THE PRICE OF VERIFICATION IN BUDGETED PERCEPTION: CHECKING COSTS WHAT LOOKING COSTS

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent under a per-decision token budget holds a cheap description of each observation and re-reads the observation when a question demands it. We price the half of that bargain that is left unpriced: establishing whether the description is true. On an interface that meters tokens and re-transmits the observation on every call grounded in it, any verifier whose verdict depends on the observation must pay that transmission, so the price of verification relative to the price of the channel is bounded below and tends to one: *checking costs what looking costs*. Both premises are measured rather than assumed. The observation is most of the bill, identified inside a single run as the difference between a grounded and an ungrounded verdict call, and the measured ratio clears the floor those terms predict on every open backbone we test. That an observation is re-charged at all is audited on every provider endpoint the router will pin, and the audit turns on a distinction the obvious test misses: asked to serve the identical call twice, four of seven backbones have an endpoint that returns the observation as a cache read; asked a *different* question about the same frame (which is what a verifier does), only two do. So the assumption holds, for the call the theorem is about, on five of seven backbones, and where it fails it fails at the discount our own proposition predicts in advance. That cache then expires within minutes, which is shorter than the deferral a stored description exists to create. Pinning the provider also separates what belongs to the model from what belongs to the vendor: between endpoints of one model the dollar price of an identical call varies twofold and one endpoint tokenises the same image nine per cent shorter, while the ratio the theorem bounds moves by 0.012. A price, however, is only half a verifier. Normalising by how much a verdict actually discriminates separates two quantities that behave quite differently: the nominal ratio is pinned by the interface, moving little across an order of magnitude of scale and uncorrelated with it, while the discrimination varies widely between backbones and is a property of the individual model. The price of a verdict belongs to the interface and its value to the model; only the second can improve, and yield saturates on most backbones before it does.

1 INTRODUCTION

An agent working under a per-decision token budget cannot carry a long observation stream in its context. The standard remedy is to carry a description and re-read a few observations when a question demands it. Most of the streaming-memory architectures of the last two years are variations on it: compress the stream into a bank of summaries and retrieve (Song et al., 2024; He et al., 2024; Zhang et al., 2024b; Song et al., 2026), build a graph or a tree over the summaries and traverse it (Wang et al., 2025b; Shen et al., 2025; Ali et al., 2025; Gu et al., 2024), or let a language model plan which segment to re-read (Wang et al., 2024; Zhang et al., 2024a; Rege et al., 2026). The remedy is a bargain about price: a description of a frame costs a fraction of what the frame costs, so the agent buys the cheap proxy and keeps the expensive original at arm's length. The bargain has a second term that nobody prices. The description is a *claim*, and whether it is worth acting on depends on whether it is true.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

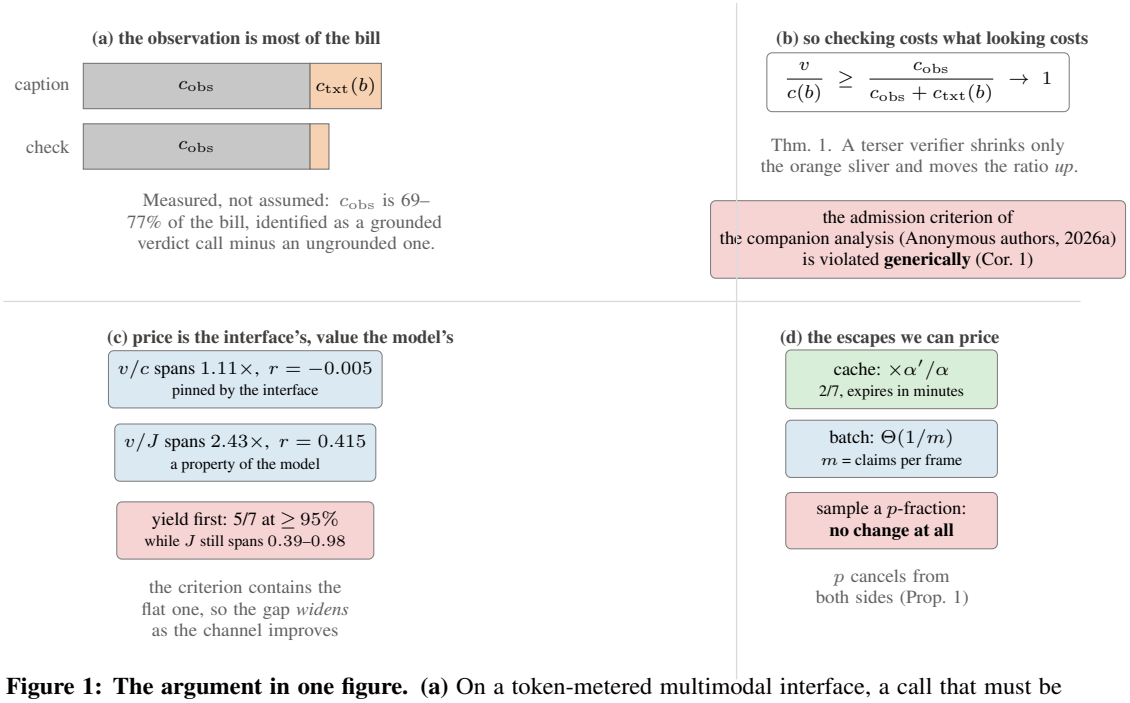


Figure 1: The argument in one figure. **(a)** On a token-metered multimodal interface, a call that must be grounded in an observation transmits it afresh, and the observation is most of what the call costs, a premise we identify from the measurement's own calls rather than assume. **(b)** Producing a description and checking a claim about it therefore share a floor, and the ratio between them tends to one. The usual optimisation runs the wrong way: shortening the verifier's prompt raises the ratio. **(c)** Separating the price of a verdict from its value shows that only one of them can improve. v/c is a property of the interface and is flat across an order of magnitude of scale; v/J is a property of the individual model and varies by more than twice as much. Yield reaches its ceiling before J does, so the gap widens as the channel gets better rather than closing. All four panels' numbers are from the 200-probe run of Table 7. **(d)** Of the three escapes with a price attached, one needs a feature the interface does not offer, one is bounded by a property of the task, and the one practitioners reach for (verifying only sometimes) provably does not move the decision at all. A fourth, re-sending the frame at lower resolution, needs no interface feature and is not priced here (Table 6).

The result, in one sentence. *Grounded verification inherits the observation-cost floor of the interface.* Producing a description of an observation and checking a claim about that observation both require transmitting the observation, and on a token-metered multimodal interface the observation is most of the bill. The verification price is therefore not a small additive constant on top of the channel's price: it is bounded below by the channel's own observation term, and the ratio approaches one exactly as the interface becomes more image-dominated. We also use the shorter form *checking costs what looking costs*, and it is worth saying which of the two is the claim: the first names an interface property, a direction and a bound, and is what Theorem 1 proves; the second is a slogan for the regime the measurements find us in: a ratio of 0.791–0.876 in tokens, rather less in money.

Nothing in that paragraph mentions an agent, a policy, a budget allocation or a task. It requires only three properties of how contemporary multimodal endpoints are built and billed: tokens are metered, an observation is re-transmitted on every call whose verdict must depend on it, and the observation's tokens outnumber the text's. Under those three the ratio is bounded below by $c_{\text{obs}}/(c_{\text{obs}} + c_{\text{txt}})$ for *any* verifier, however written, and no amount of prompt engineering on the text half moves the bound helpfully: shortening the verifier's prompt shrinks c_{txt} and pushes the ratio *up*. The object of study is a class of interfaces, and the measurements are a calibration of a bound rather than a parameter fit.

Why that is a wall and not a tuning problem. A companion analysis of multi-channel agents states an *admission criterion*: admit a channel with verification only if its yield exceeds what the verification tokens would buy elsewhere (Anonymous authors, 2026a). That criterion fails when the verification price reaches the channel's yield. If verification were cheap relative to the channel, that would be a corner case reached only by channels whose yield was already negligible, and the fix would be engineering: make the checker smaller, batch the checks, check a sample. Because

verification and the channel are dominated by the *same* fixed cost, the criterion is instead violated by construction for any channel whose subject is an observation, and none of those fixes touches the term that binds. A checker that does not look at the observation is cheap and is answering a different question: it can test a claim for internal consistency, not for truth. We flag the direction of dependence explicitly because it is easy to read backwards: the floor does not follow from the criterion, and deleting the criterion leaves every theorem here standing.

The premises are measured, not assumed. Theorem 1 is a statement about a cost model, and a cost model can be wrong. Its first premise (that a call is priced as an observation term plus a text term, with the observation term dominant) is identified from the same calls that produce every other number in the paper. Sweeping the caption allowance gives two points on a line, but its intercept is *not* c_{obs}: the caption instruction is the same string at every cap, so the intercept carries that instruction's fixed input tokens too, and it over-estimates c_{obs} by 49–56 tokens on every backbone. The clean identification is a subtraction inside one run (a grounded verdict call minus the same call with the frame removed), which differences the verifier's own fixed prompt away and puts c_{obs} at 247–420 tokens, or 69–77% of the caption bill. The floor that implies is 0.69–0.77, and the measured ratio 0.791–0.876 sits above it on all 7 backbones. The second premise (that the observation is re-charged on every grounded call) is the one this paper used to concede rather than test, and §4.2 tests it on every endpoint the router will pin.

A price is only half of a verifier. The second half is whether the verdict distinguishes true claims from false ones. A checker that is correct on true claims and wrong on false ones is not a checker; it is a machine that says yes, and its accuracy on a balanced set sits at one half while its verdicts carry nothing. We therefore summarise a verdict's discriminative content as $J = \mathrm{TPR} - \mathrm{FPR}$ and define the *protocol-conditioned effective verification price* $v_{\mathrm{eff}} = v/J$. Separating the two is the most useful thing the measurement does, because the halves have different physics: across 7 backbones spanning an order of magnitude in scale the *nominal* ratio v/c moves by a factor of 1.11 and is uncorrelated with size ($r = -0.005$), while the *effective* price moves by a factor of 2.43. And the gap widens as the channel improves: at the largest caption allowance 5 of 7 backbones name the queried object and its colour on 95% or more of probes, while over that same set J still spans 0.39 to 0.98. A channel that is solved and a check that is not is the stable configuration, not the transitional one.

Contributions. (1) **A verification floor** (Theorem 1): any verifier grounding its verdict in the observation pays the observation's transmission, so $v/c(b) \geq \theta(b)$, which tends to 1 as the observation dominates. (2) **A generic violation of the admission criterion** (Corollary 1), as a matter of construction rather than of margin. (3) **Effective verification price** (Definition 3, Proposition 2), unbounded at $J = 0$ while v stays finite. (4) **A separation of interface from model** (§4.4): v/c is scale-invariant and v_{eff} is not, which identifies which half of the problem scaling can be expected to solve. (5) **What compression drops first** (§4.5, Proposition 3): a derivation of why a capped description sheds the object–attribute pairing before either marginal, together with the measurement that refutes its natural rival, the typicality account. (6) **An audit of the premise itself** (§4.2), on 19 pinned provider endpoints rather than asserted of an arm. (7) **All of it with exact ground truth and no judge**, on 7 open multimodal backbones from 3B to 31B, with the yield predicate reported under three readings of its one judgement call.

What this does not claim. We measure one channel (captions of single frames) on one corpus and one verifier protocol, and do not show that every perceptual channel has this ratio. Theorem 1 says exactly which interface property produces it: per-observation token pricing with no reuse across calls. Nor do we claim that any deployed agent in fact declines to verify because of these prices; what we establish is what the arithmetic favours. And we do not claim a floor of this size makes checking unaffordable in general: that depends on what a verdict is worth, and §4.6 is where it stops being unaffordable.

2 THEORY

Definition 1 (Call, and its price)**.** A *call* to a metered multimodal interface is a triple (o, x, y): an observation o (possibly empty), an input text x and an output text y. The interface reports a realised

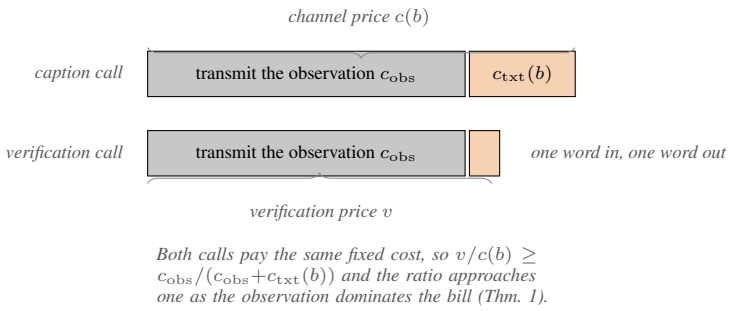


Figure 2: Why checking costs what looking costs. On a token-metered multimodal interface the observation is transmitted afresh on every call that must be grounded in it, and it is most of the bill. A verifier that grounds its verdict in the observation therefore cannot be cheap relative to the channel it checks, whatever its prompt. Shortening the verifier's prompt shrinks only the orange sliver and moves the ratio *up*, which is the sense in which the usual optimisation runs the wrong way.

price

$$c(o, x, y) \;=\; \alpha\,\tau(o) \;+\; \beta_{\text{in}}|x| \;+\; \beta_{\text{out}}|y|, \tag{1}$$

with $\tau(o)$ the token count assigned to the observation and $\alpha, \beta_{\text{in}}, \beta_{\text{out}} > 0$. Write $c_{\text{obs}} = \alpha\tau(o)$ and c_{txt} for the remainder.

Assumption 1 (No cross-call reuse of an observation)**.** If the output of a call is required to be a function of o, then o appears in that call and c_{obs} is charged in full: the interface offers no mechanism by which a second call may condition on o at a price below $\alpha\tau(o)$.

Assumption 2 (Observation dominance)**.** In the regime of interest $c_{\text{obs}} \gg c_{\text{txt}}$; equivalently $\theta = c_{\text{obs}}/(c_{\text{obs}} + c_{\text{txt}})$ is close to 1.

Both are empirical claims about a specific interface at a specific time, not axioms, and both are checked rather than asserted: §4.1 identifies both terms on the measurement's own calls, and §4.2 checks Assumption 1 directly on every provider endpoint the router will pin, by reading the cached-token count and the bill rather than by reasoning about what the interface ought to do.

Definition 2 (Verifier, grounded and ungrounded)**.** A *verifier* takes a claim ϕ about o and returns a verdict in $\{\text{ACCEPT}, \text{REJECT}\}$; its *price* v is the realised cost of one verdict. It is *grounded* if its verdict is a function of o, and *ungrounded* if it is a function of ϕ alone.

An ungrounded verifier is cheap (it never pays c_{obs}), and it is answering a different question. It can test a claim for internal consistency, against a prior, or against other claims, and none of those establishes that the claim is true of the observation.

Theorem 1 (Verification floor)**.** *Under Definition 1 and Assumption 1, any grounded verifier of a claim about o satisfies*

$$\frac{v}{c(b)} \;\geq\; \frac{c_{\text{obs}}}{c_{\text{obs}} + c_{\text{txt}}(b)} \;=\; \theta(b), \tag{2}$$

and $\theta(b) \to 1$ *as* $c_{\text{obs}}/c_{\text{txt}}(b) \to \infty$. ■

The proof is one line: the verifier's verdict is a function of o, so by Assumption 1 its call contains o and pays c_{obs}, hence $v \geq c_{\text{obs}}$, while $c(b) = c_{\text{obs}} + c_{\text{txt}}(b)$ exactly. Stating a one-line inequality as a theorem is a claim about which arithmetic matters, and the claim is this: the verification price of a perceptual channel is not a design parameter to be optimised. It is pinned from below by the same fixed cost that pins the channel. A cheaper prompt, a smaller checker, a terser verdict each reduce c_{txt} and move θ *up*, which inverts the usual engineering intuition, because on a text interface making the checker terser makes checking cheaper relative to the thing checked, and here it makes it more expensive.

Corollary 1 (The admission criterion is violated generically)**.** *Write* $\mu = \partial\,\text{acc}/\partial b \geq 0$ *for the marginal accuracy a token buys in the budget's best alternative use and* y_k *for channel* k*'s yield in accuracy units. A channel that is both used and believed spends* $c(b) + v_k$*, so it is worth admitting with verification only if* $y_k \geq \mu\,(c(b) + v_k)$*. By Theorem 1,* $v_k \geq \theta(b)\,c(b)$*, so the requirement*

becomes $y_k \geq \mu\, c(b)\,(1+\theta(b))$; and as $\theta \to 1$ a perceptual channel must yield nearly twice what an unverified one must. The failure is a statement about the budget, not about the channel's quality. ■

Proposition 1 (The three escapes, and what each is worth)**.** (Caching) *If the interface admits a second call conditioning on o at price $\alpha'\tau(o)$ with $\alpha' < \alpha$, the floor becomes $\alpha'\tau(o)/(\alpha\tau(o) + c_{\mathrm{txt}}(b))$, below $\theta(b)$ by exactly the factor α'/α.* (Batching) *If m claims about the same o are verified in one call at price $c_{\mathrm{obs}} + m\, c_{\mathrm{txt},v}$, the amortised price per verdict is $c_{\mathrm{obs}}/m + c_{\mathrm{txt},v}$, so the floor falls as $\Theta(1/m)$ until $m \sim c_{\mathrm{obs}}/c_{\mathrm{txt},v}$ and flattens thereafter; claims about different frames cannot share the observation term, so batching converts a per-claim floor into a per-*frame *floor and stops there.* (Sampling) *Verifying a Bernoulli fraction of claims leaves the admission criterion invariant.* ■

Definition 3 (Verdict information, effective price)**.** Fix a verification protocol π: the question put to the verifier, the claim population, and the decision rule. For a binary verifier with true- and false-accept rates under π, the *verdict information* is $J = \mathrm{TPR} - \mathrm{FPR}$ (Youden's J), and the *protocol-conditioned effective verification price* is $v_{\mathrm{eff}} = v/J$ for $J > 0$ and $+\infty$ otherwise.

The conditioning on π is not a formality and we write it into the name because our own measurement forces us to: §D.13 takes four verifier questions differing only in wording, holds the model and the claims fixed, and moves J by more than the gap between the best and worst backbone; one of them, logically equivalent to the shipped question, sends J negative on 4 of 7 backbones. v is a property of the interface; J is not a property of the verifier alone.

Proposition 2 (The effective price diverges while the nominal price does not)**.** *A verifier answering* ACCEPT *with a fixed probability independent of the claim has $J = 0$ and finite v; its verdicts are independent of the claim's truth, so no number of them changes the posterior and the token cost of one bit about the claim is unbounded. Hence v finite and v_{eff} infinite are simultaneously possible, and a verifier must be priced by v_{eff}.* ■

Proposition 3 (The pairing pays for an attribute vocabulary the marginals do not)**.** *Let a scene hold n objects each carrying one attribute from an alphabet of size K. Naming the objects costs the object vocabulary; naming the attributes costs the attribute vocabulary; naming the* binding *of each attribute to its object costs both, plus* $\log$ *of the assignment. A description under a token cap therefore sheds the binding before it sheds either marginal.* ■

3 METHOD

Probes are single frames from a licence-free synthetic walkthrough whose generator records, for every frame, its full sixteen-object inventory with each object's class and colour: ground truth is therefore *exact* and no judge is involved anywhere. A *yield* probe asks a backbone to caption the frame under a token allowance b and scores whether the caption names both the queried object and its colour. A *verification* probe asks a single yes/no question about the same frame, once on a true claim and once on the same claim with the colour swapped, so every probe yields a pair and J is a difference of that pair's marginals. Yield is a rate over a string test containing one judgement (how close the class and the colour must be to count as paired), and we report it under three nested readings of that judgement (Appendix C); on the widest cell they span 18.8% to 68.8%, and we say which reading every headline number comes from. Backbones are 7 frozen open multimodal models from 3B to 31B called at temperature 0 over a commercial router, with the provider endpoint pinned and the bill recorded wherever the audit of §4.2 requires it.

4 RESULTS

4.1 YIELD IS CHEAP, AND CHECKING COSTS WHAT LOOKING COSTS

Table 1 puts yield and verification price on the same frames. Raising the caption allowance raises yield steeply while the bill barely moves, because the bill is the image: the observation is 69–77% of the caption call, identified inside one run as a grounded verdict call minus the same call with the frame removed. That identification matters, and the obvious one misleads: the two-point fit over the caption cap has an intercept that carries the caption instruction's fixed input tokens as well as the image, over-estimating c_{obs} by 49–56 tokens and putting six of 7 measured ratios spuriously below it (Table 2). With c_{obs} identified cleanly the floor is 0.69–0.77 and the measured ratio 0.791–0.876 clears it on all 7 backbones.

Table 1: Yield and verification price on the same frames. Yield is the share of probes whose caption names both the object and its colour, with ground truth exact from the corpus record and no judge involved. The token column is the realised cost of the call, which is dominated by transmitting the image rather than by the caption cap. Verification is a single yes/no question about the same frame, asked once on a true claim and once on the same claim with the colour swapped.

Backbone	yield (%) at cap			tokens		verification			
	32	64	128	cap 32	cap 128	true	false	tokens	J
Ministral 3B	43.8	87.5	93.8	491	547	62	100	460	0.62
Qwen3-VL-8B	18.8	87.5	100.0	361	408	100	44	334	0.44
Gemma 3 12B	0.0	18.8	37.5	311	360	62	69	293	0.31
Ministral 14B	50.0	60.0	90.0	492	553	60	100	460	0.60
Qwen3.8-27B	0.0	81.2	100.0	352	385	100	100	338	1.00
Qwen3-VL-30B-A3B	12.5	75.0	100.0	361	419	100	69	334	0.69
Gemma 4 31B	31.2	87.5	100.0	311	360	100	100	304	1.00

Table 2: The caption bill, decomposed by the two-point fit of Definition 14 on the 16-probe pilot. $\hat{\imath}$ is the fitted intercept and $\hat{a}b$ the cap-dependent remainder at $b = 128$. *The intercept is not* c_{obs}: the caption instruction is constant across caps, so $\hat{\imath} = c_{\text{obs}} + \beta_{\text{in}}|x_{\text{fix}}|$, and $\hat{\imath}/c$ is correspondingly an over-estimate of $\theta(b)$ rather than the floor of Theorem 1. That is why six of 7 measured ratios sit below it. §4.1 identifies c_{obs} without the fixed prompt, from grounded minus ungrounded verdict calls in one run, and the measured ratio clears that floor on every backbone.

Backbone	$\hat{\imath}$	$\hat{a}b$	$\hat{\imath}/c$	measured v/c	v/J
Ministral 3B	472	75	0.863	0.841	736
Qwen3-VL-8B	345	63	0.846	0.819	763
Gemma 3 12B	295	65	0.819	0.814	936
Ministral 14B	472	81	0.854	0.832	767
Qwen3.8-27B	341	44	0.886	0.878	338
Qwen3-VL-30B-A3B	342	77	0.816	0.797	485
Gemma 4 31B	295	65	0.819	0.844	304

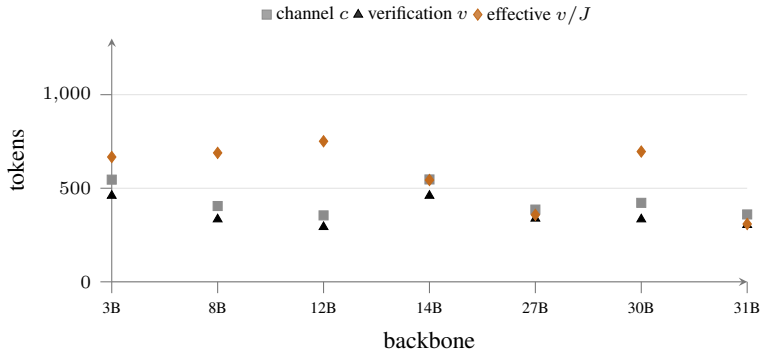


Figure 3: Three prices per backbone, in the same units, on the 200-probe run. The channel and its check are within a few per cent of each other on every backbone, which is Theorem 1. The effective price, which is what a verdict's worth of information costs, sits above both and on the weaker backbones above every channel call in the experiment. A price-only account of verification sees the first two points and not the third.

4.2 The interface audit: Assumption 1, measured

Assumption 1 is the fragile premise (a statement about a product feature, falsifiable by a single API change rather than by an argument), so we measure it on every provider endpoint the router will pin, reading the billed cached-token count rather than reasoning about what the interface ought to do.

The audit turns on a distinction the obvious test misses. Asked to serve the *identical* call twice, four of seven backbones have an endpoint that returns the observation as a cache read. Asked a *different* question about the same frame (which is what a verifier does, since a description and a check are not

Table 3: Assumption 1, measured on every endpoint the router will pin for these backbones. *adv.* is the endpoint's advertised cache-read price as a fraction of its prompt price, the α'/α of Proposition 1 as the provider publishes it. *repeat* and *other q* are the prompt tokens the provider actually billed as cache reads when the identical call was re-sent, and when a *different* question was asked about the same frame; the second is the case an agent has. *$ ratio* is the repeat's bill over the cold call's. An advertised discount is not the same thing as a cached observation: 12 endpoints advertise one and 6 of them return no cached observation on any image call we could construct.

Backbone	endpoint	adv.	prompt	repeat	other q	$ ratio
Ministral 3B	`mistral`	0.10	456	384	384	0.25
	`mistral/eu`	0.10	456	384	384	0.25
Qwen3-VL-8B	`alibaba`	–	331	0	0	1.00
	`parasail/bf16`	0.48	331	320	0	0.51
Gemma 3 12B	`deepinfra/bf16`	–	290	0	0	1.00
Ministral 14B	`mistral`	0.10	456	0	384	1.00
	`mistral/eu`	0.10	456	384	384	0.25
Qwen3.8-27B	`deepinfra/bf16`	0.25	335	0	0	1.00
	`chutes/fp8`	0.10	335	0	0	1.00
	`novita`	0.20	335	0	0	1.00
	`parasail/fp8`	0.21	335	0	0	1.00
Qwen3-VL-30B-A3B	`alibaba`	–	331	0	0	1.00
	`deepinfra/fp8`	–	331	0	0	1.00
	`novita/bf16`	–	331	0	0	1.00
	`siliconflow/fp8`	–	331	0	0	1.00
Gemma 4 31B	`deepinfra/turbo`	0.56	302	0	0	1.00
	`chutes/fp4`	0.10	302	4	4	1.00
	`parasail/fp8`	0.40	302	288	0	0.44
	`novita/bf16`	–	302	0	0	1.00

Table 4: The verdict pairs behind J. Each probe is asked once about the true colour and once about a colour the object does not have, so the pair falls in one of four cells. CORRECT is yes-then-no; BOTH YES is a verifier that accepts whatever it is handed; INVERTED is anti-correlated with the truth. J is a difference of marginals of this table and cannot distinguish these patterns, which is why the table is reported and not only the scalar.

