematic documentation of energy consumption throughout the AI development process. Several additional teams [1, 4, 10, 18, 19] have developed methodologies to estimate or track the energy consumption of AI. These approaches vary in granularity, hardware compatibility, and methodology, often yielding inconsistent results that complicate systematic comparisons between models and systems.

The hardware compatibility constraints of energy monitoring tools present significant challenges for comprehensive energy assessment. Most tools for energy monitoring of a system's internal components only support Intel CPUs and NVIDIA graphics cards [1, 4, 10], as they rely on the manufacturers' proprietary monitoring interfaces, RAPL and NVML/`nvidia-smi`, respectively. While these represent common hardware choices, this dependency limits applicability across deployment scenarios such as edge computing. Moreover, Yang et al. [45] identified significant limitations in `nvidia-smi`'s accuracy, showing a ±5% error margin (versus the claimed ±5W), which can lead to substantial measurement discrepancies on high-power GPUs. Their study also revealed that newer GPU architectures only sample power during 25% of runtime, leaving most power fluctuations unmonitored. Software-based methods additionally have the inherent limitation of relying on power models for providing their metrics [16], meaning that they provide estimates rather than direct measurements.

The AI Energy Score Leaderboard [22] is a benchmarking initiative evaluating models based on standardised power efficiency metrics. While important for promoting transparency in AI energy consumption and a useful high-level benchmark, it has notable methodological limitations. First, the AI Energy Score focuses exclusively on GPU power consumption, without accounting for significant system-wide energy factors—a shortcoming shared with other energy assessment studies [33]. The creators of the benchmark acknowledge that "CPU and RAM usage was found to be approximately 30% greater than GPU energy use," yet these components remain excluded from their primary metrics. Second, given `nvidia-smi`'s aforementioned accuracy limitations and sampling gaps, these GPU-centric measurement approaches may compound measurement errors with incomplete system coverage, potentially leading to significant underestimation of actual energy consumption. Finally, such benchmarks typically isolate energy efficiency from other critical performance indicators, and the AI Energy Score creators note that users should independently consider "throughput, accuracy, and latency" alongside efficiency metrics.[4]

Despite the growing focus on Green AI, relatively few studies have specifically addressed energy consumption in Automatic Speech Recognition (ASR) systems. Parcollet et al. [28] investigated the carbon footprint of training end-to-end speech recognisers, quantifying CO_2 emissions during model training and highlighting how minimal performance improvements often come at extremely high environmental costs. However, their work focused primarily on the training phase rather than inference.

Chakravarty [7] conducted one of the few studies examining ASR inference energy consumption, specifically for edge deployment. This work measured energy consumption for various ASR models on the NVIDIA Jetson Orin Nano, analysing the effects of model quantisation, precision levels, and noise on performance and energy efficiency. The study found that changing precision from FP32 to FP16 halved energy consumption across different models with minimal performance degradation, and that larger model size does not necessarily predict better noise resilience or energy consumption patterns.

While these studies provide valuable insights into ASR energy consumption, they lack a cohesive method evaluating both performance and sustainability. Most existing work focuses either on single-metric evaluations or fails to provide context-specific evaluations tailored to different deployment scenarios. Additionally, many studies do not adequately address the trade-offs between energy efficiency, accuracy, and other performance metrics that practitioners must navigate when deploying ASR systems.

[4] Thus, while the AI Energy Score offers meaningful ratings, our objective is to enable the level of granularity needed for deployment-specific trade-off analysis across various performance and sustainability factors. Differences in energy measurement methodology preclude direct comparison with our framework.

3 Benchmark Framework

Recent ASR models perform remarkably well, reaching accuracy levels comparable to human annotations [36]. However, these advances often come with significant computational and energy costs. Our benchmarking framework addresses this fundamental trade-off by providing a comprehensive evaluation approach that integrates both traditional performance metrics and sustainability considerations. We design the framework around four key principles: *(1)* moving beyond accuracy-only assessment to encompass efficiency metrics, *(2)* recognising that different deployment scenarios prioritise different aspects of performance, *(3)* ensuring consistent and reliable measurements across test runs, and *(4)* providing insights that are actionable for researchers and practitioners.

Moreover, our approach acknowledges that ASR deployment occurs across diverse computing environments, from cloud servers to edge devices. Each environment has unique constraints and different evaluation criteria for different priorities. For edge deployment, specifically, where battery life and thermal management are critical concerns, energy efficiency plays a crucial role.

The framework follows a modular pipeline architecture with four main components. The *Inference Engine* executes ASR models on audio inputs while measuring computational metrics such as inference time, latency, and accuracy. It ensures that the model's real-world performance is evaluated effectively. The *Power Monitoring module* collects data on the power consumption in real time during inference. By tracking energy usage, it helps to assess the efficiency of ASR models and optimise them for deployment on various hardware platforms. The *Metrics Aggregator* collects and combines performance and efficiency metrics from multiple sources. Collect data related to ASR accuracy, processing speed, and power consumption, providing a comprehensive evaluation of the effectiveness of each model. The *Weighted Scoring System* applies context-specific weightings to different performance metrics to generate deployment-optimised scores. This enables informed decision-making by prioritising models that best meet the specific requirements of a given use case.

3.1 Metrics Definition

The framework measures six key metrics that comprehensively characterise ASR system performance.

