Measuring what matters Key findings

Over-Reliance on Estimates and Proxies

- Many assessments use forecasted values rather than measured telemetry, introducing uncertainty. Hardware details, energy use, and carbon intensity are often inferred, not observed.

Inconsistent Lifecycle Boundaries and Units

- Definitions of lifecycle phases (e.g., “training” vs. “development”) vary, complicating comparison. Impact units differ across studies (e.g., per-token vs. per-inference vs. per-session), limiting interoperability and practical use.

Underreported Scope 3 and Embodied Emissions

- Supply chain impacts—such as emissions from chip fabrication, hardware transport, and e-waste— are often excluded. Embodied carbon and mining impacts are poorly tracked and inconsistently reported.

Opaque Water Use and Infrastructure Overheads

- Water usage data remains sparse and non-standardized, especially from hyperscale providers. Energy overheads like cooling (PUE) are frequently omitted from model-level assessments.

Neglect of Inference Phase and User Behavior

- Inference emissions are often overlooked despite their cumulative scale. Usage patterns, prompt retries, and inefficient deployment choices (e.g., region, hardware) are rarely factored in.

Lack of Standardization Across Tools and Methodologies

- Disparate tools and reporting formats hinder comparability. There is no consensus on amortization methods, benchmarking protocols, or lifecycle accounting models.

Carbon-Centric Metrics Mask Broader Impacts

- Most studies focus narrowly on CO_2e, ignoring water, minerals, biodiversity loss, and energy source transparency. Results may be obscured by offsets or averaged emissions factors.


ITUPublications
International Telecommunication Union
Measuring what matters: How to assess AI's environmental impact
Green Digital Action
ITU