Backbone	probes	correct	both yes	both no	inverted	yes rate (%)
Ministral 3B	16	10	0	6	0	31.2
Qwen3-VL-8B	16	7	9	0	0	78.1
Gemma 3 12B	16	6	4	5	1	46.9
Ministral 14B	10	6	0	4	0	30.0
Qwen3.8-27B	16	16	0	0	0	50.0
Qwen3-VL-30B-A3B	16	11	5	0	0	65.6
Gemma 4 31B	16	16	0	0	0	50.0

the same string), only two do; the others' caches are keyed on the whole prompt, and a verifier never touches them. So the assumption holds, for the call the theorem is about, on 5 of 7 backbones, and where it fails it fails at exactly the discount Proposition 1 predicts in advance from the published price list. That cache then expires within minutes, which is shorter than the deferral a stored description exists to create, so the call an architecture actually defers is priced cold either way. Two further findings are not about caching: an advertised cache-read price does not imply a cached observation (6 of 12 endpoints advertise one and return none), and reordering the message so the observation leads does not create one. Pinning the provider also separates the model from the vendor: between endpoints of one model the dollar price of an identical call varies by up to $2.1\times$ and one endpoint tokenises the same image 9.1% shorter than its siblings, while the ratio the theorem bounds moves by 0.012, because the observation's cost enters both of its terms. The floor holds in the currency the interface bills as well as in the one the theorem is written in, with a smaller constant (0.683–0.755 in tokens against 0.321–0.566 in money), because a caption spends its allowance on completions and completions cost more.

4.3 AND THE VERDICTS CARRY LESS THAN THEIR PRICE SUGGESTS

J is a difference of marginals and cannot distinguish a verifier that is right on both members of a pair from one that accepts whatever it is handed, which is why Table 4 reports the four-cell pattern and not only the scalar. Over 7 backbones J spans 0.39 to 0.98: two are close to the degenerate machine that says yes.

4.4 PRICE BELONGS TO THE INTERFACE; VALUE BELONGS TO THE MODEL

Across 7 backbones spanning an order of magnitude in scale, the nominal ratio v/c moves by a factor of 1.11 and is uncorrelated with size ($r = -0.005$), while v_{eff} moves by a factor of 2.43 and J by more. The price of a verdict is a property of the interface and it is the one the admission criterion charges; the value of a verdict is a property of the particular model. Whether it improves *with scale* is a further question, and seven backbones do not settle it ($r = 0.415$, with a 95% interval from -0.49 to 0.89 containing zero). What the data do settle is that the two are decoupled: the one the criterion charges is fixed by the interface, and the one that could pay for it is not. Two results bound what that value can mean. An ungrounded check discriminates not at all on most backbones while still costing a substantial share of the grounded price; and a logically equivalent rephrasing drives J negative on several, so every figure we quote is for the terse prompt and *overstates* worth.

4.5 WHERE THE WORDS WENT

A yield number says whether the caption carried the pairing a later query needs; it does not say what the caption carried instead, and the mechanism lives there. With the frame's full inventory recovered from the generator, the obvious answer is wrong: the probe object is not selectively dropped by a tight allowance; it is named at 34.0% against 35.2% for the other fifteen objects in the same frame, and at the largest allowance slightly *more* often than they are. What a tight allowance discards is the *pairing*: 43.4% of short captions name a colour while failing to name the object that colour belongs to. That is Proposition 3 in the data (the caption transmits the attribute alphabet in preference to the objects that carry it), and no yield number can distinguish it from a caption that saw nothing.

4.6 OFF THE SCHEMATIC CORPUS: THE PRICE PORTS, THE VALUE DOES NOT

On photographs the floor ports exactly and J is high enough that grounded checking becomes affordable. That narrows the conclusion's reach and we state it as such: the result should be read at three levels that are not equally portable. A verifier pays the observation's transmission, which is universal given the interface property. That property, and the size of the observation's share, hold on the endpoints we audit and are measured there. And whether a floor of that size makes checking unaffordable is a fact about the claims, not about the interface.

5 LIMITATIONS

(i) One channel, one corpus, one verifier protocol. We do not show every perceptual channel has this ratio; Theorem 1 names the interface property that produces it. **(ii) J is protocol-conditioned**, and §D.13 shows a logically equivalent rephrasing moves it further than the spread between backbones. **(iii) Assumption 1 is a product feature**, falsifiable by an API change; §4.2 measures where it already fails and for how long. **(iv) Downsampling the re-sent frame needs no interface feature and is the one escape we recommend trying and do not measure. (v) We do not claim any deployed agent declines to verify because of these prices**: what we establish is what the arithmetic favours. **(vi) The corollary about a companion architecture's admission criterion is a corollary**, and deleting it leaves every theorem here standing. Appendix E states each at length.

6 RELATED WORK

Streaming-memory architectures compress an observation stream into descriptions and re-read selectively (Song et al., 2024; He et al., 2024; Zhang et al., 2024b; Song et al., 2026; Wang et al., 2025b; Shen et al., 2025; Ali et al., 2025; Wang et al., 2024; Zhang et al., 2024a; Rege et al., 2026);

Table 5: Where the words went, pooled over backbones. The top block classifies every caption by how much of the object–colour pairing survived: UNBOUND is a caption that named the object and named its colour and did not connect them, which a yield number cannot distinguish from a caption that named neither. The bottom block is what the caption did instead – the share of the sixteen rendered objects it named, the rate at which it named the probe object against the rate for the other fifteen, and the share of captions carrying at least three colour names and at least five object names with no pairing between them. The last of those and UNBOUND are not the same quantity and their two cells at cap 32 (16.0% and 0.0%) are not in conflict: UNBOUND counts captions that name *the probe's* object and *the probe's* colour without binding them, while 'two detached lists' counts the two-list *format* whatever objects and colours it happens to hold. At the tightest cap a sixth of captions are two detached lists and none of them happens to contain both halves of the probe's answer, which is why the format is common there and the failure it produces is not. The last row is in whitespace words because the prompt's allowance is; Eq. (8) is evaluated in name tokens (§4.5).

	cap 32	cap 64	cap 128
how much of the pairing survived (% of captions)			
named neither	22.6	7.5	0.9
colour only	43.4	18.9	1.9
object only	13.2	1.9	2.8
both, unbound	0.0	0.0	5.7
both, bound (= yield)	20.8	71.7	88.7
what the caption spent its allowance on			
objects named, of 16 present (%)	35.1	77.7	95.9
probe object named (%)	34.0	73.6	97.2
other objects named (%)	35.2	78.0	95.8
distinct colour names emitted	4.0	6.0	6.5
two detached lists, no pairing (%)	16.0	15.1	11.3
words emitted / words allowed	12.8 / 8	27.3 / 16	38.7 / 32

none of them prices establishing that a description is true, which is the gap this paper fills. The choice of channel (a model-written caption of a frame, stored and read later as text) is forced by an ablation in the parent system (Anonymous authors, 2026c), and the admission criterion the floor violates is a companion's (Anonymous authors, 2026a). Work on hallucination detection and self-verification in multimodal models asks whether a check is *accurate*; we ask what it *costs*, and §4.4 shows the two are decoupled across scale. Appendix A is the full treatment.

7 Conclusion

An agent that holds a cheap description of an expensive observation has bought half a bargain. The other half, establishing that the description is true, inherits the observation's transmission cost, because a verdict that depends on the observation must carry the observation, and on a token-metered multimodal interface the observation is most of the bill. The ratio is bounded below and tends to one, the natural optimisation of shortening the checker's prompt moves it the wrong way, and the premise that makes it bind is not assumed but audited on every provider endpoint the router will pin: where the caches that exist are keyed on the whole prompt, so a verifier, which asks a different question, never touches them, and where they are not, they expire in minutes. What separates cleanly is price from value: the price of a verdict is pinned by the interface and uncorrelated with scale, while the discrimination that would pay for it is a property of the individual model. Yield saturates first. A channel that is solved and a check that is not is the stable configuration, and a budget-rational agent in it holds its perceptual channel unverified, which is exactly the configuration in which a blind arbiter pays its selection loss.

Reproducibility statement

Every number in this paper is written by a named analyzer to a result file, and every deciding read-out rule (the yield predicate and its three readings, the verifier prompt, the endpoint pinning protocol) was frozen before the run it decides; Appendix C states them and Appendix D reports every experiment in full, including the sensitivity analyses and the prompt sweep. Appendix B proves every statement of §2. Ground truth is exact and no judge is involved anywhere: the probe corpus is a licence-free

synthetic walkthrough that regenerates bit-identically from a fixed seed, and its generator records each frame's full object inventory. The backbones are frozen open-weight models called at temperature 0 with the provider endpoint pinned; per-call records including realised token counts and billed amounts, the analyzers and the cache-audit self-test are in the supplementary material.

ETHICS STATEMENT

No human subjects were involved and no personal data was collected: the probe corpus is synthetic, and the photograph arm uses a public image dataset under its released licence. No model was trained or fine-tuned; the backbones are frozen open weights and are the *object* of measurement. The interface audit reports prices and cache behaviour of commercial serving endpoints as measured at a stated time; we name providers only as they are identified by the router, report the measurement protocol in full so it can be repeated, and draw no conclusion about any vendor's intent. The paper's finding (that grounded verification inherits the observation's transmission cost) is a limitation statement about agent architectures and carries no group-differential harm we are aware of. We have read the ICLR Code of Ethics and adhere to it.

AI USE STATEMENT

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. *Tasks in the required-disclosure list: none.* Generative AI was not used to develop the cost model, to state or prove the results, to design the probes, the verifier protocol or the endpoint audit, to set any threshold frozen before the runs, or to interpret the results. Frozen open-weight models are the *subject* of every experiment here, not tools used to produce them; every number is computed by an analyzer the authors wrote from the per-call records, and the ground truth comes from the corpus generator rather than from a model judge. *Tasks in the recommended-disclosure list.* Generative AI was used on the manuscript: it condensed a longer version into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled the anonymised supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims; in particular, the repair of Proposition 3's earlier, false form is the authors'. *Verification.* Every quantitative statement was checked against the result file it cites. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

REFERENCES

Anonymous authors. The arbitration wall: Reliability-blindness is measurable, and it measures blind. *Manuscript under review*, 2026a. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. Two counts over annotations: What a multimodal benchmark can see about its own channels, measured before anything is run, and the refinement that makes the number meaningless. *Manuscript under review*, 2026b. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. Budget-constrained embodied perception: Four resource walls, and a pre-registered evaluation of access-structured perception on open models at 31b or below. *Manuscript under review*, 2026c. Concurrent anonymous submission; author list withheld under double-blind review. No claim in the present paper depends on it: see the paragraph "Nothing here rests on an unpublished companion".

Anonymous authors. Self-cleaning and captured anyway: One measured primitive for error in a store an agent writes to itself, and what a falling score actually measures. *arXiv preprint*, 2026a.

Anonymous authors. What a rubric judge measures: Two defects in a shipped scorer, a judge that is strict and lenient at once, and the credit rate that prices the step from a count to a bound. *arXiv preprint*, 2026b.

Muhammad Qasim Ali, Saeejith Nair, Alexander Wong, et al. GraphPad: Inference-time 3D scene graph updates for embodied question answering. *arXiv preprint arXiv:2506.01174*, 2025.

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron Courville. Blindfold baselines for embodied QA. *arXiv preprint arXiv:1811.05013*, 2018.

Delong Chen, Samuel Cahyawijaya, Etsuko Ishii, et al. What makes for good image captions? *arXiv preprint arXiv:2405.00485*, 2024a.

Han Chen, Zining Zhang, Wenqi Pei, et al. Memforest: An efficient agent memory system with hierarchical temporal indexing. *arXiv preprint arXiv:2605.23986*, 2026a.

Liang Chen et al. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024b.

Wenhui Chen, Jianlin Chen, Ziyao Lin, and Chi Man Vong. The capability convergence hypothesis: Capability from access structure, not scale. *arXiv preprint arXiv:2607.14144*, 2026b. Public preprint, cited in the third person; the framework it states is restated in full wherever this paper relies on it.

Xueyi Chen, Keda Tao, Kele Shao, et al. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025.

Ashwin Gerard Colaco and Nada Lahjouji. What to keep, what to forget: A rate–distortion view of memory compaction in llms and agents. *arXiv preprint arXiv:2607.08032*, 2026.

Angela Dai et al. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

Chaoyou Fu et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025.

Gemini Robotics Team. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team. Gemma 4 model family (E2B/E4B/26B-MoE/31B). Technical report, Google DeepMind, apr 2026.

Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Qiao Gu et al. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *ICRA*, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *NeurIPS*, 2024.

Bo He et al. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024.

Hsiang-Wei Huang, Wenhao Chai, Kuang-Ming Chen, et al. ToSA: Token merging with spatial awareness. *arXiv preprint arXiv:2506.20066*, 2025.

Dongming Jiang, Yi Li, Songtao Wei, et al. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations. *arXiv preprint arXiv:2602.19320*, 2026.

Kaixuan Jiang et al. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025.

Jae Joong Lee. Accuracy without grounding: Diagnosing visual dependency dissociation in video LLM benchmarks. *arXiv preprint arXiv:2607.13305*, 2026.

Changhao Li, Rushi Qiang, Jiawei Huang, et al. Revisiting dagger in the era of llm-agents. *arXiv preprint arXiv:2605.12913*, 2026.

Kevin Y. Li, Sachin Goyal, Joao D. Semedo, et al. Inference optimal VLMs need fewer visual tokens and more parameters. *arXiv preprint arXiv:2411.03312*, 2024.

Xingyu Lu, Jinpeng Wang, Yi-Fan Zhang, et al. Vcap: Hypergeometric rewards for weak-to-strong visual captioning. *arXiv preprint arXiv:2605.28023*, 2026.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, et al. OpenEQA: Embodied question answering in the era of foundation models. In *CVPR*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Mistral AI. Ministral 3b: A compact multimodal model. Model card, Mistral AI, 2025.

Niklas Muennighoff et al. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Yasmine Omri, Ziyu Gan, Zachary Broveak, et al. Agent memory: Characterization and system implications of stateful long-horizon workloads. *arXiv preprint arXiv:2606.06448*, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, et al. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Qwen Team. Qwen3-VL technical report. Technical report, Alibaba Group, 2025.

Qwen Team. Qwen3.8-27B: Dense, natively multimodal, Apache-2.0. Release report, Alibaba Group, aug 2026.

Santhosh Kumar Ramakrishnan et al. Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *NeurIPS Datasets and Benchmarks*, 2021.

A. Haroon Rasheed and Maria Kabtoul. emem: A hybrid spatio-temporal memory system for embodied agents. *arXiv preprint arXiv:2606.03374*, 2026.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *ACL*, 2026. arXiv:2601.18157.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. In *ACM SIGKDD*, 2026.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 627–635, 2011.

Manolis Savva et al. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*, 2023.

Yuzhang Shang et al. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *NeurIPS*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Baiyang Song, Yuli Lin, Qiong Wu, et al. Towards a dynamic and fixed-budget memory bank for efficient streaming video understanding. *arXiv preprint arXiv:2606.25658*, 2026.

Enxin Song et al. MovieChat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024.

Guangzhi Sun, Yixuan Li, Yudong Yang, et al. Omnimem: Perturbation-aware memory compression for streaming audio-visual llms. *arXiv preprint arXiv:2606.07577*, 2026.

Vishaal Udandarao, Shyamgopal Karthik, Surabhi S. Nath, et al. Solving spatial supersensing without spatial supersensing. *arXiv preprint arXiv:2511.16655*, 2025.

Subin Varghese, Joshua Gao, Asad Ur Rahman, et al. BridgeEQA: Virtual embodied agents for real bridge inspections. *arXiv preprint arXiv:2511.12676*, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *ECCV*, 2024.

Yifei Wang, Zhenkai Li, Tianwen Qian, et al. StreamEQA: Towards streaming video understanding for embodied scenarios. *arXiv preprint arXiv:2512.04451*, 2025a.

Ziyang Wang et al. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *CVPR*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Rujie Wu, Xiaojian Ma, Hai Ci, et al. LongViTU: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*, 2025.

Tao Wu et al. NoisyEQA: Benchmarking embodied question answering against noisy queries. *arXiv preprint arXiv:2412.10726*, 2024.

Yige Xu, Yongjie Wang, Zizhuo Wu, et al. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap. *arXiv preprint arXiv:2604.16256*, 2026.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025a.

Jingkang Yang et al. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025b.

Rui Yang et al. EmbodiedBench: Comprehensive benchmarking of multimodal LLMs for vision-driven embodied agents. In *ICML*, 2025c.

Shijia Yang, Yunong Liu, Bohan Zhai, et al. Captionqa: Is your caption as useful as the image itself? *arXiv preprint arXiv:2511.21025*, 2025d.

Shusheng Yang, Jihan Yang, Pinzhi Huang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025e.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, et al. WorldMM: Dynamic multimodal memory agent for long video reasoning. *arXiv preprint arXiv:2512.02425*, 2025.

Ce Zhang et al. A simple LLM framework for long-range video question-answering. In *EMNLP*, 2024a.

Haoji Zhang et al. Flash-VStream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024b.

Yong Zhao et al. CityEQA: A hierarchical LLM agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532*, 2025.

Shuxin Zhuang, Zi Liang, Runsheng Yu, et al. How much information can a vision token hold? a scaling law for recognition limits in VLMs. *arXiv preprint arXiv:2602.02539*, 2026.

A Related Work, in Full

The quantity this paper measures (the token price of establishing that a perceptual channel's output is true) sits in a gap between five literatures, each of which supplies one of its halves and none of which puts the halves in the same units. We organise the review around that gap rather than around topics.

The architecture this channel serves. The system under study equips a frozen multimodal backbone with two channels: a compressive $O(1)$-state summary maintained across a stream, and a scalable verbatim index over model-written captions, with a router deciding which to read under a per-decision token budget (Anonymous authors, 2026c). Its central negative finding is the reason the channel priced here is textual: the pixel half of the index recovers essentially nothing on its corpus while the caption half recovers almost everything. The companion analysis (Anonymous authors, 2026a) adds the arbiter to that formalism and shows that a channel emits a *claim* rather than a resource, so that capability becomes non-monotone in the channel set and a blind arbiter attains a reliability-weighted average instead of the best available channel. It derives the admission criterion this paper supplies a number for, and states plainly that the number is missing. The programme both belong to argues that capability is purchased by access structure under a fixed budget rather than by scale (Chen et al., 2026b), which is what makes a *price* the right object to measure rather than an accuracy.

What a description can carry. There is a mature account of how much of an observation a linguistic representation can hold. Captioning has been formalised as lossy fact-level extraction with a hypergeometric reward (Lu et al., 2026); information-theoretic treatments model a caption as a compressed selection from a latent set of describable facts (Chen et al., 2024a); whether a caption is as useful as the image it describes has been measured head to head (Yang et al., 2025d); and the information capacity of vision tokens themselves has been given a scaling law with a phase transition (Zhuang et al., 2026). Adjacent work asks the same capacity question of the visual substrate directly, from token pruning that discards most visual tokens with little loss (Chen et al., 2024b; Shang et al., 2024; Huang et al., 2025) to the finding that at a fixed inference budget the optimal configuration uses very few visual tokens and a much larger language model (Li et al., 2024).

Every one of these results assumes the content is correct and asks how much of it there is. That is the right question for a capacity theory and it is not the question here. The quantity we price is orthogonal to capacity and, as far as we can determine, unpriced anywhere: what it costs to find out whether the content is right, measured in the same units as the content. A channel with a proven capacity and an unpriced check is exactly the object the companion admission criterion (Anonymous authors, 2026a) cannot admit.

Budgeted multimodal inference and streaming memory. A large and fast-moving literature allocates a token budget across modalities, frames and time. Streaming compression schemes maintain a bounded bank under a per-step budget (Chen et al., 2025; Song et al., 2026; Sun et al., 2026); memory architectures index, consolidate and repair what they have stored (Packer et al., 2023; Gutiérrez et al., 2024; Chen et al., 2026a; Rasheed & Kabtoul, 2026; Yeo et al., 2025); long-video systems combine retrieval with an agentic read loop (Ren et al., 2026; Shen et al., 2025; Rege et al., 2026; Wang et al., 2024); and the design space has been surveyed and taxonomised (Omri et al., 2026; Jiang et al., 2026). A rate–distortion view of what a compacted memory should keep is the closest of these to a pricing argument (Colaco & Lahjouji, 2026).

These optimise the *use* of a channel and are, without exception, silent on the cost of establishing that what the channel returned is true. That silence is not an oversight of any one paper; it is the structural gap the companion analysis (Anonymous authors, 2026a) identifies, namely that the standard formalism has a slot for a channel's cost and its accuracy and no slot for its reliability as arbitrated against another channel. Our contribution to this literature is a number in its own units: on the interface these systems run on, a grounded check of one stored claim costs 0.797–0.878 of what producing the claim cost.

Verifiers, judges, and what a verdict is worth. Using a model to check another model's output is standard practice across self-correction, tool-use and agentic loops (Shinn et al., 2023; Yao et al., 2023; Schick et al., 2023), and it is the mechanism behind much of the recent test-time-compute

literature, where extra tokens are spent on generating and selecting among candidates (Snell et al., 2024; Muennighoff et al., 2025; Wei et al., 2022). The rubric-judge literature generally treats the checker as a noisy instrument whose variance can be reduced by averaging. An audit of one shipped scorer shows failure modes the variance frame cannot express, because they change what is measured rather than how noisily (Anonymous authors, 2026b); a companion result measures how errors enter and persist in a stored representation (Anonymous authors, 2026a). The interactive-imitation tradition supplies the classical version of the same accounting, where the price of an expert query is the object of study (Ross et al., 2011) and its language-model descendant asks when such a query is worth issuing (Li et al., 2026).

Our J belongs to that tradition and is deliberately cruder than a rubric: a verdict with $\mathrm{TPR} = \mathrm{FPR}$ is not noisy, it is empty, and no amount of repetition recovers information that was never in it. What the classical accounts do not have, because on a text interface it does not arise, is a floor under the query price set by the modality of the thing being checked. That floor is the specific contribution of Theorem 1, and it is why a multimodal agent cannot import the text-agent solution of checking cheaply and often.

Grounding, and channels that are not used. That multimodal systems under-use the channel they were given is now reported from several independent directions: text-only inputs outperforming multimodal ones in a controlled study of the modality gap (Xu et al., 2026); visual-dependency dissociation measured across twenty models (Lee, 2026); embodied and spatial benchmarks substantially solved by baselines that discard the visual modality (Anand et al., 2018; Udandarao et al., 2025); and spatial supersensing proposed as the capability the current generation lacks (Yang et al., 2025e;a). The embodied-QA benchmark literature that these critiques target is itself broad and still growing (Majumdar et al., 2024; Das et al., 2018; Yang et al., 2025c; Jiang et al., 2025; Wu et al., 2024; Zhao et al., 2025; Varghese et al., 2025; Wang et al., 2025a; Gemini Robotics Team, 2025), as is the long-horizon video setting it borrows from (Mangalam et al., 2023; Fu et al., 2025; Wu et al., 2025; Grauman et al., 2022; Yang et al., 2025b).

The measurement here offers a *resource* reading of that pattern, which is different in kind from the usual dataset-artefact reading. If checking a perceptual channel costs approximately what using it costs, then a system operating under a budget can neither trust the channel nor afford to verify it, and down-weighting it is the only remaining move, which is what down-weighting looks like from the outside. We offer this as a reading and not a demonstration. Establishing it would require the intervention §5 lists first, and we have not run it.

Why a price and not a benchmark. Finally, a word on method. The results below are not a leaderboard, and the corpus is deliberately schematic (Anonymous authors, 2026c) rather than natural (Ramakrishnan et al., 2021; Dai et al., 2017; Savva et al., 2019). The reason is that every number in this paper is either a ratio of two prices or a rate over probes with exact ground truth, and inserting a scored judgement into either would place an instrument with its own unmeasured reliability inside a measurement about reliability. The corpus record states which object appears at which frame in which colour, so the yield test is a string test against a record. This costs external validity and buys the one property the argument cannot do without, and §5 accounts for the trade rather than eliding it. The same reasoning governs why the backbones are open models served over a metered API (Mistral AI, 2025; Qwen Team, 2025; Gemma Team, 2025; Qwen Team, 2026; Gemma Team, 2026): the interface's own usage report is the measurement instrument, and a self-hosted deployment would require us to model the price rather than read it.

It is worth naming the genre, because it sets what the paper owes a reader. This is a measurement paper about an evaluation quantity, not a modelling paper: the object of study is the price of establishing that a perceptual claim is true, and the contribution is a definition of that quantity, an identification strategy for it, a bound it must satisfy, and a characterisation of the interface property the bound depends on. Nothing here proposes a better model or reports a better score, and a reader looking for either will not find one. What such a paper owes instead is that its instrument be auditable end to end, which is why c_{obs} is identified by a subtraction inside one run rather than fitted (§4.1), why the interface assumption is measured per endpoint rather than asserted of an arm (§4.2), why the one judgement call in the yield predicate is reported under three readings (§D.12), why two of our own registered predictions are reported as failing (§D.8, §D.10), and why an inequality we published in a

false form is republished with its counterexample (§B.6). Those are not housekeeping; on this kind of claim they are the evidence.

Relation to the companion submissions, stated rather than left to be discovered. This paper belongs to a programme with an overlapping author set, and reviewing policy treats any concurrent archival submission by such a set as prior work. We therefore state the dependency structure explicitly, including what each paper is *forbidden* to claim, so that the boundaries can be checked rather than inferred. The public preprint (Chen et al., 2026b) states the hypothesis that capability is bought by access structure under a fixed per-token budget. Four submissions follow, and no two share a population, an instrument and a question:

- a budgeted access-structure grid whose outcome is *accuracy*, testing whether reallocating a fixed budget beats growing it (Anonymous authors, 2026c);
- an analysis of arbitration between channels whose outcome is *which exit an access loop takes*, computed from round transcripts, which turns reliability-blindness from the assumption its own consequences rest on into a term a transcript settles, and settles it blind (Anonymous authors, 2026a);
- a model-free audit of what a benchmark's own annotations can see, which calls no model at all (Anonymous authors, 2026b);
- a measurement of what it costs to *check* a perceptual claim on a metered interface, whose population is caption probes rather than agent decisions (this paper).

This paper is the fourth. The direction of dependence is easy to read backwards, so we state it: the verification floor is a property of token-metered interfaces and holds with no agent, no arbiter and no budget allocation in it. Two of the nine formal results here mention the admission criterion of (Anonymous authors, 2026a): Corollary 1, which substitutes the floor into it, and Proposition 1, which is a statement about the criterion rather than about the floor. Deleting the criterion deletes those two statements and leaves Theorem 1, the other two escapes and both negative results standing. Conversely this paper reports no agent accuracy: it measures a price, not a policy.

Nothing here rests on an unpublished companion. The companions are under review and carry no archival identifier, so a claim of ours that depended on one would be a claim a reader cannot check, and we have kept the dependence to citation rather than substance. The admission criterion is not imported: Eq. (5) is written out here in this paper's own notation and derived from this paper's cost model, and Anonymous authors (2026a) is cited for having stated it first and for the definition of yield in accuracy units, which §2 repeats rather than assumes. The blind-arbiter accuracy $r^{\circ} = \sum_k \lambda_k r_k + \lambda_0 r_0$ used in Proposition 1 is likewise written out and is a law-of-total-expectation identity, not a borrowed result. Every price, probe and interface measurement reported below was run for this paper and its records are released with it. A reader who deletes every companion citation loses two attributions and no argument.

B Theory, in Full

This section builds the argument in six steps. §B.1 writes down the cost model and states, as assumptions that can be checked, the two properties of a metered multimodal interface the rest depends on. §B.2 derives the verification floor and the corollary it forces on the admission criterion. §B.3 asks what the floor survives: three natural ways to make checking cheaper are priced, and one of them turns out to leave the criterion exactly where it was. §4.3 prices the other half of a verifier: whether its verdict carries information. §B.5 shows that the cheap half of the bill is the half that buys yield, which is what makes the configuration stable rather than transitional. And §B.6 asks which part of a description a token cap destroys first, which is the question the experiments in §4.5 answer.

