

Joules-per-Solution for Stochastic and Agentic Inference: Benchmarking Frontier and Agentic LLMs Against the Human Brain

Rowan Brad Quni-Gudzin

2026-08-15 · DOI: 10.5281/zenodo.21945415

Abstract. The joules-per-solution (J/S) metric, introduced in JPCUB P0 as a universal, physics-grounded measure of computational efficiency, assumed deterministic solvers: one run yields one solution. Large language models (LLMs) violate that assumption twice over: they are stochastic samplers whose outputs must be verified, and agentic systems execute graphs of sub-queries whose total token budget dwarfs a single inference. This paper extends J/S to stochastic and agentic inference. We define the closed-form stochastic correction $J/S_\tau(n) = n E_q / (1 - (1 - p_q(\tau))^n)$, prove that — at fixed per-query energy — it is monotonically increasing in the sampling regime n (so single-shot sampling is the joules-per-solution-efficient operating point, and sampling buys accuracy at strictly increasing energy per correct solution), show that interior minima arise only through batching efficiency, and show that the J/S of agentic tasks must charge the full orchestration graph, including failed attempts. Using published measurement data (TokenPowerBench; Where Do the Joules Go?), we derive order-of-magnitude task-stratified estimates and compare them to the human brain baseline (≈ 20 W resting brain power). Result: on cheap single-shot tasks a frontier LLM is roughly one order of magnitude more energy-efficient than a human expert; on graduate-level reasoning the two substrates meet at parity ($\sim 10^4$ J per correct solution); and under self-consistency sampling or agentic orchestration the LLM is frequently less efficient. We argue the efficiency crossover is governed by task verifiability, accuracy, and token budget — not by intrinsic substrate advantage — and pre-register falsification conditions and a calibration register for independent measurement.

Keywords. JPCUB, joules per solution, LLM energy, agentic AI, energy efficiency, human brain, stochastic inference, test-time compute, benchmarking

1. Introduction

The joules-per-solution metric ($J_{\text{CUB}} = P_{\text{sys}} \cdot t_{\text{sol}}$) was introduced in JPCUB P0 [qnfo2026joules] as a universal, cross-domain measure of computational energy efficiency: the total system-level energy required to produce a *correct* solution to a computational task at a specified correctness threshold. P0 surveyed fourteen domain-specific benchmarks and found none provide cross-domain comparability; the J/S metric was offered as the first universal arbiter, with anti-gaming provisions (pre-registration, adversarial validation, Pareto-frontier reporting, component audit, living benchmark protocol). P1 applied it to seventeen quantum computing platforms [qnfo2026jpcubcl], and P2 extended it to qudit architectures [qnfo2026qudit].

This paper applies J/S to the dominant compute paradigm of the 2020s: large language models (LLMs), including frontier reasoning models and agentic systems. The application is not a routine re-benchmarking. LLMs break two implicit assumptions of P0's formulation:

1. **Determinism.** A quantum or classical solver produces a solution with bounded failure probability per run; the J/S accounting treats one run as one solution attempt. An LLM is a *stochastic sampler* over a distribution of outputs. Its "energy per query" is well defined, but its "energy per solution" is an expectation over outputs, discounted by the probability that a sampled output passes a verifier — and that probability depends on the sampling regime (temperature, number of samples, self-consistency, best-of- n), which is a free choice of the *user*, not of the hardware.
2. **Fixed cost.** P0's t_{sol} is the time of one solve. An agentic system (tool use, retrieval, multi-step planning, multi-agent orchestration) spends energy on a graph of sub-queries whose context grows monotonically; a reasoning

model spends a *user-selected* token budget on "thinking". The energy per solution is therefore a decision variable, not a constant.

The consequence is that the entire "joules per token" literature — TokenPowerBench [niu2025tokenpowerbench], the ML.ENERGY lineage [chung2026where], energy-to-token advocacy [liu2026position], energy-per-token metrics [wilhelm2026beyond] — measures the *wrong denominator* for the question that matters: how much energy does it cost to produce a correct answer? A model that emits cheap tokens but fails half the time is not cheap per solution.

