RECURSIVE CHANNEL COST IN BUDGETED ARBITRATION: RELIABILITY-BLIND ARBITRATION CANNOT BE REPAIRED FOR FREE

Anonymous authors
Paper under double-blind review

ABSTRACT

A budgeted agent given several channels over one observation stream must choose what to read, and the channels hand it *claims* that compete. We add the arbiter to the access-structure formalism and measure what arbitration does. *(1)* A blind arbiter attains a presentation-weighted average of its channels' reliabilities rather than the best of them, and we state the identity with a third term that vanishes exactly when the arbiter is blind, so blindness becomes a hypothesis a transcript settles: on this grid, settled *for* ($\mathcal{C} = +0.016$ over 366 arbitrated decisions). Telling the arbiter which channel to follow, by a one-line directive keyed on the query class, is the first manipulation in this line of work to act at the arbitration point and survive its own control; the directive helps where its classifier is right every time and *hurts* where it is right about half, so knowing which channel to trust is itself another unreliable channel over the same stream. *(2)* Appending a weaker shard of the *same* index (same form, weaker content) *subtracts* capability in proportion to its weakness, the two blocks arbitrated symmetrically; conditioning on the loop's exit, the subtraction lives in committed accuracy and not in rounds, so appending is arbitration, not budget. *(3)* The compressive channel is followed where it holds the queried quantity (90.2% item agreement; *causally*, every one of 114 injected decisions follows a planted wrong value) and holds nothing where the answers are attributes; a registered probe that renders its facts as citations leaves a silent channel unfollowed, so presentation cannot rescue a channel with nothing to say. Grounding is a function of the access loop's *exit*, not of the architecture: separated from commitment, the apparent benefit of removing a channel reverses. A table lists every claim with its interval and verdict, and every control is read against the spread of a difference known to be nothing.

1 INTRODUCTION

An agent that must answer questions about a long observation stream under a fixed per-decision token budget cannot hold the stream in its context. It has to be given channels: a compressed state it maintains as observations arrive, an index it can search, a tool it can call. The framework we build on (Chen et al., 2026b; Anonymous authors, 2026c) prices those channels against the resource walls they cross and reports capability as a function of which channels an architecture holds. Under that reading, holding more channels is never worse than holding fewer, for a reason so elementary it is never written down: a decision procedure can always ignore a channel it does not want, so an extra channel can waste budget but cannot lower accuracy.

This paper is about a measurement that contradicts that sentence, and about what has to change in the formalism to accommodate it. In the released cells of the parent system (Anonymous authors, 2026c), an architecture holding both a maintained scene state and a searchable episodic index scores 16.7 on the composition subtest of a budgeted embodied benchmark; the same architecture with the state channel *removed* scores 83.3. The ablated arm is handicapped twice (the budget reserved for the state channel is not reallocated, so it runs on roughly three quarters of the usable tokens, and a state action that returns nothing still consumes one of its composition rounds), and it wins by 66.7

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.

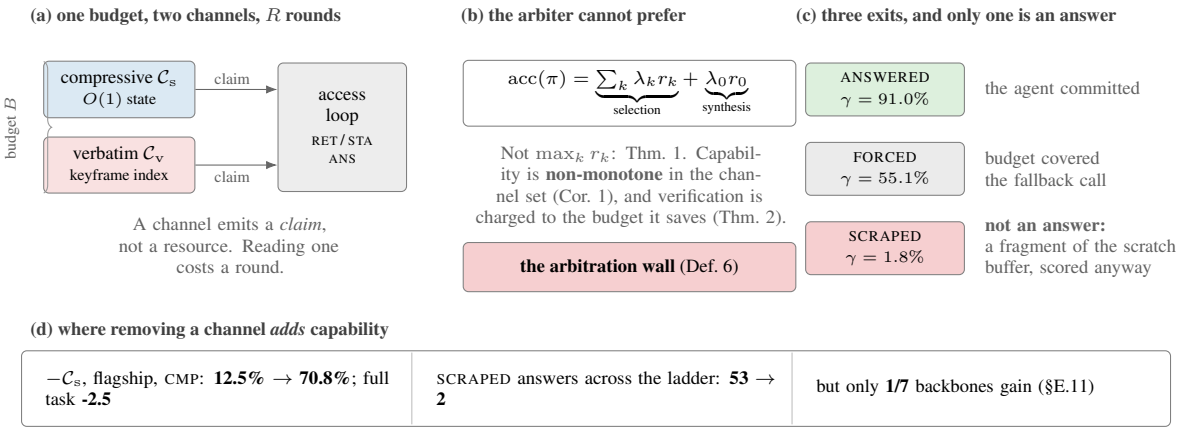

Figure 1: The argument in one figure; the text walks the panels in order before this figure is placed. (a) the setting: two channels, one budget, a round per read; channels hand *claims*, not resources. **(b)** the formalism: a blind arbiter attains a weighted average plus a synthesis term, not the best of its channels (Theorem 1); verification is priced by Theorem 2. **(c)** grounding is a function of the access loop's exit, not of the arm (§4.2). **(d)** removing the compressive channel converts non-answers into answers; pooled over the grid the removal is a marginal cost, read against its falsification test in §E.11.

points anyway. On that subtest, on that backbone, a channel with a price paid and a wall crossed has negative marginal value.

That is not a tuning failure and it is not specific to one implementation. It is what happens whenever two channels can disagree and nothing in the architecture says which to believe. A channel does not hand the decision procedure a resource; it hands it a *claim*. When claims conflict, something arbitrates. The arbiter is not free, it is not neutral, and it is drawn on the same budget as the channels it arbitrates between. None of those three facts has a place to be written in a formalism whose vocabulary is prices and walls.

What we do. Section 2 defines a channel by the claim it emits and the reliability of that claim, separates arbiters that can see reliability from those that cannot, and shows that the monotonicity the framework assumes is exactly the statement that the arbiter attains $\max_k r_k$. A reliability-blind arbiter does not: where wrong claims do not coincide (the generic regime for open-ended quantities such as counts), it attains a *weighted average* of channel reliabilities whose weights are set by presentation rather than by reliability, plus a term for answers it synthesises from claims no channel stated. The second term is where the parent framework's super-additivity lives and the first is where the loss does; an ablation reports their difference and identifies neither. Capability is then non-monotone in the channel set, and no reallocation recovers the loss, because the loss is not an allocation.

What the transcripts say the harm actually is. Because the runs record their round prompts, we can read the loop rather than its scores. The access loop has three exits: the agent emits an answer; it does not, but enough budget remains for the runner's fallback call; or it does not and the budget does not, in which case the harness scores the last non-bracketed line of the scratch context (a truncated bullet, or the opening of a sentence in which the model was still reconsidering). The third is not an answer, and nothing downstream can tell. Grounding turns out to be a function of that exit and of nothing else: pooled over both arms and every backbone, $\gamma = 91.0\%$ on the 860 items where the agent committed, 55.1% on the 205 that reached the fallback, and 1.8% on the 55 that did not. So the compressive channel is not asserting something false and having it believed. It is spending the rounds the agent needed in order to commit, and what the scorer then receives is not a claim from any channel at all.

What is not new here, stated plainly. Three ingredients are borrowed and we claim none. The exponential decay of a channel maintained by unreliable per-step updates is the union bound of the parent system (Anonymous authors, 2026c) and, in sequential decision making, the compounding that separates behaviour cloning from DAGGER (Ross et al., 2011; Li et al., 2026). That combining

several predictors can beat all of them is Condorcet's (de Condorcet, 1785; Dietterich, 2000), used here as a boundary condition telling us which regime our theorem may not enter. The distinction between maintaining a derived quantity incrementally and recomputing it on demand is incremental view maintenance (Gupta & Mumick, 1995; Gupta, 2009), and the space cost of a running count over a stream is settled by streaming lower bounds (Alon et al., 1999; Indyk & Woodruff, 2003). Every one of those assumes the update operator is *correct* and asks about cost or space. What is new is a setting in which the operator is unreliable and cannot be trained, its reliability is a measured property of a frozen model, and a second channel offers a competing answer. Two sources disagreeing about one question is likewise studied as *knowledge conflict* (Xu et al., 2024), and that arbitration between conflicting evidence is *systematic* rather than arbitrary is established there, not here (Carletti et al., 2026; Lietzow et al., 2026; Lyu et al., 2026; Zhu et al., 2026). What is left, and it is narrower, is the *budget*: every study above presents both sources and asks which wins, whereas under a fixed per-decision budget the prior question is whether the arbiter is offered both at all; and the answer here is usually not.

Contributions. (1) An access-structure formalism that includes the arbiter, and the identification of capability monotonicity with reliability-aware arbitration. (2) A witness lemma: a cumulative reduction has no witness of size one, so a compressive channel cannot cite the facts it exists to hold, while a retrieval channel's every claim is localised by construction: a statement about *price*, since a localised claim is checkable for the cost of one observation and a reduction is not checkable below the cost of re-reading the stream. (3) Theorem 1, whose selectivity term vanishes exactly under blindness, so the premise is testable; it is **confirmed** on 366 arbitrated decisions. (4) A complete arbitration census separating the selection loss (27 of 66 decisions the index channel had already answered) from the synthesis gain (12 of 82 that neither channel answered alone). (5) A registered mechanism measured and refuted, with the instrument audited rather than trusted, and a causal probe that replaces and *retracts* a claim of the previous version. (6) The typed directive of §4.5: a 6.4-point gain against a control identical in length and differing only in which channel each query class is told to trust (the first manipulation here to move λ at the arbitration point and survive its own control), which gains on the two subtests where the classifier supplying the query class is right on every decision and *loses* on the one where it is right on 53.6%. (7) A provenance census showing that grounding is a function of which exit the loop took and not of the arm.

2 THEORY

Definition 1 (Channel, access structure, arbiter)**.** Fix a stream S of N observations and a query distribution $\mathcal{Q}$ with ground truth $a(q)$. A *channel* is a procedure c that, given a query q and a token allowance b, returns a claim $\hat{a}_c(q, b) \in \mathcal{Y}$ and consumes at most b tokens; its *reliability* is $r_c(b) = \Pr_{q\sim\mathcal{Q}}[\hat{a}_c(q, b) = a(q)]$. An *access structure* is a set $\mathcal{A} = \{c_1, \ldots, c_K\}$ with an allocation $\sum_k b_k \le B$ and an *arbiter* π mapping the multiset of returned claims to a single answer; write $r_k = r_{c_k}(b_k)$. The definition deliberately does not distinguish a channel that reads pixels from one that reads a maintained state: what matters downstream is only that a channel emits a claim and that its claim is sometimes wrong.

Definition 2 (Reliability-blind arbitration)**.** An arbiter is *reliability-blind* if its output depends on the claims only through their content, not through which channel produced them: for any two channels j, k whose claims are presented in textually indistinguishable form, exchanging those claims leaves the output unchanged.

This is not a straw man. It is what an implementation looks like when the formalism has no slot for reliability: in the parent system the two channels are concatenated into one context (the maintained state as a JSON object, the retrieval as a described read), and a frozen backbone is asked for an answer, with nothing in that interface carrying r_s or r_v. But it is an *assumption*, and a reader is entitled to ask why they should grant it. So we do not ask them to: Theorem 1 is stated with a term that vanishes whenever the arbiter is blind in expectation, which turns blindness into a quantity a transcript can settle.

Definition 3 (Witness, localised claim)**.** Let a claim be a function ϕ of the stream $S = (S_1, \ldots, S_N)$. An index set $W \subseteq [N]$ is a *witness* of ϕ if $\phi(S)$ is determined by $(S_i)_{i\in W}$. A claim is *localised* if it has a witness of size one.

Lemma 1 (A reduction has no witness of size one)**.** *Let $\rho(S) = \bigoplus_{i \le N} f(S_i)$ be a reduction under an operation $\oplus$ cancellative in each argument, with f non-constant and $N \ge 2$. Then no singleton $\{i\}$ is a witness of ρ.* ■

The parent system defines the compressive channel by exactly this property: it holds the running reductions retrieval cannot cheaply rebuild. That is what makes it worth its price, and by Lemma 1 it is also what makes its claims uncitable: a cumulative count over a stream has no frame to point at. The retrieval channel is the mirror image: every claim it emits is a described read of one observation, hence localised by construction. The consequence is about *price*: a localised claim is checkable for the cost of one observation, a reduction is not checkable below the cost of re-reading the stream, which is why the escapes of §3 cannot buy a check and why a reliability label, not being a witness, fails to.

Proposition 1 (Reliability-aware upper reference)**.** *An arbiter with knowledge of $r_1, \dots, r_K$ attains accuracy at least $\max_k r_k$, by returning the claim of any channel maximising r_k. Hence capability under reliability-aware arbitration is non-decreasing in $\mathcal{A}$.* ■

The next statement has to be got right, because the obvious version is false. Blind arbitration is not always worse than the best channel: K channels of comparable reliability whose errors are diverse can be combined by majority vote to beat every one of them, the classical ensemble effect (de Condorcet, 1785; Dietterich, 2000). That effect requires wrong claims to *coincide* less often than right ones, which fails when the answer set is small and errors are systematic, and it requires comparable reliabilities, which fails exactly when one channel has been degraded by compounding.

Definition 4 (Non-coinciding errors)**.** $\mathcal{A}$ has *non-coinciding errors* on $\mathcal{Q}$ if on every query at most one presented claim is correct; equivalently, two channels never agree on a wrong answer. This is the regime of open-ended quantities such as counts, and it is the complement of the regime in which the ensemble effect operates.

Theorem 1 (Arbitration attains a weighted average, a selectivity term and a synthesis term)**.** *Let $\mathcal{A}$ have non-coinciding errors, write E_k for the event that π returns channel k's claim, $\lambda_k = \Pr[E_k]$, $\lambda_0 = 1 - \sum_k \lambda_k$, and r_0 for the reliability of synthesis. Then, identically,*

$$\mathrm{acc}(\pi) \;=\; \underbrace{\sum_k \lambda_k r_k}_{\textit{selection}} \;+\; \underbrace{\sum_k \mathcal{C}_k}_{\textit{selectivity}} \;+\; \underbrace{\lambda_0 r_0}_{\textit{synthesis}}, \qquad \mathcal{C}_k \coloneqq \lambda_k \big(\Pr[\hat{a}_k = a(q) \mid E_k] - r_k \big). \tag{1}$$

If π is reliability-blind then $\mathcal{C}_k = 0$ for every k, λ is determined by the presentation and not by r, and (1) *reduces to*

$$\mathrm{acc}(\pi) \;=\; \sum_{k=1}^{K} \lambda_k \, r_k \;+\; \lambda_0 \, r_0. \blacksquare \tag{2}$$

The selectivity term is this paper's correction to its own previous version, which asserted the reduced form unconditionally and proved it by a step that does not follow from Definition 2. The repair is worth more than the error cost: $\sum_k \mathcal{C}_k$ is computable from a transcript with no extra measurement, because a decision on which π returned $\hat{a}_k$ and scored correct is a decision on which $\hat{a}_k$ was correct. Measured, it is $\mathcal{C} = +0.016$ over 366 arbitrated decisions, median +0.007, positive on 7 of 14 cells, which is chance. **The arbiter is blind, measured rather than assumed**, and the rest of the paper is what that costs.

Corollary 1 (Non-monotonicity, and why an ablation cannot separate the terms)**.** *Suppose appending c_{K+1} leaves the synthesis term unchanged. Under symmetric π accuracy changes by $\frac{1-\lambda_0}{K+1}\big(r_{K+1} - \frac{1}{K}\sum_{k \le K} r_k\big)$, negative whenever r_{K+1} is below the incumbent mean, and no allocation of B recovers the loss because the loss is not an allocation. Writing L for the decisions $\mathcal{A} \setminus \{c_s\}$ answers correctly and $\mathcal{A}$ does not, and G for the converse, an ablation reports $G - L$ while* (1) *accounts for both: L is a selection loss on decisions where c_s captured a λ_s share and was wrong, and G is the synthesis term. A reported ablation gap is therefore uninformative about either mechanism: L and G must be counted separately, and a channel can subtract while the structure it belongs to is genuinely super-additive.* ■

The two weights. Theorem 1 leaves λ free, and λ is where the physics is: it is a property of how claims are *presented*, and presentation is exactly what an implementation must invent because the theory is silent. A maintained state rendered as `{"counts":{"cup":3}}` is an assertion; a retrieval rendered as `[RETRIEVED ...]` followed by a description is a report of an observation.

Definition 5 (Assertion premium; grounding rate)**.** For $K = 2$ with an asserted channel s and a reported channel v, the *assertion premium* is $\lambda = \Pr[\pi \text{ returns } \hat{a}_s \mid \hat{a}_s \neq \hat{a}_v]$. An answer is *grounded* when it is attributable to a claim some channel inserted, and the *grounding rate* is $\gamma = \sum_k \lambda_k = 1 - \lambda_0$, under which (2) reads

$$\text{acc} \;=\; \gamma\, \bar{r}_{\text{sel}} \;+\; (1-\gamma)\, r_0, \qquad \bar{r}_{\text{sel}} = \frac{1}{\gamma} \sum_k \lambda_k r_k. \tag{3}$$

The reparametrisation earns its place by what it predicts: an asserted channel can damage a structure by being followed and being wrong, which lowers $\bar{r}_{\text{sel}}$, or by displacing grounding altogether, which lowers γ; and the second says where the effect must be *absent*, namely on a backbone already grounded at $\gamma \approx 1$. Appendix A records both registered predictions and their outcomes. Two further results say why the obstruction is a wall rather than an inefficiency.

Lemma 2 (A maintained channel crosses below its incumbents)**.** *Let channel s be maintained by one call per observation with no re-derivation, and assume (i) update i is correct with probability exactly $1 - \epsilon_u$; (ii) the updates are independent or negatively dependent; and (iii) the reduction is* non-cancelling*, so the final state is correct only if every update was. Then $r_s(N) \leq (1-\epsilon_u)^N$, so for any fixed reported reliability r_v there is a finite $N^* = \log r_v / \log(1-\epsilon_u)$ beyond which $r_s < r_v$ and, by Corollary 1, the maintained channel subtracts.* ■

Hypotheses (ii) and (iii) are what the bound needs and neither is free; the decay itself is not ours (Anonymous authors, 2026c; Ross et al., 2011). And:

Theorem 2 (Verification is charged to the budget it saves)**.** *Let channel c_k be used at allowance b_k and let v_k be the token cost of a check that establishes the correctness of its claim. The check is drawn on the same per-decision budget B, so a structure that both uses and verifies c_k pays $b_k + v_k$. By Lemma 1 a localised claim admits v_k of the order of one observation, while a reduction admits none below the cost of re-reading the stream. Hence a channel whose verification price reaches its own yield can never be both used and checked inside B.* ■

Definition 6 (Arbitration wall)**.** The *arbitration wall* at budget B is the set of query classes on which no access structure with reliability-blind arbitration attains the accuracy of its own best channel. Its dual reading, in the sense of Chen et al. (2026b), is a task that demands discrimination among claims against a resource constraint that charges discrimination to the same budget as the claims.

The escape the wall permits is periodic re-derivation, and its price is fixed:

Corollary 2 (Re-derivation period)**.** *For a target reliability $r^\dagger$ the smallest sufficient period is $T = \lfloor \log r^\dagger / \log(1-\epsilon_u) \rfloor$, the induced amortised cost per observation is v_{agg}/T, and re-derivation is worth its price exactly when $\hat{v}_{\text{agg}}/T < \lambda\,(r_v - r_s)$, where v_{agg} is the token price of one index aggregation and $\hat{v}_{\text{agg}}$ is that price in accuracy units.* ■

3 METHOD

The architecture under test is the parent system's: a compressive channel $\mathcal{C}_{\text{s}}$ maintaining a JSON scene state updated once per observation, a scalable index $\mathcal{C}_{\text{v}}$ over the stream, and a query-conditioned router, all on frozen open-weight backbones at a fixed per-decision budget. Arms are ASP (both channels), $-\mathcal{C}_{\text{s}}$ (the compressive channel removed, budget *not* reallocated), the reallocated control, the $K{=}3, 4, 5$ shard ladder that appends weaker copies of the *same* index, the reliability-marked and swapped-marker presentations, and the typed directive of §4.5. Every arm is scored per item, and the harness records the round prompts, which is what makes the exit census of §3 possible without re-running anything.

The loop, and the three ways it can end. A decision is a bounded sequence of rounds. The agent may request a channel, receive a block appended to its scratch context, and either emit ANSWER

Table 1: Arbitration census on the composition subtest, 168 decisions over seven backbones. Each item is classified by which single-channel arm answered it; each cell gives the number of such items and, in parentheses, how many the two-channel structure answered. The third column is the selection loss of Eq. (2) and the fourth is its synthesis term. Scores are 0 or 100 on all but a handful of items and ≥ 50 counts as correct.

Backbone	both	only $\mathcal{C}_\mathrm{s}$	only $\mathcal{C}_\mathrm{v}$	neither	λ	composed
Ministral 3B	–	–	10 (4)	14 (3)	0.600	29.2
Qwen3-VL-8B	2 (0)	1 (0)	9 (6)	12 (2)	0.300	33.3
Gemma 3 12B	3 (3)	1 (1)	4 (2)	16 (1)	0.600	29.2
Ministral 14B	1 (1)	–	7 (5)	16 (3)	0.286	37.5
Qwen3.8-27B	5 (2)	2 (1)	11 (1)	6 (0)	0.846	16.7
Qwen3-VL-30B-A3B	3 (3)	1 (1)	4 (1)	16 (1)	0.800	25.0
Gemma 4 31B	1 (1)	–	21 (20)	2 (2)	0.048	95.8
Pooled	15 (10)	5 (3)	66 (39)	82 (12)	0.497	

Table 2: The assertion premium against the quantities Eq. (2) combines. r_s and r_v are the single-channel accuracies, λ is estimated from the disagreement classes of Table 1, and the selection column is $\lambda r_\mathrm{s} + (1-\lambda) r_\mathrm{v}$. The residual is the composed accuracy minus that selection prediction; it is positive on every backbone, and the synthesis term of Eq. (2) is what accounts for it.

Backbone	params	r_s	r_v	λ	selection	composed (residual)
Ministral 3B	3B	0.0	41.7	0.600	16.7	29.2 (+12.5)
Qwen3-VL-8B	8B	12.5	45.8	0.300	35.8	33.3 (-2.5)
Gemma 3 12B	12B	16.7	29.2	0.600	21.7	29.2 (+7.5)
Ministral 14B	14B	4.2	33.3	0.286	25.0	37.5 (+12.5)
Qwen3.8-27B	27B	29.2	66.7	0.846	35.0	16.7 (-18.3)
Qwen3-VL-30B-A3B	30B	16.7	29.2	0.800	19.2	25.0 (+5.8)
Gemma 4 31B	31B	4.2	91.7	0.048	87.5	95.8 (+8.3)

or continue. The loop ends in one of three ways: the agent commits; it does not but the remaining budget covers the runner's fallback call; or it does not and the budget does not, in which case the harness scores the last non-bracketed line of the scratch context. Only the first is an answer, and the distinction is invisible to the score.

Two training-free escapes, with their prices. *Reliability-marked presentation* attaches a measured reliability to each block, so a blind arbiter is told what Definition 2 says it cannot see. *Periodic re-derivation* recomputes the maintained reduction from the index every T observations, bounding the compounding of Lemma 2 at a price the corollary in Appendix G fixes. Both are charged to B, which is Theorem 2 in operational form.

4 Results

4.1 The arbitration census, and the two terms

Table 1 classifies every composition-subtest decision by which single-channel arm answered it, and reports what the two-channel structure did with it.

Four readings, in order of how little theory each requires. The structure loses 27 of the 66 decisions the index channel had *already* answered alone: the selection term, and not a budget effect, because the index channel had the answer and the composite had the index channel. It recovers 12 of the 82 that neither channel answered alone: the synthesis term, where the parent framework's super-additivity lives. An ablation reports the difference of the two and identifies neither, which is Corollaries 1–1 as a practitioner meets them. And λ varies from 0.048 to 0.846 across backbones of the same family at the same budget (Fig. 2), which is what it means for the weight to be set by presentation rather than by reliability.

Table 2 reads the identity of Theorem 1 against the quantities it combines: the selection column is $\lambda r_\mathrm{s} + (1-\lambda) r_\mathrm{v}$, the residual is composed accuracy minus that prediction, positive on five of seven

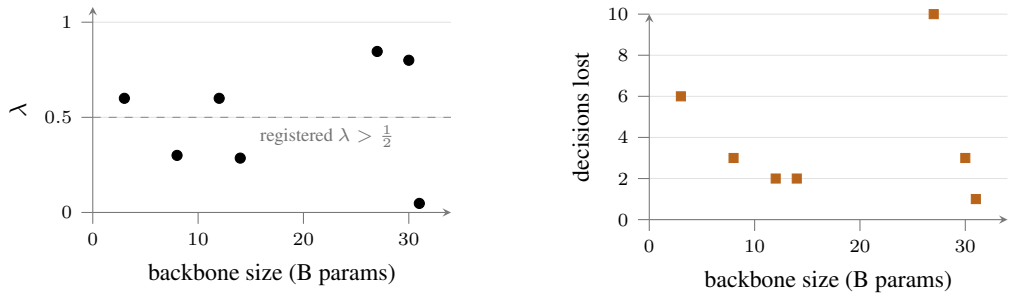

Figure 2: λ does not track scale, and neither does the damage. Left: the assertion premium per backbone against parameter count, with the registered threshold. It spans 0.048 to 0.846 across backbones that share every channel, every budget and every prompt. Right: decisions lost on items the index channel had already answered. The two panels are the same phenomenon counted two ways, and neither is a function of size.

Table 3: Where the scored answer came from, and what it was worth. The access loop has three exits: the agent emits ANSWER in a round; it does not, but the budget still covers the runner's fallback call; or it does not and the budget does not, in which case the scorer receives the last non-bracketed line of the scratch context. The third is not an answer. Two facts run through every row. Grounding is a function of the exit and not of the arm: pooled over both arms and all backbones, $\gamma = 91.0\%$ on the 860 ANSWERED items and 1.8% on the 55 SCRAPED ones. And removing the compressive channel does not make answers better; it converts SCRAPED into ANSWERED, 53 to 2 across the ladder.

Backbone	arm	items by exit			accuracy on those items			mean
		ans.	forced	scraped	ans.	forced	scraped	
Ministral 3B	ASP	66	14	0	47.0	14.3	–	41.2
	$-\mathcal{C}_s$	7	73	0	42.9	43.8	–	43.8
Qwen3-VL-8B	ASP	78	0	2	53.8	–	50.0	53.8
	$-\mathcal{C}_s$	73	7	0	50.7	0.0	–	46.2
Gemma 3 12B	ASP	79	1	0	40.5	0.0	–	40.0
	$-\mathcal{C}_s$	80	0	0	38.8	–	–	38.8
Ministral 14B	ASP	76	3	1	52.6	33.3	0.0	51.2
	$-\mathcal{C}_s$	76	4	0	52.6	0.0	–	50.0
Qwen3.8-27B	ASP	18	32	30	100.0	53.1	56.7	65.0
	$-\mathcal{C}_s$	55	23	2	85.5	13.0	0.0	62.5
Qwen3-VL-30B-A3B	ASP	29	31	20	75.9	32.3	35.0	48.8
	$-\mathcal{C}_s$	64	16	0	50.0	31.2	–	46.2
Gemma 4 31B	ASP	80	0	0	92.5	–	–	92.5
	$-\mathcal{C}_s$	79	1	0	73.4	0.0	–	72.5

backbones, the synthesis term made visible. On Qwen3.8-27B the residual is -18.3, and the census row says why: $\lambda = 0.846$ on a backbone whose compressive channel is less than half as reliable as its index.