Word Error Rate (WER) is a standard accuracy metric for ASR that measures the edit distance between the reference and hypothesised transcripts [24, 43]. The ASR model inference output is a text file, each corresponding to a single audio input file. To ensure consistency between characters in prediction and ground truth datasets, we apply a normalisation process where punctuation marks, special characters, and capitalisation are removed. WER is expressed as a percentage, and calculated as: $\text{WER} = \frac{\text{Substitutions+Deletions+Insertions}}{\text{Number of Words in Reference}}$.

Real-Time Factor (RTF) is the dimensionless ratio of the processing time to the audio duration, indicating how much faster (or slower) the system operates compared to real-time [23]. This metric is crucial for assessing whether an ASR system can keep up with live audio input.

Energy Per Audio Second (EPAS) is a metric that quantifies the energy required to process one second of audio, measured in joules per second. This metric was inspired by the energy-per-token metric used for profiling energy consumption in LLM inference [15, 33]. EPAS is calculated as the ratio of total energy consumption to total audio duration.

Inference Latency (in milliseconds) is computed as the 95th percentile of the processing times of all audio segments, providing insight into the worst-case responsiveness. The 95th percentile is more useful than the maximum latency, because it filters out rare outliers, while still representing worst-case scenarios. It is also known as tail latency [46].

GPU Memory Efficiency (GME) is the ratio of active GPU memory to total allocated GPU memory, expressed as a percentage.

Hardware Utilisation Rate (HUR) is a measure for the average utilisation of the compute resources (CPU and GPU), providing insight into balanced resource usage. The HUR is expressed as a percentage, and is calculated by taking the mean of the average GPU utilisation and average CPU utilisation.

Lower WER indicates better accuracy, and lower RTF indicates more efficient processing. When RTF < 1, the system processes audio faster than it would take to play it. Lower EPAS and Inference Latency values are desirable. Higher GME values are better, values close to 100% indicate effective use of allocated GPU memory. Balanced, high HUR (60–90%) is ideal: lower values suggest underutilisation, higher HUR may indicate bottlenecks and high energy usage.

3.2 Weights and Aggregation

We use min–max normalisation to transform the raw metrics to a 0–1 scale where 1 represents the best performance. For metrics where lower values are better (WER, RTF, EPAS, Latency), we invert the normalised values. For HUR, we use a piecewise normalisation that assigns optimal scores (0.8–1.0) to the balanced utilisation range (60–90%), with lower scores for both underutilisation (<60%) and overutilisation (>90%). This approach rewards efficient resource usage while penalising potential bottlenecks and wasteful underutilisation.

We define a green score as a weighted sum of normalised metrics:

$$\text{Green Score} = \mathbf{w} \cdot \mathbf{m}_{\text{norm}} = \sum_{i=1}^{6} w_i \cdot m_i,$$

where $\mathbf{w} = [w_{\text{WER}}, w_{\text{RTF}}, w_{\text{EPAS}}, w_{\text{lat}}, w_{\text{GME}}, w_{\text{HUR}}]$ and the components of $\mathbf{m}_{\text{norm}}$ are ordered analogously.

In our framework we use configurable weighting schemes that adapt the evaluation to different deployment contexts. Rather than providing a single score, we define four weighting schemes, shown in Table 1, that reflect common ASR deployment scenarios:

Table 1. Weighting scheme coefficients reflecting different deployment contexts

Scenario	w_{WER}	w_{RTF}	w_{EPAS}	w_{lat}	w_{GME}	w_{HUR}
Balanced	0.25	0.20	0.20	0.15	0.10	0.10
Mobile	0.15	0.20	0.30	0.05	0.25	0.05
Real-time	0.25	0.25	0.10	0.30	0.05	0.05
Server	0.25	0.10	0.35	0.00	0.20	0.10

Balanced General-Purpose Evaluation (Balanced) is designed for general-purpose applications with balanced requirements. In this scenario, accuracy is always important; processing speed matters in most cases; energy efficiency is highly beneficial across all cases; responsiveness is crucial mostly in interactive systems; and memory efficiency and efficient hardware utilisation are beneficial but less critical.

Resource-Constrained Environments (Mobile) is optimised for battery-powered and edge devices with limited resources. In this scenario, accuracy still matters, but some degradation is acceptable; responsiveness is essential despite limited resources; energy efficiency is critical for battery-powered devices; latency is less important than overall energy efficiency; memory constraints are significant; and balanced utilisation is important, but not a primary concern.

Real-Time Processing (Real-time) is optimised for applications requiring immediate responses, like voice assistants and interactive systems. In this scenario, high accuracy for user experience is important; the system must process faster than real-time; energy matters but is secondary to responsiveness; immediate response is critical; memory efficiency is less critical for devices with sufficient RAM; and hardware utilisation is a minor concern.

High-Throughput Batch Inference (Server) is optimised for large-scale cloud deployments. In this scenario, accuracy is still important; throughput matters, but time sensitivity is lower; energy efficiency is crucial for managing operational costs; per-request latency is irrelevant for batch jobs; memory efficiency determines how many models fit on a server; and balancing CPU/GPU load enhances computational performance.

4 Experimental Setup

We conducted experiments on several ASR models obtained from Hugging Face. We evaluated the models with half- and single-precision settings, and measured and aggregated metrics from the proposed benchmark framework. We used high-quality English speech recordings from public-domain readings and a diverse, crowdsourced voice collection with broad demographic coverage. To assess robustness, we also injected synthetic noise into the clean speech samples.