We formalize the stochastic and agentic corrections (Section 4), fix the system boundary and anti-gaming provisions (Section 5), define the human-brain baseline with a pre-registered attribution rule (Section 6), derive task-stratified estimates from published measurement data (Section 7), expose the training-amortization asymmetry that biases informal "AI vs brain" claims (Section 8), characterize the reasoning-budget frontier (Section 9), specify the measurement protocol extension (Section 10), and pre-register falsification conditions with a calibration register (Section 11).

The headline finding is deliberately modest and deliberately falsifiable: **on verifiable cognitive tasks, frontier and agentic LLMs are not orders of magnitude more energy-efficient than the human brain. They are roughly one order of magnitude more efficient on cheap single-shot tasks, meet the brain at parity on graduate-level reasoning, and fall behind under sampling or orchestration.** The crossover is governed by task verifiability, accuracy, and token budget — not by any intrinsic advantage of silicon or carbon.

2. Related Work

Energy-per-token measurement. TokenPowerBench [niu2025tokenpowerbench] provides the first open benchmark for LLM inference power, measuring GPU-, node-, and system-level energy across models from 1B to 405B parameters. Its headline anchors: Llama3-405B consumes approximately 40–60 J per output token in FP16 on 16 H100 GPUs (FP8 reduces this by roughly 30%), MoE models consume about as much as dense models of their *active* size, and context growth from 2K to 10K tokens raises energy per token by roughly $3\times$ for large dense models. "Where Do the Joules Go?" [chung2026where] measures 46 models, 7 tasks, and 1,858 configurations on H100 and B200, finding a $25\times$ energy-per-response spread driven by task type: GPQA problem-solving averages 4,625 J per response versus 184 J for text conversation, because reasoning emits $10\times$ the output tokens (mean 6,988 vs 717) *and* longer sequences cap batch size, raising energy per token. GPUs account for 50–70% of datacenter power; B200 reduces energy vs H100 by a median 35% at matched latency. These two studies are our empirical backbone.

Energy-per-token advocacy. Liu et al. [liu2026position] argue inference should be evaluated as energy-to-token production under joint compute, power, cooling, and PUE constraints. Wilhelm et al. [wilhelm2026beyond] advocate energy-per-token as a complement to accuracy and analyze the energy-accuracy trade-off of test-time strategies. Both stop at the token denominator; neither divides by the probability of a correct answer. Our contribution is precisely that division.

Test-time compute and token budgets. A large literature controls reasoning cost: token-budget-aware reasoning [han2024tokenbudgetaware], BudgetThinker [wen2025budgetthinker], budget-guided MCTS [miyamoto2026aligning], conformal risk control over compute budgets [wang2026conformal], sleep-time compute [lin2025sleep-time], and a survey of adaptive test-time compute [alomrani2025reasoning]. CROP reports 80.6% token reduction with nominal accuracy loss [shah2026crop]; quantization is shown to inflate reasoning tokens [lian2026quantization]. None of these convert token budgets to joules, and none report a joules-per-*solution* frontier. Our Section 9 connects this literature to J/S via the reasoning-budget frontier: accuracy rises sub-linearly in the token budget while $J/S_\tau(b)$ rises monotonically at fixed per-token cost (Proposition 1).

Carbon and lifecycle. LLMCarbon [faiz2023llmcarbon], LLMspace [jiang2026llmspace], and simulation-based inference energy studies (cf. [desislavov2021compute]) quantify operational and embodied carbon. A scoping review [kim2025toward] documents the field's "methodological inconsistencies, technology-specific biases, and insufficient attention to end-to-end system perspectives" — the exact gap JPCUB's six-component boundary was designed to close.

Efficiency and orchestration. Small models can outperform large ones on task-specific efficiency ratios [cao2026taskspecific]; multi-agent orchestration suffers $O(n \cdot S \cdot |D|)$ broadcast-induced token overhead [parakhin2026token]. Training energy baselines are anchored by Patterson et al. [patterson2021carbon].