B.1 The interface as a cost model

Definition 4 (Call, and its price)**.** A *call* to a metered multimodal interface is a triple (o, x, y): an observation o (possibly empty), an input text x, and an output text y. The interface reports a realised price

$$c(o, x, y) \;=\; \alpha\,\tau(o) \;+\; \beta_{\text{in}}\,|x| \;+\; \beta_{\text{out}}\,|y|, \tag{3}$$

where $\tau(o)$ is the token count the interface assigns to the observation, $|x|, |y|$ are token counts of the texts, and $\alpha, \beta_{\text{in}}, \beta_{\text{out}} > 0$. We write $c_{\text{obs}} = \alpha\,\tau(o)$ for the observation term and $c_{\text{txt}} = \beta_{\text{in}}|x| + \beta_{\text{out}}|y|$ for the text term.

Equation (1) is not a modelling choice we are free to vary; it is how the interfaces these systems run on actually bill, and $\tau(o)$ is a published function of the image's resolution. What is a modelling choice is the next assumption, and it is the one that carries the argument.

Assumption 3 (No cross-call reuse of an observation)**.** If the output of a call is required to be a function of o, then o appears in that call and c_{obs} is charged in full. Equivalently, the interface offers no mechanism by which a second call may condition on o at a price below $\alpha\,\tau(o)$.

Assumption 4 (Observation dominance)**.** In the regime of interest, $c_{\text{obs}} \gg c_{\text{txt}}$; that is, the ratio $\theta = c_{\text{obs}}/(c_{\text{obs}} + c_{\text{txt}})$ is close to 1.

Both assumptions are empirical claims about a specific interface at a specific time, not axioms, and both are checked rather than asserted. §4.1 identifies c_{obs} and c_{txt} on the measurement's own calls and reports θ per backbone, which is Assumption 2 measured. Assumption 1 is the fragile one (it is a statement about a product feature rather than about computation, and it is falsifiable by a single API change rather than by an argument), so §4.2 checks it directly, on every provider endpoint the router will pin for these backbones, by reading the cached-token count and the bill of each call rather than by reasoning about what the interface ought to do. It holds, for the call a verifier makes, on 5 of 7 backbones and fails on 2, by the factor Proposition 1 predicts and for a lifetime of minutes. We flag it here because a reader who disbelieves it should disbelieve everything that follows; the point of §4.2 is that the reader does not have to take our word for which of the two cases they are in.

Definition 5 (Observation channel)**.** Fix an observation o and a family $\mathcal{Q}$ of queries about it. An *observation channel* is a procedure that, given a token allowance b, emits a description $\delta(o, b)$ in one call and thereafter answers queries from $\delta(o, b)$ alone. Its *yield* $y(b) = \Pr_{q\sim\mathcal{Q}}[\delta(o, b) \text{ determines } q]$ is the probability that the description settles a query drawn from the family, and its *price* $c(b)$ is the realised cost of the emitting call.

Definition 6 (Verifier, grounded and ungrounded)**.** A *verifier* takes a claim ϕ about o and returns a verdict in $\{\text{ACCEPT}, \text{REJECT}\}$. Its *price* v is the realised cost of one verdict. It is *grounded* if its verdict is a function of o, and *ungrounded* if it is a function of ϕ alone.

The distinction in Definition 2 carries as much of the argument as Assumption 1. An ungrounded verifier is cheap (it never pays c_{obs}), and it is answering a different question. It can test a claim for internal consistency, against a prior, or against other claims, and none of those establishes that the claim is true of the observation. The admission criterion of the companion analysis (Anonymous authors, 2026a) requires the latter, because what it buys with a verdict is the right to weight one channel over another on a particular item, and an ungrounded verdict is correlated with the claim's plausibility rather than with its truth. We therefore price grounded verification throughout and record ungrounded verification as an escape whose value is not established (§5).

B.2 Two prices in the same units

Theorem 2 (Verification floor)**.** *Under Definition 1 and Assumption 1, any grounded verifier of a claim about o satisfies*

$$\frac{v}{c(b)} \;\geq\; \frac{c_{\text{obs}}}{c_{\text{obs}} + c_{\text{txt}}(b)} \;=\; \theta(b), \tag{4}$$

and $\theta(b) \to 1$ as $c_{\text{obs}}/c_{\text{txt}}(b) \to \infty$.

The proof is one line and is given in full in Appendix F: the verifier's verdict is a function of o, so by Assumption 1 its call contains o and pays c_{obs}; hence $v \geq c_{\text{obs}}$, while $c(b) = c_{\text{obs}} + c_{\text{txt}}(b)$ exactly. Stating a one-line inequality as a theorem is a claim about which arithmetic matters, and the claim is this: the verification price of a perceptual channel is not a design parameter to be optimised. It is pinned from below by the same fixed cost that pins the channel, the two converge as the observation dominates the bill, and on a token-metered multimodal interface the observation does dominate. A cheaper prompt, a smaller checker, a terser verdict: each of these reduces c_{txt}, and every one of them moves θ *up*.

Figure 2 draws the two calls side by side; the shared grey block is the whole of the argument. That last observation deserves emphasis because it inverts the usual engineering intuition. On a text interface, making the checker terser makes checking cheaper relative to the thing checked. Here it makes it *more* expensive relative to the thing checked, because both prices share a floor and only the numerator's optional part is being cut. The natural optimisation runs the wrong way.

Corollary 2 (The admission criterion is violated generically)**.** *Write $\mu = \partial\,\mathrm{acc}/\partial b \geq 0$ for the marginal accuracy a token buys in the budget's best alternative use, and let y_k be channel k's yield* in accuracy units*: the accuracy the decision procedure gains by having k's claim available, which is how the companion analysis (Anonymous authors, 2026a) defines it, and without which the two sides below are not commensurable. A channel that is both used and believed spends $c(b) + v_k$, so it is worth admitting with verification only if*

$$y_k \;>\; \mu\,\big(c(b) + v_k\big) \;\geq\; \mu\,c(b)\,\big(1 + \theta(b)\big), \qquad \theta(b) \to 1, \tag{5}$$

where the second step is Eq. (2)*. The channel must therefore return nearly* twice *what the same tokens buy anywhere else: once to pay for itself and once to pay for being believed. Since $y_k \leq 1$, the requirement is unsatisfiable as soon as $\mu > 1/\big((1+\theta(b))\,c(b)\big)$, which tends to $1/(2c(b))$ as $\theta(b) \to 1$, a condition on the budget and on the marginal value of tokens elsewhere, not on the channel's quality.*

Corollary 1 is the reason this paper exists. It converts an inequality that reads like a threshold to be engineered past into a statement about when the threshold is reachable at all, and the answer does not mention the channel. A perfect channel ($y_k = 1$, every description correct and complete) fails the criterion whenever tokens are worth enough elsewhere. Improving the channel cannot fix it; only lowering the marginal value of budget elsewhere can, which means having budget to spare, which is the case the whole architecture exists to avoid.

B.3 What the floor survives

A floor invites attempts to get under it. Three can be priced from the cost model alone, and pricing them is more informative than the theorem, because two of them work and the third (the one practitioners actually reach for) provably does not. Two more, which the cost model permits but cannot settle because they turn on what a particular interface does rather than on what the cost model says, are added at the end of this subsection. Table 6 collects those five with what each was worth when measured, and lists the ungrounded check beside them, which is not an escape but the thing an escape must not turn into.

Proposition 4 (Caching moves the floor by exactly the cache discount)**.** *Suppose the interface admits a second call conditioning on o at price $\alpha'\tau(o)$ with $\alpha' < \alpha$, violating Assumption 1. Then the floor of Eq.* (2) *becomes $\alpha'\tau(o)/(\alpha\tau(o) + c_{\mathrm{txt}}(b))$, which is below $\theta(b)$ by the factor α'/α and can be made arbitrarily small.*

Proposition 1 is the only one of the five that attacks the constant Theorem 1 pins, and it is therefore the one the paper has to establish or rule out rather than argue about. §4.2 measures it: on which of these endpoints an observation is in fact re-priced below $\alpha\tau(o)$, by how much against the α'/α they publish, and for how long: the quantity the proposition does not contain, and the one that turns out to decide the matter.

Proposition 5 (Batching divides the floor by the number of claims per observation)**.** *Let m claims about the same o be verified in one call, at price $v_m = c_{\mathrm{obs}} + m\,c_{\mathrm{txt},v}$. The amortised price per verdict is $\bar{v} = c_{\mathrm{obs}}/m + c_{\mathrm{txt},v}$, so the floor falls as $\Theta(1/m)$ while the observation term dominates (that is, until $m \sim c_{\mathrm{obs}}/c_{\mathrm{txt},v}$), and flattens at $c_{\mathrm{txt},v}$ thereafter.*

Proposition 1 is a real escape and it has a hard ceiling that is a property of the task rather than of the budget: m is the number of distinct claims an agent needs to check *about one observation*. In a stream of n frames with μ claims per frame, the total verification bill is $n(c_{\mathrm{obs}} + \mu c_{\mathrm{txt},v})$ however it is batched, because claims about different frames cannot share the observation term. Batching converts a per-claim floor into a per-*frame* floor and stops there. For the retrieval channel studied here, where a query typically concerns a single frame, μ is small and the escape is correspondingly small. §D.8 measures where that ceiling falls relative to the knee of the $1/m$ curve, and it falls on the wrong side of it: the per-claim price is still descending when the frame runs out of claims, so the escape is truncated before it saturates rather than saturating and then being truncated.

Table 6: The escapes from the verification floor, priced. Only the first and the last change the constant that Theorem 1 pins, and both do it by changing what is transmitted rather than what is asked; the second is bounded by a property of the task; the third does not move the admission criterion at all; and the fifth, the only one that costs a caller nothing to try, moves nothing either. The fourth row is not an escape but the thing an escape must not become: it is cheap because it answers a different question. The final column records what the measurements say. The batching row's $\Theta(\theta(b)/m)$ holds only while $m \lesssim c_{\mathrm{obs}}/c_{\mathrm{txt},v}$, after which the amortised price flattens (Prop. 1); §D.8 finds that a frame runs out of claims first.

Escape	New floor	Bounded by	What we measured
Observation caching (Prop. 1)	$(\alpha'/\alpha)\,\theta(b)$	the discount *and its lifetime*	§4.2: reaches a verifier on 2/7 backbones; expires in minutes
Batch m claims (Prop. 1)	$\Theta(\theta(b)/m)$, then flat	claims per obs. μ	§D.8: μ binds
Verify a p-fraction (Prop. 1)	$\theta(b)$, unchanged	— (invariant)	no run needed
Ungrounded check (Def. 2)	0	a different question	§D.14: $J = 0$
Message layout (observation first)	$\theta(b)$, unchanged	— (no effect)	§4.2: 0/18 cells move
Re-send the frame downsampled	$(\tau(o')/\tau(o))\,\theta(b)$	resolution the claim needs	not run

The fourth escape is not a feature request and is the most practical of them, so it belongs here even though we did not run it. Nothing in Theorem 1 says the verifier must be shown the *same* observation the channel was shown. $\tau(o)$ is a published function of the image's resolution (§B.1), so a grounded verifier may re-send the frame downsampled: its verdict is still a function of the observation, Assumption 1 is untouched, and c_{obs} falls by the provider's own resolution schedule: roughly with pixel count for the Qwen-VL endpoints, and not at all for the Gemma ones, which price an image at a fixed token count whatever its size. What bounds it is not the interface but the claim: a verdict about an object's colour survives heavy downsampling, a verdict about a small piece of text on that object does not, and the escape is worth exactly the resolution the claim can spare. We flag it as the escape a practitioner should try first and as the one this paper does not price; §5 lists it as the one escape in Table 6 that is still unrun, caching having since been measured (§4.2).

The fifth is cheaper still, needs neither a feature nor a re-encode, and is the reason the caching row cannot be read off a price list. If an interface caches a *prefix* rather than an image, then whether two calls about one frame share anything depends on where the caller puts the frame in the message. The shipped order is $[\text{text}, \text{observation}]$, which puts the differing question in front of the observation and shares nothing; $[\text{observation}, \text{text}]$ would make the observation the shared prefix of a caption call and every later verdict call about it. Nothing in Definition 1 forbids either order and neither costs anything to choose. Whether the swap buys a cache hit is a fact about the provider's cache key, not about the cost model, so it is measured rather than derived (§4.2), and it is the one escape in Table 6 that turns out to be worth exactly nothing.

The third escape in Table 6 is the one that matters, because it is what a budget-conscious engineer does by reflex: check only some of the time.

Proposition 6 (Bernoulli sampling leaves the criterion invariant)**.** *Let the arbiter verify each claim independently with probability p and otherwise arbitrate blind. Write $r^\star = \max_k r_k$ for the accuracy of an informed arbiter and $r^\circ = \sum_k \lambda_k r_k + \lambda_0 r_0$ for that of the blind one (Anonymous authors, 2026a). Then the expected accuracy is $p\,r^\star + (1-p)\,r^\circ$ and the expected verification bill is $p\,v$ per claim, so both the gain and the cost of verification scale by the same factor p, and the admission criterion* (5) *is independent of p.*

Figure 4 draws the first three against the quantity they are supposed to reduce. Proposition 1 is the sharpest statement in this section. Sampling is not a partial escape that trades a little accuracy for a lot of budget; it is a *no-op* with respect to the decision of whether verification is worth buying at all. If a full verification regime fails the criterion, so does every sampled one, and if it passes, so does every sampled one: the choice of p then affects only the variance. The intuition that one can "check a few and see" does not survive the arithmetic, because the thing being sampled is the very quantity whose price/benefit ratio is in question.

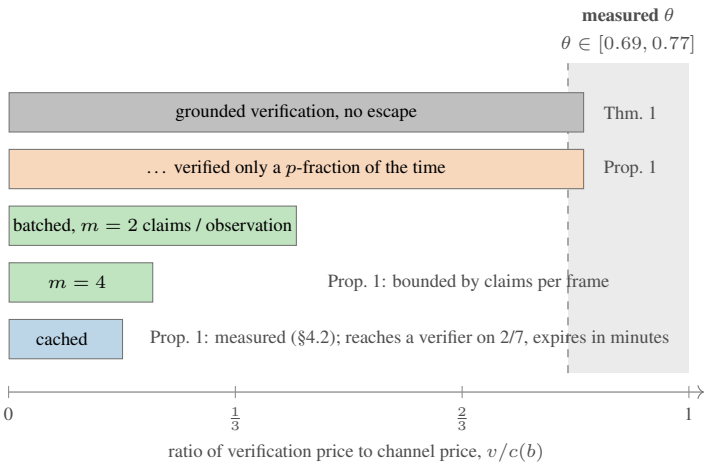


Figure 4: The three priced escapes, drawn against the quantity they are supposed to reduce. Bar length is the resulting floor on $v/c(b)$; the shaded band at the right is where the floor actually sits on the interface measured here. Sampling (orange) is drawn at full length deliberately: it lowers the expected bill and lowers the expected benefit by the same factor, so the bar that matters for the admission decision does not move at all (Prop. 1). Batching shortens the bar as $1/m$, but m counts claims about *one* observation and is a property of the task. Only caching attacks the constant, and it is a property of the interface rather than of the agent, which is why §4.2 audits it on every endpoint the router will pin rather than assuming it away. The cached bar is drawn at the discount the endpoints that offer it publish; what the figure cannot draw is that the discount lasts minutes, so the bar applies to a check issued immediately and not to the deferred one this architecture makes.

B.4 What a verdict is worth

A price is only half of a verifier. The other half is whether its verdict distinguishes true claims from false ones, and a verifier that fails at this does not fail gracefully.

Definition 7 (Verdict information, effective price)**.** Fix a verification protocol π: the question put to the verifier, the claim population it is put over, and the decision rule applied to its reply. For a binary verifier with true-accept rate TPR and false-accept rate FPR *under* π, the *verdict information* is $J = \mathrm{TPR} - \mathrm{FPR}$ (Youden's J), and the *protocol-conditioned effective verification price* is $v_{\mathrm{eff}} = v/J$ for $J > 0$ and $+\infty$ otherwise.

The conditioning on π is not a formality and we write it into the name because our own measurement forces us to: §D.13 takes four verifier questions that differ only in wording, holds the model and the claims fixed, and moves J by more than the gap between the best and worst backbone; one of them, logically equivalent to the shipped question, sends J negative on 4 of 7 backbones. v is a property of the interface and J is not a property of the verifier alone. Every v_{eff} in this paper is therefore a price *under the terse protocol*, not a constant of the model, and the quantity that travels between protocols is the price, not the value.

It is also worth being explicit about what J leaves out, because a decision-theoretic reader will ask. $J = \mathrm{TPR} - \mathrm{FPR}$ is the balanced-prevalence, symmetric-loss summary: it charges a false accept and a false reject alike and it assumes the claim population is the one measured. A deployment with skewed prevalence or asymmetric downstream loss should replace it with the expected loss its own decision incurs. We use J because this paper prices a *floor* and J is the summary that makes the floor's numerator honest without importing a prior we would then have to defend; Proposition 8 reports the information-theoretic figure alongside it for the same reason, and that figure is the one to prefer where the two disagree.

Proposition 7 (The effective price diverges while the nominal price does not)**.** *A verifier answering* ACCEPT *with a fixed probability independent of the claim has $J = 0$ and finite v. Its verdicts are independent of the truth of the claim, so no number of them changes the posterior, and the token cost of obtaining one bit about the claim is unbounded. Hence v finite and v_{eff} infinite are simultaneously possible, and a verifier must be priced by v_{eff}.*

The normalisation by J rather than by accuracy is deliberate. Accuracy on a balanced set of true and false claims sits at one half for a verifier that is right on every true claim and wrong on every false one, and such a verifier is not half-useful: it is a machine that says yes. J is zero exactly there. This is the same distinction that separates a noisy instrument from an empty one (Anonymous authors, 2026b): averaging fixes the first and cannot touch the second.

v_{eff} is a surrogate, and it is worth being explicit that it is an *optimistic* one.

Proposition 8 (Tokens per bit, and how far v/J understates it)**.** *Model the verifier as a binary channel between the truth T of a claim and its verdict A, and take the claim's prior to be even. Writing H_b for the binary entropy, the mutual information a verdict carries is*

$$I(T;A) \;=\; H_b\big(\tfrac{\mathrm{TPR}+\mathrm{FPR}}{2}\big) \;-\; \tfrac{1}{2}\big[H_b(\mathrm{TPR}) + H_b(\mathrm{FPR})\big], \tag{6}$$

which depends on the pair and not on J alone. In the symmetric *special case* $\mathrm{TPR} = 1 - \mathrm{FPR}$ *it collapses to $I(J) = 1 - H_b\big(\frac{1+J}{2}\big)$, and expanding that about $J = 0$,*

$$I(J) \;=\; \frac{J^2}{2\ln 2} \;+\; O(J^4), \qquad \textit{so} \qquad \frac{v}{I(J)} \;=\; 2\ln 2\,\frac{v}{J^2}\,\big(1 + O(J^2)\big). \tag{7}$$

The token price of one bit about a claim therefore diverges as $1/J^2$ near $J = 0$, while $v_{\text{eff}} = v/J$ diverges only as $1/J$.

We report v_{eff} as the headline because it requires no assumption about the prior, and we report $v/I(T;A)$ alongside it as the honest figure. Which of the two expressions above is used matters here, because the verifiers measured in §D.4 are conspicuously *not* symmetric (their asymmetry is one of this paper's findings), and Eq. (7) applied to their J understates Eq. (6) by up to $1.55\times$, in the direction that flatters the figure we call honest. Every per-bit number below therefore comes from Eq. (6) and the measured $(\mathrm{TPR}, \mathrm{FPR})$ pair. So computed, the surrogate understates the per-bit price by up to $3.39\times$ (§D.4), and the worst case is not the worst-priced verifier but the least informative one, Gemma 3 12B, whose verdicts carry 0.115 bits each.

B.5 Yield is cheap, which makes the ratio worse

Proposition 9 (Marginal yield per token)**.** *Under Definition 1, raising the text allowance from b to b' changes the price by $c_{\mathrm{txt}}(b') - c_{\mathrm{txt}}(b)$ and leaves c_{obs} unchanged. Hence the marginal token price of additional yield is bounded by the text term alone, while by Theorem 1 the verification price is bounded below by the observation term. The two quantities are governed by different halves of the same bill.*

Proposition 9 is what makes the configuration stable rather than transitional. If yield were expensive and checking cheap, an agent would buy less description and check what it had. Because yield is cheap and checking is expensive, the rational move is the opposite: buy a generous description and do not check it. That is the configuration in which the selection loss of the companion analysis (Anonymous authors, 2026a) is incurred, and an agent reaches it by following the prices rather than by reasoning badly. It also predicts the empirical pattern of §4.4: as backbones improve, y saturates while J still has room, so the numerator of the criterion stops growing before the price of believing it does.

B.6 What a token cap destroys first

The three propositions above treat yield as a scalar $y(b)$ and say nothing about *which* part of a description a cap removes. The experiments in §4.5 read the descriptions themselves, so it is worth deriving in advance what a cap can and cannot preserve.

Definition 8 (Scene, description, binding)**.** A *scene* is a set of n objects, each carrying one attribute value from an alphabet $\mathcal{A}$ with $|\mathcal{A}| = K$. A *description* of realised length w *name tokens* (maximal runs of alphabetic characters, which is the unit the counting argument below charges in and the unit every evaluation of it in this paper uses), *names* a set of objects (of size n_o) and a set of attribute values (of size n_a), and *binds* a pair $(\text{object}, \text{value})$ when both are named and asserted of each other. Write n_b for the number of bound pairs. A query names an object and asks for its value, so it is settled by the description only if that object's pair is bound.

Proposition 10 (The pairing pays for an attribute vocabulary the marginals do not)**.** *Naming an object costs at least one name token, naming an attribute value costs at least one name token, and the two sets are disjoint, so $w \geq n_o + n_a$. Each object carries a single attribute value, so distinct bound pairs have distinct objects and $n_b \leq n_o$. Consequently, for a description of realised length w over a scene of n objects,*

$$\Pr[\textit{a uniformly drawn object's pair is bound}] \;=\; \frac{n_b}{n} \;\leq\; \frac{w - n_a}{n}. \tag{8}$$

By contrast the object marginal admits $n_o \leq w$ with no attribute term at all, and the attribute marginal saturates at $n_a \leq \min(K, w)$, independent of n. The joint is therefore the only one of the three that pays for the attribute vocabulary as well as for the objects, and it is the first to be squeezed as w falls, by an additive n_a.

The unit is worth one more sentence, because three different lengths could be meant and the paper reports two of them. A BPE token count is at least the name-token count, since a tokeniser can split a name but cannot merge two of them into one token; so a bound proved and evaluated in name tokens is at least as tight as the same bound in BPE tokens, and stating it in name tokens is the conservative choice rather than a convenient one. The name-token count also differs from a whitespace word count, upward, wherever a backbone writes `hat(yellow),`: one word, two names. That is the whole of the gap between the 12.8 words and the 13.1 name tokens the tightest cap emits (§4.5); Table 5 reports words because the prompt's allowance is stated in words, and Eq. (8) is evaluated in name tokens because that is what it charges.

Definition 8 fixes the vocabulary Eq. (8) counts in. Equation (8) is an upper bound and it is not tight; a describer can waste tokens in many ways that the counting argument does not see. Its value is that it becomes *vacuous* at a computable point, namely $w \geq n + n_a$, and this gives the experiments something to check rather than merely to report: yield should be constrained below that length and free above it. We evaluate Eq. (8) at the realised caption lengths in §4.5: realised rather than instructed, because the backbones systematically overrun the word allowance they are given, which is itself measured.

Eq. (8) **is a repair, and the inequality it replaces was false.** The bound first read $n_b/n \leq w/(2n)$, from the chain $w \geq n_o + n_a \geq 2n_b$. The second inequality there needs $n_b \leq n_a$, every bound pair holding an attribute value of its own. The projection of the bound set onto the attribute coordinate is indeed contained in the named values, but *that projection is not injective*: one value serves arbitrarily many objects the moment a describer writes `blue sock/pen/ruler/hat/belt/tray`. Testing the forced-format grid of §D.10, one backbone binds all 16 objects of a frame in 23 tokens where $w/(2n)$ demands 32, using 7 attribute values for 16 pairs; across 2,685 descriptions $w/(2n)$ is violated 5 times. Eq. (8) as now stated is violated 0 times and is *attained with equality* in 52 of them, so it is both valid and reached. The repair is not free: the factor of two is gone, the joint is squeezed at the same rate as the object marginal rather than at twice it, and the vacuity point moves in from $2n$ to $n + n_a$, which on this corpus is 19.9 to 22.5 tokens, so the bound constrains at 1 of the three caption caps rather than at two. What survives is the asymmetry: the pairing is the only one of the three quantities charged for the vocabulary it uses, and it remains the first to go.

Proposition 3 says how many pairings survive. It does not say *which*, and two accounts of the same phenomenon differ exactly there.

The typicality account. A capped describer enumerates in order of salience or frequency, so the objects that fall off the end are the atypical ones. Because our probes are by construction the rarest objects in the corpus (classes that appear in exactly one frame of an episode), this account predicts that the probe object is named *less* often than the others present in the same frame, and that the gap widens as the cap tightens.

The binding account. A cap acts on the description's structure rather than on its selection: objects are named roughly uniformly, and what is lost is the assertion connecting an object to its value, in the manner of Proposition 3. This account predicts *no* gap in naming rate between the probe object and the others, and a large population of descriptions that name both the object and the correct value without binding them.

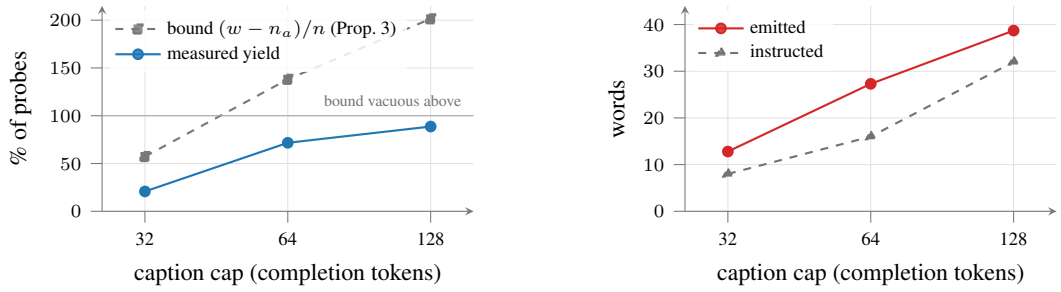


Figure 5: Left: the counting bound of Eq. (8), evaluated at the realised caption length, against the yield actually measured. The bound holds at every cap and stops constraining anything once $w \geq n + n_a$, which happens between the bottom and the middle cap: it therefore binds at one cap of three, and yield is still climbing above it. The inequality it replaces, $w/(2n)$, went vacuous a cap later and was false; see under Eq. (8). It is a bound and not a fit, and it is loose at the tight cap by construction. **Right:** why the bound must be evaluated at realised rather than instructed length. Every backbone overruns the word allowance the prompt names, running to 160.1% *of* it at the tight cap and 170.9% at the middle one, so the cap is enforced by the completion-token limit rather than by the instruction.