Table 2. Technical hardware specifications

Component	Specification
GPU	NVIDIA Ampere architecture (cores: 2048 NVIDIA CUDA, 64 Tensor)
CPU	12-core Arm Cortex-A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3, 2.2GHz max frequency
Memory	32GB LPDDR5 RAM
Storage	64GB eMMC 5.1 Flash Storage
Power	Configurable TDP from 15W to 60W
Dimensions	100mm × 79mm × 21mm

4.1 Hardware Platform

All experiments were conducted on the NVIDIA Jetson AGX Orin Developer Kit, which represents a high-performance edge computing platform, Table 2 shows its technical specifications. The Jetson AGX Orin was chosen for its ability to represent both edge and small server deployment scenarios, with sufficient computational power to run all the evaluated models, while still being constrained enough to reveal meaningful differences in efficiency metrics. Its lightweight design allows for seamless integration into smaller devices, while still providing the computational capabilities necessary to handle complex loads without relying on cloud computing resources.

The device was interfaced through wired keyboard and mouse, with the JetPack 6.2 [L4T 36.4.3] OS providing the graphical user interface. For power measurements, we utilised the built-in INA3221 power monitors [38] on the Jetson platform, which provide accurate measurements of system-on-chip (SoC) and memory subsystem power consumption.

4.2 Implementation Details

Our framework is implemented as a collection of Python modules that can be easily extended to accommodate new models and metrics. It includes *ASR engine wrappers*: custom scripts for each model architecture, providing a unified interface for model loading, inference, and metrics collection; a *power monitoring daemon*: a background process that collects power consumption data using platform-specific utilities; and a *metrics aggregator*: a data processing pipeline that combines metrics from various sources, handles normalisation, and applies weighting schemes.

The entire framework is designed to be portable across different computing environments, with currently optimised support for NVIDIA Jetson platforms. The source code is publicly available[5], allowing researchers to replicate our experiments and enhance the framework to suit their specific needs.

We employ continuous sampling of power consumption using the NVIDIA `tegrastats` utility [26] at 200 ms intervals, which provides fine-grained power

[5] `https://github.com/ulmarise/asr-green-metrics-framework`

measurements for Jetson platforms. Baseline measurements of RAM and power are first established during device inactivity, and these values are subtracted from the total consumption to calculate the specific resources allocated to ASR tasks. Moreover, we utilise PyTorch's built-in CUDA memory tracking capabilities combined with system-level memory monitoring to calculate memory efficiency. We track both allocated and active memory to understand utilisation patterns.

For consistent and reliable metrics collection, our framework implements the following methodology. For WER calculation, we use the `jiwer` library [17] with a custom normalisation procedure to handle punctuation, capitalisation, and spacing consistently. This ensures fair comparisons across different model architectures and output styles. We set up model inference to capture the exact start and end times for each audio segment, ensuring the accurate calculation of RTF and Inference Latency metrics. Each audio file is processed separately to provide per-file metrics for statistical analysis. For EPAS calculations, we isolated the incremental energy consumption attributable to the ASR model by subtracting baseline power draw. GME and HUR metrics were calculated from memory usage statistics and from CPU and GPU percentages in `tegrastats` output.

4.3 ASR Model Selection

To demonstrate the utility of our framework, we selected six state-of-the-art ASR architectures representing a diverse range of model families:

1. **Whisper**, a transformer-based encoder–decoder model trained on a massive multilingual dataset. We used the distilled version [11], which maintains performance with reduced computational requirements. Specifically, we used a compact version optimised for English speech recognition with reduced parameters yet similar performance ; a medium-sized version that balances computational efficiency and transcription accuracy ; and the largest variant in the Distil-Whisper family, offering high-quality speech recognition with significantly fewer parameters than the original Whisper Large-v2 model.
2. **Wav2Vec2** [5], a self-supervised model that learns representations directly from raw audio, using a convolutional feature encoder followed by a transformer. We used Facebook Wav2Vec2 Large 960h, a self-supervised speech recognition model pre-trained on 960 hours of data, featuring robust performance on diverse speech inputs.
3. **HuBERT** [14], a hidden-unit BERT model that learns representations from clustered audio features, with a mask prediction objective. We used Facebook HuBERT Large LS960 FT, a model pre-trained on unlabelled audio and fine-tuned on 960 hours of labelled speech for improved representation learning.
4. **WavLM** [8], an evolution of the HuBERT approach incorporating masked language modeling and denoising objectives. We used WavLM Libri-Clean 100h Base Plus, an enhanced audio representation model that builds upon the Wav2Vec framework, fine-tuned on 100 hours of clean data.

5. **UniSpeech** [42], a unified pre-training framework that combines self-supervised and supervised learning. We used Microsoft UniSpeech-SAT Base 100h Libri FT, a speech representation model leveraging self-supervised and semi-supervised training approaches, fine-tuned on 100 hours of data.
6. **SpeechT5** [2], a unified-modal encoder-decoder framework that handles both speech and text. We used Microsoft SpeechT5 ASR, a unified text-to-speech and speech-to-text transformer model based on the T5 architecture, specifically optimised for automatic speech recognition tasks.

For each model, we utilised the implementations available through the Hugging Face library, which provides consistent APIs and optimised CUDA support. Each model was tested in both FP32 (full precision) and FP16 (half precision) to evaluate the performance–efficiency trade-offs of reduced precision. Each model's inference code was configured to capture timing and resource utilisation metrics, with detailed logs generated for subsequent analysis.