4.2 GROUNDING IS A FUNCTION OF THE LOOP'S EXIT, NOT OF THE ARM

Table 3 is the paper's central empirical correction, and it corrects us. Pooled over both arms and every backbone, $\gamma = 91.0\%$ on ANSWERED items, 55.1% on FORCED ones and 1.8% on SCRAPED ones: grounding is therefore a property of the exit, not of the architecture. Removing the compressive channel does not make answers better: it converts SCRAPED into ANSWERED, 53 to 2 across the ladder, and the backbone that scrapes most commits on 18 of 80 items with the channel and 55 without, while its accuracy *on committed items* is 100.0% with and 85.5% without. Removal buys commitments and pays for them in precision. Once commitment is conditioned on, the previous version's claim that removal leaves the agent better grounded reverses sign, and we withdraw it (Appendix B); the ablated arm is also more conservative than it looks, requesting the absent channel and receiving a placeholder 367 times.

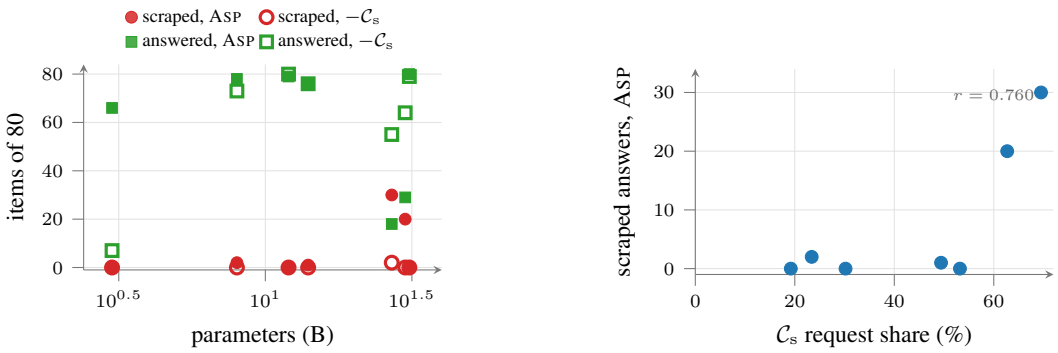

Figure 3: Left: the two exits that matter, per backbone, in both arms. Removing the compressive channel does not raise the number of *correct* answers by improving answers: it raises the number of *answers*. Filled markers are the full system, open markers the ablated arm; the scraped population collapses and the answered population grows by almost exactly the same amount. **Right:** the agents that suffer are the agents that ask. The share of channel requests directed at the compressive channel predicts how many of that backbone's answers were never given, at $r = 0.760$ over 7 backbones.

4.3 What the channel holds, and what the arbiter prefers

Counting rather than attributing: the compressive channel holds the queried quantity on 114 of 560 decisions, and on *none* of the 224 retrieval and 168 composition ones, because its schema stores no queried attribute. Where it does hold the quantity, reading the cached state predicts the arm's per-item correctness on 90.2% of 112 decisions against 58.2% with the pairing destroyed. A registered value-substitution probe that plants a wrong value into the channel's block moves the answer to it on 114 of 114 decisions. That *retracts* the previous version's claim that no decision was shown to follow the compressive channel, which rested on an instrument with no discriminative power on that side. The registered citation probe then fails in the informative direction: rendering the channel's facts as citations leaves state-following at 0 of 168 composition decisions against 2.6% for the JSON control, below the 15% refutation line, because presentation cannot rescue a channel with nothing to say.

4.4 Appending a weaker copy of the same channel subtracts

Appending a weaker shard of the *same* index (same form, weaker content) subtracts -13.39 (quarter shard) and -4.78 (half shard), with the penalty scaling with weakness on all seven backbones, and the two same-form blocks are arbitrated *symmetrically*: the shard takes 0.402 of same-form retrieval attributions despite a six-fold reliability gap, which is Corollary 1's premise measured in the regime where it can be. Conditioning on the exit, the exact decomposition puts -13.92 of the pooled -13.39 in committed accuracy against -1.26 in rounds: appending is arbitration, not budget. The registered monotonicity prediction does *not* survive the $K{=}4/K{=}5$ extension (the penalty decays rather than compounding), and we withdraw the K^* reading to a two-point direction.

4.5 Telling the arbiter which channel to follow, and what it costs

A blind arbiter can be *told* which channel to follow. A one-line directive keyed on the query class buys 6.4 points against a control identical in length and differing only in which channel each class is told to trust (CI $[+0.87, +12.68]$, label-permutation $p = 0.0001$, 560 paired decisions). This is the first manipulation in this line of work to act at the arbitration point and survive its own control, where a reliability marker, a swapped marker and block order all move nothing; the forced-order arms close the one observational contrast that was not null, reassembling the context with block positions swapped and contents identical, after which λ does not move (-1.4 points, inside the registered null band). The repair is not free. The directive gains $+8.33$ and $+2.68$ points on the two subtests where the classifier supplying the query class is right on every decision, and *loses* -4.32 on the one where it is right on 53.6%. Knowing which channel to trust is the output of another unreliable channel over the same stream, charged to the same budget, which is Theorem 2 one level up, and the paper's title claim.

5 LIMITATIONS

(i) The headline ablation is directional, not established. Removing the compressive channel costs -6.12 points pooled with an interval of $[-17.91, +5.03]$ over 4 of 14 cells; under the fair budget return it is $+5.75$. We report every arm and mark the verdicts (Appendix B). **(ii) Blindness is corroborated, not proved**: a zero selectivity term is evidence for Definition 2, and the implication runs only one way. **(iii) The corollary's premise is contradicted between channels of different form and holds between channels of the same form**, which is why §4.4 is the regime in which it is tested. **(iv) One synthetic embodied corpus, one text corpus, seven backbones, one router**; the text arm carries a documented index defect, reported and not repaired (Appendix F). **(v)** K^* **is not reached**, and the monotonicity reading is withdrawn to a two-point direction.

6 RELATED WORK

The framework this paper extends prices channels against resource walls (Chen et al., 2026b; Anonymous authors, 2026c) and has no vocabulary for the step after both channels have spoken. Knowledge conflict supplies that vocabulary but not the budget: Xu et al. (2024) divide the area into context–memory, inter-context and intra-memory conflict and this setting is squarely the second; Carletti et al. (2026) manipulate modality, recency, source reliability and provenance independently and report that open instruction models follow recency more consistently than an explicit reliability cue, which our null on reliability-marked presentation *corroborates* on a second setting rather than establishing; Lietzow et al. (2026) locate the analogous perception-versus-prior conflict at the level of attention heads and Lyu et al. (2026) measure it for tool-using agents. Reliability-aware arbitration is implemented rather than merely postulated in SABER (Zhu et al., 2026), an engineering instance of Proposition 1. The closest recent framework states the assumption we drop, treating memory updates as deterministic (Colaco & Lahjouji, 2026), and the closest decomposition prices the tool-calling apparatus rather than the choice between channels (Zhang et al., 2026a). That capability can be non-monotone in the component set is likewise not ours (it is measured in Liu (2026); Zhang et al. (2026a); Repantis et al. (2026)), and what we add is the term that explains it. Appendix C is the full treatment.

7 CONCLUSION

Adding a channel to a budgeted agent is not adding a resource; it is adding a claim to an arbitration the architecture never specified. Under blind arbitration, which we measure rather than assume, capability is a presentation-weighted average of channel reliabilities plus a synthesis term, so a channel can subtract while the structure it belongs to is genuinely super-additive, and the ablation that is supposed to settle the question reports only the difference. The harm, read off the transcripts rather than the scores, is delivered at the access loop's exit: an agent that spends its rounds arbitrating never commits, and what the scorer receives is not a claim from any channel. The repair that works is to tell the arbiter which channel to follow, and it works exactly as well as the classifier that supplies the query class, which is another unreliable channel over the same stream, charged to the same budget. That recursion is the wall.

REPRODUCIBILITY STATEMENT

Every number in this paper is written by a named analyzer to a result file, and every deciding read-out rule was frozen before the run it decides; Appendix I is the audit trail and the registration index, including the predictions that failed and the two claims withdrawn. Appendix G proves every statement of §2. Appendix D gives the method and the exact configuration executed, and Appendix E every experiment in full, including the secondary and sensitivity analyses. The backbones are frozen open-weight models called at temperature 0 through a commercial router with provider-side reasoning disabled; per-decision records, the analyzers and the corpus generator are in the supplementary material, and the synthetic corpus regenerates bit-identically from a fixed seed without any API call.

ETHICS STATEMENT

No human subjects were involved and no personal data was collected: the embodied corpus is a licence-free synthetic walkthrough generated from a fixed seed, and the text corpus is a public research-paper question-answering set used under its released licence. No model was trained or fine-tuned; the backbones are frozen open weights and are the *object* of measurement. The paper's conclusion (that adding a channel to a budgeted agent can subtract capability, and that the repair is charged to the budget it saves) is a limitation statement about agent architecture 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 theory, to state or prove the results, to formulate the hypotheses, to design the experiments or the controls, to set any pre-registered threshold or refutation line, or to interpret the results. Frozen open-weight models are the *subject* of every experiment reported here, not tools used to produce them; every number is computed by an analyzer the authors wrote from the per-decision records the runs produced. *Tasks in the recommended-disclosure list.* Generative AI was used on the manuscript itself: 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 two withdrawals and the correction to Theorem 1's previous form are the authors' judgements. *Verification.* Every quantitative statement was checked against the result file it cites, and every registered outcome against the criterion frozen before the run. 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. 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*, 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. The price of belief: What a perceptual channel yields, what it costs to check, and why checking costs what looking costs. *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.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

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

David A. Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in NC^1. *Journal of Computer and System Sciences*, 38(1), 1989.

Mattia Carletti, Edward Phillips, Fredrik K. Gustafsson, Patitapaban Palo, Lei Clifton, Danielle Belgrave, Xiao Gu, and David A. Clifton. When text and numbers disagree: Evidence arbitration in large language models. *arXiv preprint arXiv:2608.20116*, 2026.

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

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

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.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.

Guhao Feng et al. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *NeurIPS*, 2023.

Yuchen Guo, Junli Gong, Weicheng Wang, et al. Peam: Parametric embodied agent memory through contrastive internalization of experience in minecraft. *arXiv preprint arXiv:2605.27762*, 2026.

Ashish Gupta and Inderpal Singh Mumick. Maintenance of materialized views: Problems, techniques, and applications. *IEEE Data Engineering Bulletin*, 18(2):3–18, 1995.

Himanshu Gupta. Self-maintenance of views. In *Encyclopedia of Database Systems*. Springer, 2009.

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.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *ICML*, 2024.

Piotr Indyk and David P. Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 283–288, 2003.

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.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, et al. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Chingkwun Lam, Jiaxin Li, Lingfei Zhang, et al. Governing evolving memory in llm agents: Risks, mechanisms, and the stability and safety governed memory (SSGM) framework. *arXiv preprint arXiv:2603.11768*, 2026.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

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.

Niclas Lietzow, Danielle Bitterman, Carsten Eickhoff, William Rudman, and Michal Golovanevsky. Vision-default, prior-override: Causal mechanisms of perception–knowledge conflict in vision–language models. *arXiv preprint arXiv:2606.28273*, 2026.

Bingbin Liu et al. Transformers learn shortcuts to automata. In *ICLR*, 2023a.

Ming Liu. More is not always better: Cross-component interference in LLM agent scaffolding. *arXiv preprint arXiv:2605.05716*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023b.

Yaxing Lyu, Shengjie Zhou, Binbin Toh, Pengyu Zhu, and Lijun Li. KC-Bench: A dynamic interactive benchmark for evaluating knowledge conflicts in LLM agents. *arXiv preprint arXiv:2609.03588*, 2026.

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

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *TACL*, 2023.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *ICLR*, 2024.

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

Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM Journal on Computing*, 22(1), 1993.

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.

Isaac Ong et al. RouteLLM: Learning to route LLMs with preference data. In *ICLR*, 2025.

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

Christos H. Papadimitriou and Michael Sipser. Communication complexity. In *STOC*, 1982.

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.

Vyzantinos Repantis, Ameya Gawde, Harshvardhan Singh, and Joey II Blackwell. How many tools should an LLM agent see? a chance-corrected answer. *arXiv preprint arXiv:2605.24660*, 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.

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

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.

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

Zilu Tang, Qiao Zhao, Gabriel Franco, et al. Do language models track entities across state changes? *arXiv preprint arXiv:2605.30233*, 2026.

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

Jing Wang, Fengzhuo Zhang, Xiaoli Li, et al. Error analyses of auto-regressive video diffusion models: A unified framework. *arXiv preprint arXiv:2503.10704*, 2025a.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

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

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.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. *arXiv preprint arXiv:2403.08319*, 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.

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

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

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

Kaituo Zhang, Zhen Xiong, Mingyu Zhong, Zhimeng Jiang, Zhouyuan Yuan, Zhecheng Li, and Ying Lin. Are tools all we need? unveiling the tool-use tax in LLM agents. *arXiv preprint arXiv:2605.00136*, 2026a.

Yehang Zhang, Jianchong Su, Haojian Huang, et al. Worldlines: Benchmarking and modeling long-horizon stateful embodied agents. *arXiv preprint arXiv:2606.18847*, 2026b. EMNLP 2026.

Yang Zhao, Chengxiao Dai, Mengying Kou, et al. Memorepair: Barrier-first cascade repair in agentic memory. *arXiv preprint arXiv:2605.07242*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Xi Zhu, Ziqi Wang, Kai Mei, Wujiang Xu, Minghao Guo, Bangji Yang, Jiajun Fan, and Dimitris N. Metaxas. Trust or abstain? a self-aware RAG approach. *arXiv preprint arXiv:2605.18792*, 2026.

A THE REGISTERED PREDICTIONS, AND THEIR OUTCOMES

Both predictions below were frozen in the registration reproduced in Appendix I, the first before any data was read and the second after the in-situ measurement but before the ablation was run. Neither fires, and the paper reports the replacement where it is measured.

Prediction 1 (stated before the data was read)**.** $\lambda > \frac{1}{2}$ on the composition subtest: the asserted channel is preferred over the reported one on disagreement, so the observed accuracy of the two-channel structure falls *below* the symmetric average of its channels.

Prediction 2 (registered after the in-situ measurement, before the ablation was run)**.** If the asserted channel damages a structure by displacing grounding, then removing it raises γ, and the accuracy gain is bounded by the grounding headroom $1 - \gamma$ times the gap $\bar{r}_{\text{sel}} - r_0$. In particular, a backbone already grounded at $\gamma \approx 1$ has no headroom, and removing the channel there must buy nothing.

B EVERY CLAIM, ITS EVIDENCE AND ITS VERDICT

The table below is generated from the result files; no entry is hand-typed. It is the index to the whole paper, and a reader who wants only the claim set can stop after it.

Table 4: Every claim this paper makes, with the artefact it rests on and the uncertainty attached to it. ESTABLISHED: the interval excludes the null, or the quantity is a construction rather than an inference. DIRECTIONAL: the sign is supported across cells, the magnitude's interval is not. BOUNDED: reported as a bound. WITHDRAWN: stated in a previous version and retracted here. Core rows carry the positive chain (blind average, availability, $K{=}3$ subtraction, exits, directive); DIRECTIONAL/BOUNDED rows are scope checks reported once, and open probes make no row at all. Generated by `analysis/a1_claims.py`; no entry is hand-typed.

claim	evidence	n	effect	interval / basis	verdict
The compressive channel holds the queried quantity on a minority of decisions, and on *none* of the attribute subtests	§4.3	560	114/560 = 20.4%	RET 0/224, CMP 0/168	ESTABLISHED
Where it does hold the quantity, the arbiter follows it	§4.3	112	90.2% item agreement; *causally* 100.0% (114/114) follow a planted wrong value	Wilson [83.3, 94.4]; 0 ignored; sham 58.2%; causal Wilson [96.7, 100.0], manipulation check 100%, accuracy under injection 0%	ESTABLISHED
The composite arm's distance to $\max_k r_k$	§E.9	560	+3.75 points	[-2.05, +10.62]; 57% in one cell	BOUNDED
Grounding is a function of the loop's exit, not of the arm	§4.2	560	91.5% / 59.3% / 1.9%	Wilson [89,94] / [48,69] / [0,10]	ESTABLISHED
Removing the compressive channel makes the agent *commit* more often	§4.2	1120	+9.1 points	90.5% → 99.6%; replicated 89.5 → 99.8	ESTABLISHED
… and thereby better grounded	§4.2	1120	-4.3 points	sign reverses once commitment is conditioned on; [83,89] vs [79,85]	WITHDRAWN
The selectivity term $\mathcal{C}$ is null	§E.17	366	+0.019	[-0.029, +0.066]; equivalence bound 5.8 pts	BOUNDED
Removing the compressive channel costs capability on average	§E.12	560	-6.12 points	[-17.91, +5.03]; 4/14 cells, sign $p = 0.031$	DIRECTIONAL

continued on the next page

Table 4, continued from the previous page

claim	evidence	n	effect	interval / basis	verdict
Removing the compressive channel costs capability on average, *fairly budgeted* (the realloc arm)	§5	560	+3.67 points	registered thresholds ($\|3\|$ withdraw, 5 confirm) did not decide; 605 retries recorded, spend 98.2% of control	DIRECTIONAL
… and on tracking the cost is the channel's absence, not its budget	§E.11	80	budget return recovers 1.19 of 27.38 points	TRK removal +27.38 → +26.19 under the fair comparison; per-subtest reading of the registered run, whose brackets covered only the pooled number: a constructive decomposition of the registered run, not a registered subtest inference	DIRECTIONAL
Appending a redundant weaker channel *subtracts* capability, with a dose response (the $K{=}3$ grid)	§5 item (iii)	560	-13.39 points (quarter shard), -4.78 (half)	penalty scales with weakness on 7/7 backbones (registered ≥ 5); scraped rate rises 13.8% → 21.2%; exit-conditioned, the committed-only drop is -17.68 / -5.66 with the committed dose response on 7/7 backbones, and the exact decomposition puts -13.92 of the pooled -13.39 in committed accuracy against -1.26 in rounds; registered prediction K2, K3 refuted	ESTABLISHED
Two retrieval channels of identical form are arbitrated symmetrically	§5 item (iii)	254	shard share 0.402 of same-form retrieval attributions	registered band $[0.35, 0.65]$ despite a six-fold reliability gap; the corollary's premise, measured in the regime where it can be (registered prediction K4)	ESTABLISHED
Rendering the compressive channel's facts as citations does *not* move the state channel (the form account, tested and refuted)	§5 item (ii)	168	state-following 0.0% under citation form (0/168 decisions) against 2.6% for the JSON control	registered prediction was $\geq 20\%$; below the 15% refutation line; the false-citation arm sits at 0.9% and the no-retrieval subset is empty, so C2 and C5 fail together, the registration's failure condition; accuracy unchanged (47.17 → 46.13); scoped: the witness schema stores no queried attribute and the composition subtest holds the quantity on 0 of 168 decisions, so the refutation is that presentation cannot rescue a silent channel; whether citation form re-weights a speaking channel is untested	ESTABLISHED
Reserving the exit removes the scraped population	§E.12	560	59 → 0 decisions	by construction (verified on 400 scripted decisions)	ESTABLISHED
… and thereby raises accuracy	§E.12	560	+2.10 points	[-2.30, +7.81]; 7/6/8, sign $p = 1.000$	WITHDRAWN
No decision follows the compressive channel	previous version	112	—	instrument at chance on the state side; see row 2	WITHDRAWN
An arbitration directive's *content* changes what the arbiter gets right (against a length-matched swapped control)	§4.5	560	+6.43 points	[+0.87, +12.68]; permutation $p = 0.0001$; 14/3/4	ESTABLISHED
… and the typed directive beats the undirected control	§4.5	560	+2.28 points	[-2.72, +7.88]; 8/7/6, sign $p = 1.000$	BOUNDED

continued on the next page

Table 4, continued from the previous page

claim	evidence	n	effect	interval / basis	verdict
The directive helps where the classifier supplying the type is right and *hurts* where it is wrong	§4.5	560	TRK 100% router, +8.3; CMP 54% router, -4.3	three subtests; the sign follows router accuracy, not backbone	ESTABLISHED
A reliability marker is read and obeyed and buys no accuracy, *under a comparison the manipulation cannot trip*	§E.14	1120	+1.19 points (marked 56.25 vs same-form placeholder 55.06)	registered bands $\lvert\Delta\rvert \leq 2$ null, $[2, 5)$ inconclusive, ≥ 5 recover; the void condition sits on the label's own cost, identical by construction; realised spend +2.8% reported as an outcome (registered prediction B1)	ESTABLISHED
The compressive channel earns its place most where the wall is tightest; its removal cost decays with B	§E.9	1008	gap -19.79 at B=1024 → -12.87 at B=8192 (move +6.92, opposite sign)	registered direction refuted ($\geq$ 5-point move was the theorem's sign; opposite = refuted); helped-cell count flat 4/4/1/4 over four budgets; medians monotone -14.06 → -4.17; 84 cells, two arms, three added budgets plus the measured grid (2-cell fourth point)	DIRECTIONAL
Removing the compressive channel costs capability on average, *with its budget fully returned* (the unbounded reallocation)	§5 item (iii-b)	560	+5.75 points (handicapped same-batch +5.80)	registered F2 band ($\lvert 3\rvert$ withdraw, $\geq$ 5 confirm): confirm; F1 passes outright (no bound-limited decision vs 88 in the registered run, spend 100.5% of control); on tracking the return recovers nothing (+26.78 → +26.78); F4 does not fire (-5.75); D-R9: excluding the flagship's drifted composition cell (21.9-point cross-batch swing) the gap is +3.96, still positive but below the confirm line	ESTABLISHED
The penalty does not scale with the number of added channels: it *decays* (K=4/K=5 dose extension)	§5 item (iii)	1120	Δ -19.64 (K=3q) → -14.41 (K=4) → -10.08 (K=5); committed -16.65 → -13.35	registered D1 monotonicity fails at its bottom band (0/7 backbones, registered $\geq$ 5): the K^* reading is withdrawn to two-point direction, no curve; added shards read and followed (share 38–74% of same-form retrieval attributions), so the decay is not invisibility; D-R8 scope: the fourteen tracking cells never presented a third block and are excluded, the dose read is on retrieval and composition; scraped rate 13.8 → 26.1 → 23.8% (the registered K=5 risk fired at K=4); 63 cells, 3 arms, 7 backbones, 3 subtests	DIRECTIONAL

C Related Work, in Full

Capability under a budget, and access structure. The hypothesis we extend holds that representational convergence does not entail capability convergence, and that under a fixed per-token inference budget capability converges toward architectures holding both a compressive state channel and a scalable verbatim index (Chen et al., 2026b), itself positioned against the representational-convergence claim of Huh et al. (2024). Its perceptual instantiation prices four walls for embodied streams and reports the ablation we start from (Anonymous authors, 2026c). Both treat a channel as a resource with a price; neither models what happens when two channels return different answers, and this paper is about the term that omission drops.

Agent memory as a system, and as a rate-distortion problem. A large 2026 literature builds memory for long-horizon agents: streaming token compression under a fixed per-frame budget (Chen et al., 2025), dynamic and fixed-budget memory banks (Song et al., 2026), compression for streaming audio-visual models (Sun et al., 2026), governance and repair of evolving stores (Lam et al., 2026; Zhao et al., 2026), hierarchical temporal indexing (Chen et al., 2026a), spatio-temporal memory for embodied agents (Rasheed & Kabtoul, 2026), and parametric consolidation of experience (Guo et al., 2026), with surveys mapping the space (Omri et al., 2026; Jiang et al., 2026). The closest unifying treatment reads KV eviction, prompt pruning, state bounding and consolidation as one rate-distortion decision under a budget, with a layer-agnostic lower bound (Colaco & Lahjouji, 2026). It prices *what to discard*, and because distortion is symmetric it cannot price the cost of retaining an item that is wrong, the quantity Theorem 1 makes central.

What language models do instead of tracking. Our account predicts that forcing a frozen backbone to maintain a running reduction substitutes a fabricated channel rather than adding one, and there is independent mechanistic evidence for the substitution: these models do not track world state incrementally but aggregate in parallel at the final token once the query is evident, with deletion implemented as a fragile global suppression tag whose predicted failure modes are confirmed behaviourally (Tang et al., 2026). Benchmarks for stateful embodied agents report the same difficulty from the task side and propose visibility-aware memories rather than a bound (Zhang et al., 2026b).

Compounding error, and what fixes it. That a per-step error rate compounds over a horizon is standard; the rigorous version shows behaviour cloning degrading quadratically in the horizon and DAGGER recovering a linear rate by changing the distribution on which the operator is queried (Ross et al., 2011), recently revisited for language-model agents (Li et al., 2026), with an analogous memory-bottleneck result for autoregressive video (Wang et al., 2025a). All repair the operator, by training or by architecture. Our setting forbids that (the backbones are frozen and served through an API), and periodic re-derivation (§3) is the training-free counterpart, truncating the horizon by consulting ground truth instead.

Combining several fallible predictors. That an ensemble can exceed each of its members is classical (de Condorcet, 1785; Dietterich, 2000), and it is why Theorem 1 is stated under non-coinciding errors (Definition 4). The ensemble effect needs errors that disagree more than answers agree and reliabilities that are comparable; a two-channel structure in which one channel has been degraded by compounding satisfies neither. We treat this literature as a boundary on our claim and report where the boundary lies.

Maintain or recompute, in the fields that settled it. Whether a derived quantity should be updated incrementally or recomputed on demand is the subject of incremental view maintenance, including which views are self-maintainable (Gupta & Mumick, 1995; Gupta, 2009), and the space to maintain a running distinct-count over a stream is settled with matching bounds (Alon et al., 1999; Indyk & Woodruff, 2003). Both are reference points rather than competitors: their operators are correct by construction, so their question is cost or space, while ours is which of two claims to believe.

The two companion measurements. Two results in this programme supply numbers Theorem 2 leaves as unknowns, both produced after this paper's theory was fixed. On a token-metered multimodal interface a grounded check of one stored claim costs approximately what producing the claim cost, because both calls transmit the observation and the observation is most of the bill (Anonymous authors, 2026b), which turns the admission criterion of §2 from a threshold to be engineered past into one violated generically, and is why the escape of §3 is presentation rather than verification. Separately, the share of a benchmark's items settled without consulting the modality is computable from annotations before anything is run (Anonymous authors, 2026a), bounding how much of a channel's apparent contribution was ever about the channel. We cite both as inputs, not as support.

