The Token Is Not the Unit

Cost per Useful Decision, or why the falling cost of inference shifts cost toward proof and risk. Doctrinal note 1/4.

1. The thesis before the equation

Vendors bill inference by the million tokens. A reductive reading of FinOps applied to AI, dominant today, adopts this commercial unit and turns it into the primary proxy for economic efficiency. The problem is not that the token is measured. The problem is that it serves as the measure of efficiency of a system for which it is only the input, an input bound to become more substitutable and less differentiating than the institutional capabilities that surround the decision.

The thesis of this note holds in one distinction, under one condition. When a system influences or executes a trade-off that commits the institution, it is sometimes billed by the token, but it is *steered* by the decision. Outside that domain, the token remains a legitimate unit and this note claims nothing. Three units are conflated in the ambient discourse and must be separated: the unit of *billing*, the resource purchased from the vendor, the unit of *management*, what the operator monitors, and the *economic* unit, that by which efficiency is judged. In the committing domain, the economic unit is not the inference, it is the decision.

The corollary is the real object of this note. The cost of a decision reduces to three charges, gathered below under the letters O, K and H:

- Operating cost O, which sums compute (written O_tech) and human supervision (written O_human),
- The cost of governance and proof K,
- The risk charge H.

When the cost of compute tends toward zero, the cost of the decision does not disappear: it changes composition and, often, bearer. The charge recomposes toward the supervision that remains, the proof still to be produced and the risk still to be absorbed, that is toward O_human, K and H.

This proposition must be falsifiable, and its refutation condition deserves careful statement, since a lazy formulation would make it either trivial or untestable. Two statements must be distinguished.

- The first is analytical: with the rest of the cost structure held fixed, reducing the technical component O_tech alone mechanically raises the share of non-technical charges in the total. This is an identity, not a prediction, and it proves nothing.
- The second is empirical, and it is the thesis: in real deployments, a fall in O_tech does not produce a proportional fall in the non-technical charge, that is in residual supervision, proof and risk, gathered in the aggregate O_human + K + H.

These charges are governed by regulation and exposure, not by the price of the GPU. Their elasticity to the cost of compute is low. The refutation is clear: if, at constant perimeter, volume, quality and liability regime, the absolute level of O_human + K + H per decision fell at the pace of O_tech, the migration thesis would be false. This note does not claim to establish that low elasticity empirically. It proposes its measurement frame and its test condition.

A warning follows. A fall in the total cost per decision does not refute the thesis, because that total may decline for reasons foreign to compute: mutualisation of the proof base, better architecture reducing exposure, higher volume, contractual transfer of risk. The thesis bears on composition and on the residual, not on the level.

The domain of validity is narrow, and I state it before going further. The boundary is not the technical form of the task but its effect: the frame begins when the output of the system materially modifies a right, an obligation, an exposure or a trade-off borne by the institution. Batch processing or inconsequential reformatting stays outside it. A simple classification enters it, by contrast, if it denies access, prioritises an investigation, excludes a file, steers a diagnosis or raises a premium.

2. The decision as minimal unit of commitment

A firm does not turn to AI in order to accumulate generated tokens. It turns to AI to produce a result. Two objects that the ambient discourse conflates must still be distinguished: the *minimal unit of commitment* and the *ultimate unit of value creation*.

The decision is the first, not the second. It is the first institutional state to which the organisation can attach a commitment, a liability and a cost. A determining algorithmic recommendation suffices, even without the legal form of a decision. Final economic value, for its part, depends on the later execution of the business process, over which the AI system no longer has any hold. Conflating the two leads to two symmetrical errors: crediting AI with all the downstream value, or denying it any cost of its own upstream. Cost per Useful Decision sits exactly between the two, on the commitment segment.

The chain reads as follows. Compute is the raw input, inference the system resource, the decision the commitment, the action the execution, the outcome the impact, value the entry on the balance sheet. The perimeter of CUD stops at the decision, the first point at which delegation produces a measurable institutional commitment. The institution continues to control execution, remediation, appeal and reversibility, but these controls act downstream of a commitment already made.

The frame covers the full set of regimes of decisional participation, and this generality is deliberate. Whether the system executes the trade-off autonomously or formulates a determining recommendation validated by a human, it commits the institution. Where human validation is routine or weakly adversarial, it does not neutralise the economic influence of the system; it mainly displaces its formal signature. An autonomous validation, informed and endowed with the time to object, is another case, and CUD separates them instead of conflating them.