4.4 Dataset Characteristics

We used the benchmark framework to evaluate two speech datasets. The first is a smaller benchmark: a subset of the LibriSpeech dataset [27], consisting of 70 audio tracks from the test-clean partition. The second dataset is larger and more diverse, is an English subset of the Common Voice dataset [3], comprising 3995 audio tracks from the valid-test partition.

The LibriSpeech audio data is stored in FLAC-encoded (lossless), sampled at 16 kHz with a mono channel. Each segment lasts 3–5 seconds, totaling approximately 328 seconds. The recordings feature high signal-to-noise ratio speech, balanced gender distribution, and native English speakers reading public domain audiobooks. The dataset size was chosen to be large enough to provide statistically significant results while remaining manageable for repeated evaluations across multiple models and configurations.

To evaluate the model's ability to handle acoustic interference, we considered a noisy version of the dataset. It was created using Gaussian white noise with a mean of zero and a standard deviation of one, ensuring it matched the length of each audio track. The noise amplitude was reduced by 10 dB to limit its impact on the original signal. Then the noise was combined with the audio signal to produce a composite track that simulates real-world audio corruption in noisy environments. This approach allowed us to systematically evaluate performance degradation under consistent noise conditions.

In addition to LibriSpeech, we include the Common Voice dataset to introduce greater variability in speaker demographics, recording conditions, and background noise levels. This enables evaluation under more realistic, less controlled conditions. The audio files are in MP3 format, sampled at 48 kHz mono. Segments range 1–28 seconds, totalling around 5 hours of audio. Recordings are crowd-sourced with moderate noise levels, covering global English speakers with balanced gender distribution, reading from open-domain text prompts. Audio files were used as-is, but downsampled to 16 kHz.

4.5 Evaluation Details

Our evaluation followed these steps for each model and precision configuration. First, each model was loaded into memory with the specified precision (FP16 or FP32). Loading time was measured but not included in the inference metrics. Second, ten audio samples (not part of the dataset) were processed to prime the model and stabilise GPU memory allocation. Third, during the inference phase, each audio file was processed sequentially with the following measurements: processing time for each audio sample, memory utilisation tracked at 200 ms intervals, power consumption recorded continuously, and output transcriptions saved for subsequent WER calculation. Fourth, raw measurements were processed to calculate the six core metrics (see Sec. 3.1). Fifth and finally, the four weighting schemes were applied to generate context-specific Green Scores.

To ensure reliability, each experiment was conducted three times, and the median values were used for the final results. The ambient temperature was maintained at 22°C ± 1°C to minimise thermal variability. The Jetson AGX Orin was configured in MAXN power mode to ensure consistent maximum performance across all tests.

5 Results and Analysis

We first examine the overall performance trends under clean conditions, followed by an analysis of noise robustness, where artificial noise is added. This is then complemented by an evaluation under natural conditions using real-world noisy speech. Table 3 presents the evaluation metrics, and Table 4 shows the corresponding green scores computed with various weighting schemes for ASR models evaluated on clean and artificially noisy speech. Table 5 presents the metrics and green scores for ASR models evaluated on speech with real-world noise.

On clean LibriSpeech data, HuBERT yields the best transcription accuracy, while Distil Whisper Large excels with noise. Precision format (FP16 vs. FP32) minimally impacts accuracy but greatly affects performance metrics. UniSpeech is fastest overall, with Distil Whisper Small/Medium (FP16) being the most efficient among transformer models. FP16 models consistently outperform FP32 in speed and energy use, with WavLM, UniSpeech, and Wav2Vec2 (FP16) showing the highest energy efficiency. Larger models are predictably less memory-efficient, with Distil Whisper Large (FP32) achieving the most balanced hardware utilisation.

HuBERT FP32 achieves the highest balanced green score on clean LibriSpeech data, challenging the notion that FP16 models are always more efficient. Distil Whisper FP16 models offer good accuracy–efficiency balance, while SpeechT5 and WavLM rank lowest. For mobile/edge deployment, UniSpeech and HuBERT FP32 excel. In real-time scenarios, UniSpeech FP16 and HuBERT FP32 perform best. Server-side applications favour HuBERT FP32 and Distil Whisper Large FP16, with the latter providing strong WER performance. Green scores remain consistent between clean and noisy LibriSpeech data, suggesting the aggregation approaches may underweight audio quality sensitivity.

On Common Voice data, Distil Whisper Large achieves the best transcription accuracy, followed by Distil Whisper Medium and Small. UniSpeech has the fastest processing speed overall, while WavLM and traditional transformer models (Wav2Vec2, HuBERT) show moderate performance. FP16 models consistently show superior energy efficiency, with UniSpeech, WavLM, and SpeechT5 showing the lowest energy consumption per audio second. FP32 variants of Wav2Vec2, HuBERT, and WavLM achieve lower latency, while Distil Whisper models exhibit higher latency but maintain competitive speed-accuracy trade-offs. Larger Distil Whisper models predictably consume more GPU memory, with the Large variant (FP32) reaching the highest memory utilisation, while UniSpeech and WavLM maintain more balanced resource use across precisions.

HuBERT FP32 dominates on Common Voice data with the highest balanced green score, demonstrating that FP32 models can be more efficient despite their higher precision. Wav2Vec2 FP32 and Distil Whisper Medium FP16 also perform strongly in balanced scenarios. For mobile/edge deployment, HuBERT (both precisions) and Wav2Vec2 FP32 excel, while UniSpeech maintains consistent performance across precision types. In real-time scenarios, HuBERT FP32 performs the best, with Wav2Vec2 FP32 and Distil Whisper Medium FP16 close behind. Server-side applications favour HuBERT models. SpeechT5 consistently ranks lowest across all categories, with particularly poor real-time performance.