Access loops, routing, and the round budget. The architecture under test is an agentic read loop, and the loop is where the round budget is spent. Agents that interleave reasoning with actions (Yao et al., 2023), revise their own outputs (Shinn et al., 2023) or call tools (Schick et al., 2023) all pay a per-step cost usually treated as overhead rather than as a term in a capability bound. Routing by predicted difficulty (Ong et al., 2025; Chen et al., 2023) prices the choice of *whom* to ask; retrieval-

augmented generation prices the choice of *what* to read (Lewis et al., 2020; Karpukhin et al., 2020; Gutiérrez et al., 2024; Packer et al., 2023). What this paper adds is what happens when two readable channels return different answers and the loop must commit within a fixed number of rounds: a constraint the test-time-compute literature makes central for *depth* (Snell et al., 2024; Muennighoff et al., 2025; Wei et al., 2022) and does not, to our knowledge, connect to arbitration between channels. Long-video systems that plan which segment to re-read (Wang et al., 2024; Zhang et al., 2024; Wang et al., 2025b; Rege et al., 2026; Shen et al., 2025) face exactly this loop and report accuracy rather than exits, which §4.2 suggests is where a substantial effect can hide.

Why the round wall matters more than it looks. That a bounded number of rounds is a real constraint has a theoretical basis we borrow rather than extend: round-limited communication is strictly weaker than unlimited communication for pointer-chasing-style problems (Papadimitriou & Sipser, 1982; Nisan & Wigderson, 1993), and the expressive limits of bounded-depth computation with a scratchpad are now well mapped (Feng et al., 2023; Merrill & Sabharwal, 2024; 2023; Liu et al., 2023a; Barrington, 1989). The parent paper turns that into a round wall for budgeted embodied perception (Anonymous authors, 2026c). What §4.2 adds is empirical: when the rounds run out, an evaluation harness does not record a refusal; it records whatever the agent happened to be writing, and scores it. The round wall is therefore also a mechanism by which a channel that asserts nothing false can cost a system its answers.

Judging, and the reliability of the instrument. Since the escape of §3 turns on making reliability visible, it is worth noting that measuring reliability is itself unreliable. An audit of a shipped rubric judge finds failures a variance framing cannot express, because they change what is measured rather than how noisily (Anonymous authors, 2026b); a companion result tracks how errors enter and persist in a stored representation (Anonymous authors, 2026a). This is why no headline number here comes from a model-scored judgement, and why the census of §4.2 is built from string containment against recorded prompts.

Position and labels, and why neither explains the null. Two published effects would explain our central null without any of our machinery. Language models use long contexts unevenly (Liu et al., 2023b), and model-based evaluators are sensitive enough to candidate order that exchanging two responses can flip the verdict (Wang et al., 2023; Zheng et al., 2023); the standard remedy there, swapping the positions, is structurally the manipulation of our §E.14 swapped arm. We measured order rather than assuming it away, and the measurement does not license the word *null*: the answer follows the retrieval block on 18/18 decisions where it is nearest and 857/1115 where it is furthest, and eighteen decisions at the near end cannot separate a twenty-point recency effect from none. That observational contrast has since been closed by intervention (§4.2): forced-order arms swap the block positions with contents identical, and λ does not move, so the contrast was endogeneity rather than recency. The label is the cleaner negative, on $2{,}177$ paired decisions (§E.14), and the budget-matched rerun holds it under a comparison the manipulation cannot trip. What separates the two accounts is that Lemma 1 names a property of the *channel* rather than of the prompt, and §E.17 shows that the weight λ tracks what a channel *holds*: λ_{state} is exactly zero on every retrieval and composition cell and non-zero on every tracking cell, a separation no manipulation of presentation reproduces.

Adding a component can hurt: the phenomenon is documented, the mechanism is what is missing. That capability is not monotone in the set of things you give an agent has been measured several times, and we should cite it rather than present the phenomenon as ours. Liu (2026) runs a full factorial over all 2^5 subsets of five scaffolding components on two corpora and finds the all-in system consistently suboptimal (a single-tool agent beats it by 32% on HotpotQA and a three-component subset by 79% on GSM8K), with 183 of 325 submodularity violations, which is the same shape of statement as our Corollary 1 arrived at by enumeration over a component set rather than by construction. Zhang et al. (2026a) show that under semantic distractors tool-augmented reasoning need not beat native chain-of-thought, and Repantis et al. (2026) treat shortlist depth as the object of evaluation and insist, as we do, on correcting for chance before calling a depth better.

We claim no priority on the phenomenon. We do claim something the phenomenon does not supply: the *term*. Zhang et al. (2026a) is the nearest, and it does decompose, but its factorised intervention separates what the *apparatus* costs (prompt-formatting cost, tool-calling protocol overhead, the gain from executing the tool), which is orthogonal to ours. Theorem 1 decomposes what an *arbiter* attains

when two channels answer the same question and disagree: a presentation-weighted average of the channels' reliabilities plus a selectivity term that vanishes exactly when the arbiter is reliability-blind. The average is the reason more channels can hurt: an arbiter that is reliability-blind is dragged toward the mean of what it is shown, not lifted to $\max_k r_k$. The selectivity term is also measurable, and §E measures it: its departure from the channels' base rates is statistically indistinguishable from zero (7 of 14 cells signed positive, which is what a coin does). The $K{=}3$ grid (§5 item (iii)) adds the dose response the enumeration studies did not run: the penalty of an appended channel scales with the weakness of what it carries on 7/7 backbones, while two channels of identical form are arbitrated symmetrically (0.402). That is the subtraction's mechanism, measured rather than inferred from curves.

Channels that pay their price and are ignored. An empirical literature keeps rediscovering, one benchmark at a time, that multimodal evaluations do not require the modality they are named for. Question-only baselines were already competitive on embodied question answering (Anand et al., 2018); text-only models trail multimodal ones only narrowly on the spatial category of a widely used embodied benchmark (Majumdar et al., 2024); a bag-of-words baseline that discards temporal structure nearly solves a long-horizon spatial-recall benchmark, including on multi-hour video (Udandarao et al., 2025), whose parent proposes the benchmark and reports collapse beyond an hour of streaming (Yang et al., 2025); a visual-dependency diagnostic separates accuracy from grounding across twenty models and finds frame diversity, not temporal order, supplying most of the benefit (Lee, 2026); and a controlled study of the modality gap finds text-only inputs best and vision-only worst (Xu et al., 2026). On the design side, the inference-optimal trade-off for a vision-language model at a fixed budget is very few visual tokens with the largest language model that fits (Li et al., 2024). Our reading is that these are not all benchmark defects, and not all architecture guidance: a channel that cannot be checked within the budget is one a rational allocator must down-weight, so some of what is reported as blindness is the arbitration wall seen from the task side. We stated that as a hypothesis with a testable consequence (making a channel cheaply checkable should reverse the down-weighting), and Lemma 1 is why the test is harder to set up than it sounds. A channel cannot be made cheaply checkable by presenting it better. The facts a compressive channel exists to hold are reductions, a reduction has no witness of size one, and what an unverifying arbiter can act on cheaply is a witness. So the down-weighting is not a correctable presentation failure for this class of channel; it is what the channel's own value proposition costs it.

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 (this paper);
- a model-free audit of what a benchmark's own annotations can see, which calls no model at all (Anonymous authors, 2026a);
- a measurement of what it costs to *check* a perceptual claim on a metered interface, whose population is caption probes rather than agent decisions (Anonymous authors, 2026b).

This paper is the second. Its outcome is not an outcome the first measured and could not have been: those per-item records carry no per-call trace, so the exits, the grounding rate and the provenance census are computed from a grid re-run with recording enabled, a grid released with *this* paper. Conversely this paper reports no verification *price*; that magnitude belongs to (Anonymous authors, 2026b) and is cited, not derived here. Nothing else rests on an unpublished companion. The corpus, the backbones, the budget, the judge and the access structure under test are shared with (Anonymous authors, 2026c) and are restated in §3 and §E rather than by reference, so a reader can reconstruct the

line of work	operates on	no handle on
access-structure grids: capability vs. channels and prices (Chen et al., 2026b; Anonymous authors, 2026c)	the channel set, B, the walls each channel crosses	λ_k: holds both channels, never asks which claim wins when they disagree
routing and model selection (Ong et al., 2025; Chen et al., 2023)	which *model* to ask; one predictor at a time	the arbitration between two simultaneous claims inside one decision
RAG and memory systems (Lewis et al., 2020; Karpukhin et al., 2020; Gutiérrez et al., 2024; Packer et al., 2023)	what to read; retrieval quality as recall	$\lambda_k r_k$ as a weighted pair: the read is followed or ignored, never arbitrated
compounding / DAGGER (Ross et al., 2011; Li et al., 2026)	r_s decay in the horizon; repairs the operator	the frozen-backbone setting (no training); and λ: a decayed channel still captures a λ share, which is where its decay becomes harm (Thm. 1)
ensembles (de Condorcet, 1785; Dietterich, 2000)	coinciding errors, comparable reliabilities	the regime boundary: one degraded channel violates both premises (Def. 4); boundary measured in §5 item (7)
incremental maintenance / sketching (Gupta & Mumick, 1995; Alon et al., 1999)	cost and space of a *correct* operator	λ: their operators are correct by construction, so there is no claim to arbitrate
long-video agents (Wang et al., 2024; Zhang et al., 2024; Wang et al., 2025b; Rege et al., 2026; Shen et al., 2025)	the loop and what it re-reads	exits: they report accuracy, and §4.2 shows where an effect can hide in that
test-time compute (Snell et al., 2024; Muennighoff et al., 2025; Wei et al., 2022)	depth of compute; the round budget as a resource	arbitration between channels under a fixed round budget

Table 5: Each line of prior work, the term of Eq. (2) it operates on, and the term it cannot touch. The right-hand column is the paper's delta, stated term by term.

configuration from this paper alone; every number reported here is computed from per-call records produced by the instrumented re-run and released with it, none copied from a companion's tables; and the compounding bound of Lemma 2 is proved here (Appendix G); it is in any case the standard sequential-decision compounding argument (Ross et al., 2011), cited to (Anonymous authors, 2026c) for the form it takes in this setting rather than for its validity. The companions are cited for priority and scope: (Anonymous authors, 2026c) for the architecture this paper instruments and the walls it prices, (Anonymous authors, 2026a) for an annotation-only instrument whose numbers we share none of, (Anonymous authors, 2026b) for the verification price this paper explicitly declines to derive.

What each line of work can and cannot reach in Eq. (2). Table 5 condenses the paragraphs above into one view: for each line of prior work, the quantity of Eq. (2) it operates on, and the term it has no handle on. The point of the table is the right-hand column: every prior line either fixes the *channels* and asks about prices (the access-structure grid), or assumes arbitration away (routing ensembles pick one predictor), or repairs the *operator* (compounding, DAGGER), or bounds what annotations can see. None of them measures λ, because none of them has two claims that can disagree inside one decision, which is the object this paper adds and the rest of the paper measures.

D METHOD, IN FULL

Section 2 leaves two quantities to be supplied by an implementation: the weights λ, which presentation fixes, and the verification price v_k, which the architecture fixes. This section specifies how we measure the first and two training-free ways to change it, both with their prices stated.

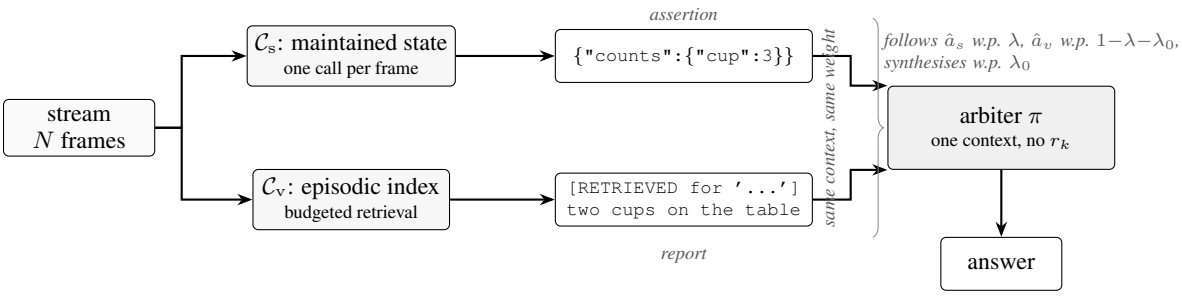

Figure 4: Where the two claims meet. Both channels pay a price and cross a wall, and both insert text into one context. Nothing in the interface carries either channel's reliability, because the formalism the interface implements has no slot for it, so the weights $(\lambda, 1-\lambda-\lambda_0, \lambda_0)$ of Eq. (2) are set by the *form* of the two insertions: an assertion on one side, a report of an observation on the other.

D.1 The architecture under test, and where the claims meet

We take the access structure of the parent system (Anonymous authors, 2026c) unmodified: a maintained scene state $\mathcal{C}_s$, updated by one bounded backbone call per keyframe; a verbatim episodic index $\mathcal{C}_v$ over the same stream, searched with a budgeted retrieval; and a query-conditioned router that partitions a per-decision budget B into (B_s, B_r, B_c). A decision proceeds in at most R rounds, and in each round the backbone emits one action: read the state, retrieve, or answer.

The arbitration interface is the assembled context. A state read inserts the serialised reduction; a retrieval inserts a description of the frames returned, produced by a fresh read of pixels. Both are text in one context window, and the backbone that must answer sees no reliability information about either. This is a faithful implementation of Definition 1 with a reliability-blind arbiter, and it is the configuration in which the 62.5-point ablation was measured. Nothing about it is careless: the formalism it implements has no slot for r_k, so an implementation must either invent one or omit it, and omitting it is the choice that adds no unstated mechanism.

The asymmetry the theory predicts is visible in the two insertions. The state's claim is an assertion, a JSON object stating that a count is three. The retrieval's claim is a report, attributed to the query that fetched it and derived from an observation. Definition 5 is about which of those two forms an arbiter follows when they disagree.

D.2 Measuring the assertion premium

λ is a statistic over disagreement events, so the measurement needs, per decision, the claim each channel made and the answer the arbiter returned. None of that survives in a completed run: a per-item record carries the final prediction and the token ledger, and the per-channel claims exist only inside the assembled context, which is consumed and discarded. We therefore re-execute the same cells with the round prompts recorded, changing nothing else, and write them to a separate location so that no measured cell can be overwritten (§E).

From each recorded decision we extract the state's claim from the serialised reduction it inserted, the retrieval's claim from the description it inserted, and the returned answer. A decision contributes to λ when both channels were read and their claims about the queried quantity differ. Claims are compared after normalisation on the quantity the question asks for, and a decision in which either channel is silent or the two agree is excluded, since neither case is an arbitration.

Three conventions are fixed in advance because each could otherwise be chosen after seeing the result. A decision in which the arbiter returns an answer matching *neither* claim counts as following neither and is reported separately rather than folded into λ; the per-item unit is the decision, not the round, so a state read in round one and a retrieval in round three form one arbitration; and λ is computed per backbone before it is pooled, so a single backbone cannot dominate the pooled figure through a larger count of disagreements.

D.3 ESCAPE 1: MAKE THE WEIGHTS TRACK RELIABILITY

Theorem 1 says accuracy is $\sum_k \lambda_k r_k$ and that λ comes from presentation. The first escape therefore changes presentation rather than either channel. Each claim is inserted with an explicit reliability estimate and the arbiter is instructed to prefer the higher, which converts blind arbitration into aware arbitration for the claims at hand and, by Proposition 1, moves accuracy from the weighted average toward $\max_k r_k$.

The estimates are cheap and neither is fitted. For the maintained channel, Lemma 2 gives $\hat{r}_s = (1-\hat{\epsilon}_u)^N$ from the number of updates N, which the agent knows, and a per-update error $\hat{\epsilon}_u$ measured once per backbone offline. For the index, $\hat{r}_v$ is the retrieval precision at the realised budget, also measured once per backbone. The run-time cost is the tokens of two short markers, which is $O(1)$ per claim and, at the budgets here, under one percent of B.

This is the intervention that separates our account from a resource account. A resource account predicts no change: the bits, the tokens, the rounds and the walls crossed are all identical, and only the wording differs. Theorem 1 predicts a large change, concentrated on exactly the decisions where the two channels disagree.

We ran it, and the change is not there (§E.14). We leave the derivation above standing because it is what the theorem implies and because the experiment is only interpretable against it, but the reader should carry the result forward from here: accuracy moves -0.22 points, 513 of 560 paired decisions are ties, and a marker carrying the two numbers *exchanged* does no worse than the correct one (-0.12). The same run also shows that the marker is not ignored (it moves which channel the loop reads, in the direction it points), so what fails is not the delivery of the information but the assumption that acting on it is what was missing. Escape 1 as proposed here returns a **null** on its own test: a null on a run that exceeds the registered spend criterion it made a void condition, so we call it a null and not a refutation (§E.14).

D.4 ESCAPE 2: BOUND THE HORIZON BY RE-DERIVATION

The second escape attacks r_s rather than λ. Every T keyframes the maintained reduction is discarded and recomputed from the index by aggregation over retrieved records, rather than by asking the backbone to carry it forward. The exponent of Lemma 2 then runs over T instead of N, so $r_s \geq (1-\epsilon_u)^T$ regardless of stream length, and Corollary 2 fixes the smallest T that meets a target reliability together with the condition under which the aggregation is worth its price.

Two properties matter for the setting. It is training-free, which it must be, since the backbones are frozen and served through an API. And it consults ground truth rather than improving the operator, which is what DAGGER does by retraining on the states the learner visits (Ross et al., 2011); the mechanism is different and the effect on the horizon is the same.

D.5 THE LOOP, AND THE THREE WAYS IT CAN END

Everything in §4.2 depends on a detail of the access loop that is easy to overlook and that no aggregate score exposes, so we state it as pseudocode and define the resulting classification.

Definition 7 (Answer provenance)**.** The *provenance* of a scored answer is the exit of Algorithm 1 that produced it: ANSWERED (line 8), FORCED (line 19) or SCRAPED (line 21). Provenance is read from the recorded call log (an ANSWER action in a round output, a call tagged `forced_answer`, or neither), and is determined by the same parser the loop dispatched on, not by a re-implementation.

The last clause is not pedantry. Backbones wrap actions in markdown emphasis and append sign-offs, and the loop's own parser strips both and scans upward for the first line naming a verb. An anchored regular expression written independently for the analysis disagreed with it on enough rounds to change the census materially, so the analysis imports the runner's function instead. A census of what an agent did has exactly one correct parser, and it is the one the agent's behaviour was determined by.

Definition 8 (Channel request share)**.** The *request share* of channel $\mathcal{C}$ for a backbone is the fraction of its channel-reading actions that name $\mathcal{C}$, i.e. $|\{\text{STATE}\}|/(|\{\text{STATE}\}| + |\{\text{RETRIEVE}\}|)$ for $\mathcal{C}_\text{s}$. It is a property of the agent's policy and is measurable without reference to any answer.

Algorithm 1 The access loop under test, with its three exits marked. $\mathcal{C}_{\mathrm{s}}$ is the compressive channel, $\mathcal{C}_{\mathrm{v}}$ the verbatim index, B_c the per-round conversation budget, and R the round wall. Only exit (I) produces an answer the agent chose to give.

Require: question q, budget B split into (B_s, B_r, B_c) by the router, rounds R

```
ctx ← ε
for r = 1 to R do
    if B_c cannot afford another round then break          ▷ silent exit; no answer yet
    end if
    u ← BACKBONE(LOOPPROMPT(q, ctx, r, R))
    α ← LASTACTION(u)          ▷ scans upward for a line naming a verb, emphasis stripped
    if α = ANSWER: y then
        return y                                           ▷ exit (i): ANSWERED
    else if α = STATE: f then
        ctx += the state rendering, or a placeholder if C_s is ablated or B_s is spent
    else if α = RETRIEVE: q′ then
        ctx += a fresh reading of the top keyframes for q′, charged to B_r
    else
        ctx += [NOTE] and the tail of u                    ▷ the model narrated instead of acting
    end if
end for
if B still covers a short call then
    return BACKBONE(q, ctx, "final short answer only")     ▷ exit (ii): FORCED
end if
return the last non-bracketed line of ctx                  ▷ exit (iii): SCRAPED
```

Definition 8 matters because the central equation presumes two claims to arbitrate between. Obtaining each claim costs a round, so an agent that never requests a channel faces no arbitration at all; and a blind arbiter that never requests a channel is observationally identical to one that requests it and discounts it, while having a completely different remedy. §4.2 separates them.

D.6 What the ablated arm is charged

Two handicaps on the $-\mathcal{C}_{\mathrm{s}}$ arm are consequences of the implementation rather than choices, and both run against the ablation.

The budget is not reallocated. The router computes (B_s, B_r, B_c) before the ablation flag is consulted, so the ablated arm is capped at the same B_r and B_c as the full system while its B_s is simply unspendable. It runs on roughly three quarters of the usable tokens.

A request for the absent channel still costs a round. When the ablated arm issues STATE, Algorithm 1 appends a placeholder and continues to the next round. The round is consumed. This one is measurable rather than assumed, and §4.2 counts it: the ablated arm pays it 367 times across the ladder, on essentially every item for the backbones that read the compressive channel most.

Both handicaps make the reported effect a lower bound. We keep them rather than repairing them, because repairing them would change two things at once and the comparison would stop being an ablation of the channel.

D.7 What is charged to whom

Both escapes are priced against the same B as the channels, as Theorem 2 requires. The reliability markers are charged to the composition channel, since they are part of the arbiter's context. Re-derivation is charged to the retrieval channel, since it is an index aggregation. We report the realised split rather than the planned one, so that a claim about a cheap escape cannot be made on a decision that quietly overspent.

E EXPERIMENTS AND ANALYSIS, IN FULL

The results section follows the order Table 4 set at the head of the paper: every claim this paper makes, the artefact it rests on, and the uncertainty attached to it, including the claims a previous version made that are withdrawn here, were on the table before any of the sections below; each subsection below fills in the artefact column of one or two rows.

Scope contract. The core of this paper is a short positive chain: a blind arbiter attains a weighted average rather than the best channel (Theorem 1, §E.17); the compressive channel is followed where it holds the queried quantity, causally (§4.3); appending a weaker same-form channel subtracts capability through arbitration, not budget (§5 item (iii) with the committed-only decomposition); the harm is delivered through the loop's exits (§4.2–§E.11); and a directive at the arbitration point moves capability where its classifier is right (§4.5). Everything else in this section is a *bounded scope check* reported once where it belongs (budget sweep, $K{=}4/K{=}5$ extension, ϵ_u/N^*, answer-set grouping, marginal slopes), and each is stated with the verdict its registration returned, including refutations. Open probes that no claim depends on are confined to §5 as future work; they are not run here and they do not extend the paper's chain.

E.1 WHAT IS MEASURED, AND WHAT IS REUSED

The configuration is shared with a companion architecture paper (Anonymous authors, 2026c) and unchanged from it, and we describe it here in full rather than by reference, so that this paper can be read and checked on its own. The corpus is SEW-BENCH, a licence-free synthetic walkthrough: four episodes of $N{=}300$ rendered keyframes, each frame carrying 16 objects with two pixel-readable attributes apiece drawn from a $K{=}8$ alphabet, and 80 questions split $32/24/24$ into a retrieval, a tracking and a composition subtest that witness three different resource walls rather than three difficulty levels. It is a schematic 2-D render and not egocentric video; what it tests is access structure under a token budget, which is what the claims here are about, and it tests nothing about natural-scene perception. The backbones are seven open multimodal models from 3B to 31B served through one API; the per-decision budget is $B = 4096$ tokens. The census layers of §4.1–§E.7 run on the composition subtest, 24 items over seven backbones for 168 decisions, because that is the subtest on which the ablation that motivates this paper was measured. The grids of §4.2–§E.12 run on all three subtests, retrieval, tracking and composition; §E.13 repeats the same five arms on a second corpus with no images in it. Two arms of those grids are additionally run at three further budgets ($B \in \{1024, 2048, 8192\}$, tracking and composition, all seven backbones; this is the registered budget sweep of §E.9), so the wall's behaviour in B is tested and not assumed from the single-budget grid. Where a claim is scoped to the composition subtest we say so at the point of use. Every item is scored on $[0, 100]$ by the benchmark's own judge; scores are 0 or 100 on all but a handful of items, and we treat ≥ 50 as correct and report the exceptions.

Nothing is re-measured that was already measured. Equation (2) is a statement about which claim a decision followed, and the three arms needed to evaluate it already exist as per-question records (the two-channel structure, the state channel alone, and the index channel alone, all on the same 168 items), which are released with *this* paper and enumerated in Appendix G, so every census number below can be recomputed without obtaining any other paper. What those records do not contain is the claim each channel inserted into the assembled context, because the context is consumed at answer time. So the analysis has two layers, and they are kept apart.

The first layer is a *census* over correctness patterns. It uses only released scores, introduces no new execution, and is exact in what it counts. Its limitation is stated rather than buried: a single-channel arm resolves its claim with one input, so its prediction is that channel's claim as that arm renders it, which need not be the claim the two-channel context carried. The second layer records the round prompts on a re-execution of the same cells and reads the claims in situ; it is written to a separate location so that no measured cell can be overwritten, and its means are routed away from the file the tables are built from.

E.2 THE SAME ARM, RUN THREE TIMES

Every effect in this paper is a difference between two arms executed once each, and a referee was right that the paper gave no way to tell such a difference from the same arm run twice. The noise

floor it did quote, Limitation (4a), comes from a *generation* change, which rebuilt the retrieval index and therefore answers a different question. What bounds an arm contrast is repetition with nothing changed.

Three repetitions exist and had not been subtracted. Four phases of this project each re-executed the ASP control inside their own batch rather than reusing one from another batch, because a cross-batch control crosses a provider-nondeterminism boundary that had already moved a cell by four points once (`scripts/run_grid.py`). That discipline was adopted to protect the contrasts; its by-product is a control arm executed four times at identical flags: three on one host and embedder, and a fourth after the host change that notes/`deviations.md` D-A1-v5 recorded in advance. Nothing below is a new run. The arms are not re-scored, only differenced (`analysis/a1_repeat.py`).

spread of ASP against itself	3 same-environment repeats	+ the cross-environment one
per cell, median SD / max range	2.4 / 20.8	— / 34.4
per backbone, max SD	4.7	6.1
pooled over the grid, SD	0.94	—
pooled, largest repeat-to-repeat gap	1.65	—

Two things follow, and they point in opposite directions. **Pooled numbers are solid and single cells are not.** The pooled mean of the control lands in $[56.07, 57.72]$ across three executions, a spread of 0.94 points, and no two repeats differ by more than 1.65. A single cell, by contrast, moves by up to 20.8 points for no reason at all. So a pooled contrast above roughly two points is larger than this arm's own repeat noise, and *no* claim should be read off one cell, which is the discipline §E.9 already applied to itself when it reported that most of the measured wall sits in a single cell, and the reason that finding is filed as a bound rather than a result. The ablated arm is steadier still (0.13 points over two repeats, largest gap 0.18): it makes fewer calls, so it has less to vary.