The gap. No existing work (i) divides inference energy by the probability of a correct answer under a stated sampling regime, (ii) charges agentic orchestration graphs their full energy, or (iii) compares LLM J/S to the human brain on a per-solution basis. Phase 1 due diligence (DUE-DILIGENCE.md, committed 2026-08-15) confirmed this gap by corpus sweep.

3. Notation and Conventions

Let τ be a task with a verifier V_τ and correctness threshold θ (exact match, unit-test pass, ground-truth match). Let q be a single inference query (prompt, model, decoding configuration), producing output $y \sim p_\theta(y \mid x)$ at full-system energy E_q (six-component JPCUB boundary). Let $p_q(\tau) = \mathbb{E}_y[V_\tau(y)]$ be the pass@1 probability. Let n be the number of independent samples; b the reasoning-token budget; G an agentic orchestration graph with sub-query set V_G .

All LLM energy figures are reported GPU-only *and* system-level ($\times 1.5$ multiplier: PUE + host + networking, given GPU share 50–70% of datacenter power [chung2026where, niu2025tokenpowerbench]). All estimates are order-of-magnitude and audit-trailed to the assumption block in ESTIMATES.md (A1–A15).

4. The Metric: J/S for Stochastic and Agentic Inference

Definition 1 (J/S, single query).

$$J/S_\tau = \frac{E_q}{p_q(\tau)}$$

the expected energy per correct solution when queries are repeated until success.

Definition 2 (J/S, n independent samples, single-answer acceptance).

$$J/S_\tau(n) = \frac{n E_q}{1 - (1 - p_q(\tau))^n}$$

Proposition 1 (monotonicity). For fixed per-query energy E_q and $0 < p_q < 1$, the function $J/S_\tau(n)$ is **strictly increasing** in n for all $n \geq 1$, with $J/S_\tau(1) = E_q/p_q(\tau)$ and $J/S_\tau(n) \sim nE_q$ as $n \rightarrow \infty$. Independent sampling never reduces the expected energy per correct solution; the joules-per-solution-efficient operating point is the single query.

Proof. With $q = 1 - p$, write $f(n) = n/(1 - q^n)$. Then $f'(n) = (1 - q^n + nq^n \ln q)/(1 - q^n)^2$. Since $\ln q < 0$, the numerator equals $1 - e^u(1 - u)$ with $u = n \ln q < 0$; and $e^u(1 - u) < 1$ because $1 - u < e^{-u}$. Hence $f'(n) > 0$ for all $n > 0$, and $f(n) \rightarrow 1/p$ as $n \rightarrow 0^+$ (Bernoulli), so $J/S_\tau(n) \geq E_q/p_q(\tau)$ with equality only in the single-shot limit. $\square$

Proposition 2 (batching restores an interior minimum). If n samples are served with batching efficiency $\eta(n) \geq 1$ (measured energy per token falls $3\text{--}5\times$ as batch size grows [chung2026where]), so that $E(n) = nE_q/\eta(n)$, then $J/S_\tau(n) = nE_q/(\eta(n)(1 - (1 - p_q)^n))$ can attain an interior minimum at finite n^* whenever early batching gains outpace the saturating acceptance term. The U-shape, when observed, is a *systems* property (batching), not an intrinsic property of stochastic inference.

Corollary (sampling is an accuracy buy). Self-consistency, best-of- n , and majority voting purchase accuracy at strictly increasing (or, under batching, non-decreasing) joules per correct solution. Any claim about LLM energy efficiency that omits the sampling regime (n, p_q) is incomplete (anti-gaming A1).

Definition 3 (majority vote). If the verifier accepts the majority answer over n samples, $P_{\text{correct}}(n) = \Pr[\text{Bin}(n, p) > n/2]$ and $J/S_\tau(n) = nE_q/P_{\text{correct}}(n)$.