Comparing results across both datasets shows consistent patterns in model performance. HuBERT FP32 demonstrates superior efficiency-accuracy trade-offs, achieving the highest balanced green scores on both LibriSpeech and Common Voice. UniSpeech FP16 maintains the lowest energy consumption and processing time across datasets, while Distil Whisper Large offers the best accuracy on Common Voice but is outperformed by HuBERT on LibriSpeech. The relative performance ranking of models remains largely consistent across datasets, suggesting that our green score metrics robustly capture model characteristics rather than dataset-specific features.

Given a specific precision format, our results may help identify the most economical model for various deployment scenarios. For FP16 precision, UniSpeech offers optimal resource efficiency, Distil Whisper models provide the best accuracy–efficiency balance for real-time applications, and HuBERT delivers superior server performance. For FP32 precision, HuBERT consistently demonstrates the best overall performance, particularly for real-time scenarios, while UniSpeech and WavLM maintain excellent resource efficiency. These findings challenge the common assumption that lower precision formats always yield more efficient models, as demonstrated by HuBERT FP32's exceptional performance across both datasets. It is important to note that these recommendations are specific to the hardware platform and dataset characteristics used in the experiments. Performance may vary with different hardware configurations, audio conditions, or application requirements.

Table 3. LibriSpeech metrics for various ASR models with different precision formats — for clean and, in parentheses, for noisy speech data

Model	Prec.	WER	RTF	EPAS	Latency	GME	HUR
Distil-Whisper-S	FP16	3.48 (17.19)	0.127 (0.125)	0.79 (0.64)	0.83 (0.79)	12.67 (12.45)	31.60 (22.68)
	FP32	3.70 (17.19)	0.182 (0.181)	3.29 (3.10)	0.91 (0.90)	12.64 (12.47)	29.45 (33.40)
Distil-Whisper-M	FP16	4.13 (14.25)	0.123 (0.120)	1.58 (1.38)	0.65 (0.64)	15.62 (15.25)	28.75 (26.86)
	FP32	3.92 (14.25)	0.248 (0.253)	11.36 (11.11)	1.18 (1.18)	18.04 (17.80)	36.94 (35.81)
Distil-Whisper-L	FP16	4.03 (12.40)	0.152 (0.152)	3.26 (3.56)	0.85 (0.86)	20.48 (20.27)	34.30 (30.68)
	FP32	4.03 (12.40)	0.417 (0.418)	23.46 (23.83)	1.97 (2.00)	27.09 (26.90)	39.92 (40.80)
Wav2Vec2	FP16	4.13 (38.96)	0.378 (0.380)	0.23 (0.25)	1.81 (1.83)	20.71 (20.32)	6.02 (16.55)
	FP32	4.13 (39.17)	0.094 (0.096)	1.91 (1.73)	0.29 (0.30)	19.74 (19.52)	18.50 (19.03)
HuBERT	FP16	2.39 (10.88)	0.387 (0.394)	0.26 (0.23)	1.86 (1.86)	20.10 (19.99)	10.02 (3.83)
	FP32	2.39 (10.88)	0.093 (0.098)	1.97 (1.85)	0.33 (0.32)	19.49 (19.24)	29.28 (21.52)
WavLM	FP16	12.5 (63.76)	0.32 (0.332)	0.18 (0.23)	1.79 (1.78)	20.70 (20.51)	5.72 (15.66)
	FP32	12.5 (63.76)	0.07 (0.063)	0.68 (0.65)	0.16 (0.15)	19.87 (19.88)	17.30 (25.06)
UniSpeech	FP16	6.31 (31.23)	0.054 (0.052)	0.26 (0.23)	0.08 (0.07)	20.10 (19.77)	12.87 (26.86)
	FP32	6.31 (31.23)	0.058 (0.061)	0.53 (0.68)	0.13 (0.12)	19.27 (19.11)	23.62 (12.71)
SpeechT5	FP16	11.9 (27.31)	0.351 (0.356)	0.34 (0.37)	2.88 (2.88)	20.59 (20.25)	14.18 (13.94)
	FP32	11.9 (27.31)	0.360 (0.361)	0.80 (0.79)	2.98 (3.00)	20.10 (19.83)	26.15 (18.13)

6 Discussion

Our evaluation across LibriSpeech and Common Voice datasets reveals significant insights for ASR model selection. While FP16 models generally offer better energy efficiency and speed, some FP32 models achieve better overall green scores due to superior accuracy and balanced hardware utilisation. HuBERT FP32 consistently demonstrates exceptional efficiency-accuracy balance across both datasets, while UniSpeech FP16 excels in resource-constrained scenarios. For practical deployments, UniSpeech FP16 is optimal for energy and latency constraints, HuBERT FP32 delivers the best balanced performance, and Distil Whisper models offer strong accuracy with reasonable efficiency. Precision-specific analysis directly addresses which model is most economical given particular constraints, challenging the assumption that energy efficiency necessarily compromises accuracy. These findings emphasise the importance of multi-dimensional ASR evaluation frameworks, as the traditional focus on WER alone fails to capture the complex trade-offs in real-world deployments, particularly for resource-constrained environments and large-scale use, where energy considerations are increasingly critical.