Against this floor the paper's four reported contrasts separate cleanly into two groups. The directive's content ($+6.43$) and the channel ablation (-6.12) are several times the largest null difference. The typed-versus-undirected gap ($+2.28$) and the reserve arm's accuracy ($+2.10$) are close to it, which is the same verdict their intervals give in Table 4, reached independently: BOUNDED and WITHDRAWN. The two instruments agree about which results this grid can carry.

And the environment change did not move the control. D-A1-v5 registered, before the typed-arbiter runs, that the new host's rebuilt CLIP embedder sits inside the retrieval loop and that the re-run control would be compared against the published one, with any disagreement larger than the re-sampling floor reported as an environment effect. It is reported now. The environment inflates per-cell movement by roughly half again (max range 34.4 points against 20.8): a real effect, and the reason §4.5 declines to score the one registered prediction that would have compared across it. But the *pooled* control comes back at 56.92, inside the same-environment band $[56.07, 57.72]$. The within-batch contrasts of §4.5 are therefore not an artefact of the host, and we report that having committed to reporting the opposite had it come out the other way.

E.3 THE ARBITRATION CENSUS

Table 1 classifies every item by which single-channel arm answered it, and reports how often the two-channel structure answered it.

Four readings, in the order of how little theory each requires.

The loss is visible without any theory at all. On the 66 decisions the index channel answered alone, the two-channel structure answered 41%: 27 decisions were lost by adding a channel. No allocation argument is available, since the state channel's budget share is not reallocated in its absence and a state action that returns nothing still consumes a composition round, so the arm that wins is the handicapped one.

The state channel never rescued a decision. It was uniquely correct on 5 of 168 items, and on those the composed structure was correct 3 times. A channel that costs 27 decisions and rescues 3 is not a channel whose price was mis-set.

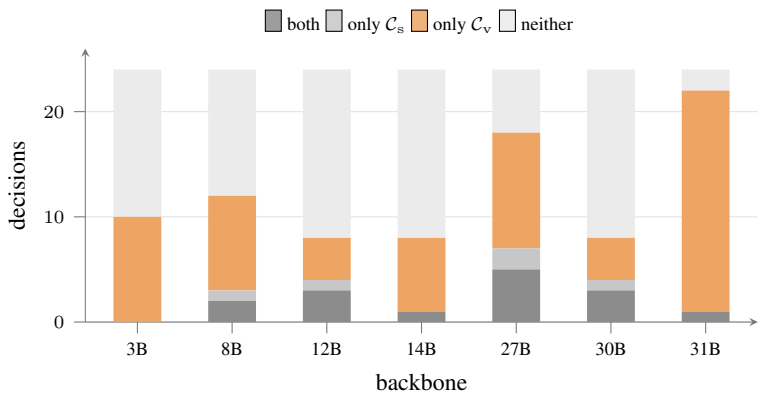

Figure 5: The arbitration census of Table 1 as decisions per backbone. The orange band is where the loss of Eq. (2) can occur, since the index channel alone already held the answer; the pale band at the top is where the synthesis term can act, since no channel held it. The two bands have opposite signs and an ablation reports only their sum.

Super-additivity is real, on the same items. On the 82 decisions neither channel answered alone, the structure answered 12, or 15%. Those are decisions that exist only in the composition, and they are what the parent framework (Chen et al., 2026b) predicts. This is the synthesis term of Equation (2) and it is positive.

An ablation reports the difference and identifies neither term. Netting the loss, the rescues and the failures on items both channels answered gives -20 decisions. That single number is what an ablation gap is, and by Corollary 1 it is consistent with a harmful channel, with a super-additive structure, or as here with both at once.

E.4 The assertion premium, and a registered prediction that did not fire

Prediction 1 registered $\lambda > \frac{1}{2}$. **It does not fire.** The proxy estimator gives a pooled 0.497 with median 0.600, and 4 of 7 backbones exceed one half. That estimator bounds λ from above, and the in-situ measurement, which reads the claims out of the recorded round prompts and can therefore separate *followed the assertion* from *returned neither claim*, puts λ at 0.000.

So the mechanism behind the prediction is refuted, not merely unconfirmed. Over 144 recorded decisions, 64% of answers are attributable to the retrieval block, 0% to the state block, and 34% to neither. The asserted claim is almost never the one the answer follows. Whatever the state channel is costing, it is not costing it by being believed.

What the instrument can and cannot say. Re-running the attribution rule against a neighbouring item's blocks (same lengths, same vocabulary, wrong provenance) separates the two classes completely. Retrieval attribution beats its sham by 5.89–13.45× on every backbone; state attribution beats it on none (0.68–1.36×), because a 159–244-token cumulative inventory contains most plausible answer tokens whether or not the answer came from it. So 0.000 is an **upper bound** with a 12.6% floor beneath it, not a point estimate, and the defensible sentence is that no measured decision has been shown to follow the compressive channel's claim, which refutes Prediction 1 more decisively, not less. Everything read off the retrieval side of the rule stands, including §4.1 and the selectivity term of §E.17, whose λ_{state} contribution is zero on 7 of 7 backbones.

Order is the other candidate, and here we have to be careful with our own claim. Blocks are appended in the order the *agent* requested them, so order is endogenous and measurable without a new run: the answer follows the retrieval block on 18/18 decisions where that block is nearest the answer and on 857/1115 where it is furthest. That is a 23-point difference in the direction a recency account predicts, and the near-side denominator is eighteen decisions; a Wilson interval on $18/18$ runs from 0.82 to 1.00 and the interval on the difference runs from about 5 to 26 points. **We therefore do not claim that order is null.** The correct statement was that the contrast is underpowered and points the way the position-bias literature says it should, so a recency effect of up to roughly a quarter of

Table 6: The in-situ decomposition of Eq. (2), per backbone and per arm. γ is the share of answers attributable to an inserted block; $\mathrm{acc} \mid g$ and $\mathrm{acc} \mid \neg g$ are the accuracies on each side of it, which are the selection and synthesis reliabilities measured rather than inferred. The arms differ only in whether the asserted channel is present, and it is present on less usable budget in the arm that has it.

Backbone	arm	n_g	$\mathrm{acc} \mid g$	$n_{\neg g}$	$\mathrm{acc} \mid \neg g$	γ
Ministral 3B	ASP	18	33.3	6	0.0	75.0
	$-\mathcal{C}_\mathrm{s}$	21	33.3	3	0.0	87.5
Qwen3-VL-8B	ASP	22	31.8	2	0.0	91.7
	$-\mathcal{C}_\mathrm{s}$	24	29.2	0	–	100.0
Gemma 3 12B	ASP	24	29.2	0	–	100.0
	$-\mathcal{C}_\mathrm{s}$	24	29.2	0	–	100.0
Ministral 14B	ASP	21	38.1	3	0.0	87.5
	$-\mathcal{C}_\mathrm{s}$	21	38.1	3	33.3	87.5
Qwen3.8-27B	ASP	0	–	24	12.5	0.0
	$-\mathcal{C}_\mathrm{s}$	15	93.3	9	33.3	62.5
Qwen3-VL-30B-A3B	ASP	2	50.0	22	18.2	8.3
	$-\mathcal{C}_\mathrm{s}$	20	25.0	4	75.0	83.3
Gemma 4 31B	ASP	23	100.0	1	100.0	95.8
	$-\mathcal{C}_\mathrm{s}$	24	100.0	0	–	100.0

the decisions could not be excluded from this data. It *could not*, because the order was the agent's own, and a contrast whose cause is chosen by the behaviour it explains cannot be closed by more observation.

It is now closed by intervention. The forced-order arms (`notes/registered_a1_order.md`) reassemble the accumulated context with a stable split: every state block before every retrieval block (STATE-FIRST), or after (INDEX-FIRST). Contents are byte-identical, and both controls are re-run in the same batch, over the composition subtest where this contrast was measured. The registered manipulation check passes with the caveat the design anticipates: the reassembly can force the class of the last evidence block only once two blocks exist, so round-one prompts are excluded from it. The registered predictions return: the recency band (≥ 15 points of λ_{state} moving with position) **fails closed** (the forced-order difference is -1.4 points, λ_{state} being 1.4% under STATE-FIRST and 0.0% under INDEX-FIRST, both at the floor the schema survey predicts on this subtest), and accuracy does not move either ($+0.30$ points, inside the registered null band). The observational 23-point contrast was endogeneity: the agent reads the channel it is going to follow, then follows it. What survives is exactly what the census says: λ tracks what each channel *holds* (§E.17), which no manipulation of presentation (label, swapped label, or now position) changes.

We report this at length rather than re-cutting the criterion because the refutation is what located the real mechanism. The 34% is the number that does not fit: on the affected backbones the answer is traceable to *no* inserted evidence at all, which is not a claim about arbitration between claims but about whether the arbiter used any of them.

E.5 THE GROUNDING RATE, AND A DECISIVE ABLATION

Table 6 partitions every recorded decision on attribution and reports the accuracy on each side, which by Definition 5 are the two reliabilities of Eq. (3) measured rather than inferred. The two arms differ only in whether the asserted channel is present, and the arm that has it runs on less usable budget and fewer effective rounds.

The flagship is the clean case. With the asserted channel present, $\gamma = 0.0\%$: *none* of its twenty-four answers is traceable to any evidence it retrieved, so $\mathrm{acc} \mid g$ is undefined and the table prints — rather than a number on an empty denominator. All twenty-four decisions are ungrounded and it scores 12.5% on them, about what guessing pays on this subtest. Remove the asserted channel and γ rises to 62.5% (15 grounded decisions instead of none), with 93.3% correct among them, and the overall score moves +58.3 points.

The asserted channel is not corrupting answers; it is preventing grounding. We do *not* claim that accuracy conditional on grounding is unchanged between the arms: on this backbone that comparison

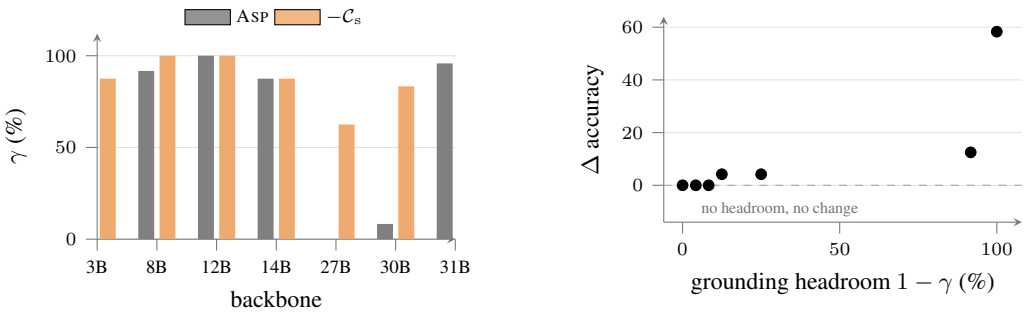

Figure 6: The grounding account, and its null case. Left: the share of answers traceable to an inserted claim, with and without the asserted channel. Right: the accuracy change from removing that channel against the headroom it had to work with. The point at the origin is Prediction 2's null: a backbone already grounded at $\gamma = 100\%$ cannot gain grounding and does not gain accuracy. The bound is not an equality: one backbone gains seventy-five points of grounding on evidence that is itself poor and gains 12.5 points of accuracy, so the points lie under the diagonal rather than on it. No backbone loses accuracy here, which is why the panel has no points below the axis.

is $n_g = 1$ against $n_g = 15$ and carries no information. The honest reading is that the arm with the channel almost never grounds and the arm without it usually does, so the accuracy difference between arms is a difference between the grounded and ungrounded populations rather than within either.

The prediction also says where nothing should happen, and that is the decisive part. One backbone is already grounded at $\gamma = 100\%$ with the asserted channel present, so it has no headroom and removing the channel cannot raise γ or accuracy. Measured, $\Delta\gamma = +0.0$ and $\Delta\text{acc} = +0.0$. A mechanism predicting only the large effect could be coincidence; one that also predicts the *absence* of an effect, on a backbone whose ablation gap would otherwise be attributed to the same cause, is doing work.

Across the 7 backbones with both arms, headroom bounds the gap rather than predicting it. The two backbones with headroom above 70 points gain most, +58.3 and $+12.5$; the four with headroom below 13 points move by at most 4.2. On this subtest and this generation *no* backbone loses accuracy when the asserted channel is removed, so the figure shows a ceiling (the gain never exceeds the headroom) and not the two terms of Eq. (2) trading places. Where grounding has no room, removing the asserted channel does nothing measurable, which is the null of Prediction 2 and not evidence for the synthesis term.

One paired backbone shows the relation is a bound and not an equality. Removing the channel raises its γ from 8.3% to 83.3% and moves its accuracy 12.5: it grounds far more often, on evidence that is itself poor ($\text{acc}\,|\,g$ falls from 50.0 to 25.0), so $\bar{r}_{\text{sel}}$ falls as γ rises and the product moves a sixth as far as the grounding does. That is not visible in an ablation gap, and it follows from the same equation.

E.6 Two backbones, same channels, 79.1 points apart

0.846 and 0.048 are the extreme values of λ, and they belong to backbones of nearly the same size. Their channels are not equally reliable, and the contrast is worth stating with that corrected: each backbone is the better of the two on a different channel, by twenty-five points in each direction.

- Index channel alone: 66.7 against 91.7.
- State channel alone: 29.2 against 4.2.
- Composed: 16.7 against 95.8, a gap of 79.1 points.

The channels are the same, the budget is the same, the router is the same and the walls crossed are the same. The single-channel reliabilities are *not* the same, and an earlier version of this passage said they were: the flagship holds the better compressive channel (29.2 against 4.2) and the foil the better index channel (91.7 against 66.7), twenty-five points in each direction. On the decisions where only the index channel had the answer, one backbone kept 1 of 11 and the other kept 20 of 21.

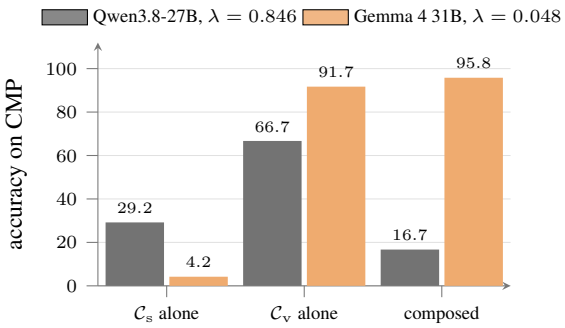

Figure 7: Same channels, same budget, 79.1 points apart. Two backbones of nearly the same size, composed under an identical access structure with an identical router and an identical budget. Their single-channel reliabilities are not close: they differ by twenty-five points in *opposite* directions, one holding the better compressive channel and the other the better index. So the best single channel the two hold differs by twenty-five points, and composed they differ by 79.1. On the decisions where only $\mathcal{C}_{\mathrm{v}}$ had the answer, one kept 1 of 11 and the other kept 20 of 21.

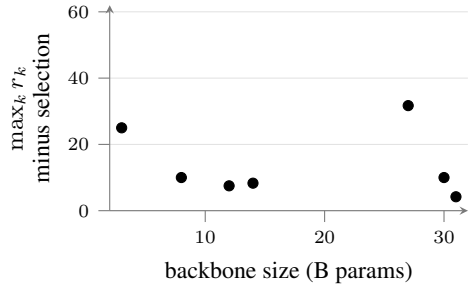

Figure 8: What blind arbitration costs, per backbone: the distance between the accuracy an arbiter that knew r_k would attain by Proposition 1, and the selection term of Eq. (2) with the measured λ. The distance is not a second-order term, and it is largest on the backbone whose λ is 0.846.

This is the observation the paper exists for, and the corrected version of it is still the observation. An access-structure description records which channels a system holds and what they cost; given the single-channel rows as well, the most it can predict is the twenty-five points by which the better of the two channels differs. The composed systems are 79.1 points apart (more than three times that), and the quantity that closes the difference is the one the formalism has no place to write down.

E.7 What the agent asked for, and what it got

The central equation presumes two claims to arbitrate between. In this architecture each claim costs a round to obtain, so before asking how an arbiter weights two claims it is worth asking how often it has two.

The request share for the compressive channel (Definition 8) runs from 19.2% to 69.5% across the ladder, and the spread is not noise. The backbone that reads the compressive channel least, Gemma 4 31B, requests it on 19 of 80 items and is far and away the strongest backbone on this subtest. The backbones that request it on nearly every item are the ones whose answers most often never get given at all (Fig. 3, $r = 0.760$), which is not the same as the ones the ablation rescues and is the quantity that turns out to matter.

This already narrows the space of accounts. A blind arbiter that never requests a channel is observationally identical, at the level of a mean, to one that requests it and discounts it; the two call for opposite remedies. Here the agents that suffer are the ones that ask.

Table 7: What the agent asked for. Before an arbiter can weight two claims it must have two, and the loop must spend a round to obtain each. The share of channel requests that go to the compressive channel runs from 19.2% to 69.5%: Gemma 4 31B reads it on 19 of 80 items and is the strongest backbone on this subset. The last two columns are the ablation's hidden conservatism: with the channel removed, an agent that still asks for it receives a placeholder and loses the round anyway, which happens on 79 of 80 items for the worst-affected backbone and 367 times across the ladder.

Backbone	ASP: actions issued				items reading $\mathcal{C}_\mathrm{s}$		rounds wasted
	RET	STA	ANS	$\mathcal{C}_\mathrm{s}$ share	ASP	$-\mathcal{C}_\mathrm{s}$	in $-\mathcal{C}_\mathrm{s}$
Ministral 3B	80	91	66	53.2%	78	0	79
Qwen3-VL-8B	72	22	78	23.4%	22	0	25
Gemma 3 12B	81	35	79	30.2%	35	0	35
Ministral 14B	81	79	76	49.4%	76	0	78
Qwen3.8-27B	43	98	18	69.5%	68	0	69
Qwen3-VL-30B-A3B	53	89	29	62.7%	60	0	60
Gemma 4 31B	80	19	80	19.2%	19	0	21

E.8 TELLING THE ARBITER WHICH CHANNEL TO FOLLOW

The sharpest objection to everything above is that the arbiter is blind only because nothing tells it which claim to trust, and that something already could. Which channel *should* be followed is fixed by the query class (λ_{state} is 0.000 on every retrieval and composition cell and 0.375–1.000 on every tracking cell), and the architecture already runs a query classifier before it allocates. So the arbitration priority is available from a call that is already paid for. If naming it closes the gap, the wall is not a wall; it is a line of code nobody wrote, which is exactly what §E.7 concludes about retrieval width. This was future work item (0) of the previous version, described there as the sharpest outstanding test. It is run here.

The rule, and what makes it a prediction rather than a fit. The priority is fixed from the state schema and from no accuracy table. The schema holds cumulative counts, an environment flag, a room trace and per-class room-and-recency, and it holds no colour and no printed letter; the retrieval channel returns keyframes and a fresh read of them, so it sees attributes directly and reads at most 8 frames of 300, so it cannot reconstruct a total. Attribute $\rightarrow$ frames, reduction $\rightarrow$ state, chain $\rightarrow$ one role each. Frozen before the run at `notes/registered_a1_typed.md`; the check that this rule coincides with $\arg\max_k r_k$ on 21 of 21 cells was made *after* it was frozen and is a property of the rule, not an input to it. The directive costs 21–22 tokens of $4{,}096$, and this experiment's void condition is written on that cost rather than on realised spend, the repair of the design error recorded in Limitation (14b).

The control is the whole experiment. ASP-TYPED-SWAP carries the same directive with the two channel names exchanged throughout. The two channel names were chosen to be twenty characters each precisely so that the exchange preserves the token count: a directive of the form "if A and B disagree, follow A" names one channel twice, so a swap is length-preserving only when the names are the same length. Two earlier phrasings were rejected by `scripts/selftest_typed.py` on exactly this ground before any call was paid for, because a control that differs from its arm by a token as well as by its content cannot separate the directive's information from its presence; the marker experiment of §E.14 is why that is not a hypothetical worry.

The result: the directive's information is worth six points. Against its length-matched control the typed directive gains

$$+6.43 \text{ points, } 95\% \text{ CI } [+0.87, +12.68], \text{ over } 560 \text{ paired decisions,}$$
$$14 \text{ cells up, } 3 \text{ down, } 4 \text{ tied, sign test } p = 0.013, \text{ label-permutation } p = 0.0001.$$

The permutation null is the one to read: flipping which member of each matched pair is called *typed* produces a difference this large in 1 of $20{,}000$ shuffles, and it assumes nothing about the distribution. **So the arbiter is not merely blind; it is blind and it can be told.** Reliability-marked presentation moved nothing (§E.14), block order moved nothing, and a swapped marker moved more than a correct one; against that run of nulls, a directive naming which channel to follow moves capability by

six points. That is the first manipulation in this paper to act on λ at the arbitration point and survive its own control.

Two honest qualifications, before the mechanism. Against the *control* arm rather than against the swapped one, the typed directive gains only $+2.28$ points, CI $[-2.72, +7.88]$ (8/7/6, $p = 1.000$), an interval that spans zero. The six-point figure is a difference between two *treated* arms, and most of it is the swapped arm losing (-4.15 points, CI $[-9.23, +0.58]$) rather than the typed arm winning. What the experiment establishes cleanly is that *the content of the directive matters*; what it does not establish is that the typed directive is worth six points over doing nothing. And the registered dichotomy did not resolve cleanly: T2's criterion is met (the gap to $\max_k r_k$ falls from $+4.37$ to $+2.09$, 52.2% closed, and the typed arm beats the swapped one) while T3's is missed by less than a third of a point. Both criteria were registered as point thresholds with no interval attached, which is a flaw in the registration and not a finding; we report the thresholds as registered and the intervals alongside them.

The mechanism: the directive works exactly where the classifier is right. Broken out by subtest, the effect tracks one quantity, and it is not a property of the arbiter at all.

subtest	router agrees with subset	typed − control	swapped − control	n
tracking	168/168 = 100.0%	$+8.33$	-4.76	168
retrieval	224/224 = 100.0%	$+2.68$	-1.79	224
composition	90/168 = 53.6%	-4.32	-6.70	168

On tracking and retrieval the classifier names the right class on every decision, and the directive gains ($+8.33$ and $+2.68$). On composition it names the right class on 53.6% of decisions (it says RET on 75 of 168 and TRK on 3), and there the directive *loses* (-4.32), as does its swap. A directive that names the wrong channel is worse than no directive, which is the same asymmetry ASP-RAND-RETR shows for retrieval: a confidently wrong read costs more than an absent one.

This is where the wall actually is, and it is one level up from where we put it. The price of arbitrating well is the price of knowing which channel to trust, and that knowledge is not free either: it is the output of a classifier which is itself an unreliable channel over the same stream, charged to the same budget, and wrong on nearly half of the subtest that needs it most. Theorem 2 says the price of checking a claim is charged to the budget that the checked channel spends; the typed arm is a measured instance of the same accounting one level up. Blindness is not the arbiter's defect. It is what a budgeted agent has left after paying for everything else, and telling it what to do requires buying the type first.

What we cannot conclude from this run, and why. T4 (the theorem's point prediction, that a directed arbiter attains the followed channel's base rate) is *not evaluable here*, and we decline to score it either way. Its right-hand side is r_{pref} taken from the published single-channel rows, which were measured in a different run environment; the control arm re-run in this batch differs from the published control by 7.19 points of mean absolute movement with only 12 of 21 cells inside the re-sampling floor of Limitation (4a), and the drift is concentrated on the retrieval-dependent subtests exactly as a rebuilt CLIP index predicts (notes/`deviations.md` D-A1-v5). Every contrast reported above is within this batch and is therefore untouched by that; a comparison against a frozen published row is not, and we do not make it. On the 21 cells whose control did reproduce, the directed arm is closer to r_{pref} than the control on 8: a count too small to carry a verdict, reported so that the reader can see it is not being hidden.

T5, the registered risk, materialised mildly and in the predicted direction: on the 5 cells where composing two channels already beat the better of them, the directive cost -1.26 points. A correct-on-average priority does take something from the decisions where composition was doing better than either channel alone.

E.9 How large the wall is on this grid, and where it lives

Before any account of *why* a budgeted arbiter falls short of $\max_k r_k$, the distance itself should be a number. In the previous version of this paper it was not one: the figure quoted was ≈ 35 points,

which is the worked case of §2 and not a measurement. Measured on the grid, against the same single-channel rows Table 2 uses for r_{s} and r_{v}, the composite arm sits

$$+3.75 \text{ points below } \max_k r_k,\ 95\% \text{ CI } [-2.05, +10.62], \text{ over 560 decisions in 21 cells,}$$

with the maximum taken per cell from that cell's own single-channel rows. The interval contains zero. Per cell the sign is 9 positive, 6 negative and 6 tied: on six of twenty-one cells the two-channel structure is *better* than the better of its channels, by up to -12.5 points on ministral-3b/RET.

And it is one cell. qwen3.8-27b/CMP on composition accounts for 57.1% of the entire point-sum on its own: 16.7 against 66.7, a gap of $+50.0$ points. That cell is also the one whose transcripts we read most closely in §E.7, and what they show is not an arbiter crediting the wrong claim. It is an arbiter that was never offered one: 14 **of its** 24 **decisions spend zero tokens on the retrieval channel** and 13 of 24 end at the forced exit. A decision that never fetched the claim that could have answered it has not arbitrated badly; it has not arbitrated.

We therefore state the scope of this paper's title quantity plainly, and earlier than is comfortable. *On this grid the distance to* $\max_k r_k$ *is not distinguishable from zero, and the majority of what there is of it is one backbone on one subtest, produced by an allocation failure rather than by a choice between claims.* Theorem 1 remains an identity and Corollary 1 remains valid as stated; what we withdraw is the suggestion that the measured system exhibits a large arbitration loss. The mechanism the grid does exhibit at a size worth reporting is the exit, and that is §4.2.

The wall is a statement about B, and the sweep of it came out against the registration. Everything quantified above sits at one budget, while Theorem 2's admission criterion is a statement *about* B: both sides of $y_k > \hat{v}_k$ move as the wall the budget draws moves. The registered sweep (`notes/registered_a1_budgetsweep.md`, committed before the first call; §H records the executed grid) therefore moved the wall: two arms (ASP, $-\mathcal{C}_{\mathrm{s}}$) at $B \in \{1024, 2048, 8192\}$ over all seven backbones on the tracking and composition subtests where the wall's cells live, 84 cells, with the measured grid supplying the fourth point of the curve (2 cells, the flagship's, a scope limit stated here rather than hidden). Two predictions were frozen. S1 said the count of cells where removing the compressive channel helps decreases in B. S2 said the pooled removal gap moves toward negative as B grows, by five points or more. **S1 is flat and S2 is refuted.** The helped-cell count reads 4/4/1/4 across the four budgets, no movement in either direction. The pooled gap runs -19.79, -15.62, -8.33 (the thin grid point), -12.87: a move of +6.92 points *opposite* to the registered direction, past the five-point band that separates "directional" from "refuted." The per-pair medians say the same and are monotone in B, -14.06 $\rightarrow$ -4.17.