Definition 4 (agentic J/S). For an orchestration graph G , total attempt energy $E_{\text{attempt}} = \sum_{v \in V_G} E_v + E_{\text{tools}} + E_{\text{overhead}}$, where E_v includes growing-context prefill, and E_{overhead} includes retries and orchestration. With success probability p_G (artifact passes the verifier),

$$J/S_{\tau_A} = \frac{E_{\text{attempt}}}{p_G}$$

Failed attempts are charged to the solution — the agentic analogue of P0's "no free cooling" rule.

Definition 5 (reasoning-budget frontier). For reasoning models, $E_q = E_q(b)$ and $p_q = p_q(b)$. The report MUST include the frontier $\{(J/S_\tau(b), p_q(b)) : b \in \mathcal{B}\}$ for a pre-registered budget set $\mathcal{B}$ — never a single point (P0 Pareto mandate). By the argument of Proposition 1 (with b in place of n), the frontier is monotonically increasing in b at fixed per-token cost; its knee is the accuracy-saturation point beyond which J/S grows linearly with zero accuracy gain.

5. System Boundary and Anti-Gaming

The six-component JPCUB boundary (computation, memory, I/O, cooling, power conversion, amortized manufacturing) is inherited unchanged. Because published LLM energy numbers are overwhelmingly GPU-only, we report both columns and use a system multiplier of 1.5 as the JPCUB lower bound. Embodied (manufacturing) energy is flagged as an open addendum, never silently included. Anti-gaming provisions extend P0: (A1) sampling transparency — every J/S number states $(n, p_q, \text{temperature})$; (A2) Pareto mandate; (A3) verifier pre-registration on a verifiable-task canon; (A4) adversarial validation by independent re-derivation; (A5) component audit (GPU-only vs system breakdown per estimate); (A6) living benchmark protocol with "as of" dates.

6. Human Brain Baseline

The resting human brain consumes approximately 20 W — roughly 20% of the ~ 100 W whole-body basal metabolism — across ~ 86 billion neurons [kety1948nitrous, raichle2006brain, herculano2009human].

$$J/S_{\text{human}} = P_{\text{brain}} \cdot t_{\text{sol}} \approx 20 \text{ W} \cdot t_{\text{sol}}$$

Pre-registered attribution rule (symmetric): primary = brain-only 20 W; sensitivity bounds = task-marginal cognitive power ($\sim 3\text{--}5$ W) and full-body basal (~ 100 W); amortization (education/development) reported separately and symmetrically with LLM training amortization (Section 8). A comparison that charges one side's amortized cost

against the other's marginal cost is biased by up to three orders of magnitude and is rejected by this protocol. P0 already remarked on the brain's $\sim 2 \times 10^{-15}$ J per synaptic operation (its Section 5.3) [qnf02026joules]; the per-solution comparison here is complementary — per-operation efficiency is not per-verified-solution efficiency.

7. Task-Stratified Estimates

Estimates are derived from the published anchors of Section 2 with the assumption block A1–A15 (ESTIMATES.md). All figures are order-of-magnitude; system-level J/S = (GPU-only $\times$ 1.5) $\div$ accuracy.

Task class	LLM J/S (system)	Human J/S (20 W)	Ratio LLM/human
Simple verifiable QA (factoid/MCQ)	$\sim 35\text{--}70$ J	400–600 J	$\sim 0.06\text{--}0.17$
Math word problems (GSM8K-class)	$\sim 90\text{--}180$ J	900–1,800 J	$\sim 0.05\text{--}0.20$
Graduate reasoning (GPQA-class, single pass)	$\sim 9,000\text{--}12,600$ J	6,000–18,000 J	$\sim 0.5\text{--}2.1$
GPQA + self-consistency (n = 8)	$\sim 55,000$ J	6,000–18,000 J	$\sim 3\text{--}9$
Agentic coding (SWE-bench-class)	$10^5\text{--}5 \times 10^6$ J	$1.4 \times 10^5\text{--}4.3 \times 10^5$ J	$\sim 0.2\text{--}36$

Sampling transparency (anti-gaming A1): n and p_q are stated per row; temperature is not applicable to these *derived* estimates (it is a measurement-time parameter, required when measurements are reported).