3. The separation, and what distinguishes it from Shannon

The founding gesture of CUD is a separation. Just as information theory separated the quantity of information from the physical medium that carries it (Shannon, *A Mathematical Theory of Communication*, 1948), the economics of AI must separate the decision produced from the compute resource that makes it possible. The cost of the medium was not the measure of the quantity of information transported; ***the token count is no more the economic unit of the decision***.

The analogy is useful. It is also asymmetric: Shannon's quantity of information could be formalised independently of its medium, in a unit, the bit, without institutional convention. CUD has no such support. The useful decision is not a natural magnitude: it is defined by a conjunction of non-compensatory quality thresholds, a maturation window and a remediation budget, all set by governance.

Where Shannon had a formal unit, CUD rests on a governed institutional convention. Its robustness therefore does not derive from a natural unit, but from the stability of its definitions, its thresholds and its imputation rules. This is not an admission of weakness slipped in out of caution, it is the very subject of this note, and everything that follows draws the consequences.

4. The equation, single, and its limit

Here is the only formal block the body of this note will carry. The reference equation and its limit property form one and the same gesture; the entire actuarial and accounting

apparatus goes down into the annexes, where it serves rigour without cluttering the argument.

$$CUD_i(t) = \frac{O_i(t) + K_i(t) + H_i(t)}{D^U_i(t, \Delta_i)}$$

$$\lim_{O_{tech} \to 0} CUD_i(t) = \frac{O_{human,i}(t) + K_i(t) + H_i(t)}{D^U_i(t, \Delta_i)}$$

The numerator gathers the whole of costs into three exclusive aggregates:

- O is the operating cost of the delegation regime, compute (O_tech) and supervising humans (O_human) taken together,
- *K* is the cost of the governance and proof capability, the architecture that maintains the attributability, the integrity and the revisability of what has been decided,
- *H* is the economic charge of risk, the financial translation of exposure.

An option cost, a vendor dependency, reversibility, exit, legal or transition cost does not escape the frame: it is imputed to one of the three aggregates according to its nature, under a stable allocation convention.

The denominator *D^U* counts *useful* decisions: institutionally closed, compliant with the quality thresholds, having not exceeded the remediation budget set for the class, and still compliant at the end of an observation window *Δ*. By way of illustration, an institution might adopt thirty days for certain credit decisions and ninety for certain claim settlements, the choice of *Δ* depending on the timing of appeals, the delay in loss materialisation and the business cycle. A decision invalidated during the chosen window cannot be counted as useful in the cohort considered.

Three precautions before any calculation. First, the numerator mixes distinct cost statuses: *observed* costs, such as billed compute, *imputed* costs, such as the share of the mutualised base allocated to the class, *estimated* losses, such as expected loss, and *capital* charges. Confusing these statuses means taking a management construct for a fully observed measure. Second, all terms are computed over the same period, the same perimeter and the same cohort, closed at the same maturation window; a numerator and a denominator that do not share their horizon measure nothing. Third, minimum quality is a *conjunction of non-compensatory thresholds*: a system does not redeem defective traceability with better latency, every threshold must pass, none compensates another.

There remains the property that justifies this note. Under the assumptions stated, at constant perimeter, volume, quality, supervision architecture and risk distribution, the other terms are held constant in order to isolate the effect of the fall in O_tech, which then leaves the numerator. This is comparative statics, not an empirical law: in practice, cheaper compute may induce more usage, more volume, new cases, and therefore move those terms. Subject to that reservation, AI in production does not become free as inference collapses. What persists is not a constant amount, the total may rise or fall, but a non-technical residual charge that becomes preponderant. Scarcity does not disappear. It moves to other pockets, and sometimes to other payers.

Box. A convergence scenario toward the same residual level

Illustrative figures, in normalised units, indexed and not empirical. The box verifies a mechanism, it does not prove a general property. Volume is held constant in order to isolate the composition effect, and the supervision, proof and risk charges are assumed sticky, an assumption which is precisely the thesis to be tested in the field.

Take a class of credit granting decisions, two delegation regimes, costs stated per useful decision.

Assisted architecture, the AI recommends and the human decides: compute 1, human supervision 5, proof 2, risk 2. Cost per decision 10.

Automated architecture, the AI decides and the human handles the exception: compute 2, residual supervision 2, proof 4, risk 3. Cost per decision 11.

First observation, counter-intuitive: in this example, before any commoditisation, automation costs more per unit, not less, even though the cost of compute is relatively small.

Now let compute tend toward free. The assisted regime moves from 10 to 9, the automated one from 11 to 9. In this scenario, the two regimes converge toward the same residual level.