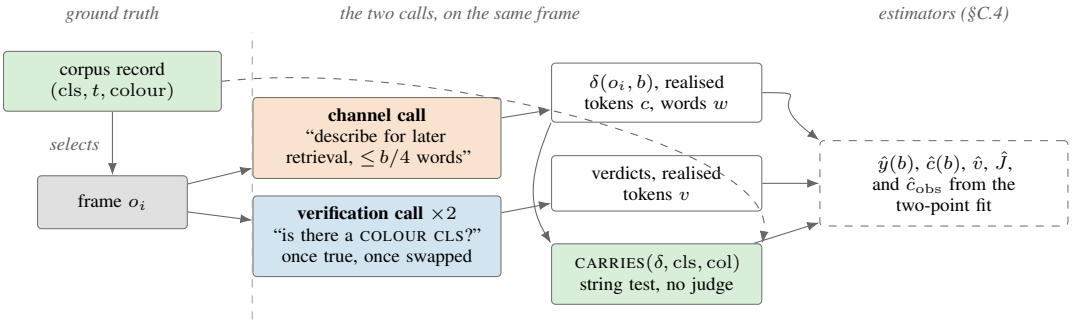


Figure 6: The protocol. One frame is selected by the corpus record, and two kinds of call are made about it: a capped description, and a pair of yes/no verifications on the same claim with the colour true and swapped. Both calls carry the frame, so both pay c_{obs}: that shared payment is what Theorem 1 is about, and it is visible as the two arrows leaving the same frame. Ground truth flows only into the string test on the right; nothing downstream of a model output is scored by another model.

The two make opposite predictions about a quantity that is directly measurable once the frame's full object inventory is known, and §4.5 measures it. We state both here because we found the first more plausible before looking, and the measurement is against it.

C Method, in Full

Everything in §3 is designed around one constraint: no number in this paper may depend on a scored judgement. Each measurement is either a ratio of two realised token counts read from the interface's own usage report, or a rate over probes whose ground truth is fixed by the corpus record. The reason is not fastidiousness. The paper's subject is what it costs to establish that something is true; putting a model-based judge inside that measurement would place an instrument with its own unmeasured reliability inside a measurement about reliability, and there would be no way to tell the instrument's failures from the result.

C.1 Probes with exact ground truth

Figure 6 sets out the whole protocol on one page and Figure 17 shows how a single probe is built from the corpus record. The corpus is the schematic egocentric render of the parent system (Anonymous authors, 2026c), used here for one property a natural corpus cannot supply. Its generator records, for every episode, which object class appears at which frame in which colour, and which classes appear in exactly one frame of the episode. Those *singletons* are the probes.

Definition 9 (Probe)**.** A *probe* is a triple $\pi = (o, \mathrm{cls}, \mathrm{col})$ where o is a rendered frame, cls is an object class appearing in exactly one frame of its episode (namely o), and col is the colour that class is rendered in at o, both read from the generator's record. The query family $\mathcal{Q}(\pi)$ contains the single question "what colour was the cls?", whose answer is col by construction.

Uniqueness is what makes the answer exact rather than adjudicated: because the class occurs nowhere else in the episode, there is no second sighting whose colour could also be defensible. The probe set is 16 probes per backbone, drawn from the corpus's episodes.

Two further facts about the corpus matter for §4.5 and are recovered rather than assumed. Each frame renders exactly sixteen objects, of which the probe class is one and the rest are drawn from a disjoint pool of common filler classes, re-randomised per frame. The per-frame inventory is not stored in the shipped corpus, so we recover it by replaying the generator's seeded random stream with rendering disabled and *verifying* the reconstruction against the shipped record field by field; a single mismatch aborts rather than emitting a plausible reconstruction. This gives exact knowledge of what was in front of the model, which is what turns "the caption did not name the probe" into "the caption named eleven of the sixteen objects present and the probe was not among them".

C.2 YIELD, AND THE ONE JUDGEMENT CALL IN ITS DEFINITION

An observation channel in the sense of Definition 5 is instantiated as follows. For each probe (Definition 9) the backbone is asked for a retrieval-oriented description of o under a completion-token cap $b \in \{32, 64, 128\}$, phrased with an explicit word allowance of $b/4$ words. The description is then tested for whether it carries the probe's pairing.

Definition 10 (Yield predicate)**.** Let δ be a description, and let $\mathrm{low}(\delta)$ be δ lower-cased with every character outside $[\mathtt{a{-}z}\ \sqcup]$ replaced by a space. The shipped predicate is

$$\textsc{carries}(\delta, \mathrm{cls}, \mathrm{col}) \;=\; \big[\mathrm{cls} \in \mathrm{low}(\delta)\big] \wedge \bigvee_{u \,\in\, \mathrm{split}(\mathrm{low}(\delta),\ \sqcup\text{“and”}\sqcup)} \big[\mathrm{cls} \in u \,\wedge\, \mathrm{col} \in u\big]. \qquad (9)$$

Yield at cap b is $\hat{y}(b) = |\{\pi : \textsc{carries}(\delta(o_\pi, b), \cdot)\}|/|\Pi|$.

Equation (9) is written out in this detail because the intent behind it and its behaviour differ, and the difference is ours to disclose. The intent was "the class and its colour must fall within the same clause, so that a caption listing every object and then every colour does not count as carrying the pairing". The implementation normalises punctuation to spaces *before* splitting, so by the time the split runs, the only separator that can still match is the literal word *and*: commas, semicolons and full stops have already been erased. The predicate that ran is therefore looser than a clause and tighter than co-occurrence, and it is the predicate every headline number in this paper comes from.

We deal with this by measuring the sensitivity rather than by quietly repairing it. All caption text is stored, so the rate can be recomputed under any predicate at no cost and with no additional calls.

Definition 11 (Three readings)**.** STRICT splits δ into clauses on sentence and list punctuation and on the connectives *and*, *with*, *in*, *before* normalisation, and requires the class and colour to share a clause. SHIPPED is Eq. (9). LOOSE requires only that both strings occur somewhere in δ. The three are nested, STRICT $\Rightarrow$ SHIPPED $\Rightarrow$ LOOSE, so the induced rates are monotone for every probe and every backbone.

One consequence of splitting STRICT on *in* should be stated rather than discovered: *a sieve in green* is a bound pair, and STRICT cuts it in two and scores it unbound. The reading is therefore biased downward on exactly the constructions that express binding with that preposition, so where STRICT is used below it understates the pairing rather than overstating it, and the bracket it forms with LOOSE is wider than the definitional uncertainty alone would make it.

Nesting is what makes the comparison informative: the three rates bracket the quantity, and the width of the bracket is the width of the definition rather than of the measurement. §D.12 reports all three, and §5 states which conclusions survive the widest reading.

C.3 VERIFICATION, IN PAIRS

On the same frame the backbone is asked a single yes/no question with the frame attached: "look at the frame; is there a ⟨colour⟩ ⟨class⟩?", capped at eight completion tokens. The question is asked twice per probe.

Definition 12 (Paired verification)**.** For probe $\pi = (o, \mathrm{cls}, \mathrm{col})$, the *true* query uses col and the *false* query uses a colour $\mathrm{col}' \neq \mathrm{col}$ drawn from the corpus alphabet and not held by that object at o. Writing $a_\pi, a'_\pi \in \{0, 1\}$ for the two ACCEPT indicators,

$$\widehat{\mathrm{TPR}} = \frac{1}{|\Pi|} \sum_{\pi} a_\pi, \qquad \widehat{\mathrm{FPR}} = \frac{1}{|\Pi|} \sum_{\pi} a'_\pi, \qquad \hat{J} = \widehat{\mathrm{TPR}} - \widehat{\mathrm{FPR}}. \tag{10}$$

Asking only the true claim would measure a verifier's willingness to agree rather than its ability to discriminate, and those are exactly the two things J separates. Pairing the queries on the same probe also controls for probe difficulty: both members of a pair concern the same object in the same frame under the same prompt, so a difference between them cannot be attributed to the item.

The verifier is grounded in the sense of Definition 2: the frame is attached, so the call pays c_{obs}. That is the configuration Theorem 1 is about, and it is the only configuration that answers the question the admission criterion asks. An ungrounded check is cheap and cannot establish truth, and we do not report it as though it could.

Every call runs in a *fresh* session with its own budget. This experiment prices single calls; under a shared budget the last cap in a sweep would be measured against a partly consumed allowance and would be cheaper than the first for reasons having nothing to do with the caption. The realised cost is read from the provider's usage report, not estimated from a tokeniser, because the estimate and the bill disagree by a margin comparable to the effects being measured.

C.4 ESTIMATORS, AND WHAT IDENTIFIES THEM

Definition 13 (The five reported quantities)**.** For each backbone: $\hat{y}(b)$ per cap (Definition 10); $\hat{c}(b)$, the mean realised cost of a caption call at cap b; $\hat{v}$, the mean realised cost of a verification call over both members of every pair; $\hat{J}$ from Eq. (10); and the effective price $\hat{v}/\hat{J}$ together with the per-bit price $\hat{v}/I(\hat{J})$ of Proposition 8.

Definition 13 is the whole of what a backbone's row contains, and Definition 12 fixes how its J is formed. The remaining quantity is the one Theorem 1's premise needs, and it is not directly observable: the interface reports a total, not a split into c_{obs} and c_{txt}. It is identified from the sweep.

Definition 14 (Two-point identification of the observation term)**.** Assume the realised cost of a caption call is affine in the cap, $c(b) = c_{\mathrm{obs}} + ab$, with $a > 0$ and c_{obs} independent of b: the content of Definition 1, since the image is unchanged across the sweep. Two caps $b_1 \neq b_2$ then identify both terms,

$$\hat{a} = \frac{\hat{c}(b_2) - \hat{c}(b_1)}{b_2 - b_1}, \qquad \hat{c}_{\mathrm{obs}} = \hat{c}(b_1) - \hat{a}\, b_1, \qquad \hat{\theta}(b) = \frac{\hat{c}_{\mathrm{obs}}}{\hat{c}(b)}. \tag{11}$$

Identification requires $b_1 \neq b_2$ and $a \neq 0$, both of which hold; we use the extreme caps of the sweep to maximise the lever arm. The estimator has two biases and both run the same way.

The first is structural and we named it wrongly at first. What Eq. (11) identifies is the part of the bill that does not move with the cap, and by Definition 1 that is $\alpha\tau(o) + \beta_{\mathrm{in}}|x_{\mathrm{fix}}|$: the caption instruction is the same string at every cap, so its input tokens sit in the intercept alongside the observation. The intercept is therefore an upper bound on c_{obs} and $\hat{\theta}$ an upper bound on θ, not a floor. We write $\hat{\imath}$ for it from here on and reserve c_{obs} for the observation term itself.

The second is the one we anticipated. Backbones do not consume the cap in full (they emit fewer completion tokens than allowed at the largest cap), so the fitted slope $\hat{a}$ understates the true per-token cost, which inflates $\hat{\imath}$ further.

Both matter for reading §4.1, and together they explain a sign the first version of this paper attributed to the second alone: measured ratios sitting marginally below $\hat{\imath}/c$. A within-run identification without

Algorithm 1 The measurement, for one backbone. Every call is priced by the interface's own usage report; nothing is scored by a model.

Require: probes Π from the corpus record, caps $B = \{32, 64, 128\}$, colour alphabet $\mathcal{A}$

$R \leftarrow \emptyset$
for all $\pi = (o, \mathrm{cls}, \mathrm{col}) \in \Pi$ **do**
  for all $b \in B$ **do**
    $S \leftarrow$ FRESHSESSION() ▷ single-call pricing: a shared budget would make late caps cheap
    $\delta \leftarrow S$.CHAT("describe for retrieval, $\leq b/4$ words", o, cap=b)
    record $\big(\delta,\ \textsc{tokens}(S),\ |\delta|_{\text{words}},\ \textsc{carries}(\delta, \mathrm{cls}, \mathrm{col})\big)$
  end for
  $\mathrm{col}' \leftarrow$ any element of $\mathcal{A} \setminus \{\mathrm{col}\}$ ▷ the false claim differs from the true one in one attribute
  for all $k \in \{\mathrm{col}, \mathrm{col}'\}$ **do**
    $S \leftarrow$ FRESHSESSION()
    $a \leftarrow \big[$"yes" $\in S$.CHAT("is there a k cls?", o, cap=8)$\big]$
    record $(k, a, \textsc{tokens}(S))$
  end for
  $R \leftarrow R \cup \{\pi$'s records$\}$
end for
return $\hat{y}(b), \hat{c}(b)$ per cap; $\hat{v}, \hat{J}$ from the pairs; $\hat{c}_{\mathrm{obs}}, \hat{\theta}$ by Eq. (11)

the first bias is available at no extra cost, because the same run asks the same verdict question with the frame removed (§D.14); the difference between the two verdict prices differences the verifier's fixed prompt away and leaves $\alpha\tau(o)$. §4.1 reports both.

C.5 THREATS TO THE MEASUREMENT, AND WHAT EACH WOULD DO

Algorithm 1 is the whole procedure for one backbone, written out so that the list below can be read against it. We list the ways this protocol could produce the reported pattern for the wrong reason, together with what each would look like, because several are checkable in the data and are checked in §4 and Appendix G.

1. *The verification prompt is unusually short, inflating v/c* (§C.3). It is short by design, because a grounded verdict needs no more; but a longer verifier prompt would raise v and raise the ratio further. The direction is against a sceptical reading, not for it.
2. *The false colour is easy.* A colour drawn from the corpus alphabet may be absent from the frame entirely, making rejection easy and inflating J. This inflates the quantity whose smallness is the finding, so again the bias runs against us. Appendix G reports the rate at which the swapped colour is present elsewhere in the frame.
3. *The cap does not bind.* If backbones ignore the word allowance, the sweep measures something other than a compression trade-off. They do ignore it, by a measured margin, which is why Eq. (8) is evaluated at realised length and why the overrun is reported (Fig. 5, right).
4. *The yield predicate carries the result.* Addressed by Definition 11 and §D.12, which report the rate under the widest and narrowest readings.
5. *Provider nondeterminism.* Realised token counts and verdicts both vary across identical calls. We report per-backbone values without pooling, and every ratio, floor and difference is formed inside one run. Three comparisons are deliberately *across* runs, and each is labelled as such where it appears: the batch-slope estimate of c_{obs} against the caption-cap estimate (§D.8 and §D.14, which is the point of those comparisons: two runs recovering the same term is the evidence), and the 200-probe replicate against the 16-probe pilot (§D.15), which is a reproducibility check and can only be made across runs. Appendix G gives the re-run spread.
6. *Missing cells.* One backbone lost six of 16 probes to provider-side errors. Its rows are reported over the probes that returned, with the denominator stated, and it is never silently pooled with the others.

Table 7: The replicate. Table 1 re-run on 200 probes per backbone instead of 16, with the same three caption caps and the same terse verifier, over 25 episodes rather than two. Yield uses the shipped predicate. *This table, not Table 1, is the source of every headline verifier number in this paper*; the 16-probe run is reported as the pilot it is. The re-run does not merely shift levels: yield moves by 5.5 points on average and 23.2 at worst, J by up to 0.245, and on 1 backbone (Gemma 3 12B) yield is non-monotone in the cap, which the 16-probe rate could not have shown. What survives is the steep rise with the allowance and the identity of the weakest backbones.

Backbone	n	yield (%) at cap			tokens		verification			
		32	64	128	cap 32	cap 128	true	false	tokens	J
Ministral 3B	200	37.5	79.0	95.5	491	546	72	98	460	0.69
Qwen3-VL-8B	200	24.0	90.0	98.5	361	405	99	50	334	0.48
Gemma 3 12B	200	1.5	24.0	21.0	311	355	63	76	293	0.39
Ministral 14B	200	52.0	63.5	88.5	492	547	86	98	460	0.84
Qwen3.8-27B	200	7.5	58.0	97.5	354	386	99	96	338	0.94
Qwen3-VL-30B-A3B	200	20.5	84.5	99.0	362	422	90	58	334	0.48
Gemma 4 31B	200	34.0	84.5	98.0	319	360	98	100	304	0.98

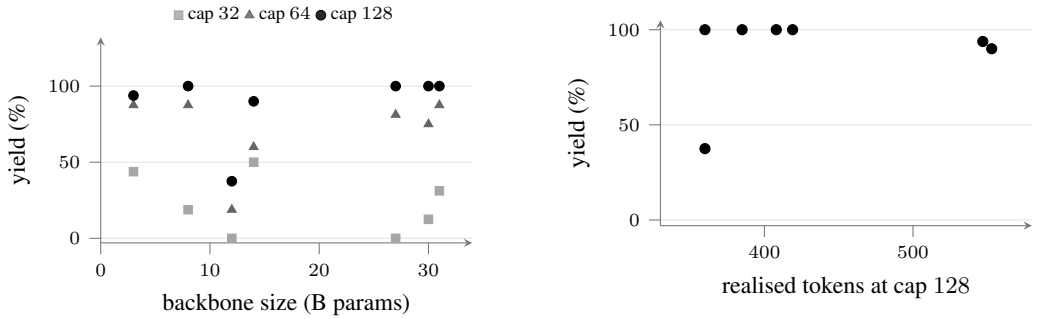


Figure 7: Yield is bought with the cheap half of the bill. Left: yield against backbone size at the three caption caps. Raising the cap moves yield by 66 points on average while the realised token cost rises 13%, because the image is paid for before the first caption token is emitted. Right: yield against the realised cost of the call it came from. The horizontal spread is dominated by how expensively a provider bills an image, not by how much description was bought.

D Experiments and Analysis, in Full

The subsections below follow the order of the threats listed in §C.5: yield and its price first (§D.1, §4.1), then what a verdict is worth, then the sensitivity of each judgement the measurement makes. Four subsections make forward references, which the ordering of the argument forces and which are marked where they occur: §4.6 and §D.8 both rest on constructions that §D.14 and §D.15 introduce.

D.1 Yield is cheap, and the bill barely moves

Raising the caption cap from 32 to 128 moves yield by 66 points on average while the realised token cost rises 13% (Figure 7). 4 of 7 backbones reach 100.0%. At the smallest cap the spread is enormous, 0.0% to 50.0%, and it closes almost entirely by the largest.

Those are the 16-probe numbers. Table 7 re-runs the same grid on 200 probes over 25 episodes, and it is the table the rest of the paper's verifier numbers come from. The re-run is not a cosmetic confirmation: yield moves by 5.5 points on average and by 23.2 at worst, on 1 backbone (Gemma 3 12B) it is non-monotone in the cap, and no backbone now reaches 100%: 5 of 7 reach 95% or better. What survives unchanged is the steep rise with the allowance and the identity of the two weakest backbones, and those are the two properties §4.4 uses. The 16-probe rate could not have resolved any of the differences just listed, which is what §5 item (3) said about it.

This is Proposition 9 measured. The image is paid for before the first caption token is emitted, so additional yield is bought with the cheap half of the bill. An agent following prices should therefore buy a generous description, and the measurements say the description it gets will be good.

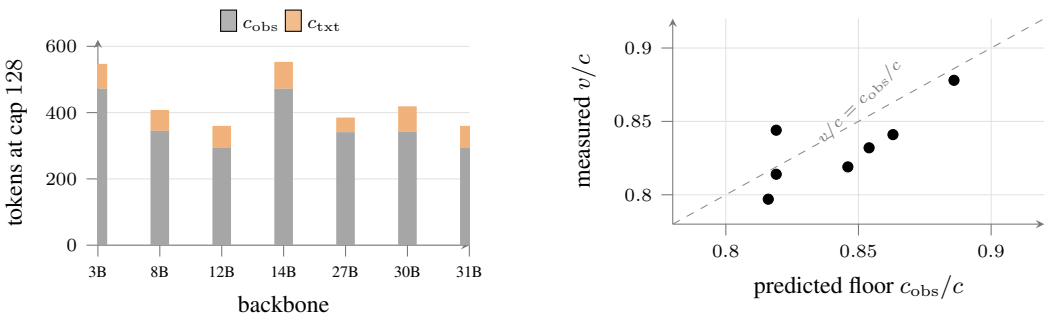


Figure 8: The two-point fit on the 16-probe pilot, and why its intercept is not the observation term. Left: realised tokens are approximately $\hat{\imath} + \hat{a} \cdot \text{cap}$, so two caps identify both; the intercept is 82–89% of the bill on every backbone. Right: the measured verification ratio against $\hat{\imath}/c$, agreeing to within 0.027 across a tenfold range of model size, but agreeing for a reason weaker than it looks. Six of 7 points sit *below* the diagonal, and a lower bound the data sits under is mis-specified: $\hat{\imath}$ holds the caption instruction's fixed input tokens as well as the observation, and separately the fit assumes a cap consumed in full when it is not. §4.1 identifies c_{obs} without the fixed prompt, from grounded minus ungrounded verdict calls in one run, and against *that* floor every measured ratio is above the diagonal.

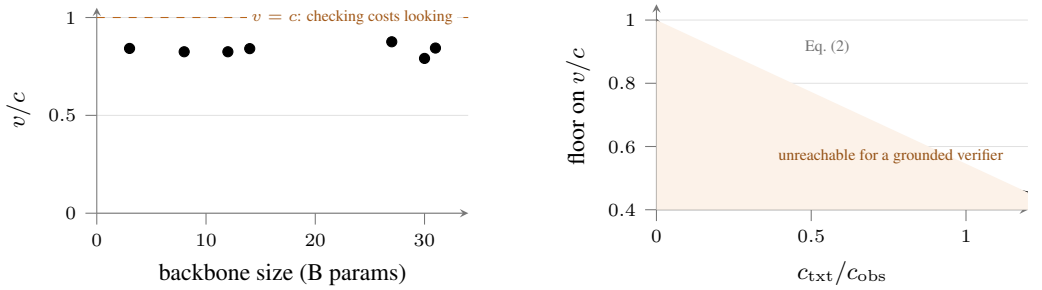


Figure 9: The measured ratio and the bound that pins it. Left: verification price over channel price, per backbone, on the 200-probe run. Every point sits between 0.791 and 0.876, a spread of a few points across a tenfold range of model size, which is what a ratio governed by the interface rather than by the model looks like. Right: Eq. (2) as a function of how much text a call buys per unit of observation. A grounded verifier cannot enter the shaded region at any prompt length.

D.2 Checking costs what looking costs

A caption call costs 360 to 553 tokens at the largest cap. A single yes/no check on the same frame costs 293 to 460. The ratio v/c runs 0.80 to 0.88 on the pilot and 0.791 to 0.876 on the 200-probe replicate (Figure 9), and it is flat across a tenfold range of model size either way.

Theorem 1 predicts a ratio approaching one as the observation dominates. The premise is an empirical claim about the interface, and identifying it took us two attempts.

What the cap sweep identifies, and what we first called it. Fitting realised tokens against the cap (Definition 14) gives an intercept of 295 to 472 tokens, 82–89% of the bill, and the ratio $\hat{\imath}/c$ it implies is 0.82 to 0.89 (Figure 8). The first version of this paper called that intercept c_{obs} and called $\hat{\imath}/c$ the floor. It is neither. The caption instruction is the same string at every cap, so its input tokens are part of what the fit holds constant, and $\hat{\imath} = c_{\text{obs}} + \beta_{\text{in}}|x_{\text{fix}}|$ over-estimates the observation term (Table 2). That is why six of 7 measured ratios sat below their own supposed floor: a lower bound the data sits under systematically is a mis-specified bound, and the unconsumed-cap story we offered for it was at best half the reason.

What identifies it, in one run and with no new calls. The expanded run asks the identical yes/no question twice, once with the frame attached and once without (§D.14). The ungrounded call carries the verifier's prompt and its one-word verdict and nothing else, so the difference between the two prices is the observation term with every fixed prompt token differenced away. It comes to 247–420

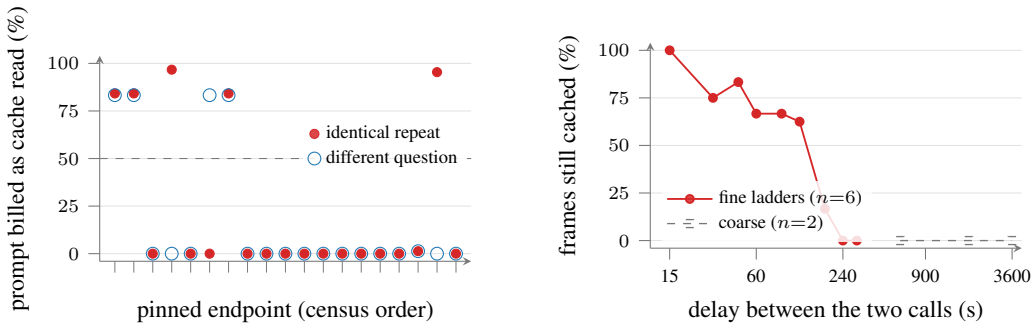


Figure 10: The interface audit. **Left:** the share of each pinned endpoint's prompt that the provider billed as a cache read, on a byte-identical repeat (filled) and on a different question about the same frame (open). The dashed line at 50% is the threshold that separates a cached observation from a cached chat template: the observation is 84–92% of one of these prompts, and the largest count anywhere below the line is four tokens. Most endpoints sit on the floor under both conditions, which is Assumption 1 holding. The two endpoints high on the filled series and on the floor of the open series are the finding a repeat-only test would have got backwards: their cache is keyed on the whole prompt, so a verifier asking a new question misses it. **Right:** how long an entry survives, on the endpoints whose cache a verifier does hit. 66 reads over 8 ladders, one previously unsent frame per delay. The escape is available to a caller that checks within about a minute of looking, and to nobody who defers, which is the whole point of holding a description instead of the frame.

tokens, against a cap-sweep intercept of 296–474 fitted on the caption calls of that same run: the intercept exceeds it by 49–56 tokens on every one of the 7 backbones, which is one caption prompt, uniformly. The observation is therefore 69–77% of the caption bill rather than 81–89%, and the floor of Eq. (2) is 0.69–0.77. Against it the measured ratio, 0.791–0.876, is above the floor on all 7 of 7 backbones, which is what Theorem 1 says and what the fitted version could not honestly claim.

The consequence is Corollary 1: verification does not add a small premium to a channel, it very nearly doubles its price, so a channel must return close to twice what the same tokens buy elsewhere before it is worth both using and believing. The gap between 69% and 81% does not touch that conclusion (a floor of 0.69 is still most of the bill), but it does dispose of the claim that the fitted floor and the measured ratio agreeing to a few thousandths was strong evidence for the decomposition. Both quantities are c_{obs} plus a fixed prompt of similar length; their agreement mostly says the two prompts are of similar length. The evidence that the decomposition is real is the batch cross-fit of §D.8, which recovers the same observation term from a different call type.

D.3 The interface audit: Assumption 1, measured

Everything above prices a call. Whether those prices compose into Theorem 1 depends on one property of the interface that the calls themselves cannot show: Assumption 1, that an observation is re-charged on every call grounded in it. An earlier form of this paper conceded that assumption: it held on our serving arm, we said, and the escape that would break it was a product feature we did not have. That was an assertion about an interface, and an interface can be asked. This subsection asks it, on every endpoint the router will pin for these seven backbones, and the concession turns out to have been wrong in both directions at once. It was too *weak*, in that on most of the arm the assumption is not something we have to be taken on trust for: it is a number the provider reports on every call, and it comes back the way the theorem needs. And it was too *strong*, in that the escape we said we did not have is on part of the arm, at the discount our own proposition predicts. What decides which of the two readings governs is a quantity neither the concession nor the proposition mentions: how long the discount lasts.

