

# LLM Billing & Benchmarking Standard (LBBS) v0.1 — Draft for Public Comment

Purpose: Establish a vendor-neutral primary unit for LLM billing and benchmarking to enable fair cross-provider comparisons.

## 1. Executive Summary

Current LLM pricing relies on vendor-defined tokens/messages that are not comparable across providers. This impairs scientific benchmarking and fair purchasing. The proposed core unit is \$-s/task — the actual monetary cost multiplied by measured wall-clock inference time required to complete a defined task to an agreed-upon quality threshold. Secondary required metrics: \$/task, s/task, and success rate at the quality threshold (success@Q).

## 2. Problem Statement

Tokenization varies by vendor and model; “messages” are not uniformly defined. This creates gaming risk (changing token boundaries to simulate efficiency gains), reduces buyer understanding, and distorts research reproducibility. Buyers and regulators need a neutral, physically meaningful unit.

## 3. Standardized Metrics

**Primary metric (required):** \$-s/task — median of (price\_paid\_in\_\$ × wall\_clock\_seconds) over N runs for a task T meeting quality Q.

**Secondary metrics (required):** \$/task; s/task; success@Q. **Informational (non-normative):** provider-reported tokens, context window, batch size, hardware class.

## 4. Measurement Protocol

**4.1 Task specification.** Public, versioned benchmark suites with fixed prompts/inputs and objective evaluation scripts defining quality Q.

**4.2 Run conditions.** Public API, declared plan tier; ≥30 runs per (model, task) over ≥24h. Wall-clock timing from first byte sent to last byte received; include tool calls if required by task. Price captured from itemized bill/usage API; allocate pro-rata for bundles using published rate cards. Fix seeds or report sampling parameters.

**4.3 Required publication.** Median and [p10, p90] of \$-s/task, \$/task, s/task; success@Q with confidence intervals; provider, plan tier, model ID, region, dates; any caching/preview features toggled.

## 5. Reporting & Transparency (Vendor Requirements)

Vendors SHOULD provide: (a) clear per-call price or credit burn for each model/tier; (b) per-call usage receipt with start/end timestamps, model ID, billing line item; (c) plain-language definition of any non-standard units (e.g., tokenizer rules).

## 6. Compliance Levels

**LBBS-C1 (Baseline):** Report \$-s/task, \$/task, s/task, success@Q, and metadata. **LBBS-C2 (Enhanced):** Add energy/task or GPU-hour equivalents if available. **LBBS-C3 (Verified):** Independent lab audit (e.g., under MLCommons).

## 7. Governance & Stewardship

Initial host: MLCommons/MLPerf for task curation and audits. Standardization target: IEEE/ISO/NIST once v1.0 stabilizes. Regulatory liaison: map LBBS to consumer billing transparency obligations (US/EU/AU) and the EU AI Act.

## 8. Transition Plan

**Phase 0 (0–3 mo):** Public comment; pilots. **Phase 1 (3–6 mo):** MLPerf experimental track publishes first \$-s/task tables. **Phase 2 (6–12 mo):** Vendors expose per-call receipts; buyers require LBBS-C1 in RFPs. **Phase 3 (12–18 mo):** Regulators cite LBBS to avoid misleading pricing.

## 9. Legal Compatibility (Nonbinding)

US: Aligns with 'truth in billing' principles (FTC/FCC) for clear, nonmisleading pricing. EU: Supports transparency requirements and procurement comparability; complements AI Act transparency. Australia: Consistent with ACCC/ACL expectations on clear, upfront digital pricing.

## 10. Risks & Mitigations

**Vendor pushback:** Address with relative reporting (publish \$-s/task without cost breakdown). **Cherry picking:** Lock benchmark suites; preregister; audits. **Regional price variance:** Report region/currency/date; convert to USD for crossvendor tables. **Subscription bundles:** Impute percall cost using published rate cards; disclose method.

## 11. Glossary

**\$-s/task:** Monetary cost  $\times$  wallclock seconds to meet quality Q. **Task (T):** Fixed input/output with objective evaluation. **Quality threshold (Q):** Pass criterion (e.g., exact match, rubric score). **Success@Q:** Fraction of runs meeting Q. **Turn (message):** One submitted interaction in chat UX (informational only).

## 12. Illustrative Example

Model	Task (Q)	Median \$/task	Median s/task	Median \$-s/task	Success@Q
Vendor A — Model X	GSM8K (exact)	\$0.018	3.2	0.0576	92%
Vendor B — Model Y	GSM8K (exact)	\$0.013	5.4	0.0702	93%

Call to action: MLCommons to host an LBBS experimental track; vendors to provide percall receipts; buyers to require \$-s/task in RFPs; regulators to reference LBBS as a recognized means to avoid misleading pricing.