Note: this equality is a parameterisation assumption, not a general consequence of CUD. It holds because the figures were set that way. And this level is a floor only under the assumptions of the scenario, at constant volume, supervision, proof and risk.

Second observation: the non-technical charge, O_human + K + H, absorbs a growing share of cost as compute recedes, and the bulk of that residual, in the automated case, lies in proof and risk.

What the box verifies: the mechanics of the model, and the fact that commoditisation of compute, on its own, does not break through the residual set by the non-technical charges. What it does not prove: the height of that residual and the gap between architectures, which depend on the distribution posited and will have to be established

on real data. What it suggests: at equal residual, automation lodges the bulk of its cost in proof and risk, pockets that no progress on inference optimises. If it is adopted, it is not for a unit saving, nil here, but for a reason that does not show up in CUD.

5. Who sets the denominator

The most serious objection against CUD is the following.

CUD, it will say, is not a measure but an instrument of policy disguised as a metric. Its denominator depends on thresholds that governance itself sets: whoever controls minimum quality, the remediation budget and the maturation window controls the count of useful decisions, hence the result. An executive who believes he is buying the true cost per decision is buying a negotiated number.

The objection lands, and deserves to be retained, but it proves less than it believes. CUD remains a measure; it is simply not a natural measure, it is a *conventional and governed* measure, whose quality depends on the governance of thresholds, perimeters and imputations. That does not disqualify it. EBITDA, RAROC, full costing and capital allocations form the same family of conventional metrics, without sharing its status: EBITDA is a weakly standardised financial indicator, RAROC an internal risk construct, full costing an allocation convention. The comparison holds as a family, not as an equivalence, and institutions are nonetheless steered with all of them. Conventionality is not the hidden vice of CUD, it is the ordinary regime of all management accounting. What CUD adds is that it makes the trade-off on thresholds explicit and quantified, where cost per token buries it under an appearance of technical neutrality.

It remains that the partition of the numerator is never purely analytical. The point is not who ordinarily pays, but that costs are distributed across several functions whose incentives differ. As an archetype, and not as an accounting truth, *O* is mostly visible in IT or platform budgets, *K* in the governance and risk functions, *H* in the business line or the central balance sheet; reality blurs that split, the business funding part of *O*, IT bearing certain proof costs, capital and insurance often being centralised. It remains that a cost per decision crossing several functions will be arbitrated at its seams, and that the power to set the thresholds determining the denominator is a real power. CUD measures. It does not align incentives. The token is commoditised, the decision is governed, and governing it presupposes naming who pays what and who defines what.

6. The seat-kilometre, subject to homogeneity

Commercial aviation offers the best precedent for a production unit, provided it is imported together with its limit. An airline does not measure its performance by the litre of kerosene burned, but by the cost of the available seat-kilometre, computed under a strict constraint of safety and continuous maintenance of fitness for use. CUD is the

equivalent of the seat-kilometre for decision systems: a production unit priced under a quality constraint, not a raw input.

The limit comes before the analogy, not after. The seat-kilometre is far more standardisable than the institutional decision, though not perfectly homogeneous: a long-haul business seat and a short-haul economy seat are not equivalent, and airlines produce variants of them. But it tolerates an aggregation that the decision does not tolerate without strict segmentation. The analogy holds for the logic of a production unit, it does not hold for a degree of homogeneity that the seat itself possesses only imperfectly. CUD therefore requires, before any calculation, a segmentation into comparable risk classes. An aggregated CUD without that segmentation is a number without a referent, and the same requirement governs the reading of any portfolio (Annex C).

This segmentation grounds the only genuinely operational use of the frame: comparing delegation regimes at equivalent quality, reversibility and risk profile. Whether an automated architecture wins or loses that unit comparison does not close the architecture decision, which may follow from a capacity or strategic dominance foreign to unit cost; that diagnosis, like the monitoring of the gap between forecast and observed CUD, is treated in Annex D.

A word finally on what CUD is not. It serves to compare architectures, arbitrate a mode of delegation, test a business case and detect a migration of costs. It does not measure the value of AI, only the cost of decisional commitment. Confusing it with a measure of value would reproduce, one notch higher, the error of unit that it corrects.

7. What the stability of the equation depends on

The economic stability of CUD rests on an essential condition that the equation does not show: the governability of the representation chain that produces the decisions, models, embeddings, guardrails, APIs and orchestration dependencies.

Let a vendor unilaterally modify its upstream model, its moderation policy or the latent space of its embeddings, and three things move together. The proofs already recorded lose part of their reach, because they attest to a system that no longer entirely exists. The cost of requalification and non-regression rises. The distribution of errors deforms, and the calibration of the risk charge with it. An exogenous variable, insufficiently framed by contract, then moves two terms of the numerator at once.