What is measured, and why it is an integer rather than an opinion. A provider that serves a cached prompt token bills it at a published discount and reports it in the response, so the question "was the observation re-charged?" has a numeric answer on every single call: `prompt_tokens_details.cached_tokens`, against the prompt length, against the bill. For each endpoint we send a frame, re-send the identical call, and then ask a *different* question about the same frame, the last being the case an agent actually has, since a description and a check are not the

same string. Table 3 reports the three counts. The observation is 84–92% of one of these prompts (identified per endpoint by the same grounded-minus-ungrounded subtraction §4.1 uses, and larger than the 69–77% it is of a *caption* bill, because a caption also buys many completion tokens), so a cached count above half the prompt is an observation that was not re-charged, and a count in the single digits is a chat template. No endpoint returns anything between the two: the largest count below the threshold anywhere in the census is four tokens.

The answer has three cases, and the middle one is the one a census would miss. Of 19 pinned endpoints across 7 backbones, 5 return a cached observation on a byte-identical repeat and 4 return one when the *question* changes; the second is the call a verifier actually makes. The two sets are not nested, which is itself informative: one endpoint missed a repeat it served a moment later under a different question, because the cache is best-effort and shared. The gap between the two counts is a distinction between two products that share a name:

- **No cache at all.** The observation is billed in full on every call, and the cost of a repeat equals the cost of the cold call to the reported precision. This is the majority of the arm.
- **A prefix cache, which a verifier misses.** 2 endpoints cache the observation when the identical call is re-sent and cache nothing when the question changes: a cache keyed on the whole prompt, not on the image inside it. Assumption 1 is violated there in its letter and satisfied in the only case this paper puts weight on: a description and a check are not the same string, so the verdict call misses.
- **An observation cache, which a verifier hits.** 4 endpoints, on 2 backbones (Ministral 3B, Ministral 14B), return the image as cache reads whatever question surrounds it and whichever order it is placed in. Here Assumption 1 is simply false, and Proposition 1 takes over.

Counted the way the theorem needs (can a *grounded verdict on a new claim* avoid paying for the observation?), the assumption holds on 5 of 7 backbones and fails on 2. It is a measurement on those five, not a concession.

The interesting negative is that an advertised discount predicts almost nothing about this. 12 of the pinned endpoints publish a cache-read price (the α' of Proposition 1, at 0.10 to 0.56 of their prompt price), and 6 of them never returned a cached *observation* on any image call we could construct, the identical repeat included; one of them returns four tokens, which is its chat template. A cache-read line in a price list is a fact about billing, not about what is cached; a paper that had read the price lists instead of the usage records would have concluded that the escape was widely available, and it is not.

The classification is a rate, not an anecdote. Table 3 sends one call of each kind per endpoint, which is enough to separate zero from four hundred but not enough to call a cell reliably one thing or the other; and one cell in the table proves the point by missing a repeat it then served under a changed question. The 16-probe price arm below supplies the replicates for free, because every verdict call in it asks a different question about a frame the caption call on the same probe has already sent: 480 such calls across the 15 cells the census calls non-caching, of which 0 were served from a cache, against 78–97% of the 128 in the 4 cells it calls caching. The classification is therefore a separation of two populations and not a reading of one call each. The same arm confirms the arrangement it depends on: the caption call is cold on 303 of 304 probes, so the verdict's cache state is a fact about the verdict's call and not an inheritance.

It also says what kind of object the cache is even where it works. Between a fifth and a fortieth of the verdict calls in the caching cells miss: inside the lifetime, on an idle frame, with nothing changed. A cache read is a rate there, not a price.

Where it is false, it is false by the factor the proposition names. On the endpoints whose cache a verifier hits, 83% of the prompt returns as cache reads and the call's bill falls to 0.25 of the cold call's, a $4.0\times$ reduction, on all 3 endpoints where the two calls differ in nothing but their cache state. Proposition 1 predicts that number rather than fitting it, and the prediction has no free parameter: the endpoint publishes $\alpha'/\alpha = 0.1$ and reports 384 of 456 prompt tokens cached, so a warm call should cost $(72 + 0.1 \times 384)/456 = 0.242$ of a cold one before completions, against a measured 0.247, the difference being the completion tokens, which are not discounted. The same check on the two prefix-cached endpoints, whose discounts and cached fractions are different again, predicts 0.428

and 0.497 against measured 0.438 and 0.506. Five endpoints, three distinct published discounts, one formula, no fitted quantity. The escape in Table 6 whose row said "not on this arm" is on part of this arm, it is worth what the proposition says it is worth, and the rows marked † in Table 8 are it, in situ.

The layout escape, meanwhile, is empty. Putting the observation *first* in the message, so that two calls about one frame share a prefix, is the one thing a caller can change for free. We ran it on all 24 endpoint cells and read it only on the 18 where it could help (those whose cache does not already survive a changed question), and it produces a cached observation for a changed question on 0 of them. On the prefix-cached endpoints it buys back the identical repeat and not the changed question, which says their cache key is the prompt rather than a prefix of it. Reordering the message does not turn a prompt cache into an observation cache.

And where it is false, the escape has a horizon the architecture cannot meet. A discount available for a few seconds is not the same product as a discount available for an hour, and the difference is exactly the deferral the channel exists to create: a streaming agent describes a frame when it arrives and checks the description when a question arrives, which is not the same moment. So we measured the entry's life directly: 66 reads over 8 independent ladders, each delay drawing a frame no other arm of the run had sent, so a hit cannot be inherited.

The survival curve is 15s:6/6 30s:6/8 45s:5/6 60s:4/6 90s:4/6 120s:5/8 180s:1/6 240s:0/6 300s:0/8 600s:0/2 1800s:0/2 3600s:0/2. Every frame is still cached at 15 seconds; half or more survive to 120 seconds; 180 seconds is the longest delay at which any frame was still cached, and by 240 seconds every frame in every ladder was cold again. The two coarse ladders confirm it does not return: at 1800 and 3600 seconds every read was cold. The decay is also not a clean threshold (one ladder reads cold at 30 seconds and warm again at 45), which is what a shared, best-effort cache under other tenants' load looks like, and it means the horizon is a distribution rather than a contract.

So the escape is real and it is available to a caller who checks within roughly a minute of looking. It is not available to the architecture this paper is about. The configuration Corollary 1 describes (describe now, and check when a later query makes the claim load-bearing) is precisely the one in which the entry has expired, and the second call pays α again.

The two regimes are worth separating carefully, because the boundary between them is not arbitrary. A check issued inside the cache window is a check issued while the agent still has the frame in hand. An agent in that position did not need a description: it could have answered the query from the observation in one call rather than describing it in one and checking it in another. So the window in which the escape is available is close to the window in which the channel whose price this paper is computing has no reason to be there. A description is bought to survive the moment the observation is discarded, and the cache does not. This is a sharper statement than the concession it replaces, and it is sharper in the direction that supports the result: the floor does not survive because caching is unavailable, it survives because caching is *short* relative to the deferral that makes a description worth holding.

The price, pinned to a provider and paid in money. The same run answers two objections that have nothing to do with caching.

The first is routing. A router load-balances a model across third-party endpoints, so "the price of Qwen3.8-27B" measured through one is a mixture of a model and whoever served it. We re-ran the price measurement with the provider pinned and fallbacks disabled, 16 probes per cell, on 19 endpoint cells covering 7 backbones of which 6 have more than one endpoint (Table 8). Across endpoints of a single model the token price of an observation moves by at most 21 tokens and v/c by 0.012, while the *dollar* price of the identical call moves by up to $2.1\times$. The quantity Theorem 1 is about is a property of the model and the interface's tokenisation; the bill is a property of the vendor.

The third column of that comparison is the one we did not expect and report as a lead rather than a result. J moves by up to 0.37 across endpoints of a single model (on Qwen3-VL-30B-A3B from 0.44 on one endpoint to 0.81 on another, with the same weights, the same probes and the same question), while v/c on those same four endpoints moves by 0.009. These endpoints differ in quantisation, and 16 probes put a wide interval on any one of those numbers, so this is not a measurement of a quantisation effect; it is an observation that the decoupling of §4.4 may run deeper than we claimed there. We said the price belongs to the interface and the value to the model. If this holds up at a serious

Table 8: The price measurement with the provider pinned and the bill recorded, 16 probes per cell. c is the caption call at the largest allowance, v a grounded verdict, c_{obs} the grounded-minus-ungrounded subtraction of §4.1; each is given in tokens and in the provider's own dollars. Two readings matter. Across endpoints of one model the dollar price of an identical call moves by up to $2.1\times$ and, on Gemma 4 31B, the observation's own token count by 9.1% (235–256 tokens for the same image), while v/c moves by 0.012 on every one of them: the same $\tau(o)$ enters both of its terms, so a server that tokenises an image differently shifts the numerator and the denominator together. The ratio the theorem is about survives a change of server that neither of the prices it is built from does. And v/c is systematically *lower* in money than in tokens, because the caption spends its allowance on completions and completions are billed above prompt tokens; the floor moves with it, and the measured ratio clears it in both currencies on 15 of 15 cells. Rows whose verdict calls hit an observation cache are marked † and are read in §4.2, not here.

Backbone	endpoint	tokens			$ per call ($\times 10^{-5}$)			J
		c	v	v/c	c	v	v/c	
Ministral 3B	`mistral`†	547	460	0.841	5.47	1.25	0.229	0.62
	`mistral/eu`†	549	460	0.838	6.03	1.61	0.267	0.62
Qwen3-VL-8B	`alibaba`	407	334	0.820	7.20	3.97	0.552	0.44
	`parasail/bf16`	406	334	0.823	13.67	8.44	0.617	0.38
Gemma 3 12B	`deepinfra/bf16`	362	293	0.809	2.49	1.49	0.599	0.31
Ministral 14B	`mistral`†	548	460	0.839	10.53	2.93	0.279	0.81
	`mistral/eu`†	545	460	0.843	12.00	4.18	0.348	0.81
Qwen3.8-27B	`chutes/fp8`	384	338	0.879	20.86	10.38	0.498	1.00
	`deepinfra/bf16`	384	338	0.879	13.56	5.41	0.399	0.94
	`novita`	385	338	0.878	27.98	14.70	0.525	0.94
	`parasail/fp8`	384	338	0.879	18.08	8.50	0.470	0.94
Qwen3-VL-30B-A3B	`alibaba`	419	334	0.796	8.73	4.42	0.506	0.44
	`deepinfra/fp8`	418	334	0.799	9.99	5.10	0.510	0.69
	`novita/bf16`	422	334	0.792	12.76	6.77	0.531	0.56
	`siliconflow/fp8`	415	333	0.801	17.73	9.72	0.548	0.81
Gemma 4 31B	`chutes/fp4`	358	304	0.850	5.54	3.65	0.658	0.94
	`deepinfra/turbo`	361	304	0.842	4.63	2.79	0.602	1.00
	`novita/bf16`	337	283	0.839	6.20	4.01	0.648	0.88
	`parasail/fp8`	363	304	0.837	6.89	4.61	0.670	1.00

probe count, the value does not belong to the model either: it belongs to the served configuration of the model, which a user of a routed endpoint does not choose and is not told. Establishing that needs a probe count this arm of the run does not have, and we list it in §5 rather than claim it.

That the routing effect is real, and not merely possible, the audit shows by accident. On the *unpinned* route two backbones billed a byte-identical repeat at $2.2\times$ and $0.49\times$ the cold call's price with no cached token on either (the router had served the two calls from different endpoints), while all 19 pinned cells billed the repeat identically to the penny. Published prices also move: between the audit and a re-check an hour later one endpoint's prompt price fell 25%. Neither fact touches a token count, which is the reason the theorem is written in tokens and the bill is reported beside it rather than in place of it.

The second is the currency. Equation (1) is written in tokens, and an interface bills in money at three different rates: prompt, completion, and cache read. The theorem does not care: it needs $\alpha, \beta_{\text{in}}, \beta_{\text{out}} > 0$ and nothing more, so it holds in whichever currency the coefficients are read. The *constant* is not currency-free, and it moves in the direction that costs us. On the 15 cells whose verdict calls are cold, v/c is 0.792–0.879 in tokens (which reproduces the routed measurement of §4.1, 0.791–0.876, to within 0.011 on every one of the 19 cells at a sixteenth of the probe count), and 0.399–0.670 in dollars. The gap is not an inconsistency: a caption spends its allowance on completion tokens, completions are billed above prompt tokens, so the denominator grows faster in money than in tokens. The floor moves with it, from 0.683–0.755 to 0.321–0.566, and the measured ratio clears its own floor in both currencies on 15 of 15 cells. What changes is the headline's size, not its sign: in money a grounded check costs roughly half of what the description costs rather than four-fifths of it, and the admission criterion of Corollary 1 demands about $1.5\times$ what the same tokens buy elsewhere rather than about $1.8\times$.

The 4 cells marked † are the cached ones, and they are the measurement of Proposition 1 in situ rather than a nuisance. In those cells the caption call is cold (it is the first time the frame is sent), and the verdict call, issued seconds later, is warm. That is the most favourable schedule the cache admits, and under it v/c in money falls to 0.229–0.348 while v/c in *tokens* stays at 0.838–0.843. The token currency is blind to the cache by construction, which is the sharpest reason to have measured both.

What was fixed before the calls. Unlike the clip and batching arms, this one was not pre-registered, and the distinction is worth drawing rather than blurring. It is a census of an interface rather than a test of a hypothesis about a model: it reads three integers off every response and there is no estimator to choose. What protects it is structural and was fixed before the run: every endpoint the catalogue lists for these backbones is called, the call sequence per cell is identical, and the threshold that separates a cached observation from a cached chat template was set at half the prompt in advance, against an observation that is 84–92% of the prompt and a largest observed template contribution of four tokens, so no cell lands anywhere near it and no threshold between 0.02 and 0.83 would change a single cell. The one quantitative prediction, Proposition 1's bill ratio, is computed from the provider's published price and its own reported token counts, so it has no free parameter and cannot be tuned after the fact. The survival curve is descriptive and is reported as an order of magnitude rather than as a confirmed prediction. What is *not* established, and would need registration if pursued, is limitation (9).

What the audit changes, stated plainly. Theorem 1 is unchanged; it was always conditional on Assumption 1 and says so. What changes is the status of the premise. It is no longer asserted of an arm: it is measured on 19 endpoints, it holds for the call a verifier makes on 5 of 7 backbones, and where it fails it fails by a factor the theory predicted in advance and for a duration too short to reach the call the architecture actually defers. The result of this paper should be read at three levels, and they are not equally portable:

1. **Universal, given the interface property.** A grounded verifier pays the observation's transmission, so $v \geq c_{\text{obs}}$ and $v/c \geq \theta$. This is Theorem 1 and it is currency-free.
2. **Interface-level, and measured here.** On contemporary hosted multimodal endpoints the observation is 84–92% of a verdict call's prompt and 69–77% of a caption's whole bill; it is re-charged in full to a verdict on 5 of 7 backbones, and where it is not, the discount expires in minutes. So θ is 0.683–0.755 in tokens and 0.321–0.566 in money on the calls an agent actually makes. These constants are the part of the result with a shelf life: they are properties of a tokenisation and a price list, both of which move.
3. **Task- and model-level, and *not* portable.** Whether a floor of that size makes checking unaffordable depends on what a verdict is worth, which is J, which varies by more than a factor of three across backbones (§D.4), with the prompt (§D.13), and with the corpus: on photographs it is high enough that checking is affordable at these very prices (§4.6).

Conflating the three is the easiest way to over-read this paper, and the second half of §4.6 is where we first caught ourselves doing it.

D.4 And the verdicts carry less than their price suggests

On the 200-probe replicate (Table 7), J runs from 0.39 to 0.98 across backbones and 3 of 7 fall below one half. 0 backbones verify perfectly in both directions. Converting to the effective price of Definition 3, one verdict's worth of information costs 309 to 751 tokens: on the worst backbone that is more than twice the price of the caption being checked, and on the best it is comparable to it. Figure 11 plots both halves (the true-accept / false-accept plane on the left and v/J on the right), and Figure 3 puts the nominal, effective and channel prices in the same units.

The pilot said something stronger, and it was the probe count talking. The 16-probe run of Table 1 gave J from 0.31 to 1.00 with 2 of 7 below one half and 2 backbones perfect in both directions, and a v/J range of 304–936 tokens. Two of those four statements do not survive a tenfold larger denominator. No backbone is perfect at 200 probes (sixteen correct pairs put an upper confidence limit nowhere near 1, and neither of the two apparently perfect verifiers reaches $J = 1$ when asked two hundred), and the count below one half rises from 2 to 3. J per backbone moves by up to

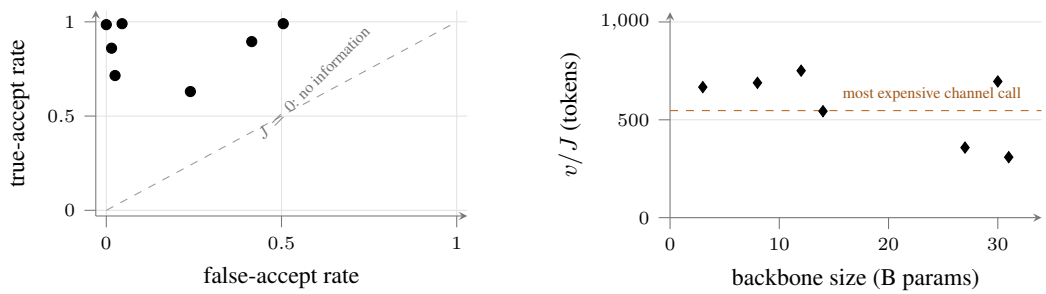


Figure 11: What a verdict is worth, against what it costs, on the 200-probe run. Left: every verifier in the true-accept / false-accept plane, with the diagonal marking $J = 0$, where a verdict carries no information about the claim. Distance above the diagonal is J. Right: the effective price v/J of Definition 3. On the weaker backbones one verdict's worth of information costs more than the most expensive caption call in the whole experiment, which is the sense in which verification is unaffordable here rather than merely expensive.



Figure 12: Verification fails asymmetrically, not noisily. A verifier that is right on true claims and wrong on false ones has an accuracy near one half on a balanced set, which is the wrong summary: it is not a coin but a machine that says yes, and its verdict carries information in one direction only. The two bars are what $J = \mathrm{TPR} - \mathrm{FPR}$ summarises, and the reason Definition 3 normalises the price by J rather than by accuracy. Rates are from the 200-probe run.

0.245 between the two runs. We report the pilot's table unchanged and take every headline from the replicate.

Two readings, and the second is the one that generalises.

The worst case is a low-yield channel with an expensive check. Gemma 3 12B reaches only 21.0% yield at the largest cap and its verdict carries $J = 0.39$, so its effective verification price is 751 tokens against a channel price of a few hundred. Nothing about that configuration is worth admitting under Corollary 1, and the two numbers fail independently.

Figure 13 places the seven backbones on the admission plane of Eq. (5), one line per value of the marginal worth of a token elsewhere.

The failure is asymmetric rather than noisy. Qwen3-VL-8B accepts 99% of true claims and rejects only 50% of false ones. Balanced accuracy would call that roughly a coin, which is the wrong summary: a coin is uninformative in both directions, while this verifier is fully informative in one and empty in the other. J is what separates them, and it is the reason Definition 3 normalises by J rather than by accuracy.

We did not find a *grounded* verifier with $J = 0$, and for grounded verifiers Proposition 2 is therefore a statement about a limit these backbones approach rather than reach. Ungrounded ones reach it exactly, on 6 of 7 backbones (§D.14). An earlier partial run, before two backbones had completed, did contain a grounded verifier at $J = 0$; the finding did not survive the full set, and we record that rather than quietly dropping it.

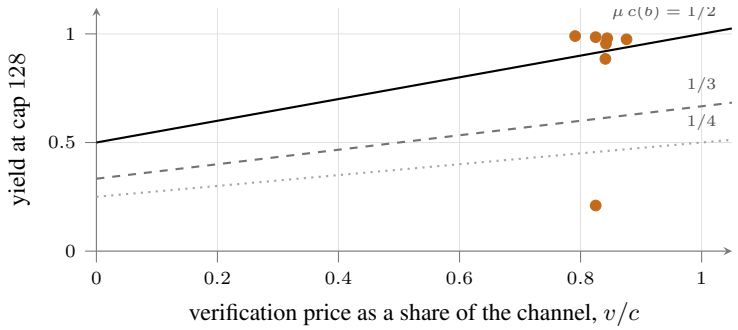


Figure 13: The admission plane of the companion analysis (Anonymous authors, 2026a) with the measurements placed on it. The horizontal coordinate is the verification price as a share of the channel's own price, which Theorem 1 pins near one; the vertical is the channel's yield in accuracy units. Eq. (5) makes the boundary $y_k = \mu c(b)\,(1 + v/c)$, which is a *family* of lines rather than a single one, because μ (what a token is worth in the budget's best alternative use) is not something this paper measures. Three members of the family are drawn. A point is admissible with checking only if it lies above the line for its own μ, so the figure says which μ each backbone could tolerate rather than sorting the backbones into admissible and inadmissible. The measured points cluster at the right-hand edge because the interface, not the model, fixes their horizontal position: whether a backbone clears the criterion is decided by its yield and by μ, and not at all by which model it is. Coordinates are the 200-probe run.

The four ways a verdict pair can come out. J is a difference of two marginals, and two backbones with similar J can be failing in opposite directions, which is what Figure 12 shows as two bars per backbone. Table 4 opens the scalar back up on the 16-probe pilot, where the individual pairs can be counted: each probe contributes one verdict on the true claim and one on the same claim with the colour swapped, so the pair lands in one of four cells. Qwen3-VL-8B accepts both members of 9 pairs: it is the machine that says yes, and its yes rate of 78.1% against a balanced set is the signature. At the other end two backbones are correct on every pair. The yes rate across the table runs from 30.0% to 78.1%, which is a wider spread than J alone suggests, and 1 pair is anti-correlated with the truth.

D.5 Price belongs to the interface; value belongs to the model

The tables so far report, per backbone, a yield, a nominal verification ratio and a J. Read across backbones rather than down them, they separate into two quantities with visibly different physics, and the separation is the most useful thing this measurement does. Every number in this subsection is from the 200-probe run of Table 7.

The *nominal* ratio v/c spans 0.791 to 0.876 (a factor of 1.11 across backbones differing by an order of magnitude in parameter count), and its correlation with scale is $r = -0.005$, which is to say none. This is what Theorem 1 predicts: the ratio is pinned by how the interface prices an image against text, and the model on the other end of the interface does not enter. A larger backbone does not get a discount on transmitting a frame.

The *effective* price v/J spans 309 to 751 tokens, a factor of 2.43, and J itself runs from 0.39 to 0.98. This is the half that belongs to the model: whether a verdict distinguishes a true claim from a false one is a capability of the particular backbone rather than of the endpoint that bills for it.

What seven backbones can and cannot say about scale. It is tempting to finish that sentence with "and capability improves with size". The correlation between J and parameter count is $r = 0.415$ on the 200-probe run, whose 95% interval on seven points runs from -0.49 to 0.89 and contains zero; on the 16-probe pilot the same correlation read 0.728, and the difference between the two is two backbones moving by more than 0.2 in J when the denominator grew. So the paper does not claim that J tracks scale, and a reader should not take the ordering of Figure 14's right panel as one. The claim that survives is the *decoupling*, and it does not need a trend: v/c varies by a factor of 1.11 with an interval on its correlation of -0.76 to 0.75 (flat, and flat at a level the interface sets), while J and v/J vary by much more across models of similar size. The half of the verifier the admission criterion

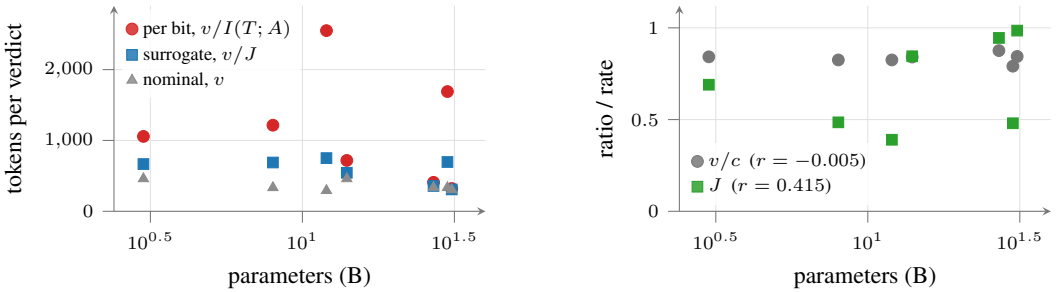


Figure 14: The two halves of a verifier behave differently. All points are the 200-probe run. **Left:** what one verdict costs, in three currencies. The nominal price (grey) is flat in scale; the surrogate v/J (blue) and the honest per-bit price $v/I(T;A)$ (red) are both far more dispersed, and the gap between them is Proposition 8: the surrogate understates the per-bit price by up to $3.39\times$ because information degrades as J^2 and not as J. The per-bit series is computed from each verifier's measured $(\mathrm{TPR}, \mathrm{FPR})$ by Eq. (6), not from J through the symmetric expansion, which would understate it by up to $1.55\times$ on these backbones. **Right:** the same point as rates. v/c is a property of the interface and does not move with scale ($r = -0.005$); J is a property of the individual model and moves a great deal, though not in a way seven points resolve into a trend ($r = 0.415$, 95% interval -0.49 to 0.89). The admission criterion contains the flat one.

charges for is fixed by the endpoint; the half that could pay for it is a property of the model, and which model is not predicted by how large it is.

Why this does not resolve itself. It would be reasonable to read the configuration as transitional: verification is expensive relative to a channel that is not yet good, and both will improve. The measurement forbids that reading. Yield reaches its ceiling first. At the largest caption allowance, 5 of 7 backbones carry the probe's pairing on 95% or more of probes, and yield's correlation with scale is only $r = 0.271$ because there is little left to buy. Over that same set, J still spans 0.39 to 0.98. The numerator of the admission criterion stops growing while the price of believing it does not fall, so the criterion becomes *harder* to satisfy as the channel improves, not easier.

The honest per-bit figure. Proposition 8 says v/J is an optimistic surrogate. Measured by Eq. (6) on each verifier's own $(\mathrm{TPR}, \mathrm{FPR})$, one verdict buys between 0.115 and one bit about the claim, and the per-bit price runs from 322 to 2550 tokens against a surrogate range of 309 to 751. The worst case is Gemma 3 12B, and it is worst not because its calls are expensive (they are the cheapest in the table) but because its verdicts are nearly empty. A cheap verifier that does not discriminate is the most expensive object in this paper.

D.6 Where the words went

A yield rate says whether the caption happened to carry the pairing. It cannot say what the caption carried instead, and the mechanism is there. Because the per-frame inventory is recovered exactly (§C.1), we can classify every caption by how much of the pairing survived and, separately, by what it spent its allowance on.

Figure 15 plots the taxonomy of Table 5 against the cap on the left and the typicality test on the right.

The counting bound binds where it should. Evaluated at realised caption length in the name-token unit of Proposition 3 (13.1, 28.1 and 38.7 against the 12.8, 27.3 and 38.7 whitespace words Table 5 reports, the difference being notations like `hat(yellow),` that hold two names in one word), Eq. (8) gives 57.3%, 138.1% and 201.4% at the three caps against measured yields of 20.8%, 71.7% and 88.7%. It holds everywhere and it constrains in only one place: with the vacuity point at $n + n_a$ (19.9, 22.0 and 22.5 name tokens against realised lengths of 13.1, 28.1 and 38.7), the bound is binding at the shortest cap and vacuous at the other two, and yield saturates above it (Fig. 5). This is weaker than what the discarded $w/(2n)$ appeared to give, which was binding at two caps; the correction is set out under Eq. (8). One constraining point and two vacuous ones is thin evidence on its own, and the argument for the binding account does not rest on it: it rests on the format manipulation of §D.10, which intervenes on the pairing instead of bounding it.

