OPEN ENERGY RECEIPT / 1

OER/1 and DRIFT/1

A receipt schema for energy claims, and a benchmark protocol for learning under drift.

Institute for Physical AI @ Bailey Military Institute

Charlot Lab ecosystem · released CC0 for comment

Draft for comment. Circulated to the MLCommons Power working group and the EU AI Office methodology process.

Abstract. Every previous alternative-computing wave died of unverifiable claims. This RFC specifies an energy receipt whose provenance axes keep modelled, stand-in and metered figures permanently distinguishable, and a companion benchmark protocol for systems that adapt while running. The composition rule is deliberately unforgiving: a composed receipt takes the weakest grade among its parts, never the average.

EFA-RFC-001: OER/1 and DRIFT/1

An Open Energy Receipt schema and a Learning-Under-Drift benchmark protocol for AI systems

Status	Draft v0.1 — for public comment
Date	2026-08-02
License	CC0 1.0 (public domain) — adopt, fork, or absorb without permission
Editors	Charlot Lab / Institute for Physical AI @ BMI ecosystem (drafted with AI assistance; see Acknowledgments)
Intended audience	MLCommons Power WG · EU AI Office methodology process (AI Act Art. 53 / Annex XI) · JEDEC-style industry consortia · hardware

vendors of any substrate · benchmark maintainers

The key words **MUST**, **MUST NOT**, **SHOULD**, **SHOULD NOT**, and **MAY** are to be interpreted as described in RFC 2119.

---

1. Motivation

Two vacuums exist where the AI industry most needs standards.

The receipt vacuum. The EU AI Act (Art. 53, Annexes XI–XII) now obliges providers of general-purpose AI models to document "known or estimated energy consumption," demandable by the AI Office without notice, with penalties up to €15M or 3% of global turnover. Yet the regulation permits estimates, omits the inference phase (where 60–90% of AI energy is actually spent), restricts disclosure to authorities, and provides **no standardized measurement methodology** — harmonized standards are not expected before ~2028. Meanwhile, vendors of alternative computing substrates publish efficiency ratios spanning 10× to 10,000× in which simulations, stand-in hardware, and task-specific results are routinely indistinguishable from audited measurements. A regulation that mandates a receipt no one knows how to write, in a market flooded with unpriceable claims, is a standardization window. OER/1 is a receipt format designed to fill it: small enough to adopt verbatim, strict enough to make a claim auditable, and paradigm-neutral enough to cover a GPU, a thermodynamic sampler, a photonic array, or a settling fabric with the same fields.

The benchmark vacuum. Every prior computing-paradigm transition was decided by a scoreboard the incumbent could not top (ImageNet for deep learning; $/kg for launch). The candidate scoreboard for energy-first, learn-at-inference AI is **performance under distribution drift, at decision latency, per audited joule** — a regime frozen-weight architectures cannot enter without becoming something else, and one no current benchmark measures. DRIFT/1 specifies it. The two specs interlock: DRIFT/1 leaderboard entries are ranked only as high as their OER conformance level allows.

2. Terminology

- **Workload** — a fully specified computational task with a quality metric (a benchmark episode, a batch of queries, a training step).
- **Receipt** — a signed record binding a workload's identity and quality outcome to the energy spent producing it, with provenance.
- **Provenance axes** — two independent labels on every energy figure: *how it was obtained* (measured / derived / simulated / projected) and *what device it describes* (target / stand-in / emulation). A figure

measured on an FPGA standing in for an unfabricated ASIC is `measured × stand_in` — never plain "measured." (Convention due to Klere's two-axis tags.)

- **Drift event** — a change to environment dynamics after deployment, drawn from a seeded schedule unknown to the system under test.
- **Regret** — cumulative shortfall of achieved return versus the best per-segment policy in hindsight.
- **Certificate** — machine-checkable evidence of behavioral guarantees (e.g., an energy function decreasing monotonically along trajectories; a verified contraction region).
- **Sealed / plastic** — whether parameters are frozen at deployment, or declared mutable at inference (fast weights, Hebbian writes, online Bayesian updates, on-chip plasticity).

3. OER/1 — The Open Energy Receipt

3.1 Design requirements

An OER receipt MUST be: **attributable** (which physical device, which software stack, which site), **quality-paired** (joules without a quality metric are meaningless and MUST NOT be issued), **provenance-explicit** (both axes, always), **composable** (receipts sum; grades do not silently upgrade), **tamper-evident** (hashed and signable), and **paradigm-neutral** (no field presumes kernels, clocks, or weights — a biological substrate is expressible).