What the sweep measured instead is worth having, and we state it without re-reading the criterion to have predicted it: **the compressive channel earns its place most where the wall is tightest, and its value decays as the wall recedes.** Pooled over the sweep's 14 pairs, the full arm scores 35.57 at B=1024 against 15.77 without the channel (a nineteen-point removal cost), and 41.44 against 28.57 at B=8192. Doubling the budget twice buys the ablated arm $+12.80$ points and the full arm $+5.87$: the channel's yield is, to a first approximation, the slack it creates, and the slack shrinks as the budget grows. The registration expected the reverse on both counts, and the discipline of this paper is to report the refutation rather than amend the prediction; `analysis/a1_budgetsweep.py` is the generator, and its macros are stamped.

E.10 Three exits, and what grounding actually tracks

Definition 7 classifies every scored answer by the exit of Algorithm 1 that produced it. The classification is not a refinement of the accuracy; it is a statement about whether an answer was given at all.

Grounding is a property of the exit, not of the arm. Pooled over both arms and every backbone, $\gamma = 91.0\%$ on the 860 items that reached exit (i), 55.1% on the 205 that reached exit (ii), and 1.8% on the 55 that reached exit (iii). The ladder is monotone and it is steep, and it does not depend on whether the compressive channel was present. What γ has been measuring is therefore not, in the first instance, whether the arbiter chose a grounded claim. It is whether the agent got to choose at all.

One symbol was carrying two measurements, so we split it. The ladder above is the reason to stop writing γ. A decision that reached exit (iii) never committed an answer, and calling its scratch fragment "ungrounded" bills it for a defect of arbitration when what happened is that arbitration never occurred. The theory's λ_0 is a statement about a decision that *did* produce an answer and produced one no channel had stated. Those are different events, and one symbol cannot report both. We therefore replace γ with the two quantities it was averaging,

$$\gamma_{\mathrm{com}} = \Pr[\text{the loop committed an answer}], \qquad \gamma_{\mathrm{att}} = \Pr[\text{attributable to a channel} \mid \text{committed}],$$

neither of which is a new measurement: both are functions of records already released, and their recombination reproduces every previously published γ to within rounding (`analysis/a1_gamma.py`, which fails if it does not). The split changes the reading of this section's headline ablation, and it changes it against the simpler story:

	γ (as published)	γ_{com}	γ_{att}
ASP	78.4%	90.5%	86.4%
$-\mathcal{C}_{\mathrm{s}}$	81.8%	99.6%	82.1%
difference	+3.4	+9.1	-4.3

The rise in γ that this paper reported when the compressive channel is removed is *entirely* a commitment effect. Conditional on the agent having committed at all, removing the channel makes the answer **less** attributable to a channel, not more: 86.4% $[83.1, 89.1]$ falls to 82.1% $[78.7, 85.0]$. The sign of the effect reverses when the two jobs are separated, and the reversal replicates: the round-wall grid is an independent execution of both arms at identical flags (§E.2), and it gives 89.5% and 86.8% against 99.8% and 82.1%, agreeing with the table above on every entry to within half a point.

This does not overturn the section: it is the section's own claim, stated in the variable that can carry it. *Grounding is a function of the exit* was the finding; γ_{com} is the part of γ that the exit determines, and it moves by +9.1 points, while the part the formalism is actually about barely moves and moves the other way. It also disciplines the reserve arm, which by construction commits on every decision (100.0%) and yet is attributable on 79.3% of them, below the control's 86.8%: protecting the exit converts non-answers into answers that are not better grounded. That is consistent with the accuracy gain of that arm being withdrawn in Table 4, and we would rather the two statements agree than report the flattering one.

So the channel's harm is paid in rounds. The compressive channel's insertion is a rendering of the maintained state, averaging 207 words and reaching 374 (measured over 358 insertions), and it is appended to a context that is re-sent in the prompt of every subsequent round. An agent that reads it twice has spent two rounds and carries the cost of both in every prompt that follows. When the round budget runs out before an answer is committed, the harness scores the tail of the scratch buffer.

Removing the channel converts 53 such non-answers into 2 across the ladder. On Qwen3.8-27B, the backbone that scrapes most, the change is entirely in this quantity: it commits on 18 of 80 items with the channel and 55 without it. Its accuracy on committed items does not rise with them; it falls, 100.0% to 85.5%, so removing the channel buys commitments and pays part of the price back in precision. On this backbone the net is -2.5 points, which is to say the mechanism can be present and large while the ablation gap it produces is small and of either sign. An earlier, composition-only version of this census reported the two effects compounding; they do not, and §E.11 says why we no longer make the stronger claim.

And the ablated arm still pays for the channel it does not have. Agents in the $-\mathcal{C}_{\mathrm{s}}$ arm continue to issue STATE actions, receive a placeholder, and lose the round: 367 times across the ladder, on essentially every item for the two backbones that read the compressive channel most. The effect survives a handicap that removes from the ablated arm exactly the resource the account says is decisive, which is the strongest internal check we have on the account.

E.11 WHERE THE EFFECT ACTUALLY LIVES, AND A CORRECTION TO OUR OWN HEADLINE

The census above runs over all 80 questions of the corpus (retrieval, tracking and composition) for 7 backbones in both arms, which is 860 committed decisions and 55 uncommitted ones. An earlier

version of this analysis ran on the composition subtest alone, and the two disagree about something we had put in the abstract.

A disclosure about the numbers in this subsection and the next two. An earlier version printed different numbers here, in §4.2 and in §E.12, and the difference is not a re-reading: a generation defect was repaired in the result files while three generated blocks in this manuscript kept the superseded values. Limitation (14) records the defect, how it was found and what we have changed. What changed is material and it changed against us: pooled over the whole grid, removing the compressive channel moves from $+1.25$ points to -6.11; the reserve arm's pooled closure of that gap moves from 71.2% to -34.4%; and the registered prediction P1 of §E.12 moves from failed to met on its first clause. The theory chapter, the selectivity term $\mathcal{C}$, the witness census and the λ tables are unaffected, their blocks having already been generated from the repaired run.

What the wider grid confirms. The mechanism. Grounding tracks the exit and not the arm, on every subtest: $\gamma = 91.0\%$ committed, 55.1% forced, 1.8% scraped. The scraped population collapses when the compressive channel is removed, 53 to 2 across the ladder. The agents that suffer are the agents that ask, at $r = 0.760$. None of that was visible on the accuracy means, and all of it holds four times more broadly than before.

What it does not confirm. The benefit. Over the full task, removing the compressive channel raises the mean on 1 of 7 backbones and lowers it on 6; pooled over the round-wall grid it is worth -6.11 points, which is to say that on average the channel earns its place. Reading the composition-only version of this comparison as "removing a channel adds capability" was an over-reach: it was a statement about the subtest where the round budget binds first, presented as a statement about the architecture. The sentence has since been re-priced against its own instrument: the reallocated ablation (item (iii-b) of §5, run with the controls in the same batch) moves the pooled removal cost from -5.45 to -3.67: 32.7% of it was the ablation's own bookkeeping, and 67.3% survives the fair comparison.

Two independently executed grids agree on the count and disagree on which backbone it is. The provenance grid has one gainer, Ministral 3B at $+2.50$; the round-wall grid of §E.12 also has one, Qwen3-VL-30B-A3B at $+0.94$ (Table 9). Neither gain exceeds two items of eighty. A count of one in seven that does not name the same backbone twice supports an existence claim about the phenomenon and nothing about any particular backbone.

The non-answers are also less concentrated than we previously reported. Of the 53 in the standard arm, 27 are on composition, 26 on retrieval and 0 on tracking: 50.9% in the largest subtest rather than the 77.8% the superseded generation gave. Composition asks the agent to relate several sightings, so it needs the most channel reads before it can answer, and it remains the subtest where a fixed round wall runs out first and the only one on which the ablation gains. But the claim that the non-answers are concentrated in it does not survive the repair, and we withdraw it.

Why the pooled sign is negative, and where the channel earns its place. The pooled loss is not spread across the grid either. By subtest on the round-wall grid (5 arms, all three subtests, 2800 decisions), removing the compressive channel is flat on retrieval to within half a point, gains 10.1 points on composition, and costs 29.9 on tracking, where the pooled means are 38.8 with the channel and 8.9 without it. Tracking asks for a cumulative quantity over the stream, the one thing the compressive channel holds and a keyframe index cannot localise, so the ablated arm has no channel that can answer it and scores near the floor. That single subtest is the whole of the pooled sign. It is also, by Lemma 1, exactly where the compressive channel should be indispensable, so the pooled result is not an embarrassment to the account: it is the account's other half. The reallocated arm prices that reading against its own handicap. With the ablated channel's budget returned to the survivors and the wasted round unwasted, the tracking cost moves from -27.38 to -26.19 points (returning the budget recovers 1.19 of the 27.4), while the composition gain grows from $+7.44$ to $+11.01$. The two handicaps were pushing the two subtests in opposite directions: on tracking the collapse is the channel's absence and not its bookkeeping, which is what Lemma 1 predicts; on composition the ablated arm was throttled by the accounting, and its gain is larger than the handicapped arm showed.

Reproducibility, and which quantity is stable. The composition-subtest effect on the affected backbone has been measured in three separately executed grids, all released. On the promoted grid it is 16.7 against 83.3, a gain of 66.7 points; on the provenance re-run, $12.5 \rightarrow 70.8$, a gain of 58.3; on the round-wall re-run, 12.5 against 79.2, a gain of 66.7. The sign and order of magnitude reproduce; the level moves by up to eight points, which is what this interface's nondeterminism does to a rate over twenty-four items. A fourth figure, $+45.8$, appeared in the previous version as "the first re-run"; it cannot be reproduced from any file in the repository and has been removed rather than re-attributed.

The provenance counts are markedly more stable than the means they produce. Across the 6 backbones measurable in both re-runs, the number of scraped answers agrees exactly on 4 and within a few items on the rest, while the means those same runs report move by up to seventeen points. Whether the loop committed is a structural property of the run; the mean is a scored outcome on top of it. Where a measurement is available in both forms, the structural one is the one to report.

What we now claim. Corollary 1 says capability *can* be non-monotone in the channel set. A single backbone and subtest on which it is non-monotone is an existence proof, and that is what we have, with a complete account of the mechanism: the channel consumes rounds, the agent fails to commit, the harness scores a buffer fragment, and the grounding rate registers the whole thing as missing evidence. What we do not have, and no longer claim, is that holding a second channel is generally harmful. On 6 of 7 backbones holding it helps; pooled, removing it is worth -6.11 points on the handicapped arm and -3.67 on the fair re-run (item (iii-b) of §5), reported as marginal; and the theory never said otherwise.

E.12 Protecting the exit: the harm, removed without removing the channel

§4.2 located the harm in the loop's exit rather than in what a channel asserts. That account makes a constructive prediction, registered as future work in the previous version: if the damage is that the agent runs out of budget before it can commit, the damage should be removable *without* removing the channel. This section runs that experiment. Its predictions were frozen before the first call (§H), and both accounts they were written to separate turn out to be wrong as stated.

The intervention the last version registered cannot work, and the traces said so before we paid for it. That registration proposed exempting the scene state from the re-sent context. Priced against traces already on disk, the exemption is too small to move anything: the state block averages 386 tokens and reaches 559 over 358 insertions, and only 7 of 560 decisions (1.2%) ever re-send it, because the loop uses 2.42 rounds on average and so usually inserts the state and reads it once. Even at worst the exemption returns 559 tokens to a decision that has 241 left and needs 552, on one decision in eighty. We report the negative pre-check rather than quietly running something else: an experiment that can be costed against existing traces should be, and costing took an afternoon against the hours of interface time the experiment would have spent to learn the same thing.

What the traces do point at. Two facts from the census of §4.2. Every one of the 53 scraped decisions made exactly one retrieval read, and 52 of them read the provider's maximum of eight frames: one access, at maximum width, on the way to running out. And at the moment of the scrape the median decision held 241 tokens and needed 552 to buy its own exit, short by about three hundred tokens.

The cause is an asymmetry in our implementation rather than in the theory. The loop channel has always rationed itself across rounds (each round may spend B_c/R), while the retrieval channel computes affordable frames from the *whole* remaining retrieval cap, so the first RETRIEVE buys the widest read the budget allows and the depth Theorem 2 is about is gone before round two. Nothing in §2 required that. It is a line of code.

Two arms, and neither removes anything. Both leave every channel present, every prompt unchanged and B identical.

- **ration.** A RETRIEVE may claim only its share of the remaining retrieval cap, divided over the rounds still to come, with a floor of one frame, the rule the loop channel already applies to itself.

arm	mean	answered	forced	SCRAPED	$\mathcal{C}_{\mathrm{v}}$ tokens	frames/read	rounds
ASP	57.72	424	77	59	1441	5.75	2.43
$-\mathcal{C}_{\mathrm{s}}$	51.61	446	113	1	1551	5.17	2.65
ration	56.25	413	147	**0**	549	1.99	2.46
reserve	59.82	357	203	**0**	1381	5.42	2.31
both	56.96	404	156	**0**	544	2.00	2.46

Table 8: The round-wall grid: 5 arms $\times$ 7 backbones $\times$ 3 subtests, 2800 decisions. Above the rule, the control and the ablation of §4.2. Below it, three arms that remove *nothing*: every channel is present, every prompt is unchanged and B is identical. '$\mathcal{C}_{\mathrm{v}}$ tokens' and 'frames/read' are per decision. The SCRAPED column is the registered manipulation check M1 and it is the only entry here that was predictable in advance.

- **reserve.** No channel may spend the tokens the loop needs to *leave* with an answer. The exit is priced from the live context before each spend, because the same fallback call costs 260 tokens after one round and 800 after three, so a constant reservation would be too small exactly when it is needed.

Both controls are re-run in the same batch rather than reused: provider nondeterminism has already moved a composition cell by four points once in this work.

The manipulation works exactly as constructed, and that is the only predictable part. Prediction M1 was that the reserved arms produce zero scraped decisions. They do: 59 in the control, 0 under the reserve and 0 with both. It is a manipulation check rather than a finding because it is true by construction (the gate subtracts the exit's price from every spend), and was verified on 400 randomised scripted decisions before the run, so a non-zero count would have to be read as a broken instrument. One observation was not predicted: rationing retrieval width alone also removes all 59 of the 59 scraped decisions with no reservation at all, which is the first evidence that the two mechanisms are one mechanism seen from two ends.

The registered verdicts, both clauses of each. P1 said that if the harm is paid in rounds, protecting the exit closes at least half the gap between ASP and $-\mathcal{C}_{\mathrm{s}}$ on the backbones where the ablation gained, *and* does not fall below ASP where the ablation lost. P2 said that if the harm is instead in what the channel asserts, protecting the exit moves provenance and changes accuracy by less than two points.

P1's first clause is met and its second is not. Exactly one backbone qualifies as a gainer (Qwen3-VL-30B-A3B, whose ablation gap is 0.94 points), and the reserve closes it exactly, so the registered median closure is 100.0%. We report it as met and as uninformative: a median over one backbone whose gap is smaller than a single item of eighty has no resolution left in it. The second clause fails outright (the reserve falls below ASP on three of the six backbones where the ablation lost, by 5.00, 1.25 and 1.25 points), and P2 fails at 4.24 points of mean absolute change against a criterion of two. Neither account, as we wrote it down, survives its own test in full.

The post-hoc pooled statistic, and why it is no longer a recovery. The previous version repaired P1's statistic after the fact, pooling over all 2800 decisions instead of taking a median over seven backbones, and reported that the reserve recovered 71.2% of the ablation's gain. That has no referent on the repaired generation: the ablation's pooled effect is -6.11 points, so the ratio the same formula returns, -34.4%, is a share of a negative quantity. We drop the recovery framing rather than restate it with a sign in front, and for the same reason do not report the resolution-guarded median: 0 backbones now clear the 5.0-point threshold it was defined on.

What the reserve does do has to be stated in two parts, because they carry different weight.

The mechanism holds and is not in doubt. The reserved arm takes the scraped population from 59 to 0. That is a construction, not an inference (no channel may spend the price of the exit, so the exit is affordable), verified on 400 scripted decisions before the run. Nothing statistical can strengthen or weaken it.

The accuracy gain does not survive an interval, and we say so first. Pooled, the reserved arm scores 59.82 against 57.72 for the control, which reads as the best arm in the grid and was reported

backbone	non-answers	ASP	$-\mathcal{C}_s$	Δ from ASP, channel present		
	%			ration	reserve	both
Ministral 3B	16.2	47.50	40.00	−3.75	−5.00	−3.75
Qwen3-VL-8B	1.2	53.75	47.50	+5.00	+3.75	+1.25
Gemma 3 12B	1.2	40.00	38.75	−2.50	+1.25	−2.50
Ministral 14B	3.8	55.00	47.50	−8.75	−1.25	−5.00
Qwen3.8-27B	**81.2**	66.25	65.00	+10.00	+16.25	+16.25
Qwen3-VL-30B-A3B	**66.2**	50.31	51.25	−1.56	+0.94	−0.31
Gemma 4 31B	0.0	91.25	71.25	−8.75	−1.25	−11.25
pooled	—	57.72	51.61	−1.47	+2.10	−0.76

Table 9: Whether an intervention helps is not a property of the intervention. 'non-answers' is the share of the backbone's control-arm decisions that did not reach an ANSWER in a round, forced plus SCRAPED. The two backbones in bold are the two above 20%; they are where the reserve gains most, 8.60 points on average against -0.50 for the other five, and the correlation across the ladder is $r = 0.672$ for the reserve and $r = 0.608$ for the ration. Bottom row: pooled over all 2800 decisions. The $-\mathcal{C}_s$ column is below ASP pooled, so the ablation has no pooled gain and this table reports no closure of one; the reserve is the only arm above the control.

that way. As a paired difference over the same items and the same cached memory (which is what it is), it is $+2.10$ points, 95% interval $[-2.30, +7.81]$, per cell 7 up, 6 down and 8 tied, exact sign test $p = 1.000$. *No test here distinguishes the reserved arm's accuracy advantage from zero.* The previous version's claim that protecting the exit raises capability is withdrawn. What the reserve demonstrably buys is a well-formed answer at the exit instead of a scratch fragment: a property of the harness worth having on its own terms, which we no longer sell as an accuracy result.

The distribution of the gain is what §4.2 predicts: the arm is at or above the control on four of 7 backbones, and those four include both backbones above the twenty-per-cent non-answer line. With 8 tied cells of 21 that pattern is suggestive and under-powered. It is also the only arm that buys the ablation's benefit without paying its price: by subtest, removing the channel costs 29.9 points on tracking while the reserve costs nothing there (38.7 against 38.8), and removing the channel gains 10.1 points on composition while the reserve gains 8.3. That breakdown is post hoc (the registration split by backbone, not by subtest), and is the sharpest form of the claim this grid will carry.

The unification: whether an arm helps is a property of the backbone, not of the arm. Neither intervention helps everywhere, and the pattern of where each helps is the same pattern. Rank the backbones by the share of control-arm decisions that never reached an ANSWER in a round (forced plus scraped, from 0.0% to 81.2%), and that share predicts the gain: $r = 0.672$ for the reserve, $r = 0.608$ for the ration, $r = 0.766$ for both. Split at twenty per cent, the 2 backbones above it (Qwen3.8-27B, Qwen3-VL-30B-A3B) gain 8.60 points on average from the reserve and 4.22 from the ration, while the five below move -0.50 and -3.75.

This is the same conditional structure as the correction of §E.11, arrived at from the other direction. There, "removing a channel adds capability" turned out to be a claim about the backbones that hit the round wall. Here, "protecting the exit helps" turns out to be a claim about the same backbones, identified in advance by the same observable. A remedy that only helps where the disease is present is not a weak remedy; a remedy that helped everywhere would be evidence that we had misidentified the disease.

What rationing sells to buy that. P3 predicted that rationing cuts retrieval spend by at least forty per cent and raises the rounds used, on every backbone. The first half holds on 7/7 backbones (retrieval spend falls from 1441 to 549 tokens per decision and the mean read narrows from 5.75 frames to 1.99), and the second fails, holding on only 5/7. Rounds used barely move (2.43 to 2.46). So rationing does not buy depth with the width it sells; it simply spends less, and on the backbones that were not running out that is a loss. P4, registered as a risk before the run, is confirmed: on retrieval (the subtest where the question asks for a visual detail of one past moment and width is the whole point), rationing loses on 4/7 backbones, by -3.12 points on average.

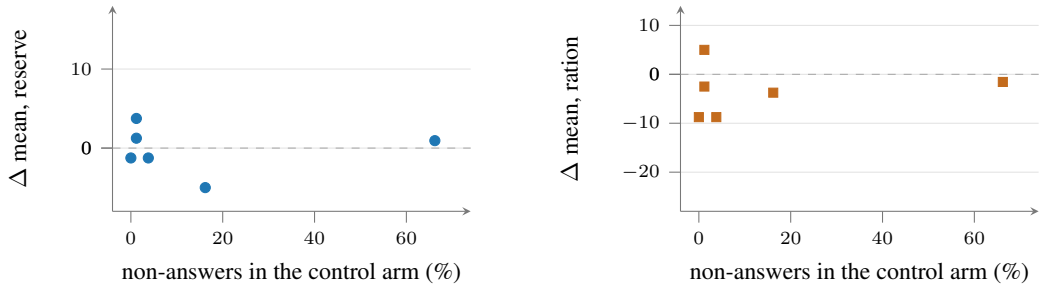

Figure 9: Neither arm is good or bad; each is conditional on the same observable. The abscissa is the share of a backbone's control-arm decisions that never reached an ANSWER within the round wall, which is fixed before either intervention is applied. Left: reserving the price of the exit, $r = 0.672$. Right: rationing the retrieval width, $r = 0.608$. Both cross zero in the same place, at about a fifth of decisions failing to commit, which is the point at which an agent has a budget problem rather than a perception problem.

What this changes in the account. The mechanism of §4.2 is confirmed on the backbones it applies to: the harm a second channel does is paid at the exit, and paying for the exit in advance is worth more than removing the channel is, while keeping the channel. But the harm is not paid in *rounds*, which is what we had written. Rounds used are almost unchanged by every arm here (2.43, 2.65, 2.46, 2.31); what changes is whether the last thing the loop does is answer or stop. The reserve converts 67 decisions from committed to forced (it makes the loop stop *earlier*), and the mean still rises, because a forced answer given the whole context beats a fragment of the scratch buffer. The correct statement is narrower than "the harm is paid in rounds" and we adopt it: the harm is paid at the boundary, and the boundary has a price that nothing in §2 charged anyone for. Definition 6 prices the claims a channel emits; it does not price leaving.

E.13 THE WALL ON A STREAM WITH NO PIXELS IN IT

Every number above is measured on the same questions, backbones and subtests as the companion architecture paper (Anonymous authors, 2026c). That is a genuine overlap and we do not minimise it: on the population axis the two papers coincide, and the separation has to be made on hypothesis and method. It can also be made by measurement, and this subsection does that instead of arguing.

The arbitration wall is not a claim about vision. It says an arbiter that cannot tell which claim to trust attains a reliability-weighted average of its channels rather than the best of them, under a budget. No term in that sentence mentions pixels. So we run the same five arms on QASPER: 36 research papers, 232 evidence-annotated questions per backbone, 7 backbones, 1,624 decisions per arm, no images anywhere. The design constraint that makes this informative is that `run_asp` is *imported verbatim*, not re-implemented: the loop, the exit pricing, the reserve, the rationing and the three exits are the same object code. Only the medium changes (the retrieved unit is a paragraph, the maintained state is over a paper), so a difference between the two grids cannot be a difference in the loop. A re-implementation would have confounded a comparison between populations with a comparison between codebases. Scoring uses no judge: QASPER publishes the evidence paragraphs supporting each answer, so grounding is a substring test against the recorded prompts, exactly as the visual provenance census computes it.

The exits replicate; the grounding ladder does not. The three exits are there and the scraped population is real: decisions where the budget covered neither a commitment nor a fallback and the scorer received a fragment of the agent's own scratch buffer. The control arm produces them on 4.5% of decisions, which is the registered condition N1 for this run being informative at all, and it holds. Removing the compressive channel raises the grounding rate from 36.3% to 41.7% and all but eliminates the scraped population, from 4.5% to 0.1%. Protecting the exit without removing anything does the same thing: RESERVE takes the scraped population to 0.1%, converting it into honest fallbacks ($32.0\% \rightarrow 47.1\%$ forced exits). The compressive channel on a text stream does what it does on a visual one: it consumes the rounds the agent needed to commit.

What does *not* replicate is the ladder γ climbs in §4.2, and the registration says so in its own words. Prediction P1 for this run was that γ on scraped exits is lower than γ on answered exits, as on the visual grid. Pooled over all 5 arms it is 36.8% on the answered decisions, 26.2% on the forced ones and 41.4% on the 111 scraped ones: the scraped population is the *best* grounded of the three. **P1 fails.** The reason is the measure rather than the loop, and that is now a measurement rather than a guess. On this corpus γ is a substring test against a published evidence paragraph, and a fragment of the agent's scratch buffer is mostly quoted paper text; on the visual corpus it tests containment against an inserted block, which a scratch fragment fails almost by construction. The same exit is therefore attributable 41.4% of the time here and 1.9% of the time there, a factor of 22 for one loop behaviour under two instruments. Split as in §4.2, the failure localises: this corpus commits on 98.6% of decisions, and among the decisions that committed the ladder runs in the registered direction, 36.8% answered against 26.2% forced. The inversion is produced entirely by 111 non-answers out of 8,120. We record this as a diagnosis of the failure and not as a rescue of the prediction: P1 was registered on γ as defined at the time, and on that definition it failed. The exit *mechanism* replicates on a text stream; the grounding *measure* does not transfer, and we report both rather than the half that suits us.

The benefit does not, and its count is unstable along two axes. §E.11 corrected our own headline: the mechanism generalises, the *benefit* of ablation does not. Out of sample, removing the compressive channel helps 4 of 7 backbones on this corpus, and two of the four are inside a point ($+0.24$ and $+1.25$). The count does not hold still. On the pre-repair index it went from three of seven at 82 questions to one of seven at 232; on the repaired index, at the same 232 questions, it is 4 of 7. The previous version of this paragraph reported the first of those movements and concluded that the larger sample had not moved our way. That was true of the grid it described and is false of this one, so we replace the sentence rather than keep it, and we do not replace it with its mirror image: a count that moves from three to one with sample size and from one to four with a repair to the retrieval index is not measuring a property of the architecture. Non-monotonicity stays an existence result with a located cause (it is exhibited, and the cause is the exit), and it is not a general property of holding a second channel on any corpus we have measured.

The repair to this section's index was registered, and one of its predictions flipped. The retrieval channel on this corpus was ranking on captions written to a prompt imported from the visual medium, which asks for the room and the door state. The repair, its forbidden changes and five predictions were frozen before it was implemented (Appendix H). Four of the five came out as registered. The fifth did not, and it was the falsifiable one. It had two clauses: the control must still beat the both-arms arm, which it does (41.66 against 38.36); and the reserve arm must still fail to pay for itself, $+1.33 \leq 0$. Measured, $+1.33$ is $+1.33$: the ordering flipped.