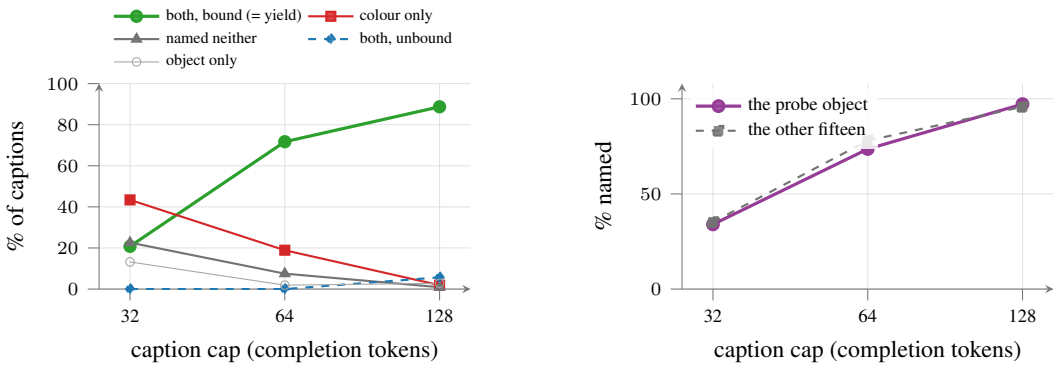


Figure 15: Left: what happens to the object–colour pairing as the allowance grows. The dominant failure at the tight cap is not that the caption saw nothing; it is *colour only*, at 43.4%: the caption names a colour and does not name the object that colour belongs to. **Right:** the test between the two accounts of §B.6. The typicality account predicts the probe object, the rarest class in the episode, is named less often than the fifteen common objects rendered beside it. The two curves are indistinguishable at every cap, and at the largest the probe is named slightly *more* often. The typicality account is refuted; the binding account is what is left.

The typicality account is refuted. The probe object is the rarest class in its episode, appearing in exactly one frame; the fifteen objects rendered beside it are drawn from a pool of common fillers. If a capped describer enumerates by salience, the probe should be the first thing dropped. It is not. Pooled over backbones the probe is named at 34.0% against 35.2% for the other fifteen at the tightest cap, and at 97.2% against 95.8% at the loosest: above, not below. Counted per backbone rather than pooled, so that two opposite effects cannot cancel, the probe is named more often than its neighbours on 4 of 7 backbones and less often on 3. That is a null, and it is the prediction the account we found more plausible made.

What a tight allowance actually destroys. The pairing, and it destroys it by dropping the *object* rather than the attribute. At the tightest cap 43.4% of captions name a colour without naming the object it belongs to, against 13.2% in the other direction, and the average caption emits 4.0 of the eight colour names in the alphabet while naming only 35.1% of the sixteen objects present. This is the asymmetry Proposition 3 points at from the counting side: the attribute marginal is bounded by the alphabet and saturates cheaply, the object marginal grows with the scene, and the pairing is the only one of the three that must pay for both: it is charged the attribute vocabulary on top of the objects. The caption spends its first words where coverage is cheapest, and coverage of the colour alphabet is very cheap.

The overrun, and why it matters for reading any cap. No backbone respects the word allowance in the prompt. Against an instruction of 8 words the captions run to 12.8 on average, which is 160.1% *of* the instruction, an overrun of sixty per cent, not of a hundred and sixty. The overrun is not monotone in the allowance either: it is largest at the middle cap (170.9% of an allowance of 16 words) and smallest at the largest (120.9%). The cap is enforced by the completion-token limit, not by the model's compliance, which is why every bound in this paper is evaluated at realised rather than instructed length. It is also a caution about the wider literature's practice of specifying description budgets in words.

D.7 Off the schematic corpus: the price ports, the value does not

Everything so far is one corpus and it is a render. We re-measure the floor's two quantities on 200 COCO photographs per backbone. The claim has to change with the corpus: an object–colour binding is not ground truth on a photograph, because a colour derived from pixels inside a hand-traced polygon is not what the annotation asserts, and three attempts at deriving one agreed with a human reading on about four fifths of probes; so the claim here is *class presence*, whose truth the annotation does assert exhaustively. The negatives are drawn from the corpus's own co-occurrence statistics, by the construction §D.15 below applies to the render corpus: the hard negative is the absent class that most often keeps company with the target.

The price half ports exactly, and tightly. c_{obs} on photographs is comparable to c_{obs} on renders, and the floor of Eq. (2) is 0.884–0.935 against a measured ratio of 0.877–0.921, and the inequality of Theorem 1 holds up to the fit's own bias on all 7 of 7. The two agree to 0.013, tighter than on the schematic corpus, and on all 7 backbones the measured ratio sits marginally *below* the fitted value, for the reason §C.4 and §4.1 give: this arm has no ungrounded call to subtract, so c_{obs} here is the two-point intercept and carries the caption prompt's fixed tokens with it, which puts the quoted floor above the true one by about a prompt's worth. Read as an upper bound on θ it is exact to 0.013; read as a floor it is the same over-estimate §4.1 corrects on the render corpus. Verification costs what looking costs on photographs too, because the interface prices a photograph the way it prices a render: per call.

The value half does not port, and it makes the result bite less. J over the 28 (backbone, prompt) cells has median $+0.828$, far above what the same verifiers achieve on the schematic corpus. Class presence in a photograph is an easier question than an object–colour binding in a dense render, and it should be. Since the effective price is v/J, the consequence is that on photographs verification is much cheaper in the currency that matters: 341–534 tokens per verdict of full information, against 313–1939 for the same quantity at a single frame on the schematic corpus, measured in the clip arm of §D.9 below. The cheap end is the same; it is the expensive end that disappears, and with it the backbones on which grounded checking was unaffordable. **This weakens the reach of the paper's conclusion and we state it as such.** The floor is a floor everywhere; how far above it a verifier sits, and therefore whether the admission criterion is actually violated, depends on how discriminable the claims are; and on natural images with coarse claims, they are discriminable enough that the floor may be affordable.

This is the clearest instance of the three-level reading §4.2 sets out, and it is worth naming the levels on this result specifically. The floor itself ports: it is a consequence of how the interface prices an observation and the interface does not know what the observation depicts. The floor's *size* ports: θ is within 0.013 of its render value here. What does not port is the third level (whether a verifier sits far enough above the floor for the verdict to be worth its price), and that is a fact about the discriminability of the claims an agent actually needs checked. A paper that reported only the first two levels would be correct and would leave a reader with the wrong practical conclusion; a paper that reported only the third would have measured a corpus.

And J remains a property of the question, not only of the corpus. Across 4 verifier phrasings on the same photographs and the same claims, J spans -0.735 to $+0.900$. One phrasing (the one that asks the verifier to consider the negation first) collapses J to below a quarter on two backbones and *inverts* it on a third, taking it to -0.735: a verifier that is reliably wrong. That is the finding of §D.13 reproduced on a different corpus with a different claim type, and it is the reason the J values above should be read as one point in a wide space rather than as a property of photographs.

D.8 The batching knee, predicted and then found on the wrong side

Proposition 1 says the amortised price of a verdict falls as $\Theta(1/m)$ when m claims about one observation are checked in one call, up to the claims available per observation, and flattens thereafter. A second, smaller batching sweep is reported in §D.14 below; it reaches only $m \leq 4$, which is entirely inside the regime where the $1/m$ term dominates and therefore says nothing about where the escape stops. This subsection is the sweep that does. A batched call costs $c_{\mathrm{obs}} + m\, c_{\mathrm{claim}}$, so the per-claim price $p(m) = c_{\mathrm{obs}}/m + c_{\mathrm{claim}}$ falls like $1/m$ only while the first term dominates, and the knee sits at $m^* \sim c_{\mathrm{obs}}/c_{\mathrm{claim}}$. We sweep $m \in \{1, 2, 4, 8, 16\}$ on 64 probes over 8 episodes per backbone, where 16 is the entire unambiguous inventory of one frame.

The void condition registered for this arm is checked before anything else is read: a truncated batch reply prices truncation rather than amortisation, so a backbone is only readable if at least 80% of its batch calls return all m parsed answers. All 7 of 7 clear it, the worst at 94%, so no row is void.

$p(m)$ is decreasing on all 7 backbones and convex in m on all 7, which is B1 as registered, convexity clause included. The straight line $\mathrm{tokens}(m) = c_{\mathrm{obs}} + m\, c_{\mathrm{claim}}$ fits it with $c_{\mathrm{claim}} = 12.7$–$14.0$ tokens (nearly independent of the backbone, while c_{obs} ranges over a factor of 1.59), and the measured $p(m)$ agrees with $c_{\mathrm{obs}}/m + c_{\mathrm{claim}}$ to within 1.4% at every m on all 7 backbones, against a registered

threshold of 5% on a majority. That is B2′, the re-specified form of a prediction whose first version we got wrong (below).

The fit's intercept is also a second, independent estimate of c_{obs}, which is B4. It agrees with the one fitted from the caption caps to 0.1–6.4%, mean 3.6%, inside the registered 10% on all 7 backbones. The two fits share no term, use different call types, different prompts and different completion allowances, and they recover the same observation cost. That agreement is the strongest evidence in this paper that the decomposition of a call into observation and text is real and not an artefact of curve fitting; and, as §4.1 notes, it is stronger evidence than the floor-versus-ratio comparison that the first version of this paper leaned on, because two similar prompts producing similar intercepts is a weaker coincidence than two dissimilar call types producing the same observation term. It is also, and this is the one place the paper deliberately breaks its own rule, a comparison across two runs: the batch sweep is 64 probes over 8 episodes and the caption sweep is the 16-probe pilot. Comparing them is the point.

And the escape does not reach its own knee. The fitted knee is at $m^* = 21$–36 on 7 of 7 backbones: *above* the 16 claims a frame supplies. That is B5, and it holds on all seven where the registration asked for a majority. So the per-claim price is still falling when the observation runs out of things to ask about: at the largest batch the price per verdict is 35.8 tokens against $2\,c_{\text{claim}} = 26.6$, so the observation cost counts as amortised away on 0 of 7 backbones. Batching is bounded by scene content, not by interface pricing. An agent with more claims per observation would get more from it, and on this corpus no agent has them.

Two of our own predictions did not survive, and both failures were arithmetic rather than empirical. B2 as first registered asserted $p(m) = c_{\text{obs}}/m + c_{\text{txt}}$ with c_{txt} taken from the *caption* fit. An eight-probe single-backbone pilot, run to check that the harness worked and before any grid or any cross-backbone number was looked at, showed it wrong by 16% at $m{=}1$ and 146% at $m{=}16$; those two figures are the pilot's and are not grid results. The caption's c_{txt} is a long instruction plus a completion allowance; a claim inside a batched call is one short line and about eight completion tokens. Those are different quantities and we wrote the wrong one into the prediction. Proposition 1 never claimed they were equal. B2 is recorded as falsified with the fault in the prediction, and B2′ above (the same law with c_{claim} in place of c_{txt}, registered at 5% before the grid) is what replaced it.

B3 asked whether $p(16)$ is within a factor of two of the per-claim text cost, and it inherits the same wrong term. It was *withdrawn* in the same amendment that repaired B2, before the grid ran, and replaced by B5. We say so because a withdrawn prediction reported as a failed one would be a nicer story than the record supports: the paper did not predict amortisation and watch it fail, it predicted that amortisation would be cut short by the frame's inventory, and that is what happened.

D.9 Widening the observation: the floor rises, the price rises faster

Every measurement so far observes one frame. Theorem 1 bounds v/c below by $c_{\text{obs}}/(c_{\text{obs}} + c_{\text{txt}})$, so widening the observation to w frames should multiply c_{obs} by about w, leave c_{txt} alone, and push the floor toward one. We run the three widths $w \in \{1, 2, 4\}$ on all 7 backbones, on the same 64 probes over 8 episodes as the batching sweep, with the clip *ending* at the probe frame because a streaming agent has the past and not the future. The predictions were registered before the run.

The premise holds first: $c_{\text{obs}}(w)/c_{\text{obs}}(1)$ is 1.83–1.90 at $w{=}2$ and 3.53–3.74 at $w{=}4$, inside the registered $[0.8w, 1.2w]$ on 7 of 7 (C2). The floor therefore rises, and the measured ratio rises with it: monotone in w on 6 of 7 backbones (C1), from 0.794 at a single frame to 0.910–0.954 at four, and never further than 0.050 from the value predicted by that width's own fitted terms (C3). C3's threshold was registered at 0.05 and the worst cell lands at 0.04995: inside it, but at the boundary to three decimals, and a reader should treat C3 as attained rather than as cleared. A bound derived for one frame keeps agreeing with the measurement as the observation grows by a factor of four. The same caveat as in §4.6 applies to what "the floor" means here: this arm has no ungrounded call, so the predicted value is the two-point intercept over the bill, an over-estimate of $\theta(w)$ by the caption prompt's fixed tokens. What C3 tests is that the intercept tracks the observation as w grows, which it does.

The effective price rises much faster than the floor does. The token price of a verification call grows as the observation does: 3.65–$3.82\times$ from $w{=}1$ to $w{=}4$, which is 0.91–0.96 of w: the shortfall is the text term, which does not grow with the clip. The *effective* price v/J grows by 2.5–$13.1\times$, and on 1 of 7 backbones it is unbounded, because J falls to 0.000, a verdict about a four-frame clip that carries no information about the claim at any price. The backbones split into two regimes and the split is a property of the backbone, not of the interface: on 4 of 7, J is stable across widths and v/J grows only as the token price does, at most $3.8\times$; on the remaining 3, J collapses and the effective price leaves the floor's account behind entirely. Widening the observation is thus a worse escape than Eq. (2) alone suggests: the bound describes what the tokens cost, and what degrades is what the verdict is worth.

One arm had to be re-run, and the reason is worth stating. The clip arm's false claim was first built as a colour absent from every frame of the clip, drawn from an eight-colour palette. One frame of this corpus renders six or seven of them, so a two-frame clip nearly exhausts the palette and a four-frame clip always does: at $w{=}4$ no probe had a false claim, identically on all 7 backbones. That costs nothing for v/c, which is a price ratio (the true and false claims differ by one word, and the repaired and original ratios agree to within 0.001 at every width), but it makes J, and therefore v/J, unmeasurable at the widest clip. The arm was re-run with the negative drawn from the classes absent from the clip instead, of which the corpus has many more than eight, holding the claim fixed across widths so that the only thing differing between the three measurements is the observation.

D.10 Forcing the pairing: the yield moves, the mechanism does not

§4.5 bounds the pairing. This subsection intervenes on it, which is the stronger move and the one Proposition 3 makes a falsifiable commitment about. Two description formats are asked of every backbone at every cap on the same 64 probes, drawn from 8 episodes: the same probe set as §D.8 and §D.9, and the one every number in those three subsections comes from. FREE is the caption prompt used everywhere else in this paper, byte-identical. FORCED asks for the same content as a list of `colour object` pairs, one per line, and forbids naming anything whose colour the describer cannot give. Nothing else differs. The full grid is $7 \times 64 \times 3 \times 2 = 2{,}688$ descriptions, of which 2,685 are usable; the 3 missing ones were truncated by the runner or returned no word count, and are dropped rather than imputed.

The comparison is made at matched *realised* length, by binning descriptions on their own token count, never at matched cap. The overrun measured above is why: FORCED and FREE do not spend the same allowance, and at 7 backbones $\times$ three caps the two formats' realised lengths differ by more than a fifth in eight cells, each of which is void for a cap-matched reading and none of which affects the binned one. The binning has a cost of its own and it should be visible: a backbone contributes only the length bins in which both formats appear, which is between 1 and 4 bins, and on 2 of 7 backbones (Gemma 4 31B, Qwen3-VL-8B) it is a single bin. A one-bin comparison is matched in length but tests the format at one length only.

What was predicted, before the run. That FORCED yield exceeds FREE yield at matched length (F1); that FORCED pays for it by naming *fewer objects* (F2), because that is the only currency Proposition 3 accepts; and that FORCED sits closer to the counting bound (F4), wasting fewer tokens on things the argument cannot see. F2 is the one that matters. "Yield can be raised" is not in dispute: more words raise it. The claim is that at fixed length the *only* way to raise it is to say less.

What happened. F1 holds on 6 of 7 backbones, and on 5 of them by more than 5 points, the step a 16-probe rate could not have resolved. The format is worth -1.3 to $+27.1$ points of yield at matched length, mean $+11.6$; the sixth holding backbone gains under a point, and the seventh loses 1.3. F4 holds on 7 of 7. **F2 holds on 3 of 7.** At matched length FORCED names between -6.51 and $+1.84$ objects relative to FREE: on most backbones the same number or more, not fewer.

This is the outcome the registration named as falsifying, and we report it as such. Yield rose, and it did not come out of the object count. Whatever FORCED bought, it did not buy by spending the budget differently between the two marginals, which is the whole content of the counting argument. The honest reading is that the shortfall the format repairs was never a counting constraint at this length: it was a choice about what to say. The pairing is cheap enough to ask for and the describer

was not asking for it: a fact about instruction following, not about the token budget. Proposition 3 still bounds what is *possible*, and 0 of 2,685 descriptions violate it while 52 attain it with equality; it is simply not what explains the gap between the possible and the realised.

The account this leaves is a two-term one, and we state it that way rather than as a single mechanism. What a description can carry is bounded above by the counting argument, which is a structural constraint and is not violated anywhere in our data. What a description *does* carry at a given allowance is set inside that bound by the describer's own policy about what to say: a model-specific output behaviour that a format instruction can move without moving perception, and that §D.11 shows differs qualitatively between backbones. The measurements in §4.5 identify the first term and refute its natural rival, the typicality account; they do not identify the second, and no experiment in this paper separates an instruction-following explanation of it from a salience one. §5 names the experiment that would.

The predicate had to be replaced first, and it would have granted F1 for free. The yield predicate used for every other number in this paper substitutes away all punctuation and then splits the result on the punctuation it just removed (§C.2), so a description reads as a single clause. That is a known looseness in prose. On a one-pair-per-line format it is not loose, it is decisive: the newlines are erased too, so any colour in the list matches any object in the list. Measured, the shipped predicate credits FORCED with up to 9.9 points of yield that a clause-respecting reading does not. Every number in this subsection therefore uses a predicate that splits before it substitutes; the shipped one is computed alongside and the gap is reported per format. Had we not changed it, F1 would have been confirmed by the layout.

D.11 THREE CAPTIONS, READ CLOSELY

The category the taxonomy adds and a yield number cannot express is *unbound*: the caption named the object, named the correct colour, and did not connect them. Pooled it is small, peaking at 5.7% at cap 128. It is not spread thinly. It is concentrated almost entirely in one backbone, Gemma 3 12B, and there it is decisive.

The three captions below are the longest unbound cases in the corpus at the largest cap, selected by that rule rather than by reading them, so the choice is not ours. All three are Gemma 3 12B's. They are quoted exactly as returned, newlines flattened to spaces and the model's own Markdown left in place; the probe's two words are set in bold, and nothing else has been altered.

> *"Here's a description for later retrieval: **Frame Description:** Bedroom scene with various objects. Items include a sock, cup, brush,* ***mortar****, ruler, fork, bowl, comb, pot, spoon, hat, and soap. Colors are red,* ***blue****, orange, purple, brown, and white."*

> *"Here's a description for later retrieval: Bedroom scene with various objects. Visible are a* ***sieve****, tray, rug, glass, belt, plate, fork, towel, comb, sock, brush, cup, bowl, pot, ruler, and hat, in* ***green****, red, blue, orange, purple, and brown."*

> *"Here's a description for later retrieval: Bedroom scene. Objects include a fan, plate, rug, comb, pot, ruler, hat, glass, candle, belt, tray,* ***tongs****, sock, cup, and bowl. Colors: green, purple, red, orange, and* ***yellow****."*

Each is an object list followed by a colour list. The first spends thirty-eight whitespace words to name twelve objects and six colours (eighteen items) where binding even the twelve objects to their colours would have cost twenty-four. Every one of them contains both halves of the answer to the query it will be asked, and none of them contains the answer. This is the two detached marginals of §B.6 in its purest form, and it is a *format*: the same structure appears at every cap for this backbone, so it is not a compression decision made under pressure.

The consequence is that Gemma 3 12B's yield deficit at the largest cap is not a perception failure. It names the probe object on 81.2% of probes and names the correct colour, and scores 37.5% because it never joins them. Counting unbound captions as successes would put it at 68.7%. Only 1 other backbone shows the pattern at all, and then weakly. We draw two conclusions. First, a yield number attributed to a channel is partly a property of the description *style* the backbone happens to have, and is therefore less portable across backbones than the literature's usage implies. Second, and against

Table 9: Yield under three readings of the one judgement call in the test. STRICT splits the caption into clauses before punctuation is normalised; SHIPPED is the predicate the measurement ran and every headline number in this paper comes from; LOOSE asks only that the class and the colour both appear somewhere. The three are nested by construction, so the row is monotone and the width of the row is the width of the definition. The steep rise in the cap survives all three on every backbone, and so does the identity of the weakest; the ordering survives except that Gemma 4 31B falls from joint first to fifth under STRICT; and *saturation does not survive at all* – four backbones reach 100% at the largest cap under SHIPPED and one does under STRICT.

Backbone	cap 32			cap 64			cap 128		
	str.	ship	loose	str.	ship	loose	str.	ship	loose
Ministral 3B	37.5	**43.8**	43.8	87.5	**87.5**	87.5	87.5	**93.8**	93.8
Qwen3-VL-8B	12.5	**18.8**	18.8	87.5	**87.5**	87.5	93.8	**100.0**	100.0
Gemma 3 12B	0.0	**0.0**	0.0	0.0	**18.8**	18.8	18.8	**37.5**	68.8
Ministral 14B	40.0	**50.0**	50.0	60.0	**60.0**	60.0	60.0	**90.0**	100.0
Qwen3.8-27B	0.0	**0.0**	0.0	68.8	**81.2**	81.2	75.0	**100.0**	100.0
Qwen3-VL-30B-A3B	6.2	**12.5**	12.5	75.0	**75.0**	75.0	100.0	**100.0**	100.0
Gemma 4 31B	31.2	**31.2**	31.2	75.0	**87.5**	87.5	62.5	**100.0**	100.0

our own interest: this particular deficit looks repairable by a prompt, which makes it the weakest link in the chain from Proposition 3 to the measurement, and §5 records it as such.

D.12 How much of the yield is the definition?

Yield is a rate over a string test, and Definition 11 gives three nested readings of the one judgement in it. Reporting one and not the others would be a choice the reader cannot audit, so we report all three.

The spread is not negligible. Pooled over backbones, the three readings give $18.2/22.3/22.3\%$ at the tight cap and $71.1/88.8/94.7\%$ at the loose one, and the widest single cell (Gemma 3 12B at cap 128) runs from 18.8% to 68.8%, a range of 50.0 points on a scale of a hundred.

What survives every reading: the steep rise of yield with the allowance, on every backbone; the identity of the weakest backbone; and the ordering, with one exception: under STRICT at the largest cap Gemma 4 31B falls from joint first to fifth. What does *not* survive is saturation. 4 of 7 backbones reach 100% at the largest cap under SHIPPED and one does under STRICT, so no claim in this paper rests on the word "saturates". §4.4 needs only that yield has little room left while J has a great deal, and STRICT makes yield lower rather than higher, which strengthens that reading rather than threatening it. Also not portable: any statement about the *level* of yield for a given backbone at a given cap. We use levels in exactly one place (the comparison against the counting bound of Eq. (8)), and there the strict reading only strengthens the bound's agreement, since it lowers the measured side of an inequality the bound already satisfies.

The disclosure in §C.2 belongs with this table. The predicate that ran is not the predicate the design intended, and the gap between the two is one of the three columns here. We report the shipped column as the headline because it is what was measured, print the other two beside it, and let the reader see how much of any number is definition. A measurement whose sensitivity to its own definitions is unreported is not a measurement of the world.

D.13 Is J a property of the verifier, or of the question?

J has so far been measured under a single verifier prompt, and a single prompt is a point rather than a ceiling. We now sweep 4 prompts on 200 probes over 25 episodes of the corpus: 6.25 times the 32-probe first pass over episodes one to four, and more than ten times the 16 probes of Table 1, which drew on two episodes. The result is worse for the verifiers than a caveat about a single prompt would have anticipated. Table 10 gives J and the realised price under each.

Deliberation does not buy discrimination. The DELIBERATE prompt asks the backbone to consider each object and its colour before committing, and gives it twelve times the completion allowance to do so, which raises the realised bill by 9.4%. By Proposition 9 those are tokens from the cheap half of the bill, so if J could be bought there the effective price would fall without the floor moving at all.

Table 10: J under four verifier prompts on the same 200 probes. TERSE is the shipped one. DELIBERATE buys 9.4% more tokens (the cheap half of the bill by Proposition 9), and does not reliably buy J: it raises it on 5 backbones and lowers it on 2, but the counts are not the finding: every gain is at most 0.085 while the worst loss is -0.830, so the mean effect is -0.109. FLIPPED asks the same claim with the question's polarity reversed, which is logically equivalent and should leave a genuine verifier's J unchanged. On 4 of 7 backbones it goes *negative*: those verdicts are anti-correlated with the truth under a rephrasing, which is the signature of a verifier answering the question's form rather than checking the claim.

Backbone	TERSE		DELIBERATE		STRICT		FLIPPED	
	J	tok	J	tok	J	tok	J	tok
Ministral 3B	0.69	460	0.55	496	0.23	471	-0.73	464
Qwen3-VL-8B	0.48	334	0.57	352	0.68	346	0.30	338
Gemma 3 12B	0.39	293	0.40	311	0.26	305	-0.01	296
Ministral 14B	0.84	460	0.01	477	0.72	472	0.79	463
Qwen3.8-27B	0.94	338	0.99	425	0.94	350	-0.61	342
Qwen3-VL-30B-A3B	0.48	334	0.53	352	0.61	346	-0.46	338
Gemma 4 31B	0.98	304	1.00	348	0.98	317	0.93	308

It cannot be relied on, and the shape of the failure matters more than its direction. J rises on 5 of 7 backbones and falls on 2, so a majority improve; but every gain is at most 0.085, while the worst loss is -0.830: Ministral 14B goes from $J = 0.84$ to 0.01 when asked to think first, which is a verifier that stops discriminating altogether. The mean effect over the seven is -0.109. What the cheap half of the bill buys is therefore a bounded gain against an unbounded loss, which is not a purchase a budget-rational agent can make.

We flag one thing about how this number was read, because it changed under a larger probe set and the reading is what changed rather than the result. On the 32-probe first pass the counts were the other way round (deliberation lowered J on more backbones than it raised it), and the conclusion was stated as those counts. At 200 they reverse while the asymmetry widens. A claim resting on which side of a tally a majority falls would have had to be withdrawn here; the claim about the distribution of the effect survived and sharpened. Whatever the second half of a verifier is, it is not something the text term purchases reliably.