3.2 Schema (normative, JSON)

```
{
  "oer_version": "1.0",
  "receipt_id": "uuid",
  "issued_at": "RFC3339 timestamp",
  "issuer": { "org": "string", "contact": "uri" },

  "workload": {
    "name": "string",
    "spec_uri": "uri",
    "input_hash": "sha256",
    "output_hash": "sha256",
    "quality": { "metric": "string", "value": 0.0, "higher_is_better":
  },
```

```
  "execution": {
    "substrate": {
      "family": "gpu | cpu | fpga_settling | tsu_pbit | ising_annealer
      "device_model": "string",
      "device_count": 1,
      "firmware": "string"
    },
    "stack": [ { "layer": "framework|compiler|runtime|kernel_lib|os",
    "site": { "grid_region": "string", "pue_applied": 1.0, "span": ["s
  },

  "energy": {
    "joules_total": 0.0,
    "breakdown": { "compute_j": null, "movement_j": null, "idle_alloc_
    "method": "wall_plug_metered | rail_metered | frequency_sweep_slop
    "meter": { "instrument": "string", "calibration_date": "date", "ur
    "sampling_hz": 0
  },

  "provenance": {
    "how": "measured | derived | simulated | projected",
    "device": "target | stand_in | emulation",
    "note": "string — REQUIRED when device != target"
  },

  "latency": { "wall_s": 0.0, "time_to_first_valid_s": null, "p99_deci

  "certificate": {
    "type": "none | energy_monotone | contraction_region | formal_proo
    "coverage": null,
    "evidence_uri": null
  },

  "attestation": { "receipt_hash": "sha256", "signature": "base64", "p
  "extensions": {}
}
```

3.3 Composition

A parent receipt MAY aggregate child receipts (`extensions.children: [receipt_id…]`). Its `energy.joules_total` MUST equal the sum of children plus declared overhead. Its provenance and conformance level MUST equal the **minimum** grade among children (min-grade propagation). Mixed-provenance aggregates MUST list the fraction of joules at each grade.

3.4 Conformance levels

Level	Requirements	Fit for
OER-L0	Schema-valid; `model_estimated` allowed; self-declared	AI Act Annex XI "estimated" compliance; internal tracking
OER-L1	`hw_counter_estimated`, `rail_metered`, or `frequency_sweep_slope`; calibration + uncertainty declared	Engineering claims; papers
OER-L2	External **wall-plug** measurement of the full system under test (MLPerf-Power/SPEC PTDaemon-compatible harness); PUE stated; failures included	Public leaderboards; procurement
OER-L3	L2 + signed attestation with independent witness + certificate populated where the workload is control/actuation	Regulatory filings; safety-relevant deployment

EU AI Act crosswalk (informative). Annex XI "known or estimated energy consumption" ? `energy.*` + `provenance.how` (known = measured/derived at L1+; estimated = L0). "Computational resources used" ? `execution.substrate` + `execution.stack`. The Act's inference-phase omission ? per-answer receipts under this spec close it voluntarily. The Model Documentation Form's energy field can carry an OER receipt URI unchanged.

3.5 Anti-gaming rules

Issuers MUST NOT: exclude idle/allocated-but-unused energy for reserved hardware (report in `idle_alloc_j`); report best-of-N runs without N receipts or an aggregate covering all N; quote a stand-in figure without the `device` axis and note; issue receipts for a quality outcome below the workload spec's declared floor; or apply PUE < 1.0. Comparisons across receipts of different conformance levels MUST state both levels.

4. DRIFT/1 — The Learning-Under-Drift Benchmark

4.1 Task classes (v0.1)

ID	Task	Drift schedule (seeded)	Episode
D-A	Cart-pole stabilization & swing-up	Pole mass/length ramps ±40%; friction shocks; actuator gain flips	10k steps, ≥6 drift events
D-B	2-DOF arm reach	Payload changes; per-joint actuator degradation; sensor bias onset	10k steps, ≥6 events
D-C	Partially observed grid navigation	Layout remaps; goal relocation; observation noise regime shifts	5k steps, ≥4 events
D-D	Non-embodied decision stream (dispatch/ portfolio-style)	Regime switches in returns/costs; constraint changes	10k decisions

Reference environments, drift-schedule generators, and oracle solvers are published with committee-held seed pools; public seeds for development, sealed seeds for the leaderboard.

4.2 Protocol

1. Entrant declares the **pre-deployment budget** (training compute, data) and whether the system is **sealed** or **plastic** at deployment —

and, if plastic, the mechanism (fast weights, Bayesian update, on-chip plasticity, etc.).
2. At t=0 the system is deployed. No human intervention, no external retraining calls, no network egress. All adaptation computation occurs **inside the metered boundary** — adaptation energy appears in the receipt or the run is void.
3. Drift events occur at times/magnitudes drawn from the sealed schedule.
4. Every episode — including failures — is receipted. Leaderboard entries require **OER-L2** minimum; L3 entries rank above L2 at equal scores; L0/L1 and any `device: stand_in` entries appear only in a separate **projected tier**, never interleaved.