Worked anchor (GPQA row, the load-bearing estimate): measured mean GPU-only response energy 4,625 J [chung2026where] $\times$ 1.5 = 6,900 J system; at pass@1 = 0.6, $J/S \approx 11,500$ J. Human expert at 20 W for 10 minutes: 12,000 J. **Both substrates sit at $\sim 10^4$ J per correct graduate-level answer.**

Three multipliers explain why J/token misleads: (1) the accuracy correction divides by p (0.6 accuracy costs 1.67 $\times$); (2) reasoning-token explosion multiplies output length by per-token cost — the measured 25 $\times$ spread between chat and problem-solving energy per response [chung2026where]; (3) agentic orchestration multiplies total tokens across the graph to $10^5\text{--}10^6$ per attempt.

The shape, not the numbers, is the claim: **the LLM advantage is largest on cheap single-shot tasks (~ 1 order of magnitude, 6–17 $\times$) and shrinks monotonically as tasks demand more tokens, more samples, or more orchestration, crossing parity inside the hard-reasoning band and going negative under sampling or agentic operation.** The result is robust to $\pm 3\times$ perturbation of any single input — not to worst-case simultaneous compounding, which can compress the cheap-task advantage toward parity (ESTIMATES.md Section 7).

8. Training-Amortization Asymmetry

System	Training energy	Amortized per solution
GPT-3	$\sim 4.6 \times 10^{12}$ J (1,287 MWh) [patterson2021carbon]	46 – 4.6×10^3 J/query (at $10^{11}\text{--}10^9$ lifetime queries)
GPT-4-class	$\sim 1.8\text{--}3.6 \times 10^{14}$ J (50–100 GWh; widely reported industry estimate, no primary disclosure)	$1.8 \times 10^3\text{--}3.6 \times 10^5$ J/query
Human brain (development)	$\sim 1.3 \times 10^{10}$ J (20 W $\times$ 20 yr, brain-only)	$\sim 1.3 \times 10^5$ J/solution (at $\sim 10^5$ lifetime professional solutions)

Frontier-model training amortizes to $\sim 10^3$ J/query only at 10^{11} lifetime queries; at 10^9 queries it *dominates* inference (up to 3.6×10^5 J/query). The human development band ($\sim 10^5$ J/solution) overlaps the agentic J/S band. Any informal "AI vs brain" energy comparison that omits this table is asymmetric by construction.

9. The Reasoning-Budget Pareto Frontier

Token-budget evidence [han2024tokenbudgetaware, wen2025budgetthinker, miyamoto2026aligning, wang2026conformal, shah2026crop] shows accuracy rises sub-linearly in the token budget, with 60–80% budget cuts achieving nominal accuracy loss. Consequently, by Proposition 1 (with b in place of n), $J/S_\tau(b)$ is monotonically increasing in b at fixed per-token cost: additional thinking tokens always raise expected energy per correct solution, and the frontier's knee is the accuracy-saturation point beyond which J/S grows linearly with zero accuracy gain. Quantization interacts pathologically: cheaper tokens can lengthen reasoning [lian2026quantization], so a per-token saving is not necessarily a per-solution saving — J/S captures this where J/token cannot. The frontier, not any point, is the reportable object.

10. Measurement Protocol Extension (J/S-LLM)

1. **Task canon:** verifiable tasks only (math exact-match, code unit-test, QA ground-truth, retrieval relevance); verifier and threshold pre-registered.
2. **Full-system energy** via the TokenPowerBench/ML.ENERGY methodology, reported GPU-only and PUE-adjusted; report J/token at stated batch, $J/\text{response}$, and $J/\text{solution}$.
3. **Sampling transparency (mandatory):** state $(n, p_q, \text{temperature})$; report $E[J]/\text{accuracy}$.
4. **Pareto frontiers:** reasoning-budget sweep and sampling sweep, reported as curves.
5. **Agentic accounting:** instrument the full orchestration graph; solution = verifier-passing artifact; report sub-call count and total tokens.
6. **Symmetric human baseline:** brain-only 20 W primary; marginal and full-body bounds; amortization reported on both sides or neither.
7. **Amortized vs operational J/S** reported as separate numbers, never conflated.