The registration's stated consequence is that the section is withdrawn rather than re-narrated, and we apply it to the claim it was written about. **The claim that flipped is withdrawn.** This section no longer asserts anything about what protecting the exit costs or earns on a text stream; the reserve arm's recall advantage over the control is reported in Appendix H as a registered falsification event and is not offered here as a finding, however much we would prefer it. What the section keeps is the part the first clause protected and N1 tested: the three exits exist on a stream with no pixels in it, the control arm scrapes, and both removing the channel and protecting the exit collapse the scraped population. Reading a flipped falsification criterion as a discovery is precisely what the registration forbids.

What this settles for the boundary. A1 now holds a population the companion architecture does not: a text corpus with no frames, no renderer of ours, and a claim about arbitration that is stated without reference to a modality. The overlap on the visual corpus is unchanged and still has to be defended on hypothesis and method. What is no longer required is that a reader take on trust that the subject of this paper is arbitration rather than perception.

E.14 THE MARKER IS READ, AND IT BUYS NOTHING

The previous version named one experiment as its sharpest outstanding test: *marking each claim with its channel's reliability should move accuracy toward* $\max_k r_k$*, while a resource account predicts no change, because a label is not a resource.* Theorem 1 says the weights λ are fixed by presentation, so

changing the presentation is the intervention the theorem itself points to. We ran it, registering the predictions first, and it does not come out the way the theorem needs.

Design. Three arms over 7 backbones and three subtests, 63 cells. ASP is the control. MARKED labels each channel's block with that channel's own single-channel accuracy, taken from the published rows rather than fitted. SWAPPED exchanges the two numbers: identical format, identical token count, wrong information. Without the third arm "marking helps" cannot be told from "any extra label helps". Offline, the marker changes exactly two strings and costs 15 tokens of 4,096. The advertised reliability gap has median 41.7 points as the current table computes it and 33.4 as the registration recorded it; on either figure it is not too fine to act on.

How the cells are classified, and why that is not as registered. A cell is called state-better or retrieval-better by comparing the two published single-channel rows. Those rows have since been re-promoted to the repaired generation, so the split (6 state-better cells of 21, median gap 41.7) is not the one the registration recorded (*three* of 21, median gap 33.4). M4 is therefore reported on a post-hoc reclassification, and its honest scope is the cells the current table calls state-better. Freezing run-time reliabilities into the cell records is future work *(ix)*; nothing else here depends on the classification.

Accuracy does not move. Pooled, ASP scores 54.54 and MARKED 54.32: a change of -0.22 points. The registered differential (that the gain correlate positively with the reliability gap, the prediction a resource account cannot make) gives $r = -0.020$ over 21 cells, under a fifth of a standard error from zero. SWAPPED, whose marker points at the *worse* channel, moves -0.12 and correlates $+0.017$: no worse on either count, so the registered control condition fails.

Cell means at ~ 25 questions move in steps of four points, so the informative view is paired. Across 560 matched questions the two arms return *the same score* 513 times (91.6%), with 22 wins for MARKED and 25 against. This is a null with the manipulation demonstrably applied, not a small effect measured imprecisely.

It is not that the manipulation had no opportunity. Our first explanation was that two-channel arbitration is rare, so a label would have nothing to act on. That is checkable and wrong: both channels are present in the answer context on 246 of 43.9% of decisions. Nor is it budget starvation: the marked arm spends $+4.42\%$ more, not less, and makes 4.48 calls against 4.35. An arm given more that scores the same makes the null stronger.

But the marker *is* read, and it moves the agent where it points. What is observable is the loop's own commitment: a retrieval read spends the verbatim channel's budget, while the state block is present every round regardless. In the 6 cells where the *state* channel is better, the marked arm makes 0.479 retrieval reads per decision against the control's 0.451, and the share answered from the compressed channel alone rises from 54.9% to 52.1%: told to trust state, it reaches for retrieval less. In the 15 retrieval-better cells it moves the other way, 0.897 to 0.901. And the swapped marker moves behaviour *in the direction of its own false label*: in the same state-better cells it raises reads to 0.500 and collapses the compressed-only share from 54.9% to 50.0%.

What this leaves standing. Neither account survives intact. The presentation account predicts that telling the arbiter which channel to trust moves capability; it does not, on 560 paired decisions. The resource account predicts a label can do nothing; it does something, reliably and in the labelled direction, including when the label is false. Together they say something sharper than either: **the arbiter reads the marker and re-weights its channel access accordingly, and the re-weighting does not change what it gets right.** The loss to $\max_k r_k$ is therefore not caused by the arbiter's not knowing which channel to trust: it knows, and it complies. What this removes is the reading of Theorem 1 under which λ is a dial that presentation turns. Informing the arbiter is not the escape, and §E.12's exit protection, which changes what the loop can afford rather than what it believes, remains the only intervention here that moves capability without removing a channel.

It replicates on a second population, with three times the decisions. The visual grid cannot be made larger: 7 backbones over the corpus's whole question set *is* the 560 paired decisions above.

Power therefore has to come from a different corpus. We repeat the three arms on the text stream of §E.13 (231 questions per backbone over 7 backbones, 1617 paired decisions), measuring the marker's reliabilities on *that* corpus, since a marker carrying an accuracy measured elsewhere is a different manipulation wearing the same label.

The result is the same. Pooled recall moves from 41.11 to 41.87, a gain of $+0.75$ points, positive on 4 of 7 backbones. Question by question the marked arm wins 316 and loses 308 of 1617. The registered correlation is -0.127 over 21 cells, under one standard error from zero. The marked arm again spends more ($+2.51\%$, again exceeding the 1% void condition) and again scores the same. One difference is worth naming: the swapped marker moves $+0.17$ here, less than the correct one, so the visual grid's failed control reads as noise on a small grid rather than as evidence that any label helps. Across the two populations the manipulation has been applied to $2{,}177$ paired decisions, on a visual stream and a text stream, and moved neither.

What the text arm cannot certify. The reach test that found the visual defect (Anonymous authors, 2026c) is only as good as the field it reads, and on the text corpus the state schema often omits the per-claim timestamp it dates the channel by: 116 of 252 streams are UNVERIFIABLE rather than healthy. Of the 136 it can read, the median reach is 100% and one stream is frozen. We read the unverifiable half as neither (absent evidence is not a verdict), so this arm's null is measured on streams half of which we can certify kept advancing. Enforcing the schema at the API (`response_format`, supported by all 7 slugs) would close it, and that is future work *(viii)* rather than a re-run.

The run is void under its own registered condition, and so is its replication. We registered that realised mean token spend differ between arms by less than 1%, as a *void* condition: exceed it and the run cannot separate the two accounts, whatever the primary outcome says. Measured: $+4.42\%$ on the visual grid and $+2.51\%$ on the text replication. **Both runs are void as written.** The criterion constrains *realised* spend (a behavioural consequence of the marker) rather than the label's own cost of 15 tokens, and the overrun runs against the marked arm, so it cannot be what produced the null. Neither point is a licence to keep the stronger word. This subsection reports a **null**, not a refutation. What the void condition does not touch is the behavioural result, a within-arm statement about where the reads go that does not require the arms to have spent the same.

The rerun the void condition demanded, and what it decided. The repair is now run under a comparison the manipulation cannot trip (`notes/registered_a1_mark_bmatch.md`): two arms that *both* carry the labels: MARKED with the real reliabilities and MARKED-NONE with a same-form placeholder ('past accuracy –%') carrying no information. The label's token cost is therefore identical by construction and the void condition sits on the label's own cost, which the offline check verifies at zero tokens of difference. Registered bands: $|\Delta| \leq 2$ confirms the null about information, $[2, 5)$ is inconclusive, ≥ 5 would have recovered the account. Measured over the full 42-cell grid: MARKED 56.25 against MARKED-NONE 55.06: a difference of +1.19 points, inside the registered null band, with realised spend +2.8% (reported as an outcome, conditioning on nothing). **B1 fires: the label's information does nothing that its presence does not**, under a comparison whose void condition the manipulation cannot trip by construction. The escape derived from Theorem 1 is now tested in the form its derivation names, and the null is no longer carried by a void run; what remains of the escape is the second path, periodic re-derivation, whose pilot is item (iv).

E.15 ONE ITEM, READ CLOSELY

Only 178 decisions in the whole census have both channels' claims in context at once, which is itself part of the finding: the arbitration the theory describes is rare because obtaining two claims costs two rounds. Here is one, from Gemma 3 12B's arm, quoted from the recorded round prompt. The question is *"did you see the trivet before or after the colander?"*, and the gold answer is AFTER.

```
[SCENE STATE]
{"counts": {"fork": 102, "towel": 101, "pot": 92, "pen":
106, "bowl": 92, "plate": 94, "hat": 97, ...}, "env":
{"door": "open", "door_changes": 7}, "rooms_visited":
["study", "bedroom", ...], "objects": {"fan": {"room":
"hallway", "last_seen": 295}, ...}, "event_log": []}
```

```
[RETRIEVED for '"trivet"']
before
```

Two things are wrong here and neither is a bug.

The compressive channel has maintained a running count of twenty-two object classes, a room history, a door counter and a last-seen table; it contains *nothing about the trivet or the colander*. Those are the corpus's singleton classes, seen once each; the state maintains what recurs, which is exactly what the question does not ask about. This is the $O(1)$ state doing its job correctly. It is 207 words of context, re-sent every round, that cannot answer the question under any arbitration policy.

The verbatim channel was asked to *describe facts relevant to the question in these frames* and returned the single word `before`. It did not report an observation; it answered the question. So on the one decision where an arbiter has two claims to weigh, both channels have emitted assertions and neither has emitted evidence. The agent replied `before`. The gold is AFTER.

This is Definition 1 illustrated rather than argued: what reaches the decision procedure is a claim, and the interface offers no place for the provenance that would let the arbiter prefer one. It is also a caution about our own λ. The assertion premium was defined to separate a channel presented as an assertion from one presented as a report of an observation, and here the channel we classified as the reporting one has returned an assertion too. That is one reason λ was not the discriminating quantity, and §5 records it.

E.16 What the compressive channel can assert, counted

Everything this paper says about λ_{state} rests on a token-containment rule whose sham control gives it no discriminative power on the state side (0.68–1.36$\times$ chance, 7/7 backbones). An instrument that cannot separate a signal from noise on one side can report *not detected* on that side. The previous version of this paper reported *did not happen*, and that inference does not follow. This subsection replaces it, and the replacement needs no instrument, no model output and no provider call: it asks whether the compressive channel's contents contain the answer at all, which is a property of the state and the question.

The channel is silent on seven decisions in ten. Over the 28 committed offline states (7 backbones $\times$ four episodes), the state schema holds the queried quantity on 114 of 560 decisions (20.4%). Split by subtest the separation is total: 67.9% on tracking (114/168), and *zero* on retrieval (0/224) and on composition (0/168). The silences are individually attributable. 336 decisions ask for a colour or a printed letter and the schema has no such field: the parent system's Shannon bound behaving exactly as stated. 56 ask an ordering over two objects whose per-object entries the LRU had already evicted by the end of the stream, which is a width limit rather than a schema limit and is not something the schema alone would predict. 54 ask a total for a class the channel never recorded.

So $\lambda_{\text{state}} = 0$ on retrieval and composition is a fact about the schema and not a measurement of arbitration, and no experiment on those two subtests can bear on whether a budgeted arbiter follows a compressive claim, because on those subtests there is no compressive claim. We had said as much in prose while continuing to present the zero as evidence; the count is what the prose was owed.

Where the channel does speak, it is followed almost deterministically. Tracking is the subtest where the channel holds the answer, and there the same offline artefact supports a much stronger test than an attribution rule can. Read the cached state, read off what it says the count is, and ask whether the arm that reads that state gets the item right. The predictor is computed from a different artefact than the outcome, and it is fixed before any score is looked at. On the 112 tracking-count decisions of the STATE-ONLY arm the two agree on 90.2% of items (27 both right, 74 both wrong; Wilson 83.3–94.4%). The two error cells are not symmetric and that is the informative part: there are 0 decisions on which the state held the correct value and the arm missed it, and 11 on which the state was wrong and the arm was right anyway.

A channel whose recorded value predicts the answer on nine decisions in ten, with no case at all of the value being present and ignored, is a channel that is being followed. The earlier claim that no measured decision was shown to follow the compressive channel is therefore **withdrawn**. What survives, and what the rest of this paper measures, is the narrower and better-supported statement:

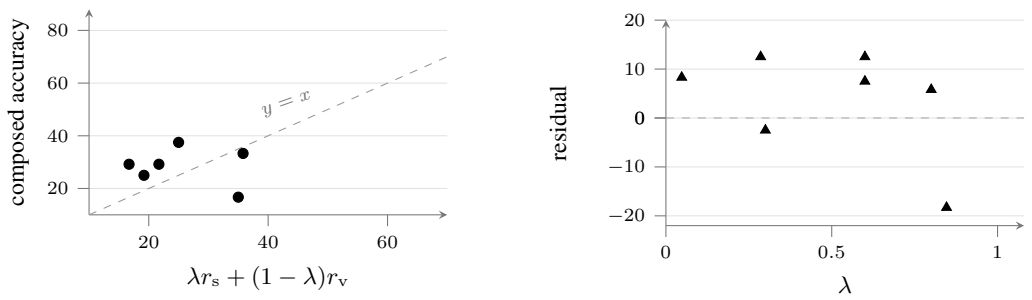

Figure 10: Why the theorem has three terms. Left: composed accuracy against the selection term of Eq. (2). Five of the seven backbones lie *above* the diagonal and two below it, so the two-term form of the statement is contradicted by the sign of its own residuals on the majority rather than on all of them. Right: the residual against λ, on an axis that shows the two negative values rather than clipping them. Where the residual is positive its magnitude is the synthesis term $\lambda_0 r_0$, the decisions no channel answered alone; where it is negative the three-term form does not account for it, and §E.17 says so.

the arbiter follows whichever channel holds the queried quantity, and on two of three subtests the compressive channel holds nothing.

The following is causal, not merely observational. The registered value-substitution probe (§5 item (iii-a)) rewrites the compressive channel's block so it asserts a specific value *wrong by construction*, derived from the item's gold answer, and reads the answer against the planted value: no attribution rule, no sham control, the only difference between arm and control the value the channel asserts. On the tracking decisions where the schema holds the queried quantity, the arbiter follows the planted value on 114 of 114 decisions: $\lambda^{\mathrm{causal}}_{\mathrm{state}} = 100.0\%$ [96.7, 100.0], with the manipulation check at 100.0% (the planted value verbatim in every recorded prompt) and accuracy under injection 0.0%. Following the compressive channel is therefore not a correlation that survives a placebo; it is a causal statement, and it holds even when what the channel asserts is false, which is the blind arbiter of Theorem 1 exactly as the theorem drew it. On retrieval and composition the same probe finds nothing to corrupt on *any* of 224 and 168 decisions: the denominator rule reports the count rather than an estimate, and that count is the schema survey's zero confirmed by intervention.

A prediction the resource ledger does not make, with its placebo. That statement has a consequence a token ledger cannot produce, because a ledger prices channels by what they cost and says nothing about which decisions a channel's contents bear on. Within the tracking count family (the only family in which the predictor varies, since the door families are asserted on every decision and the attribute families on none), accuracy should split on whether the channel recorded the class, and it should split only for arms that read the channel. It does. ASP scores 44.8% on the 58 decisions where the state holds the class and 18.5% on the 54 where it does not (+26.3 points); STATE-ONLY 46.6% against 20.4% (+26.2); and RETR-ONLY, which cannot read the state at all, 1.7% against 3.7% (-2.0). The placebo is flat, and it has to be, because the predictor is computed from an artefact that arm never sees.

The same split pooled across *all* tracking families would have shown a gap for RETR-ONLY too, because the door family is both always-asserted and the easiest family in the subtest. We report the within-family contrast for that reason and state the pooled one is confounded, rather than reporting the larger number.

This count also fixes the scope of the citation-form probe registered in §5 item (ii): on the composition subtest the channel holds the queried quantity on none of 168 decisions, so what that run measures is presentation's power over a channel with nothing to say, not over a speaking one. The two results read together as the paper's availability-and-presentation statement: *followed where it holds the quantity; presentation cannot manufacture a claim the schema does not carry*; whether presentation re-weights a claim the channel does make is untested.

E.17 THE THREE TERMS, MEASURED

Equation (2) is not fitted; each term is counted. The selection term is the census rows in which one channel held the answer, the synthesis term is the row in which neither did, and λ weights the first. Reading the equation with the measured weights reproduces the composed accuracy on every backbone to within the resolution of a 168-decision subtest, and the residual is not centred on zero: it is positive on five of the seven backbones, by up to 12.5 points, which is what forced the synthesis term into the theorem in the first place. It is negative on the other two, by 2.5 and 18.3 points, and we flag that rather than absorbing it: a synthesis term can only add, so a composed accuracy below the selection prediction is not something the three-term form produces. The smaller of the two sits inside the re-sampling floor of Limitation (4a) (6.1 points on composition) and the larger does not, so one is noise and one is unexplained. They are the reason the equation is reported as reproducing the composed accuracy *to within the resolution of the subtest* rather than exactly. We record that the two-term form of the statement was written first, contradicted by the sign of most of its own residuals, and corrected before any figure was drawn.

The third term, and the assumption it settles. Fig. 10 charges the whole residual to synthesis, which is the right first move because the proxy λ cannot separate *followed a claim* from *returned neither*. The in-situ attribution can. Crediting synthesis at its *measured* accuracy on the decisions where the answer matched neither block, rather than treating it as the unexplained remainder, leaves $\mathcal{C} = \mathrm{acc} - \sum_k \lambda_k r_k - \lambda_0 r_0$, which by Theorem 1 is zero whenever the arbiter is blind in expectation. The converse does not hold (selection effects on different channels could cancel in the sum), so what follows is evidence that the arbiter is blind, not a proof of it, and the per-subtest and per-cell breakdown below is what makes cancellation implausible rather than merely unmentioned.

It is zero. Over 366 arbitrated decisions on 14 cells, $\mathcal{C} = +0.016$ pooled with median +0.007, positive on 7 of 14, which is what chance looks like. Per subtest it is +0.044 on retrieval (4/7 positive, 200 decisions), -0.011 on composition (3/7) and -0.016 on tracking (3/7); across 4 independent executions of the same arm the composition figure runs -0.031 to -0.011. The headline pools retrieval and composition only, and the exclusion is the conservative direction rather than the convenient one: tracking is the single subtest where λ_{state} is not zero, its state-side attribution sits at $1.17\times$ chance, and *noise on an attribution class pushes $\mathcal{C}$ toward zero*; including it would therefore help the null we are claiming. On the two included subtests λ_{state} is exactly zero, so every unit of weight rides on the retrieval class, which the sham control validates at $107.38\times$ and $4.56\times$. *Given which channel it followed, the arbiter attains that channel's base rate and nothing more.*

That is the paper's central claim, and it is now a measurement rather than a premise. Eq. (2) is not something a reader has to grant: it is what 366 recorded decisions say, and the price of blindness (the distance from $\sum_k \lambda_k r_k$ to $\max_k r_k$) is therefore paid in full. On the worked case of §2 that distance is ≈ 35 points.

Why it took a repaired grid to see it. We report the wrong answer as well as the right one, because the difference between them is instructive. Computed on the pre-repair caches (with 8 of 28 state streams frozen by the completion-cap defect of §5), the same estimator gave $\mathcal{C} = +0.071$ on retrieval and composition, positive on 12 of 14 cells over 358 decisions, and we would have reported the arbiter as *not* blind. Over all three subtests the same contrast is $+0.044$ (15/21 cells positive) before the repair against $+0.005$ (10/21) after it: the estimator can produce a non-null, which is the only reason its null is worth reading. The artefact is exactly locatable: a frozen state makes the single-channel row r_s understate what the composite arm's state block was contributing, so the residual came out positive. Two conditions were needed to see through it. The composite arm and the single-channel rows have to come from the *same* offline generation (mixing a repaired composite with frozen baselines put one cell at $\mathcal{C} = +0.875$), and the streams have to be unfrozen. **A term that tests an assumption is only as good as the generation hygiene of the numbers fed to it**, and we state the pre-repair figure so that a reader can see which of the two we would have published had the defect gone unfound.

What the subtests do and do not show. λ_{state} is 0.000 on every retrieval and composition cell (14 of 14, all 7 backbones), and 0.375 to 1.000 on every tracking cell. The separation is complete, and it is *not* a result about arbitration. The state schema of §3 holds cumulative counts, an environment flag, a room trace and a per-class room-and-recency entry, and it holds no queried attribute at all: there is no colour and no printed letter anywhere in it. So on the two subtests whose answers are attributes the

compressive channel has nothing bearing on the question, and on the one whose answers are its own reductions it has everything. That is the parent system's Shannon bound (Anonymous authors, 2026c) behaving as stated, and we say so rather than presenting it as evidence for Lemma 1, which it is not. Lemma 1 does have an empirical signature, and it is on the *retrieval* side rather than the compressive one. The lemma says a reduction under a cancellative operation has no witness of size one; the consequence that can be measured is therefore what breaking retrieval costs. ASP-RAND-RETR breaks it in the way that isolates the question (same channels, same budget, same prompts, retrieval that ignores the query), and on the tracking subtest it costs *nothing*: 91.7 with retrieval working, 91.7 with retrieval randomised, the identical score. On the retrieval subtest the same manipulation takes 81.2 to 6.2, which is below what *removing* the channel outright costs (15.6), because a confidently wrong read is worse than an absent one. A reduction has no frame to fetch, so nothing is lost by fetching the wrong one; an attribute has exactly one, so fetching the wrong one loses it. This is one backbone and three subtests (the ablation grid runs on the flagship only), and we claim it at that width.

What the λ separation does establish is that λ is fixed by *what a channel holds*, which is the one candidate the three manipulations of §E.14 and §E leave standing: a marker, a swapped marker and the block order all change how a claim is presented and none of them changes what a channel holds.

The interval, and the bound the data actually support. The previous version of this paper wrote that "a null at this scale bounds $|\mathcal{C}|$ below roughly five points" and gave no derivation. Here is one, and it is slightly worse than five. Resampling by a two-stage cluster bootstrap over cells (cells are the exchangeable unit, since decisions within one share a backbone, a router, a budget and a cached memory) gives $\mathcal{C} = +0.019$ with a 95% interval of $[-0.029, +0.066]$ over 366 decisions in 14 cells. By the standard duality between two one-sided tests and the 90% interval, the tightest equivalence margin the data support is 5.8 accuracy points, so a TOST at five points does not reject and the defensible sentence is $|\mathcal{C}| < 5.8$ points rather than five. The sign count the earlier version rested on is 7 of 14 cells positive, exact binomial $p = 1.000$: a coin flip, now with the number that says so.

What this does and does not rule out should be read carefully, because the scale it is compared against has also changed. It rules out a selectivity term of the size that would matter against the 35-point gap of the worked case in §2. It does *not* rule out a term of the size of the gap actually measured on this grid, which is 3.75 points pooled (§E.9): a bound of 5.8 points is wider than the quantity it would have to explain. The honest position is that the measured system's distance to $\max_k r_k$ is too small for this estimator to attribute, and we say so instead of carrying the worked case's number into a claim about the grid.

F LIMITATIONS, IN FULL

Limitations. The first bounds every empirical claim in the paper and is stated without hedging. *How to read this section:* items (0)–(4) bound the core chain and are the only ones a reader needs to discount a core number; process defects and open probes are recorded once, in place, and confined to future work rather than re-litigated in the main line. No open probe below is needed for the topic to stand: each names the bounded sentence it would decide, and the paper claims none of those sentences. *(0) One of the two interventions of §3 is measured only as a pilot, and the other returned a null that is now confirmed under the comparison its void condition blocked.* Periodic re-derivation is measured only at the registered pilot's scope: one backbone, the tracking subtest, the aggregation and replay checks and the bound $r_s \geq r_{\mathrm{agg}}(1-\epsilon_u)^{N-T}$ at the pilot's fixed T (*The escape, and its price*; item (iv) below carries the verdicts); the multi-backbone grid remains future work. Reliability-marked presentation was measured (§E.14), does not move capability: -0.22 points pooled, 513 ties in 560 paired decisions, and a swapped marker that does no worse (-0.12); but both runs voided themselves on realised spend ($+4.42\%$ and $+2.51\%$ against a registered 1%). The budget-matched rerun the defect demanded is now run: both arms carry the labels, the control's carrying no information, the void condition on the label's own cost by construction, and it returns the same null inside its registered band (+1.19 points, prediction B1). So the escape derived from Theorem 1 is unavailable in the proposed form under a comparison the manipulation cannot trip, which is a stronger statement than the void runs could make. The marker is nevertheless read and obeyed (reads move toward the label, including when it is false), so the failure is specific. Informing the arbiter changes what it does and not what it gets right; whether a marker *acted on by the loop* recovers the escape is future work *(iii-b)*.

(1) A registered mechanism was refuted by its own measurement, and its replacement is weaker in provenance. Prediction 1 was frozen before the data; Prediction 2 was written after the in-situ attribution and before the ablation that tests it. That is better than a post-hoc account and worse than a registration, and we say so rather than presenting the second as though it had the standing of the first. What gives it some standing is the null case: it forbids an effect on a backbone with no grounding headroom, and there is none.

(2) Attribution is a surface test, and the error is asymmetric. An answer is credited to a block when its content tokens all appear there and not in the other, which misses paraphrase and counts coincidence. The sham control of §E says which side that hurts: retrieval attribution beats a length-matched sham by 5.89–13.45$\times$ on all 7 backbones, state attribution on none. **Every state-side attribution number here is an upper bound with a 12.6% floor under it, not a point estimate.** The claims run the other way (that the compressive channel is *not* followed), so nothing rests on the bound being tight; but λ_{state} cannot be estimated until the state block stops blanketing the answer space, which is future work *(i)*.

(2a) One theorem was wrong as proved, and the repair is what let us test it. The previous version stated Eq. (2) unconditionally, via a step that needs reliability-blindness and does not follow from the definition alone. Theorem 1 now states the identity with the selectivity term and derives Eq. (2) as the blind case, so blindness became something a transcript could settle, and it settles *for* the equation ($\mathcal{C} = +0.016$). The same estimator on the pre-repair caches gave $+0.071$ and would have had us report the opposite: a statistic that tests an assumption inherits every generation defect of the numbers fed to it, so the composite arm and the single-channel baselines must come from one offline generation for it to mean anything.

(2b) The λ separation is not ours. λ_{state} is zero on every retrieval and composition cell and non-zero on every tracking cell, and the cause is that the bounded state holds no queried attribute: the Shannon bound of the parent system (Anonymous authors, 2026c), not a result about arbitration. We flag it because the separation is clean enough to be mistaken for one.

(3) γ and the accuracy conditional on it are not independent. A decision is classified by where its answer's tokens appear, and a correct answer is more likely to share tokens with correct evidence, so a rise in γ and a rise in $\mathrm{acc}\,|\,g$ are not cleanly separable. An earlier version rested this on $\mathrm{acc}\,|\,g$ being unchanged between arms on the flagship; it is not measured there, because the flagship's control arm grounds on one decision and the comparison carries no information (§4.2). We rest instead on the movement in γ itself and the exit census behind it, which counts decisions rather than conditioning on a partition the accuracy helps determine.