4.3 Scoreboard

Primary result is the **tuple** ⟨**R, L, J**⟩, always reported in full (no single-scalar collapse):

- **R** — cumulative regret vs. the per-segment oracle, normalized by the frozen baseline's regret (lower is better);
- **L** — decision latency, p50 and p99 (seconds);
- **J** — receipted joules per episode, with adaptation energy itemized where separable.

Derived (optional): **Adaptation Efficiency** AE = (Regret_frozen − Regret_entrant) / J_entrant — regret bought back per joule.

4.4 Mandatory baselines

Every leaderboard release ships with maintained reference results: (a) a **frozen strong baseline** (TD-MPC2-class model-based RL, trained on the pre-drift distribution, weights sealed); (b) a **naive online fine-tune** baseline (gradient steps on the stream, same wall-clock budget); (c) each entrant's own **plasticity-off ablation**. An entrant that cannot beat its own ablation on R has not demonstrated learning under drift, whatever its J.

4.5 Reporting

Minimum 5 sealed seeds per task; mean and worst-seed reported; all receipts, configs, and logs published; seeds revealed after the leaderboard round closes. Certificate coverage (for D-A/D-B actuation) reported alongside — an entry MAY claim the **certified** flag only with `certificate.type != none` at stated coverage and 100% empirical convergence from the certified region.

5. Integration hooks (informative)

Receipts attach at existing boundaries with no new runtime required: an **ONNX Runtime** execution-provider callback; **StableHLO/IREE** module metadata; a **Triton** launch-hook wrapper; **NIR** (Neuromorphic Intermediate Representation) graph metadata for spiking targets; **Lava** process monitors; the MLPerf-Power PTDaemon harness for L2 wall-plug capture; and language-level emission from energy-typed stacks (e.g., Klere's `flowg`/Joule, Ferric runtimes). Portability layers that lower one model onto many substrates are the natural mass issuers of OER receipts — a universal stack plus a universal receipt turns every deployment into a comparable datapoint.

6. Governance

Versioning is semantic; breaking schema changes bump the major version. Change proposals via public PR; a standing group of ≥3 unaffiliated reviewers; vendor entries to DRIFT/1 do not confer governance rights. This document may be adopted in whole or part by any body without attribution (CC0). Reference implementations required before v1.0: JSON-Schema validator, PTDaemon-compatible capture harness, D-A/D-B/D-C environments + oracles, and two worked receipts on dissimilar substrates (one GPU, one non-von-Neumann).

7. Worked example (informative)

A settling-fabric micro-workload, honestly graded:

```
{
  "oer_version": "1.0",
  "receipt_id": "3f9c…",
  "issued_at": "2026-08-02T18:00:00Z",
  "issuer": { "org": "example-lab", "contact": "mailto:receipts@exampl
  "workload": {
    "name": "hopfield-recall-64x64",
    "spec_uri": "https://example.org/specs/recall-64",
    "input_hash": "sha256:ab…", "output_hash": "sha256:cd…",
    "quality": { "metric": "pattern_recall", "value": 1.0, "higher_is_
  },
  "execution": {
    "substrate": { "family": "fpga_settling", "device_model": "Virtex-
    "stack": [ { "layer": "runtime", "name": "ternary-fabric", "versio
    "site": { "grid_region": "us-east-1", "pue_applied": 1.2, "span":
  },
  "energy": {
```

```
    "joules_total": 2.12e-12,
    "breakdown": { "compute_j": 2.12e-12, "movement_j": null, "idle_al
    "method": "frequency_sweep_slope",
    "meter": { "instrument": "board telemetry + slope regression R2=0.
    "sampling_hz": 1000
  },
  "provenance": { "how": "measured", "device": "stand_in", "note": "FP
  "latency": { "wall_s": 5.3e-7, "time_to_first_valid_s": null, "p99_d
  "certificate": { "type": "energy_monotone", "coverage": 1.0, "evider
  "attestation": { "receipt_hash": "sha256:…", "signature": "…", "pubk
  "extensions": {}
}
```

Conformance: **OER-L1**, provenance `measured × stand_in` — publishable, not leaderboard-eligible; exactly the honesty the two-axis label exists to enforce.

Acknowledgments

Provenance-axis convention after Klere's receipt discipline; validation-ledger norms after the EFA program; wall-plug methodology aligned with MLCommons Power / SPEC PTDaemon; drift-task shapes after the EFA nano suite and TD-MPC2 baselines; written for the methodology vacuum documented around EU AI Act Art. 53. Drafted with AI assistance (Claude, Anthropic) at the direction of the EFA program; all normative choices remain the editors' responsibility.