Institute for Physical AI @ John Bailey Institute · The Charlot Lab

DECISIONS · ENERGY · MEASUREMENT

Technical Report TR-2026-48
Research / Position · Preprint v1.0
22 September 2026

DECISIONS · ENERGY · MEASUREMENT

What a Decision Costs

A typed-decision model answers without generating text, and the saving is real. It is also bounded by a quantity the claims do not mention: the accuracy of whatever decides which decisions go to it.

David Jean Charlot, PhD, Dean of Physical AI · The Charlot Lab

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org

Instrument: what a decision costs · Course: PAI-360

1 ceiling derived in closed form | 2 published claims read backwards | 1 quantity recorded as not located | readout metered on the Institute's own silicon

Abstract. A class of models shipped in 2026 that answers a question without writing a sentence: unstructured state in, a typed value and a calibrated probability out, in one forward pass with no autoregressive loop. The case for it is an energy case, and vendors put the saving at a hundredfold and above. This report accepts the mechanism and computes what bounds it. A decision routed to the typed engine and answered wrongly pays the typed pass *and* the generative fallback it still needs, so the composed cost is $E_2 + E_{read} + (1 - p)\cdot E_3$ and the saving tends to $1/(1 - p)$ as the typed model and its readout approach zero. The ceiling therefore belongs to the router, not the model, and no improvement behind the router can raise it. Read backwards, a hundredfold claim requires a router correct 99.00% of the time and a 444.6-fold claim requires 99.78%; the only routing accuracy this review located in the open record is 0.799, reported on held-out splits of its own training corpus by authors who state that out-of-source generalisation has not been measured. That caps the demonstrated saving near fivefold. None of this refutes the claims. It locates them:

> they rest on a quantity their own literature does not publish, and the report ends by naming the measurement that would settle them.

1. What is in scope, and what does this continue?

This report is about the energy of one decision, treated as an accounting problem across two paths rather than a benchmark of either. It does not propose a model and it does not rank vendors. It continues three lines of Institute work. *Where Is the Energy Reporting?* (TR-2026-46) established that in embodied systems the constraint is disclosure rather than instrumentation, and that a ledger pricing actuation while omitting the decision stops being adequate exactly as machines get good.[1] *The Unknown Fraction* (TR-2026-44) established the habit of computing what a system cannot see and reporting it alongside what it can.[2] *Building the Energy-Compute Future* (TR-2026-26) set the bar a novel paradigm must clear, and the shape of claim that survives it: a named task, a stated energy budget and a stated latency bound, rather than a speedup ratio that a cleverer algorithm can overturn.[3] The contribution here is to apply that bar to a class of system that arrived after it was written.

2. What does one decision cost?

The generative baseline is not mysterious: it is output tokens multiplied by joules per output token, and both halves are published. Measured figures span roughly an order of magnitude with the hardware: 0.39 J per token for a 70B model on an H100 at FP8, against 3 to 4 J per token for a comparable model on older V100 silicon.[4] A peer-reviewed 2026 analysis in *Joule* found widely circulated per-query figures overstated by four to twenty times relative to production-optimised deployments, which is a caution about the baseline before any ratio is computed against it.[5] For a 200-token answer at the measured figure, one generated answer costs 78.0 J. Every ratio in this report is stated against that number, because a ratio whose baseline is unnamed is not a measurement.

3. Why does the router set the ceiling?

Relocating a decision is not free of the thing it relocates away from. Route with accuracy p and three terms are paid: the typed forward pass E_2, the energy of reading its answer back E_{read}, and, on every miss, the generative answer E_3 that is still owed. The composed cost is therefore

$$E_{composed} = E_2 + E_{read} + (1 - p)\cdot E_3$$

and the saving is $E_3 / E_{composed}$. Now take the most generous assumption anyone could make about a model and let both E_2 and E_{read} go to zero. The saving does not diverge. It converges to

$$S_{max} = 1 / (1 - p)$$

which contains no property of the model at all. This is the report's central point and it is elementary rather than deep: the ceiling is fixed by the router before a model is chosen, and a free typed engine cannot exceed it. A router correct 79.9% of the time cannot return more than 4.98× whatever runs behind it.