11. Falsification Conditions and Calibration Register

Core claim (P6). With the stochastic correction and agentic accounting under the full system boundary, frontier and agentic LLMs are *not* orders of magnitude more energy-efficient than the human brain on verifiable cognitive tasks: they win by roughly one order of magnitude on cheap single-shot tasks, reach parity on graduate reasoning, and fall behind under sampling or orchestration.

P6-F1 (falsified if): an independent measurement, under the full J/S boundary and a standardized sampling regime, finds frontier/agentic $J/S < 10^{-1} \times$ the expert-human baseline at matched accuracy on ≥ 2 of 3 canonical hard-reasoning task classes (GPQA-grade, SWE-bench-grade, verified multi-step math).

P6-F2 (also falsified if): the $(J/S, \text{accuracy})$ frontier is not monotonically increasing with diminishing accuracy returns — i.e., if sampling or additional reasoning budget *reduces* J/S at matched accuracy (an efficiency free lunch), or if accuracy does not saturate as the budget grows. Equivalently, falsified if Proposition 1's monotonicity or the concavity of $p_q(b)$ fails under measurement.

Calibration register (pre-registered, 2026–2028):

1. C1 — open-weight frontier MoE (e.g., Qwen3-235B-A22B) on GPQA Diamond: J/S measured under full boundary, $n=1$ and $n=8$.
2. C2 — SWE-bench Verified agent (open trace): total tokens, sub-calls, verified solve rate, J/S .
3. C3 — matched human expert panel: GPQA and SWE-bench task times at 20 W attribution.
4. C4 — system multiplier measurement (GPU-only vs node vs facility) on one reference config.

5. C5 — training-amortization audit: published training energy for one open frontier model $\div$ public query counts.
6. C6 — reasoning-budget frontier: J/S(b) measured at 5 pre-registered budgets.
7. C7 — adversarial re-derivation: independent party recomputes Table 7.1 from the same source rows.
8. C8 — living update: re-run C1–C2 at 12-month intervals; J/S claims carry "as of" dates.

12. Discussion and Limitations

Limitations. (i) Estimates are derived from published GPU-only measurements; the 1.5 system multiplier is a lower bound, not a measurement. (ii) pass@1 figures are leaderboard-grade approximations, not measurements on identical task instances. (iii) Agentic token counts are assumption-graded (A12) pending instrumented traces. (iv) Frontier proprietary models disclose no joules/query; only open-weight models on known hardware are honestly measurable. (v) Human task times are convention-graded; education amortization uses an order-of-magnitude lifetime-solutions base. Every limitation is named, quantified, and audit-trailed in ESTIMATES.md.

Epistemic posture. This paper's contribution is the *denominator*, not new joules numbers: the per-correct-solution correction that the energy-per-token literature omits, and the symmetric human baseline that the informal "AI vs brain" discourse lacks. The falsification conditions are written so that a hostile measurement can kill the claim. Per JPCUB P0's ethos, if the measurements falsify P6, the finding is published as published — the framework is better served by knowing it is wrong than by not testing it.

13. Conclusion

Measured as joules per correct solution under a full system boundary, with a stochastic correction and agentic orchestration accounting, frontier and agentic LLMs are approximately one order of magnitude more efficient than the human brain on cheap single-shot tasks, at parity on graduate-level reasoning, and less efficient under sampling or orchestration. The crossover is governed by task verifiability, accuracy, and token budget — not by substrate. The J/S framework, extended here to stochastic and agentic inference, now covers the full span of JPCUB's seven computational paradigms.

References