The lesson is not that models and embeddings must be owned. An institution does not need ownership; it needs rights sufficient to identify versions, be notified of changes, reproduce decisions, reconstruct proofs and migrate uses. The key notion is *contractual and technical governability*, not ownership. A good contract organises precisely what an exogenous chain threatens: notification, versioning, retention, reversibility, minimum

stability. A stable cost per decision presupposes a governable representation chain, and that condition is not economic but architectural and contractual.

8. Conclusion

Five acquisitions close this first note. The token measures a resource, legitimate for billing and for monitoring, illegitimate for judging efficiency. The useful decision is the steering unit, because it is the first state that commits the balance sheet. CUD remains a conventional and governed measure, not a natural objectivity. The commoditisation of compute does not reduce the cost of the decision as it reduces the cost of inference: it recomposes its charge toward supervision, proof and risk. And the governability of representations conditions the very stability of the calculation.

CUD therefore promises no true cost freed of judgment. On the contrary, it requires making visible the conventions, the responsibilities and the charges that the price of the token leaves out of frame. Its interest is not to abolish the trade-off, but to place it back at the level where the institution actually commits its balance sheet: the decision. The frame is convincing; the metric, for its part, remains to be tested, and this note offers the instrument, not the proof.

There remains the condition named without being treated, the governability of the representations that support the calculation. That is the object of the next note. Note 2/4, *Sovereign Asymmetry*, deals with what a decision costs when its representations are not governed.

ANNEXES

Annex A. Decomposition of the proof base (K) and boundary with H

The governance and proof base decomposes into five items subject to an exclusive imputation convention: $K = K_g + K_p + K_m + K_a + K_\Delta$. K_g covers governance (ex-ante framing, access policies, risk mapping, alignment rules). K_p is the proof base proper (instrumentation, integrity, versioning, secure retention of logs and contexts). K_m is inline monitoring (real-time guardrails, filters, drift detection, emergency stop). K_a is ex-post audit (independent second-line reviews, sampling, periodic validation). K_Δ is change and requalification (recalibration, requalification on model updates, non-regression tests), amortised over its period of use.

K / H boundary rule, by nature of flow and not by cause of incident. K contains the costs of prevention, of proof, of control and of requalification. H contains the losses, coverage charges and absorption costs, realised or provisioned. One does not classify an incident by its cause, one classifies each economic flow by its nature. The cost of the audit apparatus is in K_a, whether or not the audit detects anything. The loss resulting from an error is in H, whether or not a better audit would have avoided it. Correction of

the apparatus after an incident is in K_Δ. A sanction for a control failure is in H, even when its cause is an insufficiency of K_a, because it is a charge suffered, not an apparatus. An H that drifts while K_a is nominal reads in the gap between observed losses and losses anticipated under the declared level of control, ΔH = H_observed − H_expected; it may signal an undersized apparatus or a risk model that has become inadequate, a change of distribution, fraud, a new environment. Re-reading that gap guides future recalibration, without double counting.

Annex B. Actuarial calibration of the risk charge (H)

To avoid any double counting between gross exposure and coverage instruments, H is built as a sum of terms with mutually exclusive perimeters:

$$H = EL_{net} + C_{ret} + C_{cap} + P_{ins} + M$$

where EL_net is the expected loss net of coverage remaining with the institution, C_ret the financial cost of retention beyond ordinary losses, C_cap the capital charge against untransferred unexpected loss, P_ins the net insurance premium for the risk effectively transferred, and M a margin for model risk. Summation rule: only the net charges effectively borne or paid by the institution add up, expected insurance indemnities are deducted, and no layer overlaps another. Without that discipline, H inflates by superposition, since the premium already incorporates a share of expected loss, capital carries the retained unexpected loss, retention may overlap expected loss, and the model margin may overlap the conservatism of the tail measure. M is therefore defined only for the uncertainty not already carried by that conservatism.

Expected loss is the expectation of loss, EL = E[L]; in the simple case of a binary event, EL = p × L. Unexpected loss is measured on the tail of the distribution at threshold α. Under the convention adopted here, it is the gap between CVaR and expected loss, $UL_\alpha = CVaR_\alpha(L) - EL$, Conditional Value at Risk being preferred to VaR for its coherence as a risk measure (Rockafellar and Uryasev, 2000); other frames use VaR or Expected Shortfall, and the choice must be declared.