4. What do the published claims require of the router?

The identity inverts. Given a claimed saving S, the accuracy it requires is $p = 1 - 1/S$. Applying that to the figures in the open record turns cost claims into statements about routing.

Published claim	Saving	Router accuracy required	Reported by its source?
"Up to 100× cheaper" [6]	100×	99.00%	no
444.6× cheaper on production workflows [6]	444.6×	99.78%	no
Kev-0.5B, accuracy read forwards [7]	4.98×	79.90%	yes, in-distribution only

The one measured accuracy in the table is the open reconstruction's, and its own model card is careful about it: 0.799 overall on held-out

splits of the six datasets it was trained on, expected calibration error 0.065 falling to 0.031 after temperature scaling, 7% option-order sensitivity, and the explicit statement that “out-of-source generalization has not been measured”.[7] The closed system's architecture is itself an inference: the most serious public reconstruction is a black-box study from latency-signature probing across roughly ten thousand API calls, which offers two readout mechanisms consistent with the evidence rather than one.[8] The open models[12] reproduce a hypothesis about the closed one, which is worth knowing before their numbers are read as measurements of it.

> The gap between 79.90% and 99.00% is not a criticism of any model. It is the observation that a hundredfold claim and a fivefold claim differ only in a quantity neither vendor nor reconstruction reports, and that the quantity is cheap to measure for anyone who has deployed one.

5. Why does readout decide the rest?

Once the router is fixed, the remaining term that moves the composed cost is reading the answer back, and it is the term energy claims omit most reliably. The Institute metered both ends of that transaction on a Kria KV260: a node update costs 9.13 ± 0.13 pJ measured run-minus-halt at 69σ with the clock tree cancelling, while reading one value back costs 581 ± 56 pJ including the host core, a ratio of 64 to 1.[9] Against the Landauer floor of $kT \ln 2 = 2.871 \times 10^{-21}$ J at 300 K,[10] the metered operation sits about 3.2×10^{9} above the floor and the metered read about 2×10^{11}. An advantage computed over operations and silent about readout is bounded by the term it omitted, which is the same failure this Institute documented in the thermodynamic-computing literature and the reason readout is a separate axis in the companion instrument rather than a component of the model cost.

6. What would a usable report contain?

Four fields, none of them expensive, and a claim that omits any one of them cannot be checked.

- **The baseline, with its hardware.** Joules per generated answer, the model, the precision and the token count. A ratio without this is not a measurement.
- **The router accuracy, out of distribution.** Measured on traffic the router was not trained on, because that is the population it will meet. In-distribution accuracy sets no ceiling anyone can rely on.
- **The readout.** The number of values that cross back per decision and what each costs, measured rather than modelled.
- **The composed figure.** Joules per decision over the whole path including misroutes, reported as a budget against a named task rather than as a ratio.

Stated that way the result becomes a resource claim in the sense of TR-2026-26: falsifiable, resistant to a cleverer algorithm because the algorithm must beat a budget rather than an exponent, and forced to name an application.[3]

7. What is this report careful not to claim?

It does not claim that typed-decision models are not worth using. A fivefold reduction in the energy of a decision is a large result and would be reported as a triumph in most parts of this field. It does not claim any vendor figure is false: a hundredfold saving is entirely achievable at a router accuracy of 99%, and this review has no evidence about what accuracy any deployed router achieves. It does not claim the ceiling is the only thing that matters; latency, and the removal of a whole class of format error, are real benefits this accounting does not price. It does not claim the two open models measure the closed one, for the reason given in §4. And it does not claim the arithmetic here is novel, since it is one line of algebra, which is rather the point. The contribution is applying it to claims that are being quoted without it.

One broader caution belongs here. A 2026 replication testing twenty post-2021 architecture modifications at 1.2B and 3B under iso-data, iso-compute and iso-recipe control with a multi-seed noise floor found that only two cleared Bonferroni correction at the smaller scale, and one of those two failed to train stably at the larger one; modifications that landed within 2–3% of baseline validation loss dropped 6–16

downstream points.[11] The base rate for a promising mechanism surviving contact with a controlled comparison is low, and a report arguing about a mechanism's accounting should say so.

8. Searched for and not located

One quantity, and the whole argument turns on it.