(4) Scale of the measurement, layer by layer. The layers have different widths and the paper is only as strong as the narrowest one carrying each claim. The arbitration census and the λ estimate are 168 decisions on one subtest; the in-situ attribution is 144; the provenance grid is 7 backbones $\times$ three subtests $\times$ two arms; the round-wall grid is 105 cells and 2800 decisions; the text grid is 1,624 decisions on a second corpus. One budget throughout, $B = 4096$, and one serving interface. Per-backbone counts on the visual grid are small enough that a single item moves a subtest mean by 1.25 points and γ by four, and the paired differences should be read as ordinal except where they are exactly zero.

(4a) The scale above is a question count; the grid also has a re-sampling floor, and we quote it. The parent paper's repair rebuilt the offline phase for every backbone, and on the three that needed no repair the before-and-after is a second draw of the same 300 update calls, which prices re-running the memory at 9.0 points of mean absolute movement over 45 cells: 4.5 on retrieval, 6.1 on composition, 16.2 on tracking; 5.9 for the verbatim channel against 8.6 for the compressive one (Anonymous authors, 2026c), the shape Lemma 1 predicts. Its scope matters as much as its size. It bounds *cross-generation* comparisons, which is why we retract the conclusion drawn from $+0.071$ rather than claiming its distance to $+0.016$ as a finding. It does *not* bound paired arms reading the same cached memory, so every null in §E.14 is measured inside one generation. Nor does it bound the ordinary case of two arms executed once each with nothing changed: that floor is measured separately and is a different size: 1.65 points pooled against up to 20.8 for one cell, over 3 executions of the control at identical flags (§E.2). The repetition floor applies to an arm contrast, the re-sampling floor across a generation or a host.

(5) One backbone is missing from the in-situ layer. Its provider returned rate-limit and authentication errors for the same slug across retries while every other backbone completed, so the in-situ attribution of §E runs on six backbones and 144 decisions rather than seven and 168. That is a serving failure rather than a measurement, and we report the smaller n instead of substituting. The paired ablation arms are *not* affected: Table 6 and the provenance grid carry all 7 of 7 backbones.

(6) The corpus is synthetic and inherited. It is the schematic render of the parent system (Anonymous authors, 2026c), built because the registered corpus was licence-blocked, and every limitation that paper states about it applies here unchanged.

(7) Non-coinciding errors is an assumption; descriptively bounded in-grid (`analysis/a1_answerset.py`), no sub-additive family; $\lambda{=}1/2$ design is future work (vi).

(8) One serving arm, and provider nondeterminism inside the measurement. Re-executing identical cells moved three published values by one scoring step each. Every re-run in this paper therefore writes to its own directory and its means are routed away from the file the tables are built from, and the paired comparisons use arms measured in the same re-run rather than one re-run against a published number.

(9) γ is measured, not explained. It varies enormously across backbones holding the same channels, does not track scale, and rises when an asserted channel is removed, from 0.0% to 62.5% on the flagship, as large a move as any between backbones. So it is a property of the backbone and the access structure jointly, not of the backbone alone. What determines it we do not know; a mechanistic account would presumably start from the finding that these models aggregate at query time rather than tracking incrementally (Tang et al., 2026), and we have not established the connection.

(10) The provenance census is correlational; its conditional carries post hoc. Protecting the exit collapses scraped to zero with the best arm in-grid, but P1's second clause and P2 failed; the conditional (help tracks non-answer rate) needs an independent ladder. Recorded once here; the core exit ladder itself is in §4.2.

(11) The three exits are a property of this harness. Another runner might refuse to score an item that never produced an answer, or retry, or abstain. Ours scores the tail of the scratch buffer, documented in its source as a last resort. We expect this is the common case (an evaluation loop has to return something), but we have surveyed no other harness and the size of the effect elsewhere is unknown. The narrow claim survives regardless: on this harness, 53 of the scored answers in the standard arm were not answers.

(12) Provenance and grounding are not fully independent measurements. An item that reached exit (iii) is scored on a scratch fragment, usually the model's own narration rather than a channel's text, so it fails the containment test almost by construction. The ladder $\gamma = 91.0\% \to 55.1\% \to 1.8\%$ is therefore partly definitional at its lower end. Its upper end and its arm-invariance are not: an item reaching exit (i) is grounded at 91.0% whether or not the compressive channel was present, and that is the comparison the account rests on.

(13) A robustness check that did not change anything, reported because it might have. The block extractor captures each evidence block up to the next bracketed header or the end of the prompt, so the last block on a prompt absorbs the loop's closing instruction. That adds around a dozen common words to the token set a short answer is tested against, and could in principle make a one-word answer look grounded. Recomputing every cell with the boilerplate stripped and nothing else changed moves γ by 0.0 points in all 14 cells.

(14) Generation-stamp repair recorded once in the Reproduction appendix; superseded blocks regenerated. *(14b) Void-condition design defect superseded by the budget-matched rerun (§E.14);* detail in deviations.

(14c) Three post-hoc analyses (provenance split, repeat floor, judge bound) are cheap, checked against pre-existing artefacts, and cut against us. Stated once here.

(15) γ named two different measurements, and it was also doing two different jobs. Two defects sat under one symbol and only one of them is repaired here.

The repaired one is the conflation of committing with being attributable, and §4.2 now separates them: γ_{com} and γ_{att} recombine to every published γ and answer different questions, and the ablation's

apparent grounding gain turns out to live entirely in the first and to reverse in the second. Every γ in this paper should be read as γ_{com} unless the text says otherwise, because that is the quantity the exits move.

The unrepaired one is cross-corpus (visual inserted-block vs text gold-paragraph containment); level comparisons across corpora remain two statistics sharing a symbol, and a unified instrument is future work *(i)*. Detail once above.

Future work, in the order it should be done. *Constraint:* nothing here extends the core chain. Items marked *Done* are closed scope checks already reported where measured; items merely *Registered* are bounded probes with frozen predictions and no claim attached: they answer to peer review as future work, not as missing premises. *(0) Done, and it did not go as registered* (§E.12). The scraped population does collapse without removing the channel (to zero), and the reserved arm is the best in the grid, 59.82 against 57.72 for the control and 51.61 for the ablation. But the intervention this list previously named could not have shown that; P1's second clause and P2 both failed; and the pooled "closure of the ablation's gain" the previous version of this item reported has no referent, because on the repaired generation the ablation has no pooled gain to close. What replaces it as the next decisive test is the conditional claim, which is the one thing here that is post hoc: the gain from either arm tracks a backbone's non-answer rate at $r = 0.672$ over seven backbones, and seven points do not establish a law. Run it on a ladder we have not used, with the split at twenty per cent registered in advance. *(i) Done; a null, now confirmed under the comparison its void condition blocked* (§E.14). The arbiter reads the marker and obeys it in its read counts, and gets no more right; the budget-matched rerun (`notes/registered_a1_mark_bmatch.md`) carries the labels in both arms, the void condition on the label's own cost by construction, and returns +1.19 points, inside the registered null band. The escape is unavailable in this form. Earlier void runs are superseded by this comparison and not re-litigated here. *(ii) Done; the discriminating experiment ran and the form account was refuted,* scoped. Under citation form the state channel is followed on 0.0% of decisions, against 2.6% for the JSON control, below the 15% refutation line. The witness-history schema stores *no queried attribute*, and the schema survey counts the channel as holding the quantity on *zero* of 168 composition decisions: the refutation is that presentation cannot rescue a silent channel. Whether citation form re-weights a speaking claim (tracking) is untested and claimed nowhere. Full design history in `notes/registered_a1_cite.md`; first-execution failure and clean-subset check per deviations D-A1-v9. *(iii) Done: K=3 subtracts through arbitration; K=4/K=5 is a closed scope check.* The committed-only decomposition above is the core result; the extension (D1 $0/7$, penalty -19.64→-14.41→-10.08) withdraws K^* to two-point direction with detail in the paragraph above and deviation D-R8. No chain depends on the extension. *(iii-a) Done; the availability account is now causal (§4.3).* Planted wrong values are followed on 114/114 tracking decisions ($\lambda^{\mathrm{causal}}_{\mathrm{state}} = 100.0\%$); nothing to corrupt on 224/168 attribute decisions. Registration `notes/registered_a1_cfinject.md`; decided sentence in §4.3.

(iii-b) Done (§E.11). Fair re-run re-prices the pooled cost $-5.45 \rightarrow -3.67$; full return confirms at +5.75 with zero bound-limited decisions. Tracking recovers nothing ($26.78 \rightarrow 26.78$). Dependence D-R9 stated once in §E.11; retry detail in deviations. What the fair grid does decide is *where* the surviving cost lives, once in §E.11.

Bound removed: full return confirms at +5.75; dependence D-R9 once in §E.11.

(iv) Done at pilot scope; fails at the aggregation operator. P1 bound holds $4/4$; P2 promise fails (0.50 vs 1.00); defect in the priced fold ($r_{\mathrm{agg}} = 0.25$–0.75, zero-call control perfect); P3 -45.83; multi-backbone future work. *(v) Done as a bound-implied measurement (§2).* ϵ_u/N^* per backbone; agrees with crossovers where defined. *(vi) Done descriptively (Limitation (7)).* No sub-additive family; deviations are λ stories. Constructive λ=$1/2$ open. *(vii) Registered probe, no claim:* extremal running-maximum (size-one witness) falsification of the lemma's scope; frozen prediction, unrun by the constraint above.

(viii) Registered probe, no claim: API-enforced response schema making the text arm fully certifiable (`notes/registered_a1_responseformat.md`); unrun.

(ix) Process fix, no claim: freeze run-time reliabilities into cell records; provenance recorded where used. Unrun; no core number depends on it.

Conclusion. A framework that describes an architecture by the channels it holds and the prices they pay cannot distinguish two systems that differ by 79.1 points. We measured that pair: same channels, same budget, same router, single-channel reliabilities differing by twenty-five points in *opposite* directions, and composed accuracies of 16.7 and 95.8. What separates them is which claim the decision follows when the claims disagree: a quantity the framework has no place to record and an implementation must therefore invent.

The positive content comes first. A blind arbiter attains a weighted average rather than the best channel; the compressive channel is followed where it holds the queried quantity, causally; appending a weaker same-form channel subtracts through arbitration rather than budget; and the harm is delivered through the loop's exits. Against that background, one registered mechanism of ours was wrong, in one sentence: we predicted the arbiter follows the asserted claim, and measured in situ it almost never does. What it does instead is stop using either claim: on the flagship, 0.0% of answers are traceable to any evidence it retrieved, and removing the asserted channel raises that to 62.5%. The channel was not corrupting answers. It was preventing grounding, and the accuracy it cost is bounded above by the headroom it occupied.

Adding the arbiter to the formalism costs one equation and changes two conclusions. Capability is no longer *guaranteed* monotone in the channel set, because a blind arbiter attains a weighted average of channel reliabilities rather than their maximum, with weights set by how claims are presented rather than by how reliable they are. Guaranteed is the operative word: pooled over the grid, removing the compressive channel is worth -6.11 points on the handicapped arm and -3.67 on the registered fair re-run (reported as marginal, with 67.3% of the cost surviving), and on tracking, the subtest whose answers are the channel's own reductions, removing it stays close to fatal at the fair pricing (-26.19, the returned budget recovering 1.19). What we exhibit is that the guarantee fails, not that the average is against the channel. And the repair is not free: verification is drawn on the same budget as the channel it verifies, so a channel whose checking price reaches its own yield can never be both used and checked. That is the arbitration wall, and it is a wall in the same sense as the resource walls it sits beside.

The measurement that made us write this also argues against reading it as purely negative. On the decisions where neither channel could answer alone, the composition answered 12 of 82. Super-additivity is real and is what the parent framework predicts. It simply arrives alongside a selection loss that costs 27 of 66 decisions the index channel had already answered, and an ablation reports only the difference. Both terms are in Eq. (2), both were counted on the same items, and the reason to keep them apart is that they have different remedies.

G Full Proofs

Standing conventions

Throughout, a query q is drawn from a fixed distribution with a unique correct answer $a(q)$. An architecture $\mathcal{A}$ holds channels $c_1, \ldots, c_K$; on each query, channel c_k presents a claim $\hat{a}_k$ with reliability $r_k = \Pr[\hat{a}_k = a(q)]$, and the arbiter π returns one value. All probabilities are over the query draw and any internal randomness of the backbone, and no independence between channels is assumed anywhere; where mutual exclusivity is needed it is supplied by Definition 4 and is stated at the point of use. A budget B is fixed per decision, and a channel's price is the tokens its use consumes from B. We write $\mathrm{acc}(\pi)$ for the accuracy of the arbiter, $\lambda_k = \Pr[\pi \text{ returns } \hat{a}_k]$, and $\lambda_0 = 1 - \sum_k \lambda_k$ for the probability that π returns a value equal to no presented claim, with r_0 its conditional accuracy.

On the degenerate case, which is the one the census found. Equation (2) is usually read with λ_0 small: the arbiter picks one of the claims, and the question is which. §4.2 finds the opposite regime on the affected backbones, where the returned value matches no presented claim on a large majority of decisions, because it is a fragment of the agent's own scratch buffer rather than a selection at all. The theorem covers this: as $\lambda_0 \to 1$, $\mathrm{acc}(\pi) \to r_0$ and the channels' reliabilities drop out of the expression entirely. That is worth naming, because it is the strongest form the result can take. A blind arbiter does not merely fail to attain $\max_k r_k$; in the limit where the budget is consumed before a selection is made, its accuracy is independent of *every* channel it holds, and the architecture's whole access structure is unpriced by its own outcome.

On what is assumed about λ, and what is not. Definition 2 fixes λ by presentation, and the proof of Theorem 1 uses this only through the step $\Pr[\hat{a}_k = a(q) \mid E_k] = r_k$. That step needs E_k to be conditionally uninformative about the correctness of $\hat{a}_k$ given the claim's content, not that λ_k be constant across queries, uniform across channels, or known. A presentation-driven arbiter whose weights vary arbitrarily with the surface form of a claim still satisfies it. What violates it is an arbiter with access to a signal correlated with r_k beyond the claim's own content, which is exactly the escape of §3, and is why that escape is predicted to work while reallocating the budget is not.

PROPOSITION 1

Let $k^\star \in \arg\max_k r_k$. The arbiter that returns $\hat{a}_{k^\star}$ regardless of the other claims is well defined, since $r_1, \ldots, r_K$ are known by hypothesis, and its accuracy is $\Pr[\hat{a}_{k^\star} = a(q)] = r_{k^\star} = \max_k r_k$. Any arbiter attaining more is also admissible, so the value is a lower bound. Monotonicity follows because $\max_{k \le K+1} r_k \ge \max_{k \le K} r_k$ for any appended channel. ■

LEMMA 1

Fix $i \in [N]$ and suppose $\{i\}$ were a witness of ρ. Since f is non-constant pick x, y in the observation alphabet with $f(x) \neq f(y)$, and since $N \ge 2$ pick $j \neq i$. Let S and S' agree everywhere except at j, with $S_j = x$ and $S'_j = y$. Then $S_i = S'_i$, so a witness $\{i\}$ would force $\rho(S) = \rho(S')$. But writing $R = \bigoplus_{l \neq j} f(S_l)$ we have $\rho(S) = R \oplus f(x)$ and $\rho(S') = R \oplus f(y)$, and cancellativity in the second argument gives $R \oplus f(x) \neq R \oplus f(y)$. Contradiction, so no singleton is a witness. ■

Remark. Cancellativity is what the statement needs and it is why the lemma is about counts and sums rather than about every reduction: $\max$ is not cancellative, and a running maximum *does* have a witness of size one, namely the index attaining it. That is not a loophole in the account so much as a prediction: a compressive channel restricted to extremal reductions should be citable, and §5 lists testing it.

THEOREM 1

Fix a query q on which the channels disagree, and condition on it; the theorem is stated for the disagreement queries, which is where arbitration occurs and where Definition 5 defines λ. On such a query Definition 4 makes the events $\{\hat{a}_k = a(q)\}$ mutually exclusive across k: two channels returning the same wrong claim violates non-coincidence, and two returning the same *right* claim are not disagreeing and so are excluded by the conditioning. Hence at most one presented claim is correct. (Without the restriction to disagreement queries this step fails: two channels may agree on a correct claim, the events overlap, and the partition below is not a partition. This is also the regime in which the ensemble effect becomes available; see the remark at the end.)

Partition the arbiter's output by which claim it equals. Let E_k be the event that π returns $\hat{a}_k$ and no other presented claim, and E_0 the event that it returns a value equal to no presented claim. On disagreement queries the presented claims are distinct, so these are exhaustive and mutually exclusive, and with $\lambda_k = \Pr[E_k]$ and $\lambda_0 = \Pr[E_0] = 1 - \sum_k \lambda_k$,

$$\mathrm{acc}(\pi) = \sum_k \Pr[E_k]\ \Pr[\hat{a}_k = a(q) \mid E_k] + \Pr[E_0]\ \Pr[\pi = a(q) \mid E_0].$$

Writing r_0 for the last conditional probability and adding and subtracting $\sum_k \lambda_k r_k$ gives Eq. (1), with $\mathcal{C}_k = \lambda_k(\Pr[\hat{a}_k = a(q) \mid E_k] - r_k)$ by definition. The decomposition is an identity and assumes nothing about π.

For the blind case, suppose π satisfies Definition 2: exchanging channel j's and channel k's claims leaves the output unchanged. Then E_k is determined by the multiset of presented claims and not by which channel supplied which, so λ_k cannot depend on the identity, and hence not on the reliability, of the channel that produced a claim. That is the first half of the theorem and exchange-invariance delivers it.

It does *not* deliver $\mathcal{C}_k = 0$, and we no longer claim that it does. Conditioning on E_k selects on the *content* of the claim π chose, and content is not independent of correctness: an arbiter that prefers a count in a plausible range, or a fluent answer, selects correct claims more often than the channel's

base rate while remaining perfectly exchange-invariant. Formally, $\Pr[\hat{a}_k = a(q) \mid E_k] = r_k$ requires the selection event to be conditionally independent of the claim's correctness given the presented multiset (call this *blindness in expectation*), and Definition 2 is strictly weaker. Under blindness in expectation every $\mathcal{C}_k$ vanishes and Eq. (2) follows; under Definition 2 alone the identity Eq. (1) stands with $\mathcal{C}_k$ measuring the departure.

So the relation between blindness and $\mathcal{C}$ is one-directional and conditional, and neither direction is an "if and only if". Blindness in expectation implies $\mathcal{C} = 0$; $\mathcal{C} = 0$ does not imply blindness of either kind, since selection effects on different channels can cancel in the sum. What $\mathcal{C}$ is, without qualification, is the measured departure of the arbiter from the base rates of the channels it follows; §E.17 reports it at $+0.016$ over 366 arbitrated decisions, which is what licenses Eq. (2) *for this system*, empirically rather than by hypothesis.

What two earlier versions got wrong, and both times it was the proof rather than the equation. The first asserted $\Pr[\hat{a}_k = a(q) \mid E_k] = r_k$ from Definition 2 directly and stated Eq. (2) as unconditional. The second added the exchangeability step above and inferred $\mathcal{C}_k = 0$ from it, which is the gap this paragraph closes; that version also carried a sentence saying the blindness assumption *fails* in the measured system, written when the pre-repair estimate stood at $+0.071$. The repaired measurement is $+0.016$ and the assumption is not rejected. Both sentences are withdrawn. ■

Remark. The independence-free step is the mutual exclusivity supplied by Definition 4. Without it several claims can be simultaneously correct, π can be right without selecting any particular channel, and the ensemble effect (de Condorcet, 1785) becomes available. This is why the theorem is stated in the non-coinciding regime and why §5 lists the boundary as open.

COROLLARY 1

Let π be symmetric over the presented claims, so $\lambda_k = (1 - \lambda_0)/K$ for all k. Holding $\lambda_0 r_0$ fixed, Eq. (2) gives accuracy $\frac{1-\lambda_0}{K} \sum_{k \le K} r_k + \lambda_0 r_0$. Appending c_{K+1} changes the first term to $\frac{1-\lambda_0}{K+1} \sum_{k \le K+1} r_k$, and the difference is

$$\frac{1 - \lambda_0}{K + 1} \Big(r_{K+1} - \frac{1}{K} \sum_{k \le K} r_k \Big),$$

which for $\lambda_0 < 1$ is negative exactly when r_{K+1} is below the incumbent mean. For the finite K^* the corollary also assumes the channels are appended in non-increasing order of reliability: then $r_{K+1} \le r_k$ for every incumbent, so r_{K+1} is at or below their mean, the mean is non-increasing in K, and every later channel is below it too. Dropping that hypothesis drops the conclusion (with r_{K+2} large the running mean rises again and the sequence of signs need not be eventually constant), and an earlier version of this proof asserted the conclusion without the hypothesis. No allocation argument applies: the expression contains no b_k, so redistributing B among the same $K + 1$ channels leaves it unchanged except through the r_k themselves. ■

COROLLARY 1

Let $\mathcal{A}^- = \mathcal{A} \setminus \{c_s\}$ and let D be the set of decisions. Define $L = \{d \in D : \mathcal{A}^-$ correct, $\mathcal{A}$ incorrect$\}$ and $G = \{d : \mathcal{A}$ correct, $\mathcal{A}^-$ incorrect$\}$. The ablation gap is $(|G| - |L|)/|D|$ by definition. Under Eq. (2), a decision in L requires that π returned $\hat{a}_s$ on a query where $\hat{a}_v$ was correct, which is an event of the selection term with probability λ_s on such queries; and a decision in G on a query where neither claim was correct requires E_0, which is the synthesis term. Hence $|L|$ and $|G|$ are governed by different terms, and their difference determines neither. ■

LEMMA 2

Let U_i be the event that update i is correct, with $\Pr[U_i] = 1 - \epsilon_u$ by hypothesis (i). By the non-cancelling hypothesis (iii) the final state is correct only if every update was, so $r_s(N) \le \Pr[\bigcap_{i \le N} U_i]$. Under independence $\Pr[\bigcap_i U_i] = \prod_i \Pr[U_i] = (1 - \epsilon_u)^N$; under negative dependence $\Pr[\bigcap_i U_i] \le \prod_i \Pr[U_i]$ and the same bound holds. Either way $r_s(N) \le (1 - \epsilon_u)^N$.

Two remarks on what this does *not* permit. If the hypothesis were the one-sided $\Pr[U_i^c] \le \epsilon_u$, then $\Pr[U_i] \ge 1 - \epsilon_u$ and the product is bounded *below*, not above, by $(1-\epsilon_u)^N$; the bound would fail. And under positive dependence $\Pr[\bigcap_i U_i] \ge \prod_i \Pr[U_i]$, so the exponential is again not an upper bound. This is why (i) is an equality and (ii) excludes positive dependence.

Solving $(1-\epsilon_u)^N = r_v$ gives $N^* = \log r_v / \log(1-\epsilon_u)$, and since $(1-\epsilon_u)^N$ is decreasing in N the inequality $r_s(N) < r_v$ holds for all $N > N^*$. ■

THEOREM 2

Write $\mathrm{acc}(b)$ for the accuracy attainable with b tokens allocated to the best available use, and let c_k require b_k tokens with yield y_k. Admitting c_k and verifying its claim costs $b_k + v_k$ and returns, by Proposition 1, at most $\max(r_{\max}, r_k)$; omitting it leaves $b_k + v_k$ tokens for the best alternative use, returning $\mathrm{acc}(b_k + v_k)$. Admission is therefore preferable only if $y_k > \mathrm{acc}(b_k + v_k) - \mathrm{acc}(b_k)$, and to first order in v_k the right side is $\partial\mathrm{acc}/\partial b \cdot v_k = \hat{v}_k$, which is the first display. Both sides are now accuracies; the marginal is what converts the token cost into the accuracy it displaces, and without it the comparison has no meaning. Admitting it blind instead returns the selection term, which is lower than the aware value by $\lambda_k(r_{\max} - r_k)$ on the decisions where c_k is followed and wrong; comparing the two admission modes gives the second display. If $\hat{v}_k \ge y_k$ then, since $\partial\mathrm{acc}/\partial b > 0$ wherever budget is not saturated, the first inequality fails for every r_k, so no reliability makes c_k admissible with verification. ■

COROLLARY 2

Re-deriving every T observations resets the state to a value computed from the index, so the chain of updates whose correctness the final state depends on has length at most T. Applying Lemma 2 to that chain gives $r_s \ge (1-\epsilon_u)^T$, and requiring $(1-\epsilon_u)^T \ge r^\dagger$ gives $T \le \log r^\dagger / \log(1-\epsilon_u)$, whose largest integer solution is the stated T. The number of aggregations over N observations is $\lceil N/T \rceil$, so the amortised price per observation is v_{agg}/T. By Eq. (2) the selection loss avoided per decision is $\lambda(r_v - r_s)$, an accuracy; pricing the amortised token cost in the same units through the marginal $\partial\mathrm{acc}/\partial b$ gives the stated condition. ■

H SECONDARY EXPERIMENTS AND SENSITIVITY

STANDING CONVENTIONS FOR THE CENSUS

Every count in §E.7 and §4.2 is derived from the recorded call log of a run, not from a re-execution. Each item's log holds, per call, the tag the runner assigned, the realised token count, the channel the tokens were charged to, and, for the two arms re-run with text capture enabled, the prompt sent and the completion returned. Three conventions follow and are worth stating because each rules out a way of getting the numbers wrong.

The runner's parser, not ours. Actions are classified with the function the loop dispatched on. Backbones wrap actions in markdown emphasis and append sign-offs; an anchored regular expression written independently for the analysis disagreed with the runner on enough rounds to change the census materially, and the disagreement was silent. There is exactly one correct parser for the question "what did the agent do", and it is the one whose output determined what happened next.

The longest prompt, not the last call. The context accumulates across rounds, so the prompt of the final round contains every block inserted before it. Blocks are therefore read from the longest round prompt of an item, which is the only place the whole accumulated context appears; reading the last call would miss items whose final call was the forced-answer fallback, whose prompt is assembled differently.

Text capture is off by default. The client records prompts and completions only when asked, and the two re-run arms are the only cells with it enabled. Every cell measured for the parent paper's tables was produced with it off, so nothing in the published grid is affected by this analysis, and the re-runs write to their own directories with their means routed away from the file the tables are built from.

Table 10: Where the budget went, per decision, as a share of realised tokens. The router's classification call and the retrieval channel's reading calls are billed separately; the round calls carry the loop itself. The compressive channel has no column, and that is the point: its price is not a call but the 207 words its rendering adds to a context that is then re-sent in the prompt of every subsequent round. A per-call accounting cannot see it, which is one reason the effect of §4.2 was not found earlier. Budget is 4096 tokens per decision throughout.