The capital charge is computed on the economic capital EC mobilised against untransferred unexpected loss, weighted by its cost r_c: C_cap = r_c × EC. The cost of capital is a financial input, distinct from a risk preference. If the institution wishes to weight its aversion explicitly, it does so in a single declared frame, for example $H_\lambda = EL_{net} + \lambda \cdot UL_\alpha$, without cumulating the two conventions, on pain of counting the same aversion twice.

The estimation of EL and UL_α rests on loss distributions revised regularly. In a strongly non-stationary environment, M materialises in H the uncertainty introduced by a non-stationary representation chain, acknowledged in section 7: an unstable chain makes any loss distribution provisional.

Annex C. Portfolio aggregation and composition indicators

Aggregation of CUD by class is valid only across classes sharing a compatible convention of utility and quality. It is performed through the observed share of useful decisions, which guarantees the accounting identity between the portfolio CUD and the ratio of the sum of costs to the sum of useful decisions, provided that all shared costs have been allocated once, across classes, under a stable convention. Failing that, platform costs are forgotten, imputed several times or left outside the portfolio, and the identity collapses.

Tail warning, raised to a rule. This aggregation is a mean weighted by the volume of useful decisions. It therefore dilutes rare classes carrying a heavy risk charge. A portfolio CUD may look healthy while a critical class, low in volume but heavy in risk, bleeds without the mean signalling it. Rule: no aggregated CUD is published without the distribution of CUD by class and without a tail indicator. Four indicators, chosen for their robustness rather than their simplicity: the distribution of CUD by class, volume-weighted and unweighted; the CUD of the critical classes identified ex-ante; the contribution of each class to H; and risk concentration, the share of H borne by the most exposed classes, together with an uncertainty interval. Maximum CUD is avoided as a cardinal indicator, being unstable on low-volume classes. Portfolio CUD must never become the single figure of the executive committee; failing that, the metric would have replaced cost per token with another average, seductive and mutilating.

Composition indicators. To track the migration, two shares are distinguished: the institutional share s_inst = (O_human + K + H) / CUD, and the proof and risk share s_pr = (K + H) / CUD, both stated relative to cost per decision. Reading them calls for one precaution: at fixed structure, these shares necessarily grow when O_tech falls, and s_inst tends toward one as compute vanishes. They are therefore composition signals, not proofs of the thesis; the thesis bears on the *absolute* level of O_human + K + H, whose low elasticity to the cost of compute remains the empirical statement to be tested.

Anti-Goodhart control. Once CUD becomes a steering target, it becomes manipulable through its definitions, and the main lever of manipulation is the denominator and its taxonomy. Governance of the metric therefore controls, at a minimum: the definition of homogeneous classes, the minimum quality thresholds, the remediation budget, the Δ windows, the allocation rules for mutualised costs, the exclusion policy for rejected decisions, the amortisation of K_Δ, the calibration of λ and of r_c, the tracking of deferred incidents, and changes of perimeter. Closing rule: any modification of the classes, the thresholds, the windows, the remediation rules or the imputation keys produces an explicitly documented break in series. A metric whose definitions are not governed ends up measuring the accounting creativity of those it evaluates.

Annex D. Dominance of regimes and ex-ante / ex-post gap

Dominance. CUD compares the unit cost of the human, assisted, hybrid and automated regimes, at equivalent quality, reversibility and risk profile. An architecture ceases to be dominant in unit terms as soon as its CUD exceeds that of the reference. The loss of unit dominance does not entail abandonment: an architecture may be justified by its *capacity* dominance, continuous availability and volumes inaccessible to humans, or *strategic* dominance, resilience, business continuity, learning. These forms of dominance do not read in CUD; they are arbitrated alongside it. CUD informs unit efficiency, it does not settle the architecture decision on its own. The box in §4 gives the limiting case: two architectures at the same residual level, whose choice can then rest only on a non-unit dominance.

Ex-ante / ex-post gap. For execution monitoring, the gap between observed CUD and forecast CUD is tracked, ΔCUD = CUD_ex-post − CUD_ex-ante. A positive gap may reveal an underestimation of the risk charge H, an undervaluation of the requalification need K_Δ or a remediation rate above the tolerated budget, but also a volume below forecast, an observed cost above forecast, a faulty allocation, an exceptional incident or a regulatory change. In order not to conclude too quickly, the gap is decomposed before being interpreted, ΔCUD = ΔO + ΔK + ΔH + ΔD^U, each term brought back to the same cohort. The gap reads only at constant perimeter and cohort; outside that constancy, it measures a change of perimeter, not a drift in the delegation regime.