And J is not stable under a logically equivalent rephrasing. The FLIPPED prompt asks, of the same claim, whether it is *false*. A verifier that evaluates the claim should answer the negation and leave J unchanged; the only thing that has moved is the polarity of the question. On 4 of 7 backbones J goes *negative*. Ministral 3B runs from $J = 0.69$ to -0.73: its verdicts are anti-correlated with the truth once the question is inverted, which is what a system that answers the question's surface form rather than checking the claim must do. Only 3 backbones keep a substantial J across the flip, Gemma 4 31B best at 0.93.

This is a stronger statement than "J is low". A quantity that changes sign under a rewriting that preserves truth conditions is not measuring the verifier's access to the observation. It follows that the v/J figures of §D.4 are, for those backbones, an *overestimate* of what a verdict is worth: the effective price is worse than reported, and Proposition 8's $1/J^2$ correction is applied to a J that is itself prompt-dependent. We report the shipped prompt's J as the headline because it is the one every other number in the paper was measured against, and this table as the reason not to trust its level.

D.14 THE TWO ESCAPES WHOSE VALUE WAS OPEN

Table 6 priced the escapes from the verification floor and left two of them open: whether an ungrounded check is worth anything, and whether batching delivers the $\Theta(1/m)$ that Proposition 1 predicts. Both are measured here, on the 200-probe run, and Table 11 reports them. This is the smaller of the two batching sweeps in the paper ($m \leq 4$ on 200 probes, against $m \leq 16$ on 64 probes in §D.8), and it is the one whose ungrounded arm makes the observation term identifiable by subtraction (§4.1).

The ungrounded check is free and empty. Asking the same yes/no question with no frame attached costs 8.6–16.0% of the grounded price, which is what Theorem 1 says it must: it never pays

Table 11: The two escapes of Table 6 whose value was open, now measured. UNGROUNDED asks the same claim with no frame attached: it costs 8.6–16.0% of the grounded price and its J is exactly zero on 6 of 7 backbones: a verifier with a finite price and no information, which is Proposition 2 realised rather than hypothesised. BATCHED verifies m claims about one frame in one call; the amortised price falls by 3.54–3.76× from m=1 to m=4, and the slope identifies c_{obs} a second time, from different calls with a different prompt than the caption sweep of Table 2. The two estimates agree to 2.2% on average over 7 backbones.

Backbone	ungrounded check			batched check			
	tok	J	vs. grounded	v_1	$v_4/4$	c_{txt}	c_{obs}
Ministral 3B	40	**0.000**	8.6%	471	127.2	12.7	458
Qwen3-VL-8B	44	**0.015**	13.1%	346	96.0	12.7	333
Gemma 3 12B	46	**0.000**	15.6%	305	86.2	13.3	292
Ministral 14B	40	**0.000**	8.6%	479	127.2	10.0	469
Qwen3.8-27B	48	**0.000**	14.1%	349	98.0	14.3	335
Qwen3-VL-30B-A3B	44	**0.000**	13.0%	346	96.0	12.7	333
Gemma 4 31B	49	**0.000**	16.0%	317	89.0	13.0	304

c_{obs}. Its verdict information is $J = 0$ on 6 of 7 backbones: not small, not noisy, exactly zero. Every backbone answers at a fixed rate independent of whether the claim is true, because it has no access to the fact that would settle it.

This is Proposition 2 as an observation rather than a construction. We wrote that proposition about a hypothetical verifier answering at a fixed rate, to establish that a finite v and an infinite v_{eff} can coexist. The hypothetical is the cheapest verifier available on this interface, and it is what an agent gets if it tries to escape the floor by not looking. The row of Table 6 that read "answers a different question" can now be priced: the different question has no answer worth having, and the escape that appears to cost nothing costs everything it was supposed to save.

Batching delivers what the proposition predicts, and identifies c_{obs} twice. Verifying m claims about one frame in one call drops the amortised price by 3.54–3.76× between $m = 1$ and $m = 4$, against the $4\times$ that a pure $\Theta(1/m)$ would give: the shortfall is the per-claim text term $c_{\text{txt}} \in [10.0, 14.3]$ tokens, which is the floor Proposition 1 says the amortisation flattens at.

The more useful consequence is methodological. Fitting $v_m = c_{\text{obs}} + m\, c_{\text{txt}}$ identifies the observation term from an entirely different set of calls than the caption sweep of §4.1: different prompt, different completion length, different task. The two estimates agree to 2.2% on average and 3.5% at worst across 7 backbones (292–469 tokens from the batch slope of this run against 295–472 from the cap sweep of the 16-probe pilot). The larger sweep of §D.8 does the same comparison over $m \leq 16$ on a different probe set and gets 3.6%: two runs, two agreement figures, both reported rather than the smaller one alone. Theorem 1's premise is an empirical claim about the interface, and it has two independent measurements that agree rather than one fit that could be an artefact of its own regression.

One caveat on what these intercepts are, since §4.1 makes a point of it: a batch intercept carries the batching prompt's fixed tokens exactly as a caption intercept carries the caption instruction's, so neither is c_{obs} on its own. What the agreement establishes is that both call types recover the same cap-invariant quantity to within a few per cent, which is the decomposition; what fixes the level is the grounded-minus-ungrounded subtraction in the first column of Table 11.

D.15 Is the false claim too easy? Measured, and mixed

Rejecting the shipped false claim ought to require binding rather than colour detection, because the named colour is usually present in the frame on some other object. That is an argument from the corpus's construction, and it is the kind of argument a limitations section makes when it cannot measure. With the per-frame inventory recovered exactly (§C.1) it can be a measurement instead: we ask the same verifier about a colour absent from the frame entirely and about one present on a different object, on the same probes. Table 12 is the result.

The gap runs the predicted way (easier to reject a colour that is not there) on 5 backbones, and the other way on 2, with a mean of 10.4 points. So the argument was not wrong, and it was not the whole

Table 12: How hard the false claim is, split by whether the named colour is in the frame at all. EASY names a colour absent from the scene, so rejecting it needs only colour detection; HARD names one carried by some other object, so rejecting it needs the verifier to bind colour to object. The paper's limitation (6) argued on difficulty grounds that the shipped false claim was of the hard kind; this measures it instead, and the answer is mixed: the gap runs the predicted way on 5 backbones and the other way on 2, mean 10.4 points. The denominators differ because a colour absent from the frame does not always exist: these frames render sixteen objects over an eight-colour alphabet, so the EASY arm is defined on 132 of 200 probes and the excluded ones are exactly the colour-saturated frames, which is not a random exclusion. The last two columns put the two arms back together at the weights a uniformly drawn false colour would give them (§D.15): J_{rand} is what the shipped J would have been had the false colour been sampled rather than fixed, and it differs from it by at most 0.06 on any backbone.

Backbone	n_{EASY}	reject (%)	n_{HARD}	reject (%)	gap	J	J_{rand}
Ministral 3B	132	97.0	200	94.5	2.5	0.69	0.67
Qwen3-VL-8B	132	78.0	200	43.5	34.5	0.48	0.48
Gemma 3 12B	132	90.9	200	68.0	22.9	0.39	0.35
Ministral 14B	132	96.2	200	91.5	4.7	0.84	0.78
Qwen3.8-27B	132	90.9	200	96.5	-5.6	0.94	0.95
Qwen3-VL-30B-A3B	132	75.8	200	61.0	14.8	0.48	0.52
Gemma 4 31B	132	99.2	200	100.0	-0.8	0.98	0.98

story either. On three backbones the binding case is genuinely harder, by 34.5 points at the widest. On the other four the split does almost nothing: two gaps are small and positive, and the two that run the other way are -5.6 and -0.8 points, which on 200 probes is a handful of items and not a reversal worth an account. We report the whole distribution rather than fitting a story to seven points.

One caution on reading the first five columns. The EASY arm exists only where some colour is absent from the frame, and these frames render sixteen objects over an eight-colour alphabet, so it is defined on 132 of 200 probes; and the excluded probes are exactly the colour-saturated ones, which is a selected rather than a random exclusion.

What the shipped false claim would have been worth had it been sampled. The deeper objection to the shipped construction is not difficulty but *determinism*: the false colour is the first alphabet entry differing from the true one, so it is the same colour on almost every probe, and an earlier form of this paper conceded that no comparison of these J values against a randomised false claim was available. It is available, and it needs no further calls, because the two arms above are already random draws *within their strata*. EASY samples uniformly from the colours absent from the frame and HARD uniformly from those present on another object, and between them those two sets partition the seven colours a uniform draw could name. So for a probe whose frame shows k of the 8 colours, a uniformly drawn false colour is absent with probability $p = (8-k)/7$ and present-elsewhere with probability $1-p$, and

$$\widehat{\mathrm{FPR}}_{\mathrm{rand}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big[\, p_i\,\mathbf{1}\{\text{ACCEPT} \mid \text{EASY}_i\} \;+\; (1-p_i)\,\mathbf{1}\{\text{ACCEPT} \mid \text{HARD}_i\}\,\Big] \tag{12}$$

is unbiased for the false-accept rate the randomised construction would have produced. The partition is exact rather than approximate: the true colour is always present, so the seven remaining colours split into the $8-k$ absent ones and the $k-1$ present on something else, and those are precisely the two pools the arms sample from. The one apparent gap closes for the same reason. The EASY arm is undefined on the colour-saturated frames (200 probes carry it, 132 the other), and on exactly those frames $k=8$, so $p_i = 0$ and the missing arm carries no weight. Verified on the data rather than argued: every probe lacking an EASY draw has all eight colours rendered, on every backbone. The last two columns of Table 12 report $J_{\mathrm{rand}} = \mathrm{TPR} - \widehat{\mathrm{FPR}}_{\mathrm{rand}}$ beside the shipped J.

They differ by at most 0.06 and by 0.026 on average, in the pessimistic direction on 5 of 7 backbones, and 20 of the 21 pairwise orderings between backbones survive the substitution. The one that does not is a pair whose deterministic J values differ by 0.005, which is a swap between two backbones the deterministic measurement had already failed to separate. The range moves from 0.39–0.98 to 0.35–0.98: the bottom falls by 0.04, so the randomised construction is slightly *harder* at the weak end rather than easier, which is the direction that matters: had it gone the other way, the shipped J would have been the flattering one. The determinism was therefore not inflating J, and the reason

is visible in the same table: only 13% of a uniform draw's mass falls on the EASY stratum (a frame rendering sixteen objects over eight colours rarely omits one), so a randomised false claim is, in this corpus, mostly a hard one, which is what the deterministic choice was already producing. This closes the comparison that limitation (6) declared invalid, and it closes it in the direction that leaves every J in the paper standing. What it does not close is the corpus: the argument above turns on the colour alphabet being small relative to the object count, and a corpus with a wider alphabet would put more mass on the easy stratum and raise J_{rand} above the shipped J rather than below it.

The expanded probe set reproduces the published one, and says how much of a spread two episodes hide. All of the above runs on 200 probes over 25 episodes, against the 16 from two that Table 1 reports. Two different comparisons are available here, and they are easy to conflate because at four episodes they nearly coincide.

The *reproducibility* check is the shared probes across the two runs: the same 16 probes of episodes one and two, re-executed a day later against the same endpoints, so what it measures is the interface's own nondeterminism and not provider drift. Yield agrees to 3.6 points on average and 12.5 at worst over 21 model–cap comparisons, with the ordering of backbones unchanged and yield at the largest cap still spanning 21.0–99.0%. The backbone that lost six probes to provider errors in the first run returned all 200 in this one, so every row here has the full denominator.

The *episode spread* is a different quantity: yield on episodes one and two against yield over all 25, which differs by 5.7 points on average and 22.8 at worst. That is not a reproducibility failure: it is how much episode-to-episode variance two episodes cannot show, and it is the reason the probe count was raised. Read as a reproducibility number it would have been badly misleading, and it grew from 9.4 to 22.8 on expansion for precisely that reason: the two shared episodes went from half the corpus to a small fraction of it.

D.16 READING THE THREE TOGETHER

The prices point one way and they point consistently. Yield is cheap to buy and the measurements say a generous caption is a good caption. Checking is expensive, because it pays the same fixed cost the channel does. And on most backbones at this scale the verdict obtained for that price carries little information about the claim.

A budget-rational agent facing these three numbers uses the channel and does not check it. That is not negligence, it is the arithmetic, and it is the configuration in which the selection loss of the companion analysis (Anonymous authors, 2026a) is incurred. The remedy that analysis proposes (convert blind arbitration into aware arbitration by establishing which claim to trust) is priced here for the first time, and at this scale it is not affordable. Whether it becomes affordable is a question about interfaces rather than about models, and §5 says which interface change would move it.

E LIMITATIONS, FUTURE WORK AND CONCLUSION, IN FULL

E.1 LIMITATIONS

(0) One channel, one interface, and now two corpora and three observation widths. The channel is a caption, the interface prices every call as text plus observation. Two of the three limits in the original form of this limitation have since been measured rather than conceded: §4.6 repeats the floor on photographs with a claim type the corpus can verify exhaustively, and §D.9 repeats it on clips of up to four frames. What did not survive that widening is the *reach* of the conclusion rather than the floor: on photographs J is high enough that grounded checking is affordable at the very prices the floor predicts, so the admission criterion is violated by construction only where claims are hard to discriminate. The single remaining point was the interface, and §4.2 has since widened that too, from one routed arm to 19 pinned endpoints across 7 backbones and from one currency to two. Theorem 1 names the interface property that produces the result, so the honest scope is: on interfaces that price observations per call, for channels whose subject is an observation. That covers the hosted multimodal APIs used here, and §4.2 says which of their endpoints price that way and which do not. It does not cover a self-hosted deployment, where the marginal cost of re-encoding an image is compute rather than a billed token and the ratio would have to be re-derived in whatever

currency that deployment actually pays; and that remains the widest untested case, because it is the one where the cost model of Definition 1 is not what anyone is charged.

(1) Observation caching was the first thing to check; it has been checked, and what remains is a boundary rather than an unknown. If an interface admits a frame once and prices subsequent references to it below $\alpha\tau(o)$, the floor of Eq. (2) falls by exactly the cache discount (Proposition 1). That is Assumption 1, it is a statement about a product feature rather than about computation, and an earlier form of this paper could only report that no such feature was available on our arm.

§4.2 replaces that report with a measurement over 19 pinned endpoints, and the result has three parts, each of which bounds this limitation differently. On 5 of 7 backbones no endpoint returns a cached observation to a call that asks a different question about the frame, under either message layout, so on those the assumption is verified for every call a verifier makes rather than assumed; and verified at a rate rather than once, since the 16-probe price arm issues 480 further verdict calls across those cells, each of them a different question about a frame the caption call has already sent, and 0 of them are served from a cache. On 2 it is false: the observation caches at the advertised $\alpha'/\alpha = 0.1$, and the bill for a warm call falls to 0.25 of a cold one on all 3 endpoints where the two calls differ in nothing but their cache state, within a percentage point of the value the proposition predicts from the published price and the cached token count. And the entry lives minutes: every frame is still cached at 15 s, none at 240 s.

The residual limitation is therefore precise, and it has two halves.

The first is that a cache lifetime is not a guarantee. Ours is measured on one provider, at one time, on a cache shared with other tenants: 78–97% of the verdict calls in the price arm's caching cells hit, so even inside the window the discount is a rate and not a contract. A provider offering a long-lived or explicitly managed image cache would invert the design advice of §D.8 on the backbones it served. We should be plain that this is not hypothetical: the major closed-model providers document prompt caching for their own models with cache reads billed at a fraction of the base input rate, at least one documents caching multimodal inputs, and the managed variants come with a stated time-to-live rather than a best-effort one. What we could not obtain was that feature on the serving arm this paper measures (open-weight backbones behind a router), and what §4.2 adds is that on that arm the feature's presence is now a measured fact per endpoint rather than an assumption over the whole of it.

The second half is a limit on how much the first would matter, and it is worth stating because it is not obvious. The cache window is roughly a minute. A check issued inside it is a check issued while the agent still has the frame in hand, and an agent in that position did not need the description: it could have answered the query from the observation directly, at one call rather than two. The regime in which the escape is available is therefore close to the regime in which the channel this paper prices has no reason to exist. That is an argument and not a measurement (we do not measure how long real agents defer), but it says why a short cache and a long one are different products rather than the same product at different sizes.

(2) The verifier prompt was swept, and the sweep makes J less portable rather than more. A single yes/no question is the cheapest grounded check and therefore the right one for a *floor*, but a better-designed check might raise J at the same price. That would not affect Theorem 1, which is about price, and it would affect every v/J figure directly. §D.13 runs 4 prompts on 200 probes and finds no prompt that reliably buys J: deliberation's gains are bounded at 0.085 while its worst loss is -0.830, and a logically equivalent rephrasing sends J negative on 4 of 7 backbones. So the limitation is not that J was measured once; it is that J is prompt-dependent to a degree that makes any single value, including the terse one we headline, a point in a wide space. Every v/J in this paper should be read as an *overestimate* of what a verdict is worth. What we have not established is whether some prompt outside the four raises J uniformly; §4.4's decoupling claim would survive a uniform improvement and not necessarily a differential one.

(3) Probe count: the pilot's 16 probes are not enough, and the replicate says by how much. Probes are drawn from the corpus's singletons because those are the items with exact ground truth. A rate over 16 probes moves in steps of several points, so the pilot's per-backbone yields should be read as ordinal and its J values as coarse. This is not a hypothetical concern: the $J = 0$ episode recorded in §D.4 is what an under-powered rate does to a headline, and it is the clearest argument we have against writing prose over an incomplete run.

Table 7 re-runs the whole grid at 200 probes over 25 episodes, and every headline verifier number in the paper is taken from it rather than from Table 1. That was the cheapest improvement left and it is done; what it shows is that the pilot was optimistic in two specific ways, not merely noisy. Yield moves by up to 23.2 points, J by up to 0.245, the count of backbones below $J = 1/2$ rises from 2 to 3, and the 2 apparently perfect verifiers stop being perfect. On the 16 probes the two runs share, yield agrees to 3.6 points on average and 12.5 at worst over 21 model–cap comparisons, so the difference is the larger sample rather than an unstable interface. Two further readings did not survive the expansion and are reported where they occur: the direction of the DELIBERATE counts reversed while the asymmetry behind the conclusion widened, and a spread we had read as reproducibility turned out to be episode-to-episode variance. What remains is that the pilot's tables are still printed, because the per-pair taxonomy of Table 4 and the caption taxonomy have no 200-probe counterpart.

(4) The yield predicate is not the predicate the design intended. Stated in full in §C.2: the implementation normalises punctuation before splitting into clauses, so the split that runs is on the word *and* alone. Every headline number comes from the predicate that ran; §D.12 reports two nested alternatives on the same stored text, and they span up to 50.0 points on the widest cell. Levels of yield are therefore not portable out of this paper. Orderings and the saturation pattern are, and those are what §4.4 uses.

(5) One backbone's yield deficit looks repairable by prompting. Gemma 3 12B writes an object list and a colour list as a matter of format and never binds them, which costs it 31.2 points of yield at the largest cap (§D.11). If a prompt change repaired it, that backbone's yield would move without anything about its perception changing. This is the weakest link between Proposition 3 and the measurement, it argues against treating yield as a stable property of a channel, and we did not test the repair.

(6) The false claim's difficulty is measured, and the answer is mixed. The swapped colour is drawn from the corpus alphabet and may be absent from the frame altogether, which would make rejection easier and inflate J. §D.15 measures the split directly rather than arguing it from the corpus's construction: rejecting an absent colour is easier than rejecting a misbound one on 5 backbones, by up to 34.5 points, and marginally harder on 2, mean 10.4 points. So the effect is real on some backbones and absent on others, and the EASY arm is itself a selected subset: it exists only where some colour is absent from the frame, which excludes exactly the colour-saturated frames. Appendix G reports how often the swapped colour is nonetheless present somewhere in the frame.

The determinism was the residual concern (the swapped colour is the first alphabet entry differing from the true one), and the comparison this paper previously called invalid is now made: because the two arms are uniform draws within complementary strata, they recombine at the weights a uniform draw would give them into an unbiased estimate of J under a randomised false claim (Eq. (12)). It differs from the shipped J by at most 0.06 and by 0.026 on average, and 20 of the 21 pairwise orderings between backbones are preserved, the exception being a pair the deterministic measurement separated by 0.005, which is to say not at all. No conclusion in this paper turns on the determinism. What that argument does turn on is the corpus: only 13% of a uniform draw's mass falls on an absent colour here, because sixteen objects over eight colours rarely leave one out, and a corpus with a wider attribute alphabet would put more weight on the easy stratum and raise J_{rand} above the shipped J rather than leave it beside it.

(7) We do not show that any system reasons about these prices, and that is a boundary rather than an omission. The reading in §4 (that a budget-rational agent would use the channel unverified) is a statement about what the arithmetic favours, not a claim about a mechanism inside a model. It is worth being precise about the two different things one might want here, because they have different prices and only one of them is this paper's.

The first is a *normative* claim: given the measured v, $c(b)$ and μ, an agent maximising accuracy under a budget should not buy verification. That claim needs no further experiment: it is Corollary 1 applied to numbers this paper measures, and its content is entirely in whether those numbers are right, which is what §4.2 and §4.1 are for. The second is a *descriptive* claim: that some particular deployed agent, given a budget, in fact allocates it that way. That is a statement about an agent's policy and not about an interface's prices, its answer would vary with the agent, the task and the planner, and measuring it would tell us about that agent rather than about the quantity this paper is trying to establish. We therefore do not make it, and a reader should not read our prices as evidence for it.

The experiment that would test the descriptive version (a budgeted agent given a blind and a reliability-aware policy over the same channel, scored on accuracy, spend and verification rate) is the reliability-marked-presentation intervention that the companion analysis (Anonymous authors, 2026a) lists first, on the population where an arbiter's behaviour is the object of study. Neither paper has run it, and it belongs in the one whose subject is the arbiter.

(8) Five escapes are settled; one is not. Batching is measured in §D.8 and §D.14, the ungrounded check in §D.14, sampling is a derivation (Proposition 1) that needs no run because it leaves the decision invariant, observation caching is measured in §4.2 (it reaches a verifier's call on 2 of 7 backbones at the predicted discount and with a lifetime of minutes), and the message-layout escape that would make the observation a shared prefix moves 0 of the 18 endpoint cells where it could have helped, which is to say it is empty. One remains. Re-sending the frame downsampled (§B.3) needs no interface feature at all and we did not run it; it is the cheapest thing a practitioner could try, its size is set by the provider's resolution schedule and by how much resolution the claim can spare, and this paper gives neither number. It is now the only unrun escape, and it is the one we would run next.

(9) The verdict's value may belong to the served configuration, not to the model, and 16 *probes cannot tell.* §4.2 pins four endpoints of one model and finds J spanning 0.37 between them while v/c spans 0.009. Those endpoints differ in quantisation; the probe count is a sixteenth of the one every J in the main tables is measured at, so the spread is within what that sample could produce by chance and we draw no conclusion from it. It matters because of what it would mean if it survived: every J this paper reports is measured through a router that chooses the endpoint, so a J attributed to a backbone would in part be a property of whoever served it that day, and the decoupling of §4.4 would need restating: price to the interface, value to the served configuration. Re-running the 200-probe verifier grid pinned to each endpoint would settle it, and is the measurement we would add next after the downsampling escape.

E.2 Future work

(i) was run and is reported in §D.13: sweeping 4 verifier prompts finds no phrasing that reliably buys J, and finds one that is logically equivalent to the shipped question and sends J negative on 4 of 7 backbones. What remains open is whether a prompt outside the four raises J *uniformly*, which is the only kind of improvement that would leave §4.4's reading intact. *(ii)* was run and is reported in §D.14: an ungrounded check costs 8.6–16.0% of the grounded price and its verdict information is exactly zero on 6 of 7 backbones. The escape whose value was unknown is now bounded, and the bound is that it has none. What remains open is whether a consistency check against *other stored claims*, rather than against nothing, carries anything: a different experiment with a different population. *(iii)* was run and is reported in §4.2: the interface was audited for observation caching rather than assumed not to have it. What remains open is not whether the escape exists (it does, on 2 of 7 backbones, at the predicted discount) but whether any provider offers an image cache with a lifetime long enough to reach a *deferred* check, which is the only version of the escape the architecture could use. A managed cache with an explicit time-to-live, of the kind the major closed providers document for text, would settle it, and would invert the design advice of §D.8 wherever it reached. Re-sending the frame downsampled is the cheaper cousin of the same experiment, needs no interface feature, and is not run here; it is now the one unrun escape of limitation (8). *(iv)* was run and is reported in §D.8: the $\Theta(1/m)$ prediction holds, the knee is at $m^* = 21$–36, and a frame supplies only 16 claims, so the escape stops before it saturates. What remains open is whether an agent with genuinely many claims per observation (a dense scene-graph consumer rather than a query-driven retriever) reaches the knee, which is a question about the agent and no longer about the interface. *(v)* was run and is reported in §D.9: c_{obs} scales with the clip width, the floor rises, and the measured ratio tracks it to 0.050. The finding that was not predicted is that the *effective* price rises far faster than the floor, because J degrades with width on 3 of 7 backbones. Whether that degradation is a property of long-context attention or of these particular verifier prompts is the obvious next question, and §D.13's method answers it directly. What is *not* open in the way a reader might expect is the width ladder itself: the interesting asymptote is $v/c \to 1$ with $J \to 0$, and reaching it would need widths well past four, which this model class cannot buy: the parent system's registered $w = 16$ is unaffordable at these per-frame token costs for every backbone in the grid (Anonymous authors, 2026c). Extending the ladder is therefore a question for a budget or a model class this paper does not have, not a measurement we declined to make. *(vi)* was run and is reported in §D.10: forcing the pairing raises yield at matched realised length on 6 of 7 backbones and does *not* cost objects,

which is the outcome that separates a counting constraint from a choice about what to say. What remains open is which of the two it is: an instruction-following account predicts that the gain should survive any format that names the pairing explicitly, and a salience account predicts that it should not survive a format that forces pairs in an order the describer did not choose. Until that is run, the account of §4.5 is the two-term one stated there: a structural upper bound that our data never violate, and a model-specific output policy inside it that we do not identify. *(vii)* A family of claim types on natural images rather than more photographs of one type. §4.6 shows the price porting and the value not porting on a single claim form, class presence, and the reading it forces (that affordability is a property of claim discriminability) predicts a specific two-dimensional picture: v/c roughly flat across claim types, J varying widely along them. Presence, attribute, spatial relation, count and text reading would test it on one corpus with one interface, and it is the measurement that would most enlarge what the third level of §4.2's reading can say. We do not run it here because the paper's object is the price and the price is already measured on two corpora and three widths; enlarging the claim taxonomy enlarges the value half, which is the half this paper reports as variable rather than as a result. *(viii)* Not run, and deliberately out of scope: a budgeted agent given a blind and a reliability-aware policy over the same channel, to show that the allocation the prices favour is the allocation an agent reaches. The reasons are in limitation (7). The short form is that the answer would be a fact about that agent's planner rather than about the interface's prices, and this paper's population is calls, not agents.

E.3 Conclusion

A perceptual channel is admitted to save budget. Checking what it says costs what it cost to look in the first place, because both calls pay to transmit the observation, and on the interfaces in use the observation is 69–77% of the bill, identified inside one run as a grounded verdict call minus an ungrounded one, which is the identification that survives scrutiny; the two-point fit we first used over-estimates it by a caption prompt and its apparent agreement with the measured ratio was largely that coincidence. Measured across 7 backbones the ratio runs 0.791 to 0.876 and clears the floor those terms imply, 0.69–0.77, on every one. Verification is not a small premium on a channel; it very nearly doubles the channel's price, and the admission criterion of the companion analysis (Anonymous authors, 2026a) is violated generically rather than at the margin.