Quantity	Why it decides the claim
A published out-of-distribution router accuracy for any deployed typed-decision system	It sets the ceiling. Without it, every cost claim in this class is conditional on an unreported number.

This review also did not locate a composed joules-per-decision figure, typed pass, readout and misroutes together, for one named task, for any system in this class. The absence is consistent with the pattern the companion reports found in embodied energy [1] and in agricultural sensing [2]: the quantities exist and are cheap, and the composed figure is the one none of these reviews located.

9. Conclusions

The saving from moving a decision off the generative path is bounded above by $1/(1 - p)$, a quantity containing no property of the model. The published hundredfold and 444.6-fold claims therefore require routers correct 99.00% and 99.78% of the time, and the only accuracy in the open record is 0.799 measured in-distribution, which caps the demonstrated saving near fivefold. Readout, metered here at 64 times the operation it reads, is the term that decides what remains once the router is fixed. The constraint on this subject is measurement: not instrumentation, which is trivial, and not the models, which work.

10. The forcing function

The engineering change that moves this bound is specific and available now: measure the router out of distribution and publish the composed figure with it. One table, joules per decision across a real traffic mix,

including readout and misroutes, against a named task and a stated budget, converts every claim in this class from an assertion into something a reader can check in a line. It requires no new hardware and no new model. Until it exists, a hundredfold claim is a statement about a router, and the quickest way to make it a true one is to measure the router. That is a week of work against a result the whole class is waiting on.

References

1. Institute for Physical AI @ JBI, *Where Is the Energy Reporting in Physical AI?*, Technical Report TR-2026-46. COMPANION REPORT, READ IN FULL
2. Institute for Physical AI @ JBI, *The Unknown Fraction*, Technical Report TR-2026-44. COMPANION REPORT, READ IN FULL
3. Institute for Physical AI @ JBI, *Building the Energy-Compute Future*, Technical Report TR-2026-26, the dequantization bar and the resource-claim form. COMPANION REPORT, READ IN FULL
4. Published per-token inference energy: 0.39 J/token for a 70B model on H100 at FP8, against 3–4 J/token for a comparable model on V100. MEASURED, SECONDARY COMPILATION
5. Microsoft Research, energy use of AI inference, efficiency pathways and test-time scaling, *Joule*, 2026: widely cited per-query figures overstated 4–20× against production-optimised deployments. MEASURED, PEER-REVIEWED
6. TypeSafe AI, *Introducing System One Models and Jev*, product announcement: 70–500 ms end-to-end, "up to 100× cheaper", 444.6× on production workflows. VENDOR-REPORTED, NOT INDEPENDENTLY AUDITED
7. J. Palmer, *kev-0.5b* model card: 0.799 accuracy, ECE 0.065 → 0.031 at T=1.47, 7% option-order sensitivity, ~1.75 h training on one Apple M5; states out-of-source generalisation not measured. MEASURED, PRIMARY, IN-DISTRIBUTION ONLY
8. A. Hume, *Jev's Architecture Unmasked*: black-box reconstruction from latency-signature probing across ~10,000 API calls, offering two readout mechanisms consistent with the evidence. INFERRED, NOT CONFIRMED BY THE VENDOR
9. Institute for Physical AI @ JBI, metered energy ledger, Kria KV260: 9.13 ± 0.13 pJ per node update at 69σ, 581 ± 56 pJ per value read back. MEASURED ON INSTITUTE SILICON
10. R. Landauer, irreversibility and heat generation in the computing process, 1961; $kT \ln 2 = 2.871 \times 10^{-21}$ J at 300 K. ESTABLISHED PHYSICS
11. *Most Transformer Modifications Still Do Not Transfer at 1–3B*, arXiv:2605.20798: of 20 modifications, two clear Bonferroni at 1.2B and one of those fails to train stably at 3B. MEASURED, PREPRINT
12. Convai Innovations, *Laya*, Apache-2.0: ModernBERT encoder with a decision head, Jev-compatible API, open weights. PRIMARY, OPEN WEIGHTS; REPOSITORY PUBLISHES NO BENCHMARK

© 2026 Institute for Physical AI @ John Bailey Institute.
Survey and position. No novel experimental results are
reported; the metered figures in §5 are the Institute’s
own and are cited as such.