Limitations While our evaluation provides valuable insights into ASR model performance, some limitations should be noted. Firstly, while our green scoring system provides a useful aggregated metric for model comparison, we acknowledge that such aggregation approaches may potentially mask poor performance in individual metrics. The relative importance of different metrics also varies according to specific application requirements. Users should always examine individual metric values alongside green scores to ensure that models meet specific

Table 4. LibriSpeech green scores — for clean and, in parentheses, for noisy speech data

Model	Prec.	Balanced	Mobile	Realtime	Server
Distil-Whisper-S	FP16	0.73 (0.72)	0.64 (0.64)	0.76 (0.76)	0.69 (0.67)
	FP32	0.67 (0.68)	0.58 (0.58)	0.70 (0.71)	0.62 (0.64)
Distil-Whisper-M	FP16	0.74 (0.76)	0.68 (0.70)	0.77 (0.80)	0.70 (0.72)
	FP32	0.59 (0.61)	0.53 (0.54)	0.61 (0.63)	0.56 (0.59)
Distil-Whisper-L	FP16	0.74 (0.76)	0.73 (0.74)	0.75 (0.78)	0.74 (0.77)
	FP32	0.41 (0.45)	0.42 (0.44)	0.39 (0.42)	0.46 (0.50)
Wav2Vec2	FP16	0.55 (0.47)	0.61 (0.56)	0.49 (0.40)	0.69 (0.61)
	FP32	0.78 (0.69)	0.76 (0.71)	0.84 (0.74)	0.74 (0.66)
HuBERT	FP16	0.59 (0.58)	0.62 (0.62)	0.52 (0.51)	0.72 (0.72)
	FP32	0.84 (0.82)	0.79 (0.78)	0.88 (0.87)	0.80 (0.79)
WavLM	FP16	0.37 (0.39)	0.51 (0.52)	0.31 (0.32)	0.49 (0.51)
	FP32	0.61 (0.62)	0.67 (0.68)	0.66 (0.68)	0.56 (0.58)
UniSpeech	FP16	0.77 (0.79)	0.78 (0.79)	0.84 (0.85)	0.72 (0.74)
	FP32	0.77 (0.76)	0.76 (0.75)	0.83 (0.82)	0.72 (0.70)
SpeechT5	FP16	0.33 (0.48)	0.49 (0.58)	0.20 (0.36)	0.51 (0.66)
	FP32	0.33 (0.47)	0.48 (0.57)	0.19 (0.35)	0.51 (0.66)

requirements for their deployment scenarios. The green scores are intended as a complementary decision-making tool rather than a replacement for detailed metric analysis. Future work could explore the incorporation of minimum performance thresholds or weighted penalty functions to address scenarios where certain metrics are considered critical for specific applications.

Secondly, in our experiments, we utilised power measurements obtained from the built-in sensors of NVIDIA Jetson devices. Although the accuracy of these measurements may pose a threat to construct validity, previous studies demonstrated that such measurements are quite accurate and can be further calibrated to develop more realistic energy consumption models [35]. Moreover, we illustrated our framework on a limited selection of eight ASR models and two speech recording datasets. Implementing our framework on other hardware with more models and data would reduce this threat to external validity, and is a matter for future work.

Finally, we emphasise that our framework does not cover the total energy cost over the full AI system lifecycle, and that improved AI efficiency may still lead to higher overall energy consumption due to increased demand (Jevons paradox). Also, the framework does not consider potential biases in the ASR training data and models, which may disproportionately affect underrepresented user groups.

7 Conclusions and Outlook

We presented a multi-metric framework for ASR systems that extends beyond traditional accuracy metrics to incorporate energy efficiency and deployment considerations. Evaluation across both controlled (LibriSpeech) and real-world

Table 5. Common Voice metrics and green scores

Model	**Prec.**	**WER**	**RTF**	**EPAS**	**Latency**	**GME**	**HUR**	**Balanced**	**Mobile**	**Realtime**	**Server**
Distil-Whisper-S	FP16	6.16	0.074	0.514	433.07	14.20	36.96	0.77	0.67	0.81	0.71
	FP32	6.15	0.106	2.995	558.52	13.94	41.70	0.73	0.62	0.77	0.67
Distil-Whisper-M	FP16	5.78	0.071	1.558	386.33	17.79	41.06	0.80	0.73	0.84	0.76
	FP32	5.77	0.192	11.147	933.02	19.83	41.66	0.63	0.56	0.64	0.61
Distil-Whisper-L	FP16	4.81	0.096	3.232	516.20	22.70	20.72	0.78	0.76	0.81	0.78
	FP32	4.82	0.360	23.972	1671.79	28.82	48.12	0.46	0.45	0.43	0.51
Wav2Vec2	FP16	9.84	0.053	0.307	1799.57	24.25	15.19	0.69	0.78	0.63	0.78
	FP32	9.84	0.030	1.816	190.39	21.75	48.70	0.82	0.79	0.86	0.78
HuBERT	FP16	6.19	0.056	0.412	1812.83	25.24	16.99	0.74	0.82	0.67	0.84
	FP32	6.17	0.033	1.945	212.09	22.46	50.24	0.87	0.83	0.90	0.83
WavLM	FP16	22.36	0.048	0.229	1371.85	25.66	18.90	0.58	0.73	0.54	0.65
	FP32	22.39	0.018	0.709	93.74	23.78	42.18	0.69	0.75	0.73	0.66
UniSpeech	FP16	18.45	0.011	0.197	64.14	24.09	42.46	0.75	0.80	0.79	0.72
	FP32	18.46	0.014	0.654	76.47	23.26	45.64	0.74	0.78	0.78	0.71
SpeechT5	FP16	24.59	0.292	0.237	2407.59	25.05	18.87	0.33	0.53	0.19	0.54
	FP32	24.60	0.286	0.452	2369.41	24.58	22.12	0.34	0.53	0.20	0.54