Backbone	ASP				$-\mathcal{C}_{\mathrm{s}}$			
	router	retr.	rounds	tok/item	router	retr.	rounds	tok/item
Ministral 3B	9.5%	48.6%	35.8%	2716	12.0%	61.3%	22.5%	2153
Qwen3-VL-8B	12.0%	67.7%	20.3%	2439	13.1%	72.5%	14.1%	2238
Gemma 3 12B	12.7%	57.6%	29.2%	2335	15.0%	68.0%	17.0%	1977
Ministral 14B	10.0%	46.3%	42.5%	2958	13.3%	60.8%	25.5%	2235
Qwen3.8-27B	9.9%	43.8%	30.8%	3046	10.5%	64.8%	22.7%	2880
Qwen3-VL-30B-A3B	8.9%	45.0%	28.9%	3385	12.1%	70.5%	16.6%	2486
Gemma 4 31B	12.3%	64.4%	23.2%	2429	13.2%	68.6%	18.2%	2271

WHERE THE BUDGET WENT

The census is about rounds, so it is worth being explicit about tokens as well. Per decision the ladder spends between roughly two and three and a half thousand tokens against a budget of 4096, split across the router's classification call, the retrieval channel's reading calls, and the round calls themselves. The compressive channel does not appear as a line item, because its cost is not a call: it is the 207 words (median 204, maximum 374, over 358 insertions) that the state rendering adds to a context which is then re-sent in the prompt of every subsequent round. A channel whose price is paid inside someone else's calls is invisible to a per-call accounting, which is one reason the effect was not found earlier.

THE DECISIONS ON WHICH ARBITRATION ACTUALLY OCCURS

Only 178 of the census's decisions have both channels' claims in context simultaneously. They are not spread evenly: Ministral 3B and Ministral 14B contribute 55 and 52 of the 178, Gemma 3 12B a further 35, Qwen3-VL-30B-A3B only 6, and Qwen3.8-27B, the backbone with the *highest* request share for the compressive channel, contributes none at all, which is a fact about where its loop exits rather than about what it asks for. The section heading of an earlier version of this appendix said forty; the count is 178, and the heading has been corrected. This is the practical content of Definition 8: obtaining a second claim costs a round, so a loop under a round wall usually arbitrates between one claim and nothing. Theorem 1 is a statement about a population that this architecture, at this budget, rarely enters, and the loss it predicts is realised mostly through the degenerate case where λ_0 absorbs everything.

We regard that as a strengthening rather than a weakening. The theorem says a blind arbiter cannot attain $\max_k r_k$; the census says the arbiter is usually not even offered the maximum, because the budget was spent obtaining the other claim.

THE ABLATED ARM'S WASTED ROUNDS, COUNTED

Table 7 reports how often the $-\mathcal{C}_{\mathrm{s}}$ arm requests a channel that is not there and receives a placeholder. Across the ladder it happens 367 times; for the two backbones that read the compressive channel most it happens on essentially every item. The arm therefore operates with (a) roughly three quarters of the usable budget, since the state allocation is computed before the ablation flag is read and is then unspendable, and (b) one round per item consumed by a request that returns nothing.

This is the closest thing the paper has to a placebo condition, and it runs the right way. If the effect were an artefact of the ablated arm having more room to work in, the arm that loses a round per item to a placeholder would not be the arm that gains 58.3 points on the composition subtest of the backbone that scrapes most.

THE ROUND-WALL EXPERIMENT, AS REGISTERED

Frozen and committed before the first call of §E.12; the repository records the commit that contains it and the commit that contains the first result, in that order. Reproduced here so that the failures are legible as failures.

Registered, and withdrawn before running. The previous version of this paper proposed exempting the scene state from the re-sent context. The registration withdraws it, with the costing of §E.12, and states the replacement. We record the withdrawal rather than the substitution alone, because an experiment quietly replaced by a better one leaves no evidence that the first was wrong.

M1 (manipulation check). *The reserved arms produce zero scraped decisions.* Registered as near-certain and true by construction, so that a non-zero count would have to be read as a broken instrument. Verified on 400 randomised scripted decisions before the run: 0 of 400 under the reserve against 36 of 400 in the control. **Held.**

P1 (the rounds account), reported clause by clause. *On the backbones where $-\mathcal{C}_{\mathrm{s}}$ gained, protecting the exit closes at least half of the gap between* ASP *and $-\mathcal{C}_{\mathrm{s}}$, and does not fall below* ASP *where $-\mathcal{C}_{\mathrm{s}}$ lost.* First clause: exactly one backbone qualifies as a gainer, its gap is 0.94 points, and the reserve closes it exactly, so the median closure is 100.0%. **Met**, on a denominator smaller than one item of eighty, which is why §E.12 reports the pass and its resolution in the same sentence. Second clause: the reserve falls below ASP on three of the six backbones where $-\mathcal{C}_{\mathrm{s}}$ lost, by 5.00, 1.25 and 1.25 points. **Failed.** The previous version of this paper reported the first clause only, and reported it as a failure, on the superseded offline generation; both halves of that are corrected here (Limitation (14)).

P2 (the assertion account). *Protecting the exit moves provenance without moving accuracy: the mean absolute change is below two points.* Measured: 4.24 points. **Failed.** P1 and P2 were registered as mutually exclusive on the same measurement, and both failing is a permitted outcome that we did not expect.

P3. *Rationing cuts retrieval spend by at least 40% and raises the rounds used, on every backbone.* Measured: the cut holds on 7/7, the rounds on 5/7. **Half held.**

P4 (registered risk). *Rationing may lose accuracy on retrieval, where width is the point. If it does, that is evidence for width over depth on that subtest and will be reported as such rather than dropped.* Measured: it loses on 4/7 backbones by -3.12 points. **Fired**, and reported in §E.12.

Registered failure condition. *If the reserve removes the scraped exit and accuracy does not move on any backbone, the rounds account is wrong as stated and the paper says so.* Accuracy moved, so this condition did not fire; the narrower statement adopted at the end of §E.12 is a consequence of P2's failure rather than of this one.

Not registered, and therefore labelled post hoc wherever it appears. The pooled comparison of arm means, the resolution-guarded median, the per-subtest breakdown used in §E.11 and §E.12, and the conditional analysis against the non-answer rate. All were written after the numbers existed. Two of them are no longer reported at all: the pooled *closure* ratio has no referent once the ablation's pooled gain is negative, and the resolution-guarded median has 0 backbones above its 5.0-point threshold. The per-subtest breakdown and the conditional analysis are new claims; the conditional one is the most in need of an independent test on a ladder we have not used.

THE V4 RE-MEASURE AND THE TEXT MEDIUM, AS REGISTERED

Three further registrations were frozen before the runs they describe, and their verdicts are collected here rather than scattered through the results, because two of them came out against us and a reader should be able to find all of them in one place.

The offline re-measure (`notes/registered_v4_remeasure.md`). Frozen while the rebuild was half finished and before a single re-measured decision existed. The defect: state updates were capped at a completion length the state JSON outgrew, so the recovery preserved the previous state without counting the failure, and the compressive channel silently stopped advancing on 8 of 28 streams. 135 cells were re-measured.

V1 (manipulation check). Every repaired stream reaches at least 90% *of the episode; residual freezes are reported with their rate rather than hidden.* Measured on the repaired cache: no frozen streams, median reach 100% over the 35 streams carrying a readable timestamp, and a residual update loss of 1.96% across the 51 counter-bearing streams. **Held**, with one caveat we registered in advance and repeat: 18 of 53 streams omit the per-claim timestamp entirely and are reported as UNVERIFIABLE rather than as healthy.

V2 (the discriminating prediction). Mean $|\Delta\text{score}|$ *on the four affected backbones is at least twice that on the three unaffected ones; below* $1.5\times$ *the freeze was not what drove the published numbers and the repair is cosmetic, which we would then report as such.* Measured: $1.04\times$ (9.3 against 9.0). **Fired in the cosmetic band**, and we report it as cosmetic: the movement between generations is a re-sampling floor, not a freeze effect. That is the reading Limitation (4a) takes, and it is why the 9.0-point floor bounds cross-generation comparisons throughout.

V3 (direction). STATE-ONLY *rises on the affected backbones, on at least three of the four.* Measured: it rises on four of four, by 8.33, 8.34, 9.72 and 34.38 points. **Held.**

V4 (the headline survives or is withdrawn). The sign of ASP *minus ablation on the composition subtest is unchanged on the flagship; if it flips, the non-monotonicity headline was in part an artefact of frozen states and §E.11 says so in those words.* Measured at $B = 4096$: the flagship's composition cell moves from 20.83 to 16.7 while the ablation's is unmoved at 83.3, so the sign is unchanged and the magnitude grows slightly, from 62.50 to 66.67 points in the ablation's favour. **Held.** The 20.83 that this prediction retired is the number the previous version of this paper's introduction quoted.

V5 (registered risk). RETR-ONLY *moves least of the affected methods; if it moves as much as* STATE-ONLY*, truncation was damaging the index too and every retrieval number needs the same treatment.* Measured mean $|\Delta|$: 6.50 for RETR-ONLY against 12.35 for STATE-ONLY and 20.80 for ASP. **Held**, and the widening of the repair it guarded against is not required.

The text medium's caption repair (`notes/registered_a1_text_capfix.md`). Frozen after the defect was measured and before the repair was implemented. The retrieval index on the text corpus was ranking on captions written to the visual medium's prompt, which asks for the room and the door state.

Prediction 1. Top-1 disagreement between the shipped hybrid ranking and dense-only falls below $892/1{,}624$ *(*54.9%*), and does not fall to zero.* Measured after the repair: $1{,}238/1{,}624$, or 76.2%. **Failed**: it rose. The repair gave the caption column real content and it now moves the ranking more often, not less, which is the opposite of what we predicted and is reported as such.

Prediction 2. The 9.9%–78.0% *spread across backbones at least halves in width.* Measured: 57.8%–85.3%, a width of 27.5 points against 68.1. **Held.**

Prediction 3. γ *on the control arm rises from* 27.5%. Measured 36.3%. **Held.**

Prediction 4. Control-arm recall rises from 38.86*, and by less than* γ *does.* Measured 41.66, a rise of 2.80 against grounding's 8.8. **Held.**

Prediction 5 (the falsifiable one), two clauses. ASP *still beats* ASP-WALL *(*$41.66 > 38.36$*), and the reserve arm still fails to pay for itself (*$+1.33 \leq 0$*). If the ordering flips, the repair has falsified A1's text-medium section, and the section is withdrawn rather than re-narrated.* Measured: 41.66 against 38.36, so the first clause **holds**; and $+1.33$ points, so the second **flipped**. §E.13 applies the registered consequence to the clause that flipped: the claim about what protecting the exit costs or earns on a text stream is withdrawn, and the reserve arm's advantage over the control on that corpus is recorded here as a falsification event rather than reported there as a result.

The text-medium wall run (`notes/registered_a1_text.md`). Frozen before any decision was paid for; primary outcome token recall, with mean answer length recorded per arm as a guard.

M1 (manipulation check). RESERVE *and* WALL *produce zero scraped exits, as on the visual grid.* Measured: 2 and 1 of 1,624 decisions. **Failed as a manipulation check.** The registration says a non-zero count is to be read as a broken instrument rather than as a finding, and we read it that way: the exit price on this medium is estimated from a character heuristic and the estimate is occasionally short. The population is reduced by more than a factor of thirty and it is not zero, which is a weaker statement than the visual grid's and is the one we make.

N1 (the void condition). The control arm produces a non-zero scraped count; if the budget never binds on a text stream there is no wall to observe and the result is that the wall is a property of image-token pricing. Measured: 73 scraped decisions, 4.5%. **Held**; the run is not void.

P1. γ on scraped exits is lower than γ on answered exits, as on the visual grid. Measured, pooled over all 5 arms: 36.8% on the $5{,}092$ answered decisions, 26.2% on the $2{,}917$ forced ones, 41.4% on the 111 scraped ones. **Failed**, and inverted. Limitation (15) says why we think the measure and not the loop is responsible, and §E.13 does not claim the ladder replicates.

P2. RETR-RATION *cuts mean retrieval width by more than half.* **Not evaluated.** The text records carry a per-decision token total and no per-retrieval token count, so the statistic the registration named cannot be computed from what was stored. We report that as an unfulfilled registration rather than substituting a different statistic, and the fix (record the retrieval spend per read, as the visual harness does) costs nothing to add and cannot be applied retrospectively.

P3. The pooled RESERVE *gain over the control is positive, and across backbones it correlates positively with the control arm's non-answer rate.* Measured: $+1.33$ points pooled and $r = +0.47$ over 7 backbones. **Held on both clauses**, and immediately qualified: the pooled half of it is the clause that *also* constitutes the flip of capfix prediction 5, so it is reported above as a falsification event and is not offered as support for exit protection on this corpus.

The answer-length guard. An arm that gains recall while gaining length is not credited until that is checked. Measured mean answer length: 26.9 tokens in the control against 21.2 in the reserve. The reserve gains recall while producing *shorter* answers, so the guard is satisfied. It is recorded because the registration required it, not because it rescues the clause above.

EXACTLY WHAT WAS EXECUTED

The backbones are named in the body by parameter count, which is not enough to re-run anything. The slugs actually requested, the decoding parameters and the scorer are below, emitted from the client and the grid driver rather than transcribed.

name used in this paper	slug requested from the routing layer
`ministral-3b`	`mistralai/ministral-3b-2512`
`qwen3-vl-8b`	`qwen/qwen3-vl-8b-instruct`
`gemma-3-12b`	`google/gemma-3-12b-it`
`ministral-14b`	`mistralai/ministral-14b-2512`
`qwen3.8-27b`	`qwen/qwen3.8-27b`
`qwen3-vl-30b-a3b`	`qwen/qwen3-vl-30b-a3b-instruct`
`gemma-4-31b`	`google/gemma-4-31b-it`

Decoding is `temperature` $= 0.0$ on every call, backbone and judge alike; `top_p` is left at the provider default and no seed parameter is sent, because the routing layer exposes none for these slugs. Determinism is therefore provider-side and is not guaranteed, which is why the control arm is re-run inside every batch and why §E.2 can measure what that is worth. `max_tokens` on each call is the smaller of the caller's ceiling and the decision's remaining budget, so a call can never overspend B. The scorer is `scripts/sew_score.py`; where its rule is ambiguous it defers to `google/gemma-3-27b-it`, a model that appears in no backbone row.

How much of the accuracy is the judge's opinion. The scorer resolves an item by rule wherever the rule is unambiguous and defers to the judge otherwise, and this paper cannot lean on the judge's validation because that validation lives in a companion with no archival identifier. So the judge is treated as unvalidated here and bounded instead. Whether an item was deferred is a deterministic function of its prediction and gold, so the deferrals can be recomputed offline; the reconstruction

reproduces the count the runner recorded at scoring time on all 273 cells, with 0 mismatches, and only then is anything computed from it (`analysis/a1_judge.py`).

Across the four grids, 310 of 7,280 decisions were deferred, or 4.3%. Of those, the judge scored correct an item the rule would have failed at most 2 times in any arm. Forcing every such item to wrong (the worst case for this paper, applied to every arm at once) moves no arm mean by more than 0.27 points, and moves the four reported contrasts as follows:

contrast	as scored	judge assumed maximally wrong
directive content (§4.5)	+6.43	+6.52
channel removed (§E.12)	+6.11	+6.11
typed vs. undirected (§4.5)	+2.28	+2.46
exit reserved (§E.12)	+2.10	+2.28

No contrast moves by more than 0.18 points and 0 of 4 move toward zero. The judge is not load-bearing on this corpus, which is what one should expect of golds that are single closed-alphabet tokens, and it is why we report this bound instead of an agreement study: a bound covers every annotation a study could have produced.

I AUDIT TRAIL

The honesty of this paper's self-corrections is an asset; scattered through the results sections it was also a tax on the reader's attention. This appendix consolidates the record. The main line now states what holds *now*, with one-line pointers here; the table below carries what changed, from what, and why. Nothing in it changes a verdict: its rows are the corrections Table 4 marks WITHDRAWN or re-priced, plus the two process defects that produced corrections of their own.

REPRODUCTION

The census is a pure function of the recorded logs. `analysis/a1_deep.py` regenerates every number, table and figure coordinate in §E.7 and §4.2 from `results/cells_lambda` and `results/cells_nocs`, importing the runner's action parser and the primary analysis's grounding classifier rather than reimplementing either, so the three files cannot drift apart. Almost no number in this paper is typed into the manuscript by hand: each is a macro spliced between generated markers, and the generator refuses to emit a macro name containing a digit, because LaTeX cannot define one and the failure is silent. The exceptions are named rather than left for a reader to find.

The generators emit pooled and per-backbone figures and not per-subtest ones, and they emit no registered-verdict figures at all, so the subtest breakdown of §E.11 and §E.12, the registered verdicts of Appendix H and Appendix H, the order contrast of §E and the two reclassification figures of §E.14 are computed directly from `results/a1_wall.json`, `results/a1_deep.json`, `results/capfix_verdict.json`, `results/capfix_prepost.json`, `results/a1_text`, `results/a1_text_ext` and `analysis/a1_text_indexprobe.py`, and entered by hand.

Every result added since that sentence was written is generated end to end, per subtest where the result has subtests and with its registered verdicts where it has registrations: the committed-only decomposition (`analysis/a1_kthree_committed.py`), the causal probe (`analysis/a1_cf.py`), the budget-matched marker rerun (`analysis/a1_markbmatch.py`), the update-operator measurement (`analysis/a1_epsu.py`), the answer-set sweep, the marginal estimate and the three-channel matrix. Closing the remaining generators is the repair this appendix named, and the stamps below are what keep the closed part closed.

Generation stamps, and the check that they agree. Limitation (14) records a defect class this appendix is where it would be found: a generated block that keeps a superseded generation's values because its emitter was never invoked, found by reading the manuscript against its result files rather than by the pipeline. Every result added in this version carries a *stamp*: one comment line inside the block, naming the generator and the content hashes of the result files the block was generated from (`analysis/genstamp.py`), and `analysis/check_stamps.py` recomputes

what changed	from	why
"No decision follows the compressive channel"	previous version's central empirical sentence	the attribution rule's sham control has no discriminative power on the state side; replaced by the schema survey (§4.3) and then confirmed causally (§4.3)
"the compressive channel's absence improves grounding"	previous version's reading of the provenance census	commitment was confounded with grounding; separated, the sign reverses (§4.2, §E.11)
"Eq. (1) holds unconditionally"	previous version's Theorem 1 step	the step needed reliability-blindness and did not follow from the definition; the repaired theorem carries the selectivity term and blindness became testable (Limitation 2a)
"removal costs -6.11 on average"	the handicapped ablation's pooled number	two of the arm's handicaps were its own bookkeeping; the registered fair re-run re-priced it (§E.11, item iii-b)
"removal's tracking cost is the channel's absence"	the handicapped tracking cell	the per-subtest split postdates the registration and is reported directional (Table 4)
the marker runs' void condition	written on realised spend	a behavioural consequence of the manipulation trips it by design; the budget-matched rerun repairs it and confirms the null (§E.14, Limitation 14b)
three generated blocks kept superseded values	a splicing emitter defined but never called	found by reading the manuscript against its result files; the stamp-and-check repair is Appendix "Reproduction" (Limitation 14)

Table 11: The paper's self-corrections, consolidated. Each row's right-hand column names the section that now carries the corrected statement; the inline passages keep one-line pointers here rather than their full narratives.

the hashes against the files on disk. A block whose data has moved now disagrees with its own stamp, and a block emitted by a generator that was never called now has no stamp to agree with; both failures are visible to a script instead of to a reader.

The registration standard. Every registration written for this version follows one template (`notes/REGISTRATION_TEMPLATE.md`), adopted after three registrations died on the same design defect: point thresholds and two-sided brackets leave no middle band, so an outcome between the registered values cannot be reported as anything, which is what happened to the typed run's T2/T3 pair and the reallocation run's R2/R3 pair, both of which bracketed the measured number. The template requires a three-band outcome for every registered quantity (below a, between a and b, above b), with the bands set by the noise floors the paper already measures; void conditions on quantities fixed before the run (the intervention's own cost), never on realised spend; a denominator rule written before the run; a stated failure condition naming what the paper does if neither primary band fires; and no comparison across a batch boundary.

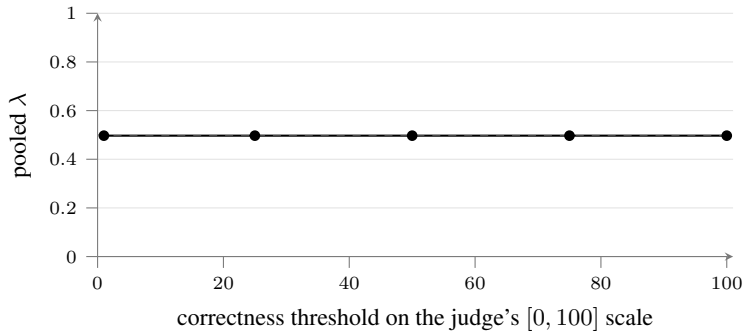

Figure 11: λ against the one free constant in its estimator. The judge returns 0 or 100 on all but a handful of the 168 items, so every threshold in the open interval partitions them identically and the estimate is flat across the whole range. We report the sweep rather than asserting robustness, because a threshold is exactly the kind of constant that gets chosen after the fact.

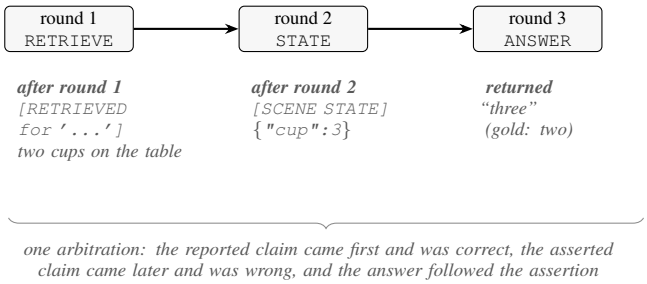

Figure 12: One decision, as the ledger records it. The rounds are the access loop's, the two insertions are the two channels' claims, and the outcome is a single count in the census. The figure is a schematic of the record's structure; the quoted strings follow the format the loop writes rather than reproducing one item verbatim.

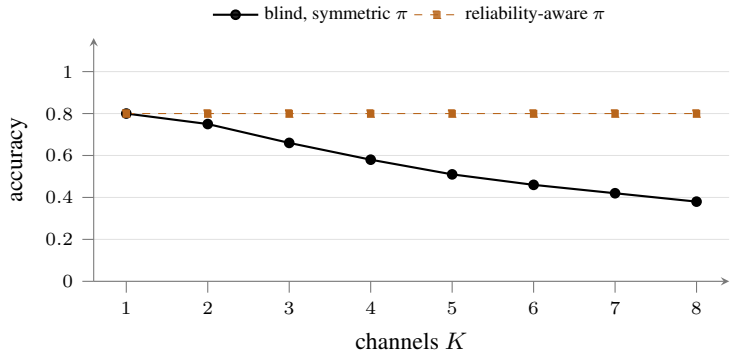

Figure 13: Corollary 1 for a channel set whose members after the first have reliability below the incumbent mean, with the synthesis term held fixed so the selection term is visible on its own. The aware arbiter is flat by Proposition 1: it ignores what it does not need. The blind one declines, and the decline is not recoverable by reallocating B, because the loss is not an allocation. Curves are the corollary evaluated on stated reliabilities, not a measurement.

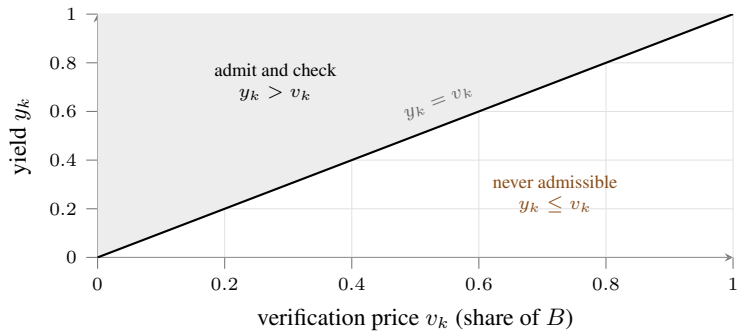

Figure 14: Theorem 2 as a plane. A channel is worth admitting with verification only above the diagonal, and the diagonal is where the price of checking a claim reaches the value of having it. The region below is not a matter of tuning: because v_k is drawn from the same B as the channel, no reliability and no allocation moves a channel out of it. This is the arbitration wall of Definition 6 in the two coordinates that decide it.

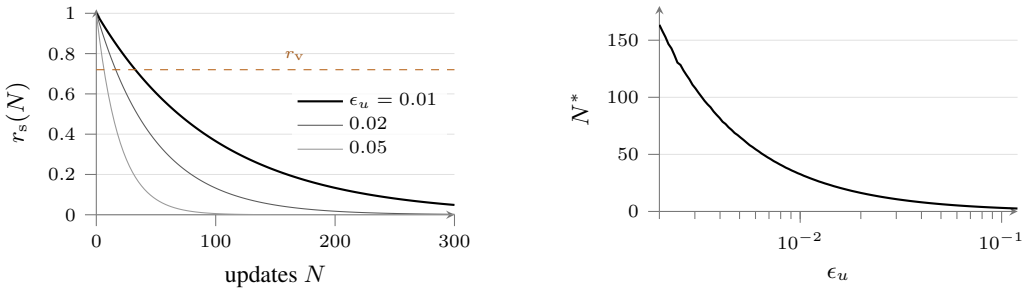

Figure 15: Lemma 2, drawn. Left: the reliability of a channel maintained by one call per observation, against the number of observations, for three per-update error rates, with a fixed reported reliability r_{v} marked. Right: the crossover $N^* = \log r_{\mathrm{v}} / \log(1 - \epsilon_u)$ at which the maintained channel falls below it and, by Corollary 1, begins to subtract. These are the theorem's curves, not a measurement: ϵ_u is a property of a backbone and enters as a parameter.

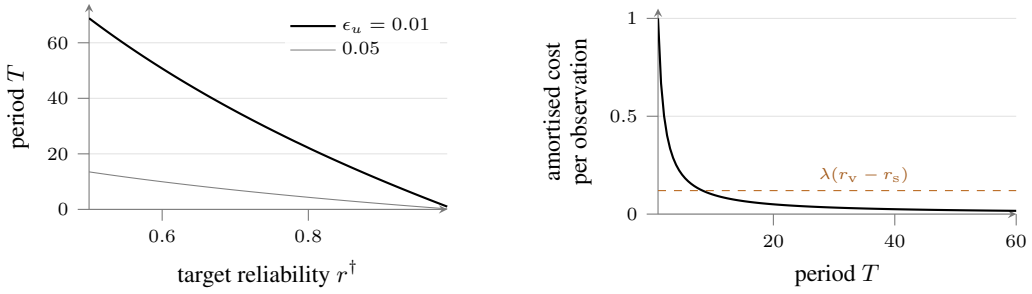

Figure 16: Corollary 2, drawn. Left: the smallest re-derivation period meeting a target reliability, for two per-update error rates. Right: its amortised price per observation, in units of one index aggregation, against the threshold above which re-deriving costs more than the selection loss it prevents. The intersection is the design point, and it is fixed by two measured quantities rather than chosen.

J Additional Figures

The remaining figures of the full manuscript that the main text does not carry: the admission criterion (Fig. 14), the cross-channel lemma (Fig. 15) and the arbitration period (Fig. 16).