Three further measurements make the situation worse rather than better. Yield is bought with the cheap half of the bill, so the channel an agent gets is good and a better one is nearly free; it reaches 95% or more on 5 of 7 backbones, while J still spans 0.39 to 0.98. The verdict bought with the expensive half carries less than its price: J below one half on 3 of 7 backbones, an effective price of 309 to 751 tokens, and an honest per-bit price up to $3.39\times$ higher than that. And the obvious ways to escape the floor are now priced rather than listed. Observation caching, the escape that would change the sign of the advice, reaches a verifier's call on 2 of 7 backbones at the discount Proposition 1 predicts; and its entry expires within minutes, which is shorter than the deferral a stored description exists to create, so the calls an architecture actually defers are priced cold. Batching is bounded by a property of the task. Verifying only sometimes does not move the admission decision at all. Reordering the message so the observation leads, the one change a caller can make for free, moves 0 of the 18 cells where it could have. What is left unpriced is a downsampled re-send, which needs no interface feature and which we did not run.

Two objections to the measurement rather than to the theory are also settled there. With the provider pinned, the token price of an observation is a property of the model and the interface's tokenisation rather than of whoever served the call, while the dollar price of the identical call varies by up to $2.1\times$ between vendors of one model. And the floor holds in the currency the interface bills as well as in the one the theorem is written in, with a different constant: θ is 0.683–0.755 in tokens and 0.321–0.566 in money, because a caption spends its allowance on completions and completions cost more than prompt tokens. In money a check costs about half of what looking costs rather than four-fifths of it: a smaller number, the same structure, and we report the one the vendor charges beside the one the theorem uses.

Reading the descriptions rather than only their scores tightened the account and cost us the mechanism we expected to find. The obvious story, that a tight allowance drops the rare object first, is refuted: the probe is named as often as the fifteen common objects rendered beside it, on 4 of 7 backbones more often. What a tight allowance destroys is the *pairing*, and it destroys it by dropping the object

while keeping the attribute: the alphabet of colours saturates in a handful of tokens, the objects do not, and binding is charged for both. A caption in that state names everything and settles nothing.

The design conclusion is not that verification should be made cheaper. It is that on a token-metered interface the price of belief is structurally tied to the price of perception, and prompting cannot separate them, because every prompting change moves the term that is not binding. Whether they can be separated at all is a question about interfaces rather than about models, and Equation (2) names the single term that would have to change.

F Full Proofs

Standing conventions

Throughout, a call is priced by Definition 1, $c(o, x, y) = \alpha\tau(o) + \beta_{\text{in}}|x| + \beta_{\text{out}}|y|$, with $\alpha, \beta_{\text{in}}, \beta_{\text{out}} > 0$ and $\tau(\varnothing) = 0$. We abbreviate $c_{\text{obs}} = \alpha\tau(o)$ and $c_{\text{txt}} = \beta_{\text{in}}|x| + \beta_{\text{out}}|y|$, both non-negative, with $c_{\text{obs}} > 0$ whenever the call carries an observation. All statements are about a fixed observation o and a fixed query family, and b denotes the caption allowance. Assumption 1 is in force except where Proposition 1 explicitly suspends it. No probabilistic statement below requires independence across probes; where an expectation is taken, it is over the randomness of a single draw.

Theorem 1 (verification floor)

Let V be a grounded verifier, so its verdict is a function of o. By Assumption 1 the call producing that verdict transmits o, hence its price satisfies

$$v \;=\; \alpha\tau(o) + \beta_{\text{in}}|x_V| + \beta_{\text{out}}|y_V| \;\geq\; \alpha\tau(o) \;=\; c_{\text{obs}},$$

the inequality being strict unless the verifier's prompt and verdict are both empty. The channel's own call transmits o and additionally pays for the instruction and the description, so exactly

$$c(b) \;=\; c_{\text{obs}} + c_{\text{txt}}(b).$$

Dividing the two displays gives $v/c(b) \geq c_{\text{obs}}/(c_{\text{obs}} + c_{\text{txt}}(b)) = \theta(b)$, which is Eq. (2). For the limit, write $\theta(b) = \big(1 + c_{\text{txt}}(b)/c_{\text{obs}}\big)^{-1}$; as $c_{\text{obs}}/c_{\text{txt}}(b) \to \infty$ the denominator tends to 1 from above, so $\theta(b) \uparrow 1$. Note the direction: θ is *decreasing* in $c_{\text{txt}}(b)$, so any reduction of the verifier's text (a terser prompt, a one-word verdict) raises the bound. ■

Corollary 1 (generic violation of the admission criterion)

Write $\mu = \partial\,\text{acc}/\partial b \geq 0$ for the marginal accuracy per token in the budget's best alternative use, and let channel k have verification price v_k and yield $y_k \in [0, 1]$ *expressed in accuracy units*, as the companion analysis (Anonymous authors, 2026a) defines it: y_k is the accuracy the decision procedure gains from having k's claim available. That unit matters and is easy to lose. μ is accuracy per token and $c(b)$ is tokens, so $\mu c(b)$ is an accuracy increment; comparing it against a bare probability would be dimensionally undefined.

The criterion is a comparison of a total spend against a total gain, and it must be stated once rather than assembled from two separate inequalities. Using channel k *and* believing it costs $c(b) + v_k$ tokens, whose best alternative use is worth $\mu\,(c(b) + v_k)$ accuracy. Hence k is worth admitting with verification only if

$$y_k \;>\; \mu\,\big(c(b) + v_k\big).$$

By Theorem 1, $v_k \geq \theta(b)\,c(b)$, so the right-hand side is at least $\mu\,c(b)\,(1 + \theta(b))$ and

$$y_k \;>\; \mu\,c(b)\,\big(1 + \theta(b)\big)$$

is necessary. Since $y_k \leq 1$ always, this is unsatisfiable as soon as

$$\mu \;>\; \frac{1}{(1 + \theta(b))\,c(b)} \;\xrightarrow[\theta(b)\to 1]{}\; \frac{1}{2\,c(b)},$$

a condition involving only the marginal value of budget elsewhere and the channel's price. No property of the channel other than its price appears, which is the sense in which the violation is generic rather than marginal.

The step this proof no longer takes. An earlier version wrote the two requirements separately ($y_k > \mu v_k$ for the verification and $y_k > \mu c(b)$ for the channel) and then added them. A conjunction of $y > a$ and $y > b$ is $y > \max(a, b)$, not $y > a + b$, so that route does not yield the factor $(1 + \theta)$; it yields $y_k > \mu \max(c(b), v_k)$, and since $\theta(b) \leq 1$ gives $v_k \geq \theta(b)c(b)$ without giving $v_k \geq c(b)$, the maximum is not even determined. The factor of two is real, but it comes from the total-spend reading above, which is the reading the companion's criterion has: one budget, spent once. ■

PROPOSITION 1 (CACHING)

Suspend Assumption 1 and suppose a call may condition on a previously transmitted o at observation price $\alpha'\tau(o)$ with $0 \leq \alpha' < \alpha$. The channel's call is the first reference to o and still pays $\alpha\tau(o)$; the verifier's call is a subsequent reference and pays $\alpha'\tau(o)$. Hence

$$\frac{v}{c(b)} \;\geq\; \frac{\alpha'\tau(o)}{\alpha\tau(o) + c_{\mathrm{txt}}(b)} \;=\; \frac{\alpha'}{\alpha} \cdot \frac{\alpha\tau(o)}{\alpha\tau(o) + c_{\mathrm{txt}}(b)} \;=\; \frac{\alpha'}{\alpha}\,\theta(b),$$

so the floor is scaled by the cache discount and tends to 0 as $\alpha' \to 0$. The bound is attained by a verifier whose prompt and verdict are of negligible length. ■

PROPOSITION 1 (BATCHING)

Let m claims $\phi_1, \ldots, \phi_m$ about the same o be verified in one call, whose input carries o once and m claims and whose output carries m verdicts. Its price is $v_m = c_{\mathrm{obs}} + m\, c_{\mathrm{txt},v}$, where $c_{\mathrm{txt},v}$ is the per-claim text cost, so the amortised price per verdict is

$$\bar{v}(m) \;=\; \frac{v_m}{m} \;=\; \frac{c_{\mathrm{obs}}}{m} + c_{\mathrm{txt},v} \;=\; \Theta(1/m) \quad \text{until} \quad m \gtrsim c_{\mathrm{obs}}/c_{\mathrm{txt},v},$$

after which it flattens at $c_{\mathrm{txt},v}$. For the aggregate claim, consider a stream of n observations with μ claims each. Claims about distinct observations cannot share an observation term, so any batching schedule pays at least $n\, c_{\mathrm{obs}}$ in observation terms, and the total is bounded below by $n(c_{\mathrm{obs}} + \mu\, c_{\mathrm{txt},v})$, attained by batching all μ claims of each observation together. Batching therefore converts a per-claim floor into a per-observation floor and no further. ■

PROPOSITION 1 (BERNOULLI SAMPLING IS A NO-OP)

Let each claim be verified independently with probability $p \in [0, 1]$. Conditional on verification the arbiter is informed and attains $r^\star = \max_k r_k$; conditional on no verification it is blind and attains $r^\circ = \sum_k \lambda_k r_k + \lambda_0 r_0$ (Anonymous authors, 2026a). By the law of total expectation the accuracy is

$$\mathrm{acc}(p) \;=\; p\, r^\star + (1 - p)\, r^\circ \;=\; r^\circ + p\,(r^\star - r^\circ),$$

and the expected verification bill per claim is $p\, v$. The admission criterion compares the accuracy *gained* by verification against what its tokens would buy elsewhere:

$$\underbrace{p\,(r^\star - r^\circ)}_{\text{gain}} \;>\; \underbrace{\mu \cdot p\, v}_{\text{opportunity cost}} \quad\Longleftrightarrow\quad (r^\star - r^\circ) \;>\; \mu\, v,$$

where $p > 0$ cancels from both sides. For $p = 0$ both sides are zero and the comparison is vacuous. Hence the decision of whether to verify at all is independent of p, and p affects only the variance of the realised accuracy and bill, not their ratio. ■

PROPOSITION 2 (THE EFFECTIVE PRICE DIVERGES)

Let the verifier return ACCEPT with probability q independent of the claim. Then $\mathrm{TPR} = \mathrm{FPR} = q$ and $J = 0$. For any prior π on the claim's truth, the likelihood ratio of either verdict is $q/q = 1$ (or $(1 - q)/(1 - q) = 1$), so the posterior equals the prior; by induction the same holds for any finite number of conditionally independent repetitions, whose joint likelihood ratio is a product of ones. The verdict sequence is therefore independent of the truth and carries zero bits about it, so no finite token expenditure yields a positive amount of information and the price per bit is $+\infty$. The nominal price v of a single call is finite by construction. Hence $v < \infty$ and $v_{\mathrm{eff}} = v/J = \infty$ hold simultaneously, and a finite v is not evidence that verification is affordable. ■

PROPOSITION 8 (TOKENS PER BIT)

Model the verifier as a binary channel from the claim's truth $T \in \{0,1\}$ to the verdict $A \in \{0,1\}$, with $\Pr[A{=}1 \mid T{=}1] = \mathrm{TPR}$ and $\Pr[A{=}1 \mid T{=}0] = \mathrm{FPR}$, and take $\Pr[T{=}1] = 1/2$. Then $\Pr[A{=}1] = (\mathrm{TPR} + \mathrm{FPR})/2$ and $H(A \mid T) = \frac{1}{2}[H_b(\mathrm{TPR}) + H_b(\mathrm{FPR})]$, so $I(T;A) = H(A) - H(A \mid T)$ is Eq. (6). That expression is the one evaluated on every measured verifier in this paper, and it is not a function of J: two verifiers with the same J but different asymmetry carry different amounts of information, and on the backbones of §D.4 the difference reaches $1.55\times$.

For the symmetric case $\mathrm{TPR} = 1 - \mathrm{FPR} = q$ this is a binary symmetric channel with crossover $1 - q$, whose capacity, attained at the uniform input, which is the prior assumed, is

$$I(T;A) \;=\; 1 - H_b(q), \qquad q = \tfrac{1+J}{2},$$

since $J = \mathrm{TPR} - \mathrm{FPR} = q - (1-q) = 2q - 1$. Write $q = \frac{1}{2} + \varepsilon$ with $\varepsilon = J/2$. The binary entropy is twice differentiable on $(0,1)$ with $H_b'(\frac{1}{2}) = 0$ and $H_b''(p) = -1/\big(p(1-p)\ln 2\big)$, so $H_b''(\frac{1}{2}) = -4/\ln 2$ and Taylor's theorem gives

$$H_b\big(\tfrac{1}{2} + \varepsilon\big) \;=\; 1 - \frac{2\varepsilon^2}{\ln 2} + O(\varepsilon^4), \qquad \text{hence} \qquad I(J) \;=\; \frac{2(J/2)^2}{\ln 2} + O(J^4) \;=\; \frac{J^2}{2\ln 2} + O(J^4),$$

which is Eq. (7). Dividing the price of one call by the information it yields, $v/I(J) = 2\ln 2 \cdot v/J^2\,(1 + O(J^2))$. Since v/J diverges as $1/J$ and $v/I(J)$ as $1/J^2$, the ratio of the honest price to the surrogate is $\Theta(1/J)$ and unbounded: the surrogate is optimistic, and increasingly so exactly where the verifier is worst. The same conclusion holds for the asymmetric channel, since Eq. (6) is likewise $O\big((\mathrm{TPR} - \mathrm{FPR})^2\big)$ as the two rates approach each other; only the constant changes, and Eq. (6) rather than the constant is what the measurements use. ■

PROPOSITION 9 (MARGINAL YIELD PER TOKEN)

Immediate from Definition 1: c_{obs} is a function of o alone and does not depend on b, so $c(b') - c(b) = c_{\mathrm{txt}}(b') - c_{\mathrm{txt}}(b)$ for any b, b'. Meanwhile $v \geq c_{\mathrm{obs}}$ by Theorem 1. Hence the price of buying more yield is governed entirely by the text term and the price of checking is bounded below by the observation term, and the two are separately controlled halves of the same bill. ■

PROPOSITION 3 (THE PAIRING BOUND)

Fix a description of realised length w tokens over a scene of n objects with attribute alphabet $\mathcal{A}$, $|\mathcal{A}| = K$. Let O be the set of objects it names, A the set of attribute values it names, and B the set of pairs it binds. Naming an object costs at least one token and naming an attribute value costs at least one token, and these token sets are disjoint by construction (a token is either an object name or an attribute name), so

$$w \;\geq\; |O| + |A| \;=\; n_o + n_a.$$

A bound pair requires both its object and its value to be named, so the projections of B onto each coordinate are contained in O and A respectively. The projection onto the *object* coordinate is injective, because each object carries a single attribute value and therefore appears in at most one pair of B; hence

$$n_b \;=\; |B| \;\leq\; n_o \;\leq\; w - n_a.$$

A query names an object uniformly at random from the n present, and is settled only if that object's pair is bound, so

$$\Pr[\text{settled}] \;=\; \frac{n_b}{n} \;\leq\; \frac{w - n_a}{n},$$

which is Eq. (8). The two marginals obey different bounds: $n_o \leq w$ with no dependence on K *and no attribute term*, and $n_a \leq \min(K, w)$ with no dependence on n. So the joint alone is charged for the vocabulary, and it is the first of the three to be squeezed as w falls. The bound is vacuous exactly when $w \geq n + n_a$. ■

The step this proof no longer takes. An earlier version continued from the two projections to $n_b \leq \min(n_o, n_a)$, hence $n_o + n_a \geq 2n_b$, hence $n_b/n \leq w/(2n)$. The move to $n_b \leq n_a$ is invalid:

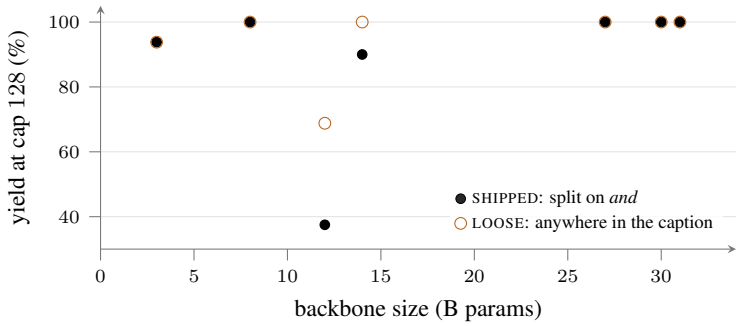
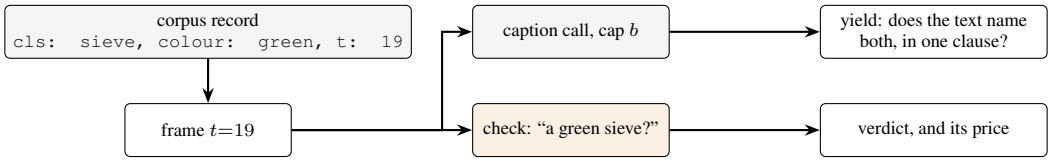


Figure 16: The one judgement call in the yield test, and what relaxing it does. SHIPPED is the predicate the measurement ran; it is not a clause test, because it normalises punctuation before splitting and so cuts only on the word *and* (§C.2). LOOSE asks only that the class and its colour both appear somewhere in the caption. The two differ on exactly one backbone by more than a rounding step, and it is the weakest one; the mean gap is 5.9 points. SHIPPED is kept as the headline because it is what ran, and because a caption listing objects in one sentence and colours in another does not give a later query the pairing it needs. Table 9 adds the third, genuinely clause-based reading.



The record makes both tests exact: a singleton appears once, at a known frame, in a known colour, so neither measurement needs a judge. The false claim is the same question with the colour replaced.

Figure 17: How one probe is built. The exactness is the reason this corpus is used: every number in the paper is a ratio of prices or a rate over probes, and a scored judgement in either would place an instrument with its own reliability inside a measurement about reliability.

containment of the attribute projection in A bounds $|\mathrm{proj}_A(B)|$, not $|B|$, and proj_A is not injective: a single value can be bound to every object in the scene. A describer exploits this with grouped notation, and one did: 16 pairs bound with 7 attribute values in 23 tokens, against the 32 the discarded bound demanded. Over 2,685 measured descriptions the discarded bound fails 5 times and the bound proved above fails 0, attaining equality in 52. The correction weakens the result (the factor of two was the whole of the "twice the rate" claim), and it is stated here rather than in a limitations section because a false inequality is a defect and not a caveat.

Remarks on tightness. The bound counts tokens, not content, and a describer can be arbitrarily wasteful (boilerplate, repeated attributes, syntax), so it is an upper bound on n_b/n and not an estimate of it. Two consequences are used in the main text. First, the bound may be checked without knowing anything about the describer, which is why §4.5 can evaluate it against a measurement rather than fit it. Second, its vacuity point $w = n + n_a$ is computable from the scene and the description's own vocabulary, which turns "the bound stops binding here" into a prediction about where yield should stop improving.

G Secondary Experiments and Sensitivity

How hard the false claim is

The verdict pairs of Table 4 are only informative if rejecting the false claim requires the capability the paper is about. Two things could make it easier than intended, and both are measurable against the recovered frame inventory (§C.1).

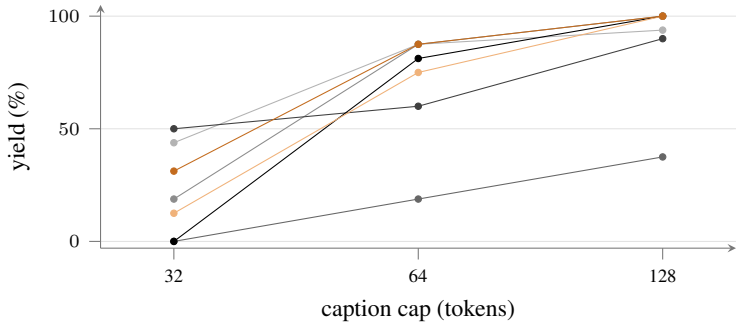


Figure 18: The cap sweep per backbone on the 16-probe pilot, one line each, ordered by size from light to dark. Every backbone gains, none reaches its ceiling below the largest cap, the weakest never reaches one at all, and the token cost along each line rises by 13% on average. The lines are what Proposition 9 predicts: additional yield is bought from the text term while the observation term is already paid.

The first is a real weakness. The runner constructs the false colour as the first entry of the corpus alphabet differing from the true one, which makes it red on 14 of 16 probes. That is a deterministic choice, not a random draw, and a backbone with any systematic disposition toward or against that one colour would show it as a J artefact. We disclose it rather than correct it after the fact; any comparison of these J values against a differently constructed false claim is invalid, and a randomised swap is the first change we would make on a re-run.

The second is not a weakness, and this is the point of the check. If the named colour were absent from the frame entirely, a correct rejection would need only colour detection ("there is nothing red here"), and J would be measuring something strictly easier than binding. It is not absent. On 15 of 16 probes the false colour is present in the scene, carried by 2.06 other objects on average. Rejecting "is there a red ⟨class⟩?" therefore requires the verifier to bind the colour to the right object while the colour itself is plainly visible on the wrong ones, which is precisely the pairing that Proposition 3 says a capped description cannot afford to transmit. The verifier and the channel are being asked for the same thing, in the same units, which is what makes the two prices comparable.

THE TAXONOMY PER BACKBONE

Table 5 pools over backbones, which is the right default and hides two things.

BOILERPLATE, AND WHERE A CAP'S BUDGET ACTUALLY GOES

At the tightest cap no caption opens with a meta-preamble: the rate is 0.0%, and it is 0.0% at the middle cap too. At the loosest, 25.5% of them do, at a cost of 8.1 words: "Here's a description for later retrieval:" and its variants. The absolute cost is small and the relative cost is not the point; the point is where it appears. Preamble is emitted exactly when the allowance is loose enough to afford it, which means the backbones are budgeting: 8.1 words is more than the entire allowance the tightest cap instructs, and the tightest cap accordingly emits none. So there is no preamble to forbid where the budget is scarce, and a prompt that forbade it would recover nothing at the cap where the recovery would matter and about a fifth of the emitted length at the cap where it would not. §4.5 shows that the tight cap's problem is binding rather than budget in any case, so we predict such a prompt would not move yield. We did not test this.

RE-RUN SPREAD, AND WHAT IS NEVER POOLED

Realised token counts and verdicts both vary across identical calls on this interface. Every number in the paper is therefore reported per backbone from a single run, and no two numbers that were not measured in the same run are compared. Where a run failed, it failed loudly: one backbone lost six of 16 probes to provider-side errors across four retry attempts, and its rows are reported over the probes that returned with the denominator stated in Table 13. It is never pooled into a mean with the others, and no cell was substituted from a different backbone or a different attempt.

Table 13: The caption taxonomy per backbone, which pooling hides. Two columns carry the appendix's weight. UNB (named both, bound neither) is not spread across backbones but concentrated in one, whose entire yield deficit at the largest cap is a two-list description format rather than a failure to see. And the last two columns are the typicality test read per backbone: the probe object is the rarest class in its episode, the other fifteen are common fillers, and if a tight allowance dropped rare things first the first column would sit below the second everywhere. It does not.

Backbone	cap	n	none	col	obj	unb	bound	probe / others named (%)
Ministral 3B	32	16	0.0	50.0	6.2	0.0	**43.8**	50.0 / 37.9
	64	16	0.0	6.2	6.2	0.0	**87.5**	93.8 / 80.4
	128	16	0.0	0.0	6.2	0.0	**93.8**	100.0 / 97.9
Qwen3-VL-8B	32	16	18.8	25.0	37.5	0.0	**18.8**	56.2 / 55.4
	64	16	0.0	12.5	0.0	0.0	**87.5**	87.5 / 96.7
	128	16	0.0	0.0	0.0	0.0	**100.0**	100.0 / 100.0
Gemma 3 12B	32	16	50.0	43.8	6.2	0.0	**0.0**	6.2 / 21.3
	64	16	25.0	50.0	6.2	0.0	**18.8**	25.0 / 51.3
	128	16	6.2	12.5	12.5	31.2	**37.5**	81.2 / 81.2
Ministral 14B	32	10	30.0	20.0	0.0	0.0	**50.0**	50.0 / 49.3
	64	10	10.0	30.0	0.0	0.0	**60.0**	60.0 / 56.7
	128	10	0.0	0.0	0.0	10.0	**90.0**	100.0 / 91.3
Qwen3.8-27B	32	16	18.8	81.2	0.0	0.0	**0.0**	0.0 / 8.3
	64	16	0.0	18.8	0.0	0.0	**81.2**	81.2 / 77.5
	128	16	0.0	0.0	0.0	0.0	**100.0**	100.0 / 99.2
Qwen3-VL-30B-A3B	32	16	6.2	43.8	37.5	0.0	**12.5**	50.0 / 51.3
	64	16	6.2	18.8	0.0	0.0	**75.0**	75.0 / 84.6
	128	16	0.0	0.0	0.0	0.0	**100.0**	100.0 / 100.0
Gemma 4 31B	32	16	37.5	31.2	0.0	0.0	**31.2**	31.2 / 28.3
	64	16	12.5	0.0	0.0	0.0	**87.5**	87.5 / 90.8
	128	16	0.0	0.0	0.0	0.0	**100.0**	100.0 / 99.6

WHY A FRESH SESSION PER CALL

This experiment prices single calls. Running the cap sweep inside one budgeted session would make each successive call cheaper in the only sense that matters here, because the earlier calls would have consumed the allowance the later ones are being priced against. Each call therefore runs in its own session and the realised cost is read from the provider's usage report rather than from our own estimate, which is deliberately pessimistic and would bias the ratio.

THE YIELD PREDICATE, AND WHAT THE THREE READINGS DO

The main text states the discrepancy in full (§C.2): the predicate that ran normalises punctuation to spaces before splitting into clauses, so the only separator that can still match is the word *and*. It is looser than the clause test the design intended and tighter than co-occurrence.

Figure 18 shows the pilot's cap sweep per backbone, one line each, and Figure 16 shows the shipped predicate against the loosest reading at the largest cap, where the mean gap is 5.9 points and one backbone accounts for almost all of it. Table 9 gives the full three-way comparison across every cap. The nesting STRICT $\Rightarrow$ SHIPPED $\Rightarrow$ LOOSE holds probe by probe, so each row is monotone by construction and the row's width is the width of the definition.

Three consequences, stated so a reader can check them against the table rather than take them on trust. *(a)* Every claim in the paper that uses yield through its *ordering* across backbones survives all three readings, because the orderings largely agree; the one exception is under STRICT at the largest cap, where Gemma 4 31B falls from joint first to fifth. *(b)* The rise of yield with the allowance survives all three readings on every backbone, and so does the identity of the weakest backbone. *Saturation does not.* Under SHIPPED at cap 128 4 backbones reach 100%; under STRICT one does, and the rest run from 18.8% to 93.8%. Any sentence in this paper that leans on "yield saturates" is therefore a statement about the shipped predicate and not about the world, which is one more reason the 200-probe replicate's weaker formulation (5 of 7 backbones at or above 95%) is the one the main text uses. What §4.4 actually needs is that yield has little room left while J has a great deal, and that

survives every reading, because under STRICT yield is lower still. *(c)* No claim about the *level* of yield at a given cap is portable, and the only place the main text uses a level is against the counting bound of Eq. (8), where the strict reading lowers the measured side of an inequality the bound already satisfies and therefore strengthens the agreement rather than threatening it.