(Common Voice) speech datasets demonstrates that our findings are robust across different speech conditions. Our results highlight that ASR model selection involves complex trade-offs between accuracy, speed, energy consumption, and hardware utilisation. We found that HuBERT and UniSpeech models achieve the best overall efficiency across different deployment scenarios, while Distil Whisper models offer an excellent balance between accuracy and efficiency, particularly in noisy environments. Importantly, our analysis challenges the assumption that energy efficiency necessarily compromises accuracy, as evidenced by models like HuBERT FP32 that excel in both dimensions. The green scoring system introduced in this work provides stakeholders with a practical tool for making informed decisions based on their specific deployment requirements. By quantifying the environmental impact of ASR models, we contribute to the growing effort to develop more sustainable AI systems without sacrificing performance.

Several limitations and promising directions for future research emerge from this work. Our evaluation was conducted on read speech datasets, which do not fully represent real-world conditions. Future work should extend this analysis to more challenging datasets, such as those involving conversational speech with overlapping speakers (e.g., CHiME [6]), realistic noise and reverberation conditions (e.g., Rev16 [32]), semi-spontaneous speech (e.g., TED-LIUM [13]), and multilingual domain-specific content (e.g., VoxPopuli [41]). Additionally, our model selection could be expanded to include non-transformer architectures like KALDI [31], which employs lightweight Hidden Markov Models. These models may offer competitive accuracy on clean speech, while potentially delivering superior efficiency metrics compared to transformer-based approaches. Beyond traditional metrics, future evaluations should consider incorporating semantic similarity measures as complementary accuracy metrics. Semantically-aware evaluation would provide a more nuanced understanding of model performance, especially in applications where precise wording is less critical than conveying the correct meaning. Finally, as ASR systems are deployed in increasingly diverse settings—from edge devices to large data centres—research on domain-specific optimisation techniques will be increasingly important. This includes exploring quantisation methods beyond FP16, model pruning, and architecture-specific optimisations that can further improve the balance between accuracy and efficiency. By advancing holistic evaluation approaches for ASR systems, we hope to encourage the development of models that not only recognise speech accurately, but do so in an environmentally responsible manner across the spectrum of deployment contexts.

Acknowledgments. This research is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No. 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

References

1. Anthony, L.F.W., et al.: Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. arXiv (Jul 2020). https://doi.org/10.48550/arXiv.2007.03051
2. Ao, J., et al.: SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. ACL Anthology pp. 5723–5738 (May 2022). https://doi.org/10.18653/v1/2022.acl-long.393
3. Ardila, R., et al.: Common Voice: A Massively-Multilingual Speech Corpus. ACL Anthology pp. 4218–4222 (May 2020), https://aclanthology.org/2020.lrec-1.520
4. Argerich, M.F., et al.: Measuring and Improving the Energy Efficiency of Large Language Models Inference. IEEE Access **12**, 80194–80207 (Jun 2024). https://doi.org/10.1109/ACCESS.2024.3409745
5. Baevski, A., et al.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS **33**, 12449–12460 (2020), https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html
6. Barker, J., et al.: The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In: Interspeech 2018. pp. 1561–1565 (2018). https://doi.org/10.21437/Interspeech.2018-1768
7. Chakravarty, A.: Deep Learning Models in Speech Recognition: Measuring GPU Energy Consumption, Impact of Noise and Model Quantization for Edge Deployment. arXiv (May 2024). https://doi.org/10.48550/arXiv.2405.01004
8. Chen, S., et al.: WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE J. Sel. Top. Signal Process. **16**(6), 1505–1518 (Jul 2022). https://doi.org/10.1109/JSTSP.2022.3188113
9. Chen, S.: How much energy will AI really consume? The good, the bad and the unknown. Nature **639**, 22–24 (Mar 2025). https://doi.org/10.1038/d41586-025-00616-z
10. Courty, B., et al.: mlco2/codecarbon: v2.4.1 (May 2024). https://doi.org/10.5281/zenodo.11171501

11. Gandhi, S., et al.: Distil-Whisper: Robust Knowledge Distillation via Large-Scale Pseudo Labelling. arXiv e-prints (Nov 2023). https://doi.org/10.48550/arXiv.2311.00430
12. Henderson, P., et al.: Towards the systematic reporting of the energy and carbon footprints of machine learning. J. Mach. Learn. Res. **21**(1), 10039–10081 (Jan 2020). https://doi.org/10.5555/3455716.3455964
13. Hernandez, F., et al.: TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In: Speech and Computer, pp. 198–208. Springer, Cham, Switzerland (Aug 2018). https://doi.org/10.1007/978-3-319-99579-3_21
14. Hsu, W.N., et al.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (Oct 2021). https://doi.org/10.1109/TASLP.2021.3122291
15. Husom, E.J., et al.: The Price of Prompting: Profiling Energy Use in Large Language Models Inference. arXiv (Jul 2024). https://doi.org/10.48550/arXiv.2407.16893
16. Jay, M., et al.: An experimental comparison of software-based power meters: focus on CPU and GPU. In: Proc. IEEE/ACM CCGrid 2023. pp. 106–118. IEEE (2023). https://doi.org/10.1109/CCGrid57682.2023.00020
17. Jitsi: JiWER (Jan 2025), https://github.com/jitsi/jiwer
18. Lacoste, A., et al.: Quantifying the Carbon Emissions of Machine Learning. arXiv (Oct 2019). https://doi.org/10.48550/arXiv.1910.09700
19. Lannelongue, L., et al.: Green algorithms: quantifying the carbon footprint of computation. Advanced Science **8**(12), 2100707 (2021)
20. Luccioni, A.S., et al.: Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. arXiv (Feb 2023). https://doi.org/10.48550/arXiv.2302.08476
21. Luccioni, S., et al.: Power Hungry Processing: Watts Driving the Cost of AI Deployment? In: FAccT '24 Proceedings, pp. 85–99. ACM (Jun 2024). https://doi.org/10.1145/3630106.3658542
22. Luccioni, S., et al.: AI Energy Score Leaderboard - February 2025. https://huggingface.co/spaces/AIEnergyScore/Leaderboard (2025)
23. Microsoft Corporation: Measuring the real-time factor on your device, https://learn.microsoft.com/en-us/azure/ai-services/speech-service/embedded-speech-performance-evaluations
24. Morris, A.C., et al.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: Interspeech 2004. pp. 2765–2768 (2004). https://doi.org/10.21437/Interspeech.2004-668
25. Naidu, R., et al.: Towards Quantifying the Carbon Emissions of Differentially Private Machine Learning. arXiv (Jul 2021). https://doi.org/10.48550/arXiv.2107.06946
26. NVIDIA Corporation: NVIDIA DRIVE OS 5.2 Linux SDK Developer Guide: tegrastats Utility (Jan 2023), https://docs.nvidia.com/drive/drive-os-5.2.0.0L/drive-os/index.html#page/DRIVE_OS_Linux_SDK_Development_Guide/Utilities/util_tegrastats.html
27. Panayotov, V., et al.: Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 19–24. IEEE (2015). https://doi.org/10.1109/ICASSP.2015.7178964

28. Parcollet, T., et al.: The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. Interspeech 2021 pp. 4583–4587 (2021). https://doi.org/10.21437/Interspeech.2021-456
29. Patterson, D., et al.: Carbon Emissions and Large Neural Network Training. arXiv (Apr 2021). https://doi.org/10.48550/arXiv.2104.10350
30. Patterson, D., et al.: The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. Computer **55**(7), 18–28 (Jun 2022). https://doi.org/10.1109/MC.2022.3148714
31. Povey, D., et al.: The Kaldi Speech Recognition Toolkit. In: Proc. IEEE ASRU 2011. IEEE Signal Process. Soc. (Dec 2011), IEEE Catalog No.: CFP11SRW-USB
32. Radford, A., et al.: Robust Speech Recognition via Large-Scale Weak Supervision. In: ICML, pp. 28492–28518. PMLR (Jul 2023), https://proceedings.mlr.press/v202/radford23a.html
33. Samsi, S., et al.: From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. In: 2023 IEEE High Performance Extreme Computing Conference (HPEC), pp. 25–29. IEEE (2023). https://doi.org/10.1109/HPEC58863.2023.10363447
34. Schwartz, R., et al.: Green AI. Commun. ACM **63**(12), 54–63 (Nov 2020). https://doi.org/10.1145/3381831
35. Shalavi, N., et al.: Accurate Calibration of Power Measurements from Internal Power Sensors on NVIDIA Jetson Devices. In: Proc. IEEE EDGE 2023. pp. 166–170. IEEE (2023)
36. Srivastav, V., et al.: Open Automatic Speech Recognition Leaderboard. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard (2023)
37. Strubell, E., et al.: Energy and Policy Considerations for Modern Deep Learning Research. AAAI **34**(09), 13693–13696 (Apr 2020). https://doi.org/10.1609/aaai.v34i09.7123
38. Texas Instruments: INA3221 data sheet, product information and support | TI.com (Mar 2016), https://www.ti.com/product/INA3221
39. Verdecchia, R., et al.: A systematic review of Green AI. WIREs Data Min. Knowl. Discovery **13**(4), e1507 (Jul 2023). https://doi.org/10.1002/widm.1507
40. de Vries, A.: The growing energy footprint of artificial intelligence. Joule **7**(10), 2191–2194 (oct 2023). https://doi.org/10.1016/j.joule.2023.09.004
41. Wang, C., et al.: VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. ACL Anthology pp. 993–1003 (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.80
42. Wang, C., et al.: UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data. In: ICML, pp. 10937–10947. PMLR (Jul 2021), https://proceedings.mlr.press/v139/wang21y.html
43. Woodard, J., et al.: An information theoretic measure of speech recognition performance. In: Workshop on Standardisation for Speech I/O Technology, Naval Air Development Center, Warminster, PA (1982)
44. Wu, C.J., et al.: Sustainable AI: Environmental implications, challenges and opportunities. PMLS **4**, 795–813 (2022)
45. Yang, Z., et al.: Accurate and Convenient Energy Measurements for GPUs: A Detailed Study of NVIDIA GPU's Built-In Power Sensor. SC '24 Proceedings pp. 1–17 (Nov 2024). https://doi.org/10.1109/SC41406.2024.00028
46. Yang, Z., et al.: Quality at the Tail of Machine Learning Inference. arXiv (Dec 2022). https://doi.org/10.48550/arXiv